Российская академия наук

Институт экологии Волжского бассейна

В.К. Шитиков

Экотоксикология и статистическое моделирование эффекта с использованием R

Тольятти 2016

Шитиков В.К. Экотоксикология и статистическое моделирование эффекта с использованием R. Тольятти: ИЭВБ РАН , 2016. – 149 с.

В книге представлено описание статистических методов, как широко используемых, так и не нашедших пока должного применения при обработке данных экотоксикологического мониторинга. В центре внимания – построение и последующий анализ различных зависимостей "доза-время-эффект" для отклика, представленного в альтернативной, категориальной и метрической шкалах наблюдений. Сюда вошли пробит- и логит-анализ, модели выживания, методы построения различных нелинейных зависимостей, модели сглаживания и т.д. Отдельная глава посвящена сравнительной оценке чувствительности видов к токсикантам и оценке экологического риска для биоценоза. Здесь не ставилась цель подробно описать теоретические аспекты всех этих методов, но широко проиллюстрирована методику их применения на примерах биологического характера.

Описываемые методы моделирования эффекта ориентировались исключительно на статистическую среду R, которая постепенно становится общепризнанным мировым стандартом при проведении научно-технических расчетов. В тексте большинства разделов представлены тексты несложных скриптов в кодах R, дающие возможность читателям легко воспроизвести самим технику выполнения расчетов.

Книга может быть использована в качестве учебного пособия по статистическим методам для студентов и аспирантов высших учебных заведений медицинского и биологического профиля.

Табл. 1, ил. 51. Библиогр. 81 назв.

445003, Россия, Самарская обл.,
г. Тольятти, ул. Комзина, 10
Институт экологии Волжского бассейна РАН
В.К. Шитиков, 2016 г.
E-mail: stok1946@gmail.com
Сайт автора: http://www.ievbras.ru/ecostat/Kiril

СОДЕРЖАНИЕ

Предисловие

Как и любая другая наука, экологическая токсикология включает две основные составные части: фактологическую и методологическую. В первой из них подробно описываются наблюдаемые феномены взаимодействия различных групп веществ (и прочих негативных факторов) с теми или иными функциональными структурами живого организма, популяциями или компонентами экосистемы. В свете этого направления издано много фундаментальных работ как на английском (Bruin, 1976; Timbrell, 2008; Woolley et al., 2008; Wilson, 2012; Fan et al., 2015), так и на русском языке (Лазарев, 1938; Рашевский, 166; Толоконцев, Филов, 1976; Курляндский, Филов, 2002; Куценко, 2004). Разумеется, лучшие из перечисленных книг не ограничиваются одними описаниями наблюдаемых фактов и содержат предположения о механизмах воздействия токсикантов или их метаболизме, обсуждение общих концепций проведения исследований и т.д.

Другое направление, представленное, например, литературными источниками (Беленький, 1969; Finney, 1971; Голубев и др., 1973; Безель и др. 1994; Klein et al, 2003; Безель, 2006; Каплин, 2006; Криштопенко и др., 2008; Newman, 2012; Гелашвили и др., 2014), с разной степенью внятности пытается дать ответ на вопрос: «как правильно поставить эксперимент и обработать результаты наблюдений, чтобы корректно обосновать ту или иную научную гипотезу о характере токсического действия?». К сожалению, многие методические руководства в плане математической обработки носят, в основном, дидактический характер с претензией на общую концептуальность, в то время как практические биологи и токсикологи часто продолжают оставаться в затруднении, сталкиваясь с конкретными проблемами статистического моделирования.

Более 35 лет назад ваш автор имел удовольствие исполнять должность программиста при токсикологической лаборатории. Классический пробит-анализ, метод наименьших квадратов и работы М.Л. Беленького были тогда в центре внимания. Впоследствии мои научные интересы нашли иную стезю, но, когда по прошествии многих лет потребовалось снова оценить изоэффективные дозы, мне неожиданно опять порекомендовали посчитать ″как у Беленького″. Я был в полном недоумении: как такое оказалось возможным после революционных изменений основных концепций статистического моделирования, прошедших в конце XX века и связанных с использованием принципа максимального правдоподобия, обобщенных моделей, рандомизации, бутстрепа и байесовских методов?

К счастью, результаты, полученные мной с использованием лог-нормальной модели почти не отличались от расчетов ″по Беленькому″ на основе макроса Excel (Бабич и др., 2003). Но полученный урок вызвал

непреодолимое стремление показать на примерах возможности использования современных методов статистической обработки.

Выраженность регистрируемого эффекта, соответствующего конкретным проявлениям токсического процесса на организменном или экосистемном уровне, является функцией от силы оказываемого внешнего воздействия. Эти количественные зависимости, представленные в форме статистических моделей ″доза-эффект″, во многом составляют сущность токсикометрии как науки и являются предметом рассмотрения настоящего методического пособия.

Глава 1 является продуктом мучительных компиляций и нужна только как некоторая отправная точка для последующего изложения (своего рода ″макулатурный лист″). Она содержит обзор понятий экотоксикометрии как науки и толкования некоторых общеизвестных терминов.

Главы 2-4 представляют общую методологию построения статистических моделей ″доза-эффект″ и конкретные примеры их реализации. Рассматриваются процедуры подбора спецификации функции регрессии, оценки параметров моделей, проверки гипотез относительно их адекватности экспериментальным данным, сравнения двух или нескольких кривых, использования различных критериев и т.д. В качестве отклика при построении моделей в гл. 2 рассматриваются альтернативные и категориальные признаки, а в гл. 4 – метрические или ″градированные″ показатели токсического эффекта. Гл. 3 рассматривает развитие процесса интоксикации во времени и описывает реализацию моделей выживания.

Здесь предполагается, что читатель уже знаком с методами прикладной статистики (теоретическими распределениями, регрессионным анализом и тестированием статистических гипотез). Было бы бессмысленным писать очередное руководство по математическим основам биометрии, поскольку к настоящему времени в этой области насчитывается сотни тысяч прекрасно изданных книг.

В главе 5 рассматривается оценка вероятности возникновения отрицательных изменений в окружающей среде (или отдалённых неблагоприятных последствий этих изменений) с использованием такого экосистемного показателя как снижение видового богатства. Смысл метода заключается в выявлении критических точек зависимости доза-эффект, которые интерпретируются «как начало наиболее быстрой трансформации экосистемы, либо как точка, после которой начинается выпадение основных компонентов биоценоза или разрушение системных связей» (Воробейчик и др., 1994).

Необходимым условием современного статистического анализа данных является применение различных компьютерных программ, от функциональной полноты и алгоритмической продуманности которых зависит итоговая интерпретация результатов исследования и правильность выводов. Наиболее полной, надежной и динамично развивающейся

статистической средой считается свободно распространяемая система R, объединяющая язык программирования высокого уровня и мощные библиотеки программных модулей для вычислительной и графической обработки данных (http://www.r-project.org/).

Сегодня R является безусловным лидером среди некоммерческих систем статистического анализа и постепенно становится незаменимой при проведении научно-технических расчетов в большинстве западных университетских центров и многих ведущих фирмах. Широкое преподавание статистики на базе пакетов этой среды и всемерная поддержка научным сообществом обусловили то, что приведение скриптов с операторами языка R постепенно становится общепризнанным мировым "стандартом" как в журнальных публикациях, так и при неформальном общении ученых всего мира.

Расширение библиотек программных модулей за счет усилий множества разработчиков привело к возникновению распределенной системы хранения и распространения пакетов R, то есть "CRAN" (Comprehensive R Archive Network – http://cran.r-project.org), которая обладает также развитой системой информационной поддержки. Список только признанных фундаментальных монографий типа «Используем R!» насчитывает уже несколько сотен томов (http://r.psylab.info/library/).

До недавнего времени основным препятствием для русскоязычных пользователей при освоении R являлось, безусловно, то, что почти вся документация по этой среде существовала на английском языке. Однако в последние годы эта проблема потеряла актуальность в связи с появлением весьма полных пособий как отечественных авторов (Заряднов, 2010; Шипунов и др., 2012; Мастицкий, Шитиков, 2015), так и прекрасных переводов лучших зарубежных монографий (Венэбльз, Смит, 2014; Кабаков, 2014). Именно поэтому в настоящем пособии не приводится развернутого описания языка R и его стандартных функций, поскольку все это достаточно полно представлено в перечисленных литературных источниках.

Для статистического анализа приводимых примеров в книге представлены подробно комментируемые перечни команд в кодах R и результаты их выполнения. Напоминаем читателям, что лучший способ освоения материала – самостоятельное выполнение на своем компьютере всех приведенных расчетов. Для этого отдельно файлы скриптов с командами R и исходными данными по всем примерам, представленным в главах 2-5, могут быть загружены из ресурса Интернет: http://www.ievbras.ru/ecostat/Kiril/R/Ecotox.

Настоящее пособие адресуется массовому слою студентов, аспирантов и молодых ученых – биологам, экологам, медикам.

1. ОСНОВНЫЕ ПОНЯТИЯ И ТЕРМИНОЛОГИЯ

1.1. Экотоксикология как научное направление

Экологическая токсикология, или *экотоксикология*, представляет собой междисциплинарное научное направление, изучающее действие вредных загрязняющих веществ, находящихся в окружающей среде, на широкий спектр видов живых организмов, а также их популяций и сообществ, входящих в состав экосистем. В самостоятельную науку экотоксикология выделилась в конце 60-х годов, когда три дисциплины – экология (наука о взаимоотношениях, которые определяют распространение и соотношения живых существ), токсикология (учение о токсичности) и химия, – были объединены в единое научное направление (Зеленин, 2000). На самом деле, эта область знаний помимо указанных включает в себя элементы и других наук, таких, как биохимия, физиология, популяционная генетика, фармакология и т.д.

Химическая экотоксикология изучает развитие токсических эффектов, проявляющихся при распространении в окружающей среде антропогенных химикатов и продуктов их трансформации. Для экотоксикологии интерес представляют такие вещества, которые обладают биодоступностью, т.е. способностью взаимодействовать с живыми организмами в ходе их метаболизма. Часть биодоступных соединений подвергается биотрансформации живыми организмами, участвуя в процессах их материального и энергетического обмена с окружающей средой, т.е. выступают в качестве ресурсов среды обитания. Другие же, чужеродные для организмов химические вещества, называемые *ксенобиотиками* или *экотоксикантами*, могут накапливаться в организмах до опасных уровней концентраций и/или оказывать вредное воздействие на течение нормальных физиологических процессов. Кроме громадного спектра химических веществ, экотоксикология изучает также воздействие таких потенциально вредных факторов, как радиация, шум, аномальные термические потоки и др.

В отличие от медицинской токсикологии, основной задачей экотоксикологии является, как правило, изучение влияния вредных факторов не только применительно к индивидуальным живым организмам, но и на уровне популяций (включая человека), сообществ или экосистемы в целом. Разумеется, в процессе изучения эффектов химических веществ, присутствующих в окружающей среде, экотоксикология основывается на уже устоявшихся категориях, понятиях и теоретической базе классической медицинской и промышленной токсикологии. Однако, в связи с необходимостью учета многочисленных аспектов экосистемного и популяционного уровня, в экотоксикологии осуществляется разработка специфических экспериментальных и методических подходов.

Теоретической основой экотоксикологии является учение об *экотоксичности*, а базовыми проблемами: характеристика *ксенобиотического профиля* среды обитания, вопросы *экотоксикокинетики*, *экотоксикодинамики*, *экотоксикометрии* (Куценко, 2004; Безель, 2006; Гелашвили, 2015).

1.2. Базовые проблемы экотоксикологии

Ксенобиотический профиль. Совокупность ксенобиотиков, содержащихся в окружающей среде и взаимодействующих с биологическими объектами экосистемы, составляют ксенобиотический профиль (КБП) биогеоценоза. Ксенобиотический профиль следует рассматривать как один из важнейших факторов внешней среды (наряду с температурой, освещенностью, влажностью, трофическими условиями и т.д.), который может быть описан качественными и количественными характеристиками. Ксенобиотики, содержащиеся в органах и тканях живых существ, являются элементами КБП, поскольку все они рано или поздно потребляются другими организмами в процессе питания (т.е. обладают биодоступностью). Напротив, химические вещества, фиксированные в твердых объектах, нерастворимых в воде и не переносимых в форме пылевидных частиц, хотя и не обладают биодоступностью, однако являются источниками формирования КБП.

Ксенобиотические профили среды, сформировавшиеся в ходе эволюционных процессов, являются естественными. Они различны в разных регионах Земли. Очевидно, что популяции и сообщества организмов, существующие в разных регионах (биотопах), в той или иной степени адаптированы к привычным для них естественным КБП. Естественный КБП может меняться под воздействием природных и антропогенных факторов, к числу которых относятся химические вещества, накапливающиеся в окружающей среде в результате хозяйственной деятельности человека в несвойственных ей количествах, выступающие в качестве экополлютантов (загрязняющих веществ).

Экополлютант, накопившийся в среде в количестве, достаточном для инициации токсического процесса в популяции или биоценозе (на любом уровне организации живой материи), может быть обозначен как экотоксикант. Определение количественных параметров, при которых экополлютант трансформируется в экотоксикант, является насущной проблемой экотоксикологии. При её решении необходимо учитывать, что в реальных условиях на популяцию или биоценоз действует весь ксенобиотический профиль среды, модифицируя при этом биологическую активность отдельного поллютанта. Поэтому в разных регионах, характеризующихся специфическим биоценозами и КБП, количественные параметры реальной опасности каждого экотоксиканта могут быть различны.

Экотоксикокинетика. Судьбу ксенобиотиков (экополлютантов) в окружающей среде изучает экотоксикокинетика. В круг ее задач входит идентификация и оценка мощности источников появления токсикантов; наблюдение за характером перераспределения вещества под действием многочисленных абиотических и биотических процессов в среде обитания; прогноз интенсивности самоочищения, биотрансформации и других химических превращений в различных разделах экосистемы. В результате разрабатываются компоненты материального баланса экополлютантов с расчетом констант скорости поступления, элиминации и накопления в различных компартментах.

Многие ксенобиотики, попав в воздух, почву, воду приносят минимальный вред экосистемам, поскольку время их воздействия мало. Вещества, оказывающиеся устойчивыми к процессам разрушения, и, вследствие этого, длительно *персистирующие* в окружающей среде, как правило, являются потенциально опасными экотоксикантами. Постоянный выброс в окружающую среду таких поллютантов приводит к их накоплению и, в конечном счете, поражению наиболее уязвимого (чувствительного) звена биосистемы. После прекращения выброса персистирующего токсиканта он еще длительное время сохраняется в среде. К числу веществ, длительно персистирующих в окружающей среде, относятся тяжелые металлы (свинец, медь, цинк, никель, кадмий, кобальт, сурьма, ртуть, мышьяк, хром), полициклические полигалогенированные углеводороды (полихлорированные дибензодиоксины и дибензофураны, полихлорированные бифенилы и т.д.), некоторые хлорорганические пестициды (ДДТ, гексахлоран, алдрин, линдан и т.д.) и многие другие вещества.

Абиотическое разрушение химических веществ обычно проходит с малой скоростью. Значительно быстрее очищение среды от ксенобиотиков происходит при участии биоты (*биотрансформация*), особенно вследствие энзиматического разрушения микроорганизмами (главным образом бактериями и грибами), которые используют их как питательные вещества. В основе биотрансформации загрязняющих веществ лежат процессы окисления, гидролиза, дегалогенирования, расщепления циклических структур молекулы, отщепление алкильных радикалов и т.д.

Процесс, посредством которого организмы накапливают токсиканты, извлекая их из абиотической фазы (воды, почвы, воздуха) и из пищи (трофическая передача), называется *биоаккумуляцией*. Результатом биоаккумуляции являются пагубные последствия как для самого организма (достижение поражающей концентрации в критических точках), так и для организмов, использующих данный биологический вид в качестве пищи. Водная среда обеспечивает наилучшие условия для биоаккумуляции вредных соединений.

Склонность экотоксикантов к биоаккумуляции зависит от ряда факторов. Первый – персистентность ксенобиотика, поскольку степень накопления загрязняющего вещества в организме, в конечном счете, определяется его содержанием в окружающей среде. Наибольшей способностью к биоаккумуляции обладают жирорастворимые (липофильные) вещества, медленно метаболизирующие в организме. Жировая ткань является, как правило, основным местом длительного депонирования ксенобиотиков. Биоаккумуляция может лежать в основе не только хронических, но и отсроченных острых токсических эффектов. В экологически неблагополучных регионах это может сопровождаться массовой гибелью животных при достижении ими половой зрелости. Стойкие поллютанты могут также передаваться потомству, у птиц и рыб – с содержимым желточного мешка, у млекопитающих – с молоком кормящей матери. При этом возможно развитие у потомства таких эффектов, которые не проявляющихся у родителей.

Химические вещества могут перемещаться по трофическим цепям от организмов–жертв, к организмам–консументам. Для высоко липофильных веществ это перемещение может сопровождаться увеличением концентрации токсиканта в тканях каждого последующего организма – звена пищевой цепи (*биомагнификация*). Типичный пример: в планктоне содержание метил-ртути составляет примерно 0.01 мкг/г, в мышечной ткани хищных рыб достигает 1.5, а у птиц-рыболовов – 3-14 мкг/г.

В ответ на воздействие токсиканта в биосистеме формируются и развиваются реакции, которые могут привести к нарушению отдельных её функций, потере жизнеспособности и, в конечном итоге, гибели. *Экотоксикодинамика* – раздел экотоксикологии, рассматривающий конкретные механизмы развития и формы токсического процесса, вызванного воздействием экотоксикантов на отдельный организм, весь биоценоз или составляющие его виды. Эти механизмы, посредством которых вещества могут вызывать неблагоприятные эффекты, многочисленны и, вероятно, в каждом конкретном случае уникальны.

Токсические процессы могут развиваться по *пороговому* или *беспороговому* принципу. Для процессов, формирующихся по пороговому принципу, причинно-следственная связь между фактом воздействия вещества и развитием процесса носит безусловный характер: при действии веществ в дозах ниже определенных уровней токсический процесс не развивается; при достижении определенной дозы процесс развивается непременно. Зависимость "доза-эффект" прослеживается на уровне каждого отдельного звена биосистемы, при этом, чем больше доза, тем значительнее проявления токсического процесса.

Для процессов, развивающиеся по беспороговому принципу, причинно-следственные связи между фактом действия вещества и развитием процесса носят случайный характер: вероятность формирования

эффекта сохраняется при действии на организм даже одной молекулы токсиканта, вместе с тем у отдельных организмов процесс может и не развиться, несмотря на значительное увеличение дозы вещества (близкое к смертельным). К таким токсическим процессам относятся некоторые аллобиотические состояния, специальные токсические процессы (канцерогенез, тератогенез), отчасти нарушение репродуктивных функций и т.д.

Экотоксичность – это способность данного ксенобиотического профиля среды вызывать неблагоприятные эффекты в соответствующем биоценозе. В тех случаях, когда нарушение естественного ксенобиотического профиля связано с избыточным накоплением в среде лишь одного поллютанта, можно условно говорить об экотоксичности только этого вещества.

Можно выделить прямое, опосредованное и смешанное действие экотоксикантов. Прямое действие – это непосредственное поражение организмов определенной популяции или биоценоза некоторым экотоксикантом или совокупностью экотоксикантов данного ксенобиотического профиля среды. Опосредованное – это действие ксенобиотического профиля на окружающие биотические или абиотические элементы, в результате чего некоторые условия среды обитания или модифицированные ресурсы перестают быть оптимальными для существования популяции. Многие токсиканты способны оказывать как прямое, так и опосредованное, т.е. смешанное действие.

Одним из важнейших путей достижения оптимальных взаимоотношений человека и природы является нормирование техногенной нагрузки на окружающую среду. Необходимо отметить, что нормирование должно учитывать оба аспекта загрязнения биосферы токсическими соединениями: *а*) установление предельно-допустимых концентраций ксенобиотиков в объектах окружающей среды и *б*) выяснение степени трансформации биотических компонентов в условиях техногенного загрязнения среды обитания. При этом многочисленными исследованиями установлена целесообразность нормирования токсикантов не по их содержанию в среде, а по реакции экосистемы на это загрязнение.

Современная методология оценки техногенной нагрузки и обоснования экологического риска основывается, в первую очередь, на натурных исследованиях природных экосистем, находящихся в градиенте воздействия. Неблагоприятные экотоксические эффекты целесообразно рассматривать на различных уровнях организации живой материи:

- *аутэкотоксические* эффекты на уровне организма проявляются как снижение резистентности к другим действующим факторам среды, понижение активности, канцерогенез, нарушение репродуктивных

функций и другие заболевания, приводящие в экстремальных случаях к гибели организма;

- *демэкотоксические* эффекты на уровне популяции проявляются как рост заболеваемости и смертности, уменьшение рождаемости, увеличение числа врожденных дефектов развития, нарушение демографических характеристик (соотношение возрастов, полов и т.д.), культурная деградация и, возможно, деградация и гибель популяции;
- *синэкотоксический* эффект на уровне биогеоценоза заключается в изменении популяционного спектра ценоза, вплоть до исчезновения отдельных видов и появления новых, не свойственных данному биоценозу, нарушении межвидовых взаимоотношений и трофических связей, скудности необходимых ресурсов и, наконец, опустынивании территории.

В зависимости от продолжительности действия экотоксикантов на экосистему можно говорить об острой и хронической экотоксичности. Острое токсического действия веществ на биоценоз может явиться следствием аварий и катастроф, сопровождающихся выбросом в окружающую среду большого количества относительно нестойкого токсиканта, или неправильного использования химикатов. С хронической токсичностью веществ, как правило, ассоциируются сублетальные эффекты. Часто при этом подразумевают нарушение репродуктивных функций, иммунные сдвиги, эндокринную патологию, пороки развития, аллергизацию и т.д. Однако хроническое воздействие токсиканта может приводить и к смертельным исходам среди особей отдельных видов.

Повышенное содержание токсичных веществ в абиотических компонентах (воде, почвах, воздухе) неизменно ведет к повышенным концентрациям этих веществ в растительных и животных организмах. Казалось бы, дело обстоит предельно просто: достаточно знать содержание токсичных веществ в объектах внешней среды, чтобы прогнозировать их дальнейшее накопление в трофической цепи, определяя тем самым токсическую нагрузку на отдельные компоненты биоты.

Однако в реальных условиях множество трудно учитываемых механизмов влияют на эти процессы, в том числе:

- Пространственная неоднородность и различие уровней загрязненности территории, определяемые спецификой техногенного воздействия, локальными почвенно-климатическими и физико-химическими условиями среды.
- Особенности экологии растительных и животных сообществ, включающие видовую и сезонную специфику пищевых рационов, разнокачественность сред обитания, миграционные потоки и т.д.

Если рассматривать экотоксичность лишь одного вещества в отношении представителей только одного вида живых существ, то в полной мере могут быть использованы качественные и количественные характеристики, принятые в классической токсикологии (величины

острой, подострой, хронической токсичности, дозы и концентрации, вызывающие мутагенное, канцерогенное и иное действие и т.д.). Однако для более сложных систем экотоксичность нельзя оценить одним параметром токсикометрии, и она характеризуется через понятия *вероятность опасности* или *экологический риск* посредством сопоставления целого ряда качественных или полуколичественных показателей.

Важнейшей характеристикой ксенобиотиков с позиции экотоксикологии является их *экотоксическая опасность,* под которой понимают потенциальную способность вещества в конкретных условиях вызывать повреждение биологических систем (организмов, популяций, сообществ) при попадании в окружающую среду. Потенциальная опасность вещества, определяется его стойкостью в окружающей среде (персистентность), способностью к биоаккумуляции, величиной токсичности для представителей различных биологических видов.

Обоснование экологических рисков в общем случае – это выявление отрицательных изменений в окружающей природной среде с оценкой *вероятности* проявления отдалённых неблагоприятных эффектов в структурно-функциональной организации биогеоценозов, возникающих вследствие негативного воздействия факторов среды. Экологический риск, например, применительно к деградации почвенного покрова является вероятностной мерой потери способности определенного вида почвы функционировать в пределах, обеспечивающих необходимую продуктивность и безопасность сопряженных экосистем естественного или искусственного типа.

Анализ риска включает три взаимосвязанных элемента: оценка риска, управление риском и информирование о риске.

Оценка экологического риска – это процесс определения вероятности развития (включая популяции человека) в результате изменений различных характеристик среды. Величина риска в стоимостном выражении определяется как произведение величины ущерба I на вероятность W события i, вызывающего этот ущерб: $R = I\ W_i$.

Управление риском – это анализ самой рисковой ситуации, разработка и обоснование управленческого решения, как правило, в форме нормативного акта, направленного на уменьшение риска, поиск путей сокращения риска.

Информирование о риске представляет собой процесс распространения результатов определения степени риска среди специалистов и заинтересованной части населения.

Методология оценки популяционных нормативов безопасного воздействия и экологического риска находится в самом начале своей разработки. В подавляющем большинстве случаев её выводы носят качественный или описательный характер, а попытки внедрить методы

количественной оценки сталкиваются с серьезными трудностями. Это обусловлено сложностью экосистем, комплексностью воздействия на среду стрессоров (не только химической, но и физической, и биологической природы), недостаточной изученностью характеристик экотоксической опасности огромного количества ксенобиотиков, циркулирующих в природе. Именно поэтому проблемы оценки популяционного экологического риска являются актуальными.

Целью последующего изложения является теоретическое и практическое представление методов статистического моделирования зависимостей "доза-эффект", оценки показателей токсикометрии и обоснования экологического риска.

1.3. Зависимость "доза-эффект" и показатели токсикометрии

Экотоксикометрия – раздел экотоксикологии, в рамках которого рассматриваются методические приемы, позволяющие *количественно* оценить (перспективно или ретроспективно) экотоксичность ксенобиотиков. Одной из ключевых позиций в алгоритме оценки токсикометрических параметров является изучение токсического процесса и установление зависимости "*доза-эффект*" (в англоязычной терминологии "*доза-отклик*" или dose-response).

Количественная оценка токсического эффекта возможна при *квантованной* (альтернативной или категориальной) *и метрической* формах учета реакций биологического объекта – организма, популяции, биоценоза. Наиболее отработанной считается методика оценки параметров токсикометрии при учете реакций в альтернативной форме, т.е. когда исследователь отмечает лишь наличие или отсутствие того или иного проявления токсического процесса. В качестве такой реакции принимается обычно вполне очевидное состояние, например, смерть животного, судороги, обездвижение, симптом "бокового положения" и т.д. Тогда целью статистической обработки результатов опыта является построение зависимости между испытанными дозами и оценками вероятности (относительными частотами, процентами) проявления изучаемого эффекта.

Учет реакций в альтернативной форме дает возможность сопоставлять между собой количественно токсичность различных веществ, оказывающих качественно одинаковое действие. Однако использование бинарной шкалы отклика в свою очередь порождает ряд проблем. Во-первых, построение корректно интерпретируемых статистических моделей возможно при относительно небольших различиях апостериорных вероятностей обеих альтернатив (не более чем на порядок). Например, если в выборочных уловах доля рыб, пораженных гельминтами, составляет менее 5%, то прогнозирующая сила такой модели будет невелика. Во-вторых, для оценки вероятностей необходимо знать

точный объем полной выборки, что не всегда доступно. Например, получив в регистратуре поликлиники данные о числе сердечнососудистых заболеваний, построить модель нельзя, поскольку отсутствует информация о числе не заболевших реципиентов. Наконец, если регистрируемые реакции выражаются в порядковых, счетных или метрических шкалах, то перевод их в более "грубую" альтернативную шкалу часто связан с какими-то субъективными предположениями о разбиении на классы, что может серьезно исказить внутреннюю структуру исходных данных .

Если токсический эффект представлен в метрической ("градированной") форме, то результаты эксперимента выражаются уже не в частотах, а произвольных единицах измерения (миллиметрах, минутах, граммах, градусах и т.д.). Как отмечалось выше, вводимые в экологической токсикологии показатели проявления эффекта токсического воздействия, учитывающие физиологические, функциональные или биохимические реакции организма, во многом аналогичны тем, которые применяются в гигиене. В дополнение к этому, зависимости, отражающие, например, численность объектов, занимаемую площадь, выживаемость, видовое разнообразие, интенсивность процессов деструкции, продуктивность биоценозов и др., применимы лишь в экологической токсикологии.

Общее количество учитываемых проявлений эффекта на организменном и ценотическом уровне может достигать многих десятков, однако к настоящему времени не прослеживаются внятные попытки унифицировать принципы формализации биологической значимости показателей и параметров. В то же время, нет сомнений в целесообразности и необходимости всесторонней оценки возможных отклонений экологического состояния от нормы, а важность того или иного параметра может быть оценена только при анализе всего многообразия связей и отношений как для индивидуального организма, так и с учетом специфичности изучаемой экосистемы.

Токсикометрия тесно связана с токсикологическим экспериментом и требует реализации трех исследовательских этапов: *а*) планирования эксперимента, *б*) его проведения и *в*) статистической обработки эмпирических данных и интерпретации результатов с получением количественных оценок. В общем случае, массивы исходной информации для нормирования воздействий могут быть составлены в процессе полевых биоиндикационных наблюдений *in situ* (пассивная выборка), в ходе лабораторного биотестирования, проведенного *ex situ*, либо по результатам острого или хронического токсикологического эксперимента.

Перечисленные методы получения исходных данных имеют различные технические условия исполнения, степень регламентации и уровень неопределенности. В частности, стандартные токсикологические исследования проводятся на правильно подобранных группах здоровых

лабораторных животных, содержащихся при соответствующих условиях, которые подвергаются воздействию точно градуированных доз исследуемого вещества, включая плацебо для контрольной группы. При проведении лабораторных испытаний по биотестированию различных многокомпонентных смесей (отходов, осадков, почв) конкретный химический состав пробы в значительной степени случаен, а токсичность образца оценивается по степени его разведения и продолжительности экспозиции. Если в первом случае "доза" имеет реальный смысл количества вещества, введенного в организм реципиента, то при биотестировании речь уже идет об общей оценке токсичности субстрата, не связанной ни с дозой, ни с концентрацией ксенобиотика в окружающей среде. В связи с этой неопределенностью, под "*дозой*" мы будем понимать любое (химическое, термическое, биологическое, радиационное) воздействие на экосистему, измеряемое в непрерывной шкале и выражаемое количественно в произвольных единицах измерения.

Токсический эффект определяется не только количеством экотоксиканта, но и временем его воздействия. Связь между дозой вещества, временем и эффектом может быть представлена в виде поверхности в трехмерном пространстве. При сечении этой поверхности плоскостями, параллельными координатным осям, получаются три семейства кривых, попарно связывающих дозу яда, время и эффект, которые и являются предметом исследования токсикологов.

В соответствии с этим возможны три типа токсикологических экспериментов и соответствующих им искомых статистических зависимостей:

- эксперименты по установлению связи между дозой (концентрацией) ксенобиотика и токсическим эффектом;
- эксперименты по установлению зависимостей между временем воздействия яда и эффектом (имеют значение, например, при установлении предельно допустимых концентраций вредных веществ для атмосферного воздуха);
- эксперименты по установлению связей между дозой (концентрацией) яда и временем наступления фиксированного токсического эффекта (опыты по изучению выживания или кумулятивных свойств токсикантов).

Если рассматривать токсические процессы, развивающиеся по пороговому принципу, то чаще всего зависимость "доза-эффект" имеет форму характерной S-образной кривой. На рис. 1.1 приведены результаты обработки эксперимента по оценке смертности морских свинок от кессонной болезни в зависимости от глубины погружения (в футах над уровнем моря). Здесь мы лишний раз показываем, что понятие “*доза*” может иметь иногда весьма специфическое наполнение.

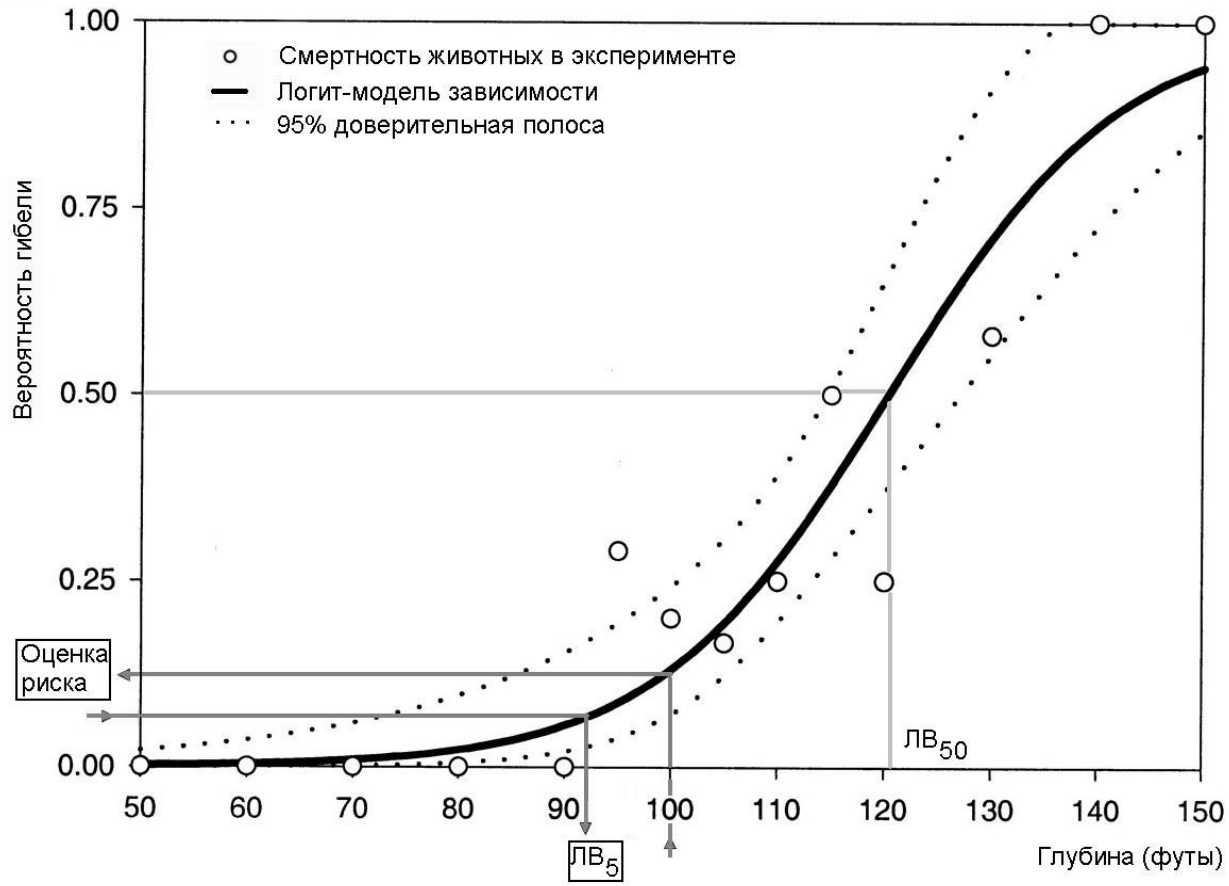


Рис. 1.1. Пример модели "доза-эффект"

Поскольку истинные координаты кривой "доза-эффект" неизвестны, ее положение в приведенном примере оценивается путем аппроксимации эмпирических данных с использованием модели логистической регрессии. Отметим, что, даже несмотря на строгий подбор животных в группах, статистический разброс экспериментальных точек, обусловленный изменчивостью внутривидовой чувствительности и другими погрешностями опыта, обычно достаточно велик. Объем выборки, необходимый для построения регрессионной зависимости, изменяется очень широко (от нескольких групп до многих сотен), но очевидно, что репрезентативность данных очень важна как для построения модели, так и для достоверности заключений, основанных на ней.

Построенная модель "доза-эффект" дает возможность оценить уровни воздействия, приводящие к заданной величине эффекта. В токсикометрии наиболее часто употребляется величина LD_{50} ($ЛД_{50}$) – среднестатистическая разовая доза, которая вызывает гибель 50% особей. В иных случаях можно говорить о среднелетальной концентрации (LC_{50} или $ЛК_{50}$ при биотестировании) или, как в примере на рис. 1.1, среднелетальной глубине (в общем случае, *воздействии*) $ЛВ_{50}$ = 120.7 фута. Если регистрируется иной эффект, кроме смертельного, то оцениваются среднеэффективные воздействия ED_{50} или EC_{50}.

Любые p -эффективные дозы { p = 1, 5, 10, 16, 50, 84, 95, 99%} носят вероятностный характер и нуждаются в оценке точности и надежности. Традиционным способом является построение *доверительных интервалов* регрессионной модели, позволяющих с заданной надежностью указать, в каких пределах может находиться случайное значение найденной выборочной характеристики. Так на рис. 1.1 с 95% доверительной вероятностью значение $ЛВ_{50}$ находится на интервале от 115.5 до 126 футов.

Стрелки на рис. 1.1 указывают, что зависимость "доза-эффект" может использоваться "прямым", так и "обратным" способом. В первом случае задаются, например, наблюдаемой концентрацией токсикантов на загрязненном участке (ось X) и оценивают экологический риск появления анализируемого эффекта на оси Y. В представленном примере при помещении животных на глубину 100 футов смертельное развитие кессонной болезни возможно в 11% случаев (с учетом 95% доверительного интервала – от 5.5 до 21%).

"Обратный" способ используется для обоснования критериев качества окружающей среды: на оси Y выбирается некоторый допустимый уровень p и на оси X находится пороговая концентрация EC_p, которая обеспечит отсутствие эффекта с вероятностью (1 - p). Например, если принять приемлемым риск p = 5% смерти от кессонной болезни, то погружение на глубину более 92 футов следует считать небезопасным (с учетом 95% доверительного интервала – более 72 футов).

Теоретически любое воздействие начинается с некоторого токсического порога (threshold), ниже которого не обнаруживается влияние токсиканта. Важным итоговым результатом эксперимента является оценка концентрации *NOEC* (No observed effect concentration), ниже которой воздействие уже не наблюдается. Некоторым "противовесом" ей служит величина экспериментально определяемого порога *LOEC* (Lowest observed effect concentraton) – минимальная концентрация, при которой наблюдается влияние токсиканта. Третий параметр *MATC* (Maximum acceptable toxicaut concentration) трактуется как максимально допустимая концентрация вредного вещества и находится, например, как среднее геометрическое между *NOEC* и *LOEC*. В отечественной литературе некоторым аналогом *MATC* является термин ПДК – предельно допустимая концентрация.

В настоящее время используются два общих подхода, чтобы оценить *NOEC* и *LOEC* по экспериментальным данным: *а*) исследования регрессионной модели "доза-эффект" и *б*) проверка гипотез в ходе дисперсионного анализа. Определенное преимущество оценки показателей токсикометрии по регрессионной модели состоит в том, что *NOEC*, также как и LC_{50} или EC_{50}, легко вычислить и интерпретировать, поэтому этот метод является важным компонентом многих экологических и

статистических процедур оценки риска (Scholze et al., 2001). Однако найти стартовую точку перегиба кривой в явном виде весьма трудно, а статистическая неопределенность *X* в области низких концентраций, как правило, весьма велика. Кроме того, дать обоснованные рекомендации по выбору конкретной пороговой величины эффекта *p*, соответствующего *NOEC*, часто невозможно. Это решение относится к сфере политики, а не науки.

С точки зрения дисперсионного анализа, *NOEC* – это самая высокая концентрация для подопытной группы со средним откликом, статистически значимо не отличающимся от среднего отклика для контрольной группы. Тогда *LOEC* – самая низкая концентрация в опыте, имеющая средний отклик, который действительно значимо отличается от контроля. Это легко установить при апостериорном анализе ANOVA с использованием, например, множественных сравнений в тесте Даннета. Однако такой подход также вызывает ряд вполне аргументированных возражений (Crane, Newman, 2000). В частности, если нулевая гипотеза не отклоняется, то нельзя утверждать, что между группами нет статистически значимых отличий (без оценки вероятности ошибки II рода), поэтому вполне возможно, что *NOEC* не является безопасной неэффективной концентрацией. Таким образом, можно констатировать, что для оценки широко декларируемого показателя токсикометрии *NOEC* отсутствуют как стандартизованные условия проведения эксперимента, так и адекватные статистические процедуры, обеспечивающие точность, несмещенность и надежность его оценки.

Построение моделей "доза-эффект" осуществляется с учетом следующих соображений. Сам вид этой функциональной зависимости далеко не всегда может быть постулирован, исходя из эколого-медицинских предположений. «Биологический смысл зависимости "доза-эффект" заключается в ее изначальной неопределенности для каждого исследуемого агента… Задавать априорно какие-либо модели, а затем статистически проверять их соответствие исходным данным возможно в довольно редких случаях.» (Криштопенко и др., 2008, с. 18).

Экспериментально показано, что многие специфические эффекты (мутагенез, канцерогенез) при воздействии токсичных соединений и низкоинтенсивного радиоактивного облучения могут быть описаны самыми разнообразными нелинейными функциями: бимодальной, инвертированной, U-образной, включающей отрезки с суперлинейностью, сублинейностью и линейностью. Такое явление получило в токсикологии название "парадоксальная токсичность". Поскольку избежать парадоксов, как правило, не удается, подбор функциональной зависимости, наиболее правдоподобно объясняющей экспериментальные данные – важная и сложная область токсикометрии.

Большинство объектов в природе и технике могут быть отнесены к линейным системам (рис. 1.2а), свойства которых не зависят от их состояния. Используя эту аналогию, линейная форма уравнения "доза-эффект", например, для альтернативных показателей будет иметь вид:

$$ED_i = ED_{50} + S_{ED}\, x_i\ ,$$

где S_{ED} – стандартное отклонение, x_i – стандартизованное значение дозы.

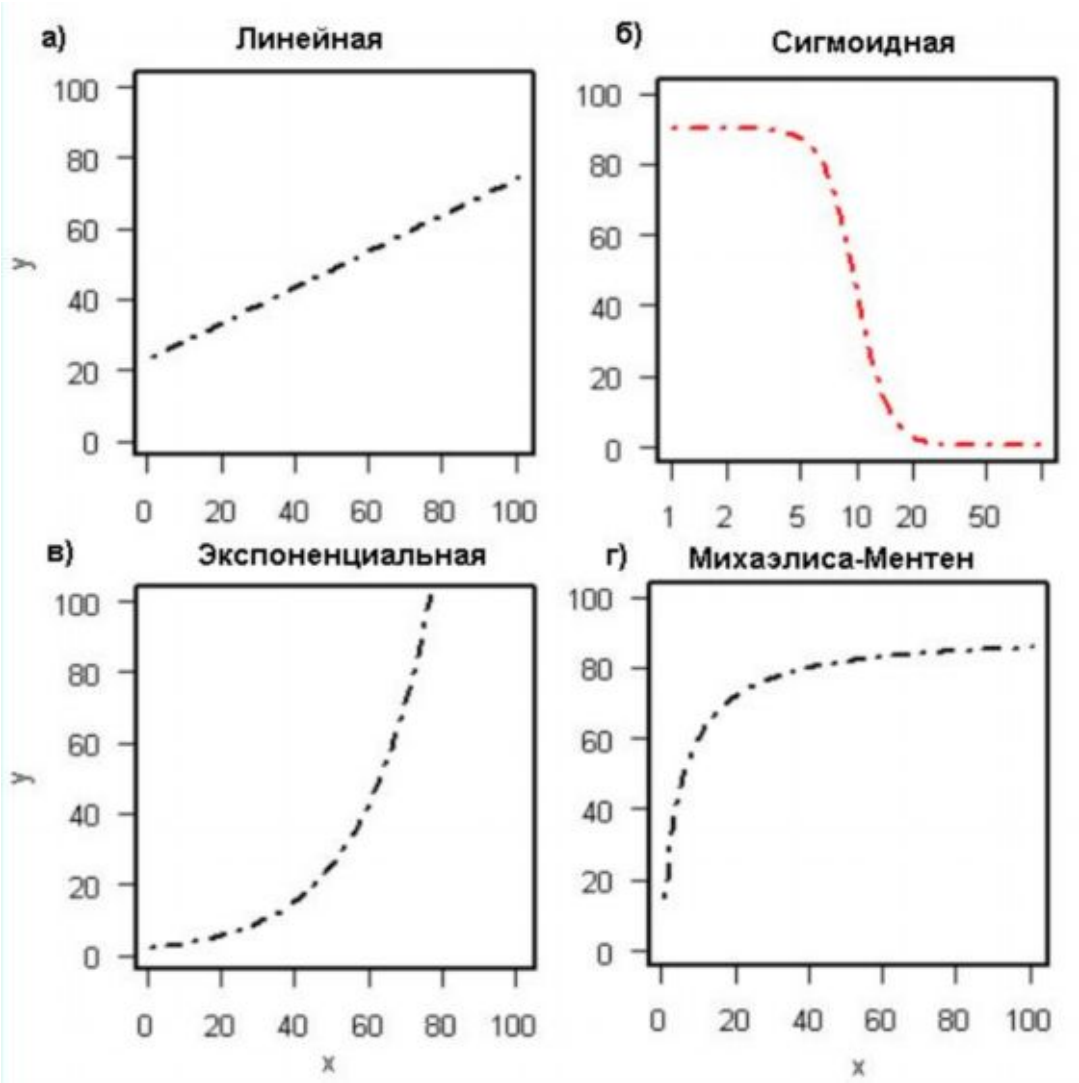


Рис. 1.2. Модели большинства динамических явлений в биологии

Однако, принимая гипотезу линейности, мы вносим заведомо нереальные предположения о механизме взаимодействия вещества с ксенобиотиком. Во-первых, что наблюдаемый отклик биосистемы неограниченно и пропорционально растет с увеличением воздействия, не приближаясь к естественному порогу насыщения. В реальности динамический процесс обычно асимптотически сходится к одному или двум критическим значениям абсциссы или ординаты (рис. 1.2б-г).

Другое нереальное предположение линейности – гомогенность дисперсии, т.е. считается, что разброс индивидуальной чувствительности

биологических объектов приблизительно одинаков на всех уровнях воздействия. Легко показать, что это далеко не так: вариация реакций организма весьма мала как при низких концентрациях, когда эффект еще никак не проявляется, так и высоких уровнях, когда "пациент скорее мертв, чем жив". Как будет показано далее в разделе 2.2, учет пороговости и гетероскедастичности дисперсии в рамках, фактически, все той же линейной модели, осуществляется с использованием двух операций: введением функции связи $q(ED)$, где $q()$ – логит или пробит-трансформация, и предварительным логарифмированием независимой переменной x.

При промежуточных значениях интенсивности могут иметь место нестационарные и "лавинообразные" (бифуркационные по И. Пригожину) механизмы: отклик организмов на токсическое воздействие часто обусловлен реакцией сразу нескольких различных разделов патогенетической системы, степень активации которых имеет индивидуальные особенности. Поэтому для большинства случаев является справедливой S-образная (сигмоидная) форма кривых "доза-эффект" с ярко выраженной зоной резкого перехода в средней части (рис. 1.2а). Существуют попытки найти истоки "сигмоидности" в кинетике развития отравления, соотношениях между скоростями поступления и выведения ядов в организме, особенностям "транспорта" веществ к рецепторам и т.д., однако вполне корректного объяснения для эмпирически полученной S-образности пока не найдено (Рашевский, 1966; Справочник…, 1989).

Кривые "доза-эффект" позволяют сделать важные сравнительные выводы о протекании токсического процесса в различных условиях, которые могут быть основаны на таких свойствах найденных зависимостей как сила воздействия, максимальный эффект, угол наклона и изменчивость – рис. 1.3.

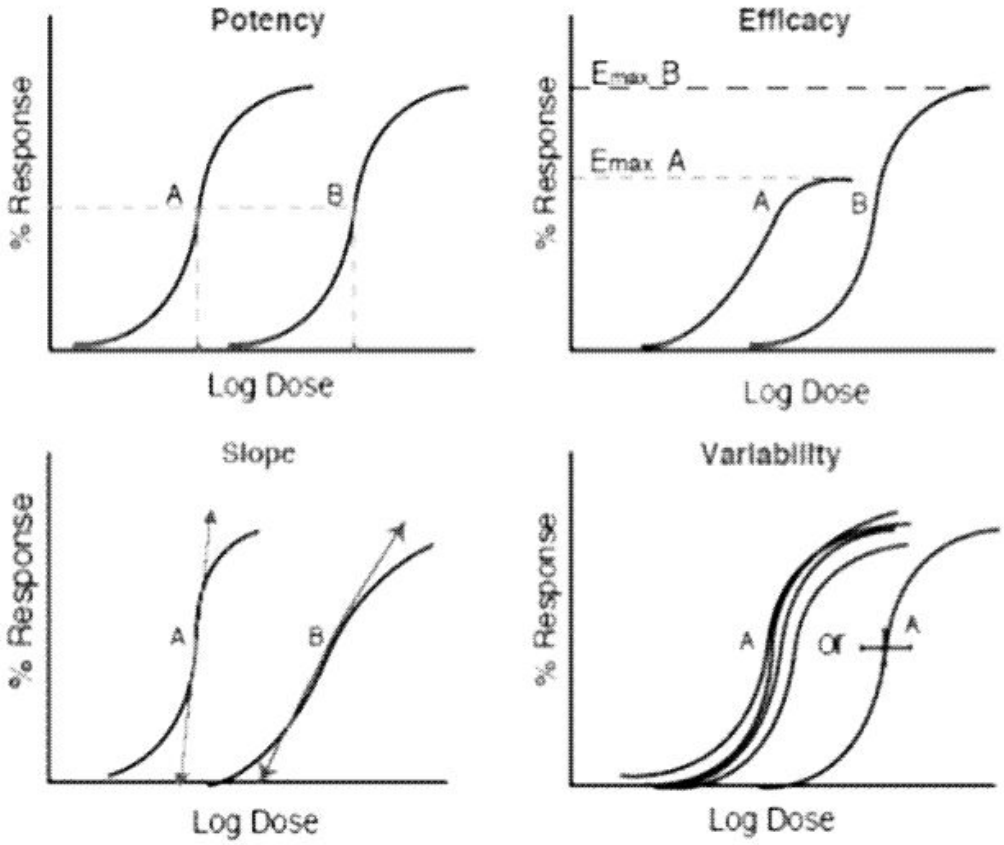


Рис. 1.3. Сила воздействия (Potency), максимальный эффект (Efficacy), угол наклона (Slope) и изменчивость (Variability)

Сила воздействия (или мощность) обычно связывается с изоэффективными концентрациями (EC_{50}) или дозами (ED_{50}) ксенобиотика. Сравнивать токсическую мощность двух ядов можно, если их кривые "доза-эффект" в некотором смысле параллельны. Иногда даже введение больших доз токсиканта может не привести к развитию полного эффекта (например, 100% смерти), поэтому важно оценить максимальное воздействие, оказываемое веществами. Скорость развития эффекта при увеличении дозы определяется углом наклона прямолинейного участка кривой. Изменчивость связана с воспроизводимостью результатов, полученных в различных условиях эксперимента (например, с различными группами подопытных животных). Обоснованные выводы в отношении перечисленных свойств токсического процесса могут быть сделаны в ходе статистического анализа параметров построенных регрессионных моделей.

1.4. Учет комбинированного действия токсикантов в смесях

Современные токсикокинетика и токсикодинамика сравнительно легко дают возможность понять, смоделировать и предсказать эффект воздействия одного изолированного ядовитого вещества. Однако в реальном мире живые организмы сталкиваются с многокомпонентными

смесями самого различного состава и оценить их воздействие на отдельные особи и экосистемы – одна из самых важных проблем экотоксикологии.

В середине XX-го столетия сформулированы два основных принципа эффектов смеси, основанные на понятиях *аддитивности* и *взаимодействия* (Altenburger et al., 2013; Ritz, Streibig, 2014). Аддитивность эффекта выражается в двух типах: суммирование концентраций (CA) и суммирование отклика (IA independent action).

Схема CA принимается для смесей веществ с одним и тем же механизмом токсического действия, например, смесь однотипных гербицидов, отличающихся только по ядовитому потенциалу. Эффект такой смеси может быть получен как сумма концентраций веществ, масштабированная по силе их токсического действия:

$$\sum_m \frac{x_i}{EC_i} = \frac{x_{mix}}{EC_{mix}} \qquad \text{или} \qquad EC_{mix} = (\sum_m \frac{d_i}{EC_i})^{-1}$$

где $d_i = x_i/x_{mix}$ – доля концентрации i-го компонента в смеси, EC – токсикометрические показатели, приводящие к заданному эффекту (например, LC_{50}).

Для схемы IA характерно, что механизмы интоксикации различными веществами отличаются между собой, и они воздействуют на структуры или функции организма независимо друг от друга. Эффект смеси может быть получен как сочетание индивидуальных эффектов от действия всех ингредиентов в смеси, но каждая из составляющих эффекта не может превышать максимально возможный эффект для каждого индивидуального вещества, т.е.

$$E(x_{mix}) = 1 - \prod_m (1 - E(x_i)),$$

где $E(x_{mix})$ и $E(x_i)$ – прогноз эффекта (на шкале от 0 до 1) для смеси с общей концентрацией x_{mix} и величина эффекта для i-го компонента с концентрацией x_i при его индивидуальном воздействии.

В отношении схемы IA независимого (или, в отечественной литературе, разнонаправленного) действия высказываются достаточно обоснованные сомнения как с позиций понимания организма как единого целого, так и в более узком токсикологическом смысле (Курляндский, Филов, 2002, с. 499). Поэтому ниже мы будем говорить только об аддитивности доз.

Взаимодействие токсичных компонентов в организме может привести к отклонениям от условий аддитивности, потому что биологическая активность одного вещества может зависеть от присутствия другого. При этом, по сравнению с аддитивными принципами, может происходить как увеличение (синергизм, потенцирование), так и уменьшение (антагонизм) общего отрицательного воздействия.

Классическим графическим методом обнаружения и характеристики отклонений от аддитивности для различных парных комбинаций токсикантов в смеси является изобологрaмма (Fraser, 1870). Пусть δ_1 и δ_2 – изоэффективные дозы (например, LC_{50}) двух веществ при индивидуальном воздействии. Тогда изобола показывает (см. рис. 1.4а) как изменяется соотношение концентраций x_1 и x_2 при условии достижения эффекта при воздействии смеси этих компонентов, идентичного δ. Если точки на осях, соответствующих величинам δ_1 и δ_2 соединить прямой линией, то речь идет о полной аддитивности компонентов на всем диапазоне их изменчивости, или $\frac{x_1}{\delta_1}+\frac{x_2}{\delta_2}=1$.

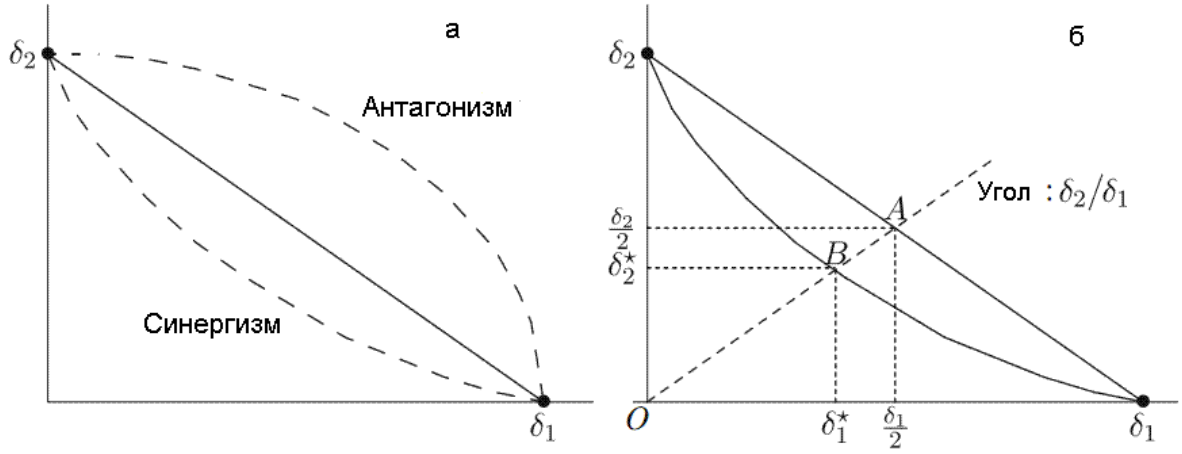


Рис. 1.4. Вид изоболических кривых при отклонении от аддитивности (а), интерпретация изоэффективных доз при аддитивном воздействии и при синергизме – подробности в тексте (б)

Если кривая, построенная по экспериментальным данным, будет проходить выше прямой аддитивности, то она является отражением процесса антагонистического взаимодействия, а кривая, целиком расположенная ниже ее – синергизма токсикантов. Для аппроксимации изобол Хьюлит (Hewlett, 1969) предложил модель $\left(\frac{x_1}{\delta_1}\right)^{1/\lambda}+\left(\frac{x_2}{\delta_2}\right)^{1/\lambda}=1$, где параметр взаимодействия λ равен 1 при аддитивности, $\lambda > 1$ – при синергизме и $\lambda < 1$ при антагонизме.

Смысл параметра λ показан на рис. 1.4 при соотношении концентраций компонентов 50:50 %. После преобразований получаем изменение величины изоэффективных концентраций $\delta_1^* = 2^{1-\lambda}\delta_1, \delta_2^* = 2^{1-\lambda}\delta_2$, или для смеси с равным соотношением компонентов $\frac{\delta_1^*}{\delta_1}+\frac{\delta_2^*}{\delta_2}=2^{1-\lambda}$. Тогда, например, при λ = 1.4 получим $2^{1-\lambda}$ = 0.76, т.е. тестовый токсический

эффект наблюдается при уменьшении воздействующей дозы обоих компонентов на 24% (Soerensen et al., 2007).

Индекс взаимодействия *II* (interaction index – Berenbaum, 1981) является расширением выражения для изобологграммы на *m*-компонентные смеси ($m > 2$) и также характеризует отклонения от аддитивности для комбинации из 3 и более компонентов. Однако, ввиду сложности оценки механизмов взаимодействия токсикантов и других комбинаторных факторов, на практике почти исключительно делается заключение о парных схемах суммирования.

2. ОЦЕНКА ПАРАМЕТРОВ ТОКСИКОМЕТРИИ ДЛЯ ЭФФЕКТОВ В ДИСКРЕТНОЙ ФОРМЕ

2.1. Учет эффекта в альтернативной форме

В общем случае количественный учет наблюдаемых реакций организма или экосистемы осуществляется в двух возможных основных формах числовых шкал: *метрической* (непрерывной или "градированной") и *дискретной* (квантифицированной, quantal). Дискретные шкалы, являющиеся предметом рассмотрения настоящей главы, объединяются по общности законов статистического распределения наблюдаемых случайных величин и, в свою очередь делятся на *альтернативные* (бинарная шкала), *номинальные* (шкала наименований) и *порядковые* (шкала баллов). Для альтернативных и номинальных шкал из всех возможных математических свойств величин используется лишь одно: отдельные измерения нечисловой природы могут отличаться между собой. Для наблюдений в порядковой шкале также не применимы математические операции типа сложения или деления, но эти числа могут строиться в вариационный ряд (от наименьшего – к наибольшему).

Учет многих реакций организма возможен лишь в альтернативной форме, т.е. когда исследователь для каждой особи может отметить лишь наличие или отсутствие того или иного проявления токсического процесса. Другой источник формирования альтернатив видится в следующем (Безель и др., 1994). В медицинской и экологической токсикологии обычно широко применяется анализ различных количественных показателей (физиологических, функциональных, биохимических и др.), регистрируемых при разных уровнях воздействия. Часто эти показатели сами по себе еще не определяют однозначно статус организма, т.е. не позволяют сделать итоговое заключение о наличии поражения. И здесь можно отметить важнейший в токсикологии факт перехода количественных изменений к новому качественному состоянию при достижении некоторого критического уровня воздействия EC_{crit}.

Если известны значения количественных показателей, характеризующихся как *норма M* (см. например, Трахтенберг и др., 1991), то за ее допустимые пределы в простейшем случае можно принять интервал, равный двум стандартным отклонениям, т.е. $EC_{crit} = (M \pm 2S_m)$. Тогда переход к зависимости "доза-эффект" на основе альтернатив (т.е. выход или нет за пределы критического диапазона) по существу подразумевает диагностику состояния "норма-патология", основанную на концепции пороговости поражающего действия токсиканта. Такая процедура дискретизации отклика по некоторым свидетельствам может иметь ряд преимуществ: *а*) к анализу дополнительно привлекаются ранее

установленные или литературные мета-данные, *б*) зависимость отражает реакцию организма на более высоком информационном, основанном на критических точках, *в*) появляется возможность оценивать отклик сразу по нескольким количественным показателям и *г*) модели для альтернативного отклика часто статистически более устойчивы.

Все альтернативные эффекты по своей природе характеризуются пороговыми зависимостями, которые описывают скачкообразный переход биологических систем из одного состояния (норма) в другое (патология или гибель). Теоретически – это прямая линия с изломом, перпендикулярным оси X (см. рис. 2.1): если мысленно предположить, что интоксикации подвергается группа совершенно идентичных объектов в идеально одинаковых условиях, то регистрируемый эффект для всех особей произойдет при введении одной и той же дозы.

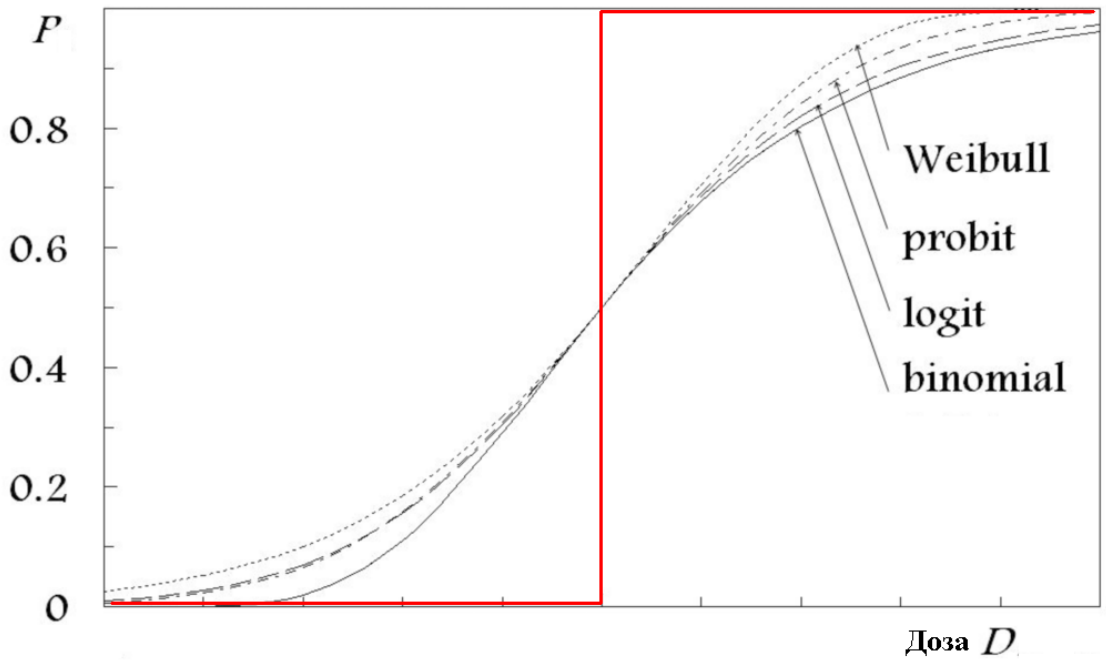


Рис. 2.1. Различные модели доза-эффект в альтернативной форме; прямая линия со скачком соответствует теоретическому развитию процесса

Разумеется, в реальных условиях наблюдается внутривыборочная изменчивость доз проявления эффекта, обусловленная ошибкой эксперимента. Ломаная линия трансформируется в плавную S-образную кривую с переходной фазой в области перегиба, ширина которой определяется главным образом индивидуальной чувствительностью тест-объектов. В основе индивидуальной чувствительности могут лежать многие факторы: генетическая гетерогенность, вариабельность токсикокинетики и токсикодинамики, зависимость между концентрацией агента в месте приложения и особенностями его фиксации биосубстратом и др. (Криштопенко и др., 2008).

В общем виде ошибка токсикометрического эксперимента включает две составляющие, первая из которых является систематической и обусловлена внешними факторами приготовления и введения в организм агента строго определенного количества однородного вещества. Вторая составляющая определяется различной реакцией отдельных организмов на одну и ту же введенную дозу, т.е. вышеупомянутой индивидуальной чувствительностью.

Безусловно, необходимо применять все необходимые меры к снижению систематической ошибки, но думается, что известное требование к токсикометрическому эксперименту по подбору строго однородных групп подопытных животных является несколько надуманным и методически сомнительным. Формально, снижая таким образом потенциальную статистическую погрешность эксперимента, делается другая серьезная ошибка – вычисленным токсикометрическим нормативам придается общепопуляционный характер. Действительно, если равные дозы исследуемого токсиканта вводить выборке однородных тест-объектов, то частота эффекта будет отражать лишь частный случай проявления индивидуальной чувствительности, например, у белых крыс-самцов одного возраста и массой 200 ± 15 г. При включении в выборку беременных самок, престарелых или ювенильных особей (о которых также следует позаботиться) мы получим более объективную картину и, возможно, ощутимый сдвиг величины нормируемого показателя. Таким образом, половозрастное распределение тест-объектов в выборке должно соответствовать половозрастной структуре популяции. Далее будет показано, что вообще группировать объекты нет никакой необходимости.

Если альтернативный показатель выступает в качестве отклика функции доза-эффект, то статистический анализ сводится, по сути, к восстановлению по эмпирическим данным вероятности P проявления эффекта в зависимости от уровня воздействия x. В течение всего XX-го века отрабатывались и уточнялись методики оценки такой зависимости, основанные на двух направлениях: построении сглаженной кривой накопления частот (методы Беренса, Рида-Менча, Кербера, Першина, Беренса-Шлоссер) и анализе одномерной линейной регрессионной модели, чаще всего, с использованием пробит-трансформации (методы Миллера и Тейнтер, Литчфилда-Вилкоксона, Финни, Прозоровского, Фрумина и др.). Большинство из них было основано на применении миллиметровой бумаги и принципов "экономных и быстрых расчетов", что потеряло всякую актуальность в результате компьютерной революции. В XXI веке после широкого распространения идеологии обобщенных регрессионных моделей, использующих принцип максимального правдоподобия, перечисленные методы представляют, вероятно, исторический интерес. Однако знакомство с ними, например, по монографиям (Беленький, 1969; Криштопенко и др., 2008) весьма полезно для понимания сути дела.

2.2. Обобщенные линейные модели "доза-эффект"

Сформулируем условия и способ построения линейной модели для альтернативного отклика, которая позволила бы количественно оценить степень влияния независимого предиктора x и/или "объяснить" внутренние механизмы моделируемого процесса. Возникает вопрос, можем ли мы использовать для этого обычную модель регрессии, основанную на методе наименьших квадратов с коэффициентами β_0 и β_1:

$$P = p(Y = 1) = \beta_0 + \beta_1 x + \varepsilon \ ?$$

Как отмечалось выше, кривая зависимости $P = f(x)$ обычно носит ярко выраженный нелинейный S-образный характер, асимптотически приближаясь к значениям $P_0 = 0$ и $P_1 = 1$, что напоминает интегральную функцию нормального распределения. При этом обычно делается предположение, что доля альтернатив $p(Y = 1)$ является случайной величиной, подчиненной биномиальному закону с параметрами (n, π). Очевидно, что при подгонке (fitting) такой модели методом наименьших квадратов (МНК) нарушаются основные предпосылки линейной регрессии и возникают следующие проблемы, связанные с особенностями биномиального распределения отклика:

1. Проблема наличия *гетероскедастичности* (т.е. неоднородности дисперсий): поскольку для бинарной переменной $\sigma^2 = pq$, то итоговую дисперсию вероятности для модели с одним предиктором можно представить как $\sigma^2_{\varepsilon_i} = (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i)$, т.е. с изменением X меняется и $\sigma^2_{\varepsilon_i}$;

2. Проблема *ненормального распределения ошибок*: остатки модели равны $\hat{\varepsilon}_i = 1 - \beta_0 - \beta_1 X_i$ при $Y_i = 1$ и $\hat{\varepsilon}_i = -\beta_0 - \beta_1 X_i$ при $Y_i = 0$, т.е. вариация $\hat{\varepsilon}$ является ограниченной и ненормально распределенной, а значит стандартная ошибка регрессии и t-статистика оценены неверно.

3. Проблема правильной спецификации модели: предсказываемые вероятности должны "уместиться" в интервале [0, 1], а обычная регрессионная модель может вполне предсказывать значения далеко за пределами этого интервала.

Обобщенные модели (McCullagh, Nelder, 1989) расширяют класс общих линейных и нелинейных моделей регрессии, связывая зависимую переменную с факторами и ковариатами посредством задаваемой *функции связи* (link function), причем допускается наличие у отклика произвольного распределения, отличающегося от нормального. При этом охватываются широко используемые статистические модели, такие как линейная регрессия для откликов с нормальным распределением, логистические модели для двоичных данных, лог-линейные модели для счетных данных, модели с дополняющим двойным логарифмированием для интервал-цензурированных данных выживания и многие другие статистические модели.

Обобщенная линейная модель GLM (Generalized Linear Model) имеет следующий вид:

$$y = g^{-1}\left(\sum_{i=1}^{n} \beta_i x_i\right),$$

где y и x_i – отклик и независимые переменные, β_i – коэффициенты регрессии, идентичные классической линейной модели, $g(y^{-1})$ – произвольная функция связи, преобразующая результат вычисления левой части уравнения в прогнозируемое значение отклика y.

С использованием такого подхода логика преодоления вышеперечисленных ограничений состоит в том, чтобы трансформировать значения частот (долей) P в некоторые непрерывные количественные величины Z, принимающие значения в диапазоне (- ∞, + ∞), а затем полученные значения регрессии преобразовать обратно в вероятности.

Одно из самых распространенных предположений (Справочник … , 1989, т.1, с. 337), хорошо подтверждаемых экспериментальными данными, состоит в том, что при воздействии дозы D вероятность гибели животного, случайно отобранного из всей совокупности, равна:

$$P(x) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{x} \{e^{-0.5(y-\mu)^2/\sigma^2}\} dy = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

где $x = \ln(D)$, $\Phi(\ldots)$ – интегральная функция стандартного нормального распределения $N(\mu, \sigma)$.

Если принять $\beta_0 = -\mu / \sigma$ и $\beta_1 = 1 / \sigma$, то получим выражение для обобщенной линейной модели:

$P = \Phi(\beta_0 + \beta_1 x + \varepsilon)$, или $\Phi^{-1}(P) = (\beta_0 + \beta_1 x + \varepsilon)$.

Таким образом, с использованием обратной функции связи $\Phi^{-1}(\ldots)$ величина $\hat{P}(x)$ определяется единственным образом по вычисленному значению $(\hat{\beta}_0 + \hat{\beta}_1 x)$.

Значение квантиля $q = \Phi^{-1}(P)$, используемое в моделях в качестве аргумента функции связи, называют *пробитом* (буквально *probit* происходит от *probability unit* или вероятностная единица). Пробиты для наиболее характерных вероятностей имеют следующие значения:

Вероятность P	0	0.05	0.25	0.5	0.75	0.95	1
Пробит $\Phi^{-1}(P)$	-∞	-1.64	-0.67	0	1.64	0.67	∞

Во времена, когда графики строили на миллиметровой бумаге, во избежание отрицательных значений к значению квантиля добавляли число 5 и *пробитом* называли полученный результат. Сейчас эта операция считается излишней.

Другая возможность преобразования альтернативного отклика в непрерывную шкалу заключается в расчете так называемого "отношения шансов" ("*odds ratio*"). Если событие происходит с вероятностью π, то шанс представляет собой отношение $O(\pi) = \pi/(1 - \pi)$. К сожалению,

$O(\pi) = 0$ при $\pi = 0$ и $O(\pi) = +\infty$ при $\pi = 1$, т.е. подобного рода трансформация является неполной, не охватывая отрицательные значения отклика. Поэтому функцию связи обычно задают в виде *логита* (logit link function) или логарифма отношения шансов: $g(y) = \log(\frac{\pi}{1-\pi})$.

Обобщенная логит-линейная модель, или логистическая регрессия на m предикторов, имеет вид

$$g(y) = \log \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \sum_{k=1}^{m} \beta_k x_k \ .$$

В целях интерпретации коэффициентов часто предпочтительнее получать не логит-значения, а непосредственно предсказанные вероятности π для каждой комбинации независимых переменных. Например, в случае с одним предиктором логит можно преобразовать в вероятность следующим образом:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X \quad \Rightarrow \quad P = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \ .$$

По сравнению с моделью пробита, логистическое распределение не столь круто дрейфует в область очень низких и очень высоких доз, т.е. имеет более пологую форму, однако в средней части графика (в области LD_{50}) различия в форме кривых практически отсутствуют (см. рис. 2.1).

На большом фактическом материале показана необходимость предварительного логарифмирования концентраций токсических компонентов $x = \ln(D)$, т.е. мы на самом деле пользуемся лог-пробит- или лог-логит-распределениями. Интересное теоретическое обоснование этому дает В. Отт (цит. по Безель и др., 1994), рассматривая простую модель, имитирующую поступление загрязнителей в биологические системы, представляющие n последовательно соединенных камер. Легко показать, что распространение токсиканта по камерам подчиняется логнормальному закону. В то же время, если речь идет не о химических загрязнениях, то предварительное логарифмирование далеко не обязательно, и его необходимость должен подтвердить статистический анализ.

Логит- и пробит-распределения являются симметричными относительно центральной точки ($p = 0.5$, $\ln x = LD_{50}$). К настоящему времени многими авторами замечено, что кривая доза-эффект часто не может быть аппроксимирована этими моделями по причине асимметрии ее левой и правой ветвей. Это может быть обусловлено сложностью механизмов развития и проявления результативного признака (эффекта), но гораздо важнее то обстоятельство, что аппроксимация несимметричными функциями позволяет точнее отследить зависимость в области низких уровней воздействия (в районе LD_5), что, собственно, и является основной задачей гигиенического нормирования.

Для описания времени работы объектов на отказ традиционно используется несимметричное распределение Вейбулла (Weibull, иногда Вейбулла-Гнеденко), функция распределения которого имеет вид:

$$P = 1 - e^{-(x/\lambda)^m}, \quad x = \left[\frac{1}{\lambda}\ln\left(\frac{1}{1-P}\right)\right]^{1/m},$$

где λ – коэффициент масштаба, m – коэффициент формы.

Выражая свое критическое отношения к повсеместному использованию линеаризованных форм пробит-анализа, Криштопенко с соавторами (2008, с. 10) пишут: «Весьма прискорбно осознавать, что за долгие годы многие тысячи экспериментально найденных зависимостей "доза- эффект" были выстроены "в прямую линию" с явным искажением истинных проявлений фармакокинетики и фармакодинамики уникальных веществ и препаратов, а вычисленные среднеэффективные дозы обладали совершенно непредсказуемой степенью неопределенности.». Однако, используя современные компьютеры и надежные статистические программы, построить любую, достаточно сложную модель технически не составляет труда.

Гораздо большую сложность составляет осуществление селекции оптимальной модели, обеспечивающей минимальную ошибку аппроксимации эмпирических данных, экономной по числу параметров и непротиворечивой относительно накопленного опыта в данной предметной области. Например, в работе (Scholze et al., 2001) представлены результаты подробной сравнительной оценки пула из 10 различных сигмоидальных функций регрессии для нахождения пороговых параметров токсичности в области низких концентраций. Показано, что при тестировании различных токсикантов и их смесей лучшими моделями могут оказаться примерно в равной мере все три функции (пробит, обобщенный логит и Вейбулла), но проведение трансформации Бокса-Кокса для исходных данных часто улучшает критерии качества найденных зависимостей.

2.3. Использование функции `glm()` статистической среды R

Стандартным средством построения обобщенной линейной модели в среде R является функция `glm()`, имеющая вид:

```
glm(formula, data = data.frame,
      family = family.generator),
```

(Венэбльз У.Н., Смит, 2014; Кабаков, 2014; Мастицкий, Шитиков, 2015). Здесь оператор `formula` задает формулу связи между независимыми (предикторными) переменными модели и откликом, `data` – имя таблицы с исходными данными, а параметр `family` служит инструментом для

описания типа модели и вызывает функцию, которая генерирует заданное распределение отклика. Генератор `family.generator` имеет вид:

<наименование распределения>

`(link =` *<наименование функции связи>*`).`

Если принять предположение о биномиальном распределении данных, то параметр `family` будет иметь вид:

`family = binomial(link = "probit"))` для модели пробита;

`family = binomial(link = "logit"))` для модели логита.

Рассмотрим построение обобщенной линейной модели на следующем примере (исследования лаборатории герпетологии и токсинологии ИЭВБ РАН). Группам мышей одинаковой массы (20 ± 1.0 грамм) подкожно вводили возрастающие дозы ядов гадюк из разных местообитаний и через сутки подсчитывали число погибших и выживших животных. Группы формировались отдельно из мышей-самцов и мышей-самок. Поставим задачу построить зависимости "доза-эффект", оценить величину средней полулетальной дозы LD_{50} и проверить гипотезы: *а*) об отсутствии различий в токсичности яда гадюк из разных регионов `местооб` и *б*) об одинаковой чувствительности самцов и самок к действию яда.

```
# Загружаем данные из файла (можно скачать с сайта пособия)
df <- read.table(file = "Гадюки.txt",
                 header = TRUE, sep = "\t")
head(df)
# Список наименований регионов, где велся отлов змей
levels(df$местооб)
 dead <- cbind(df$погибло, df$выжило)
 эффект = df$погибло/(df$погибло+df$выжило)
```

```
   пол местооб доза численность погибло выжило
1    M    Перм 0.15           3      0      3
2    M    Перм 0.30           3      0      3
3    M    Перм 0.75           3      1      2
14   F    Перм 1.48           3      1      2
15   F    Перм 2.23           3      3      0
16   F    Перм 2.97           3      3      0
Список уровней фактора местооб
[1] "Моск"  "НжНов" "Новг"  "Перм"  "Самар" "Татар" "Хвал"
```

Необходимо отметить, что при построении современных моделей GLM исходные данные интерпретируются как результаты независимых единичных испытаний с соблюдением того принципа, что каждый подопытный объект рассматривается как самостоятельная единица статистического анализа. Однако для удобства пользователя данные могут быть представлены в двух возможных формах :

o "сырые" данные без предварительного расчета долей, т.е. в каждой строке записывается информация об одном объекте (особь, испытанная доза и наблюдавшийся при этом эффект, выраженный в альтернативной форме 0/1);

o в случае формирования однородных по дозам групп бинарный отклик задается парой столбцов матрицы исходных данных: числом выживших и погибших животных или, как вариант, долей и общим числом особей в группе.

Все эти формы представления исходных данных дают совершенно идентичные результаты (функция `glm()` сама преобразует сгруппированные данные в "длинный формат").

Рассмотрим "каноническую" модель пробита на основе прологарифмированных значений доз яда с использованием всего имеющегося материала:

```
Mpl <- glm(dead ~ log(доза),
           family = binomial(link = "probit"), data = df)
summary(Mpl)
```

```
Коэффициенты:
              Оценка  Ст.ошибка z-крит. Pr(>|z|)
(Св.член)    -0.8443     0.1229   -6.87 6.42e-12 ***
log(доза)     2.3773     0.2033   11.70  < 2e-16 ***
---
Коды значимости:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Null-девианс: 335.48  для 124  степеней свободы
Девианс остатков: 105.93  для 123  степеней свободы
AIC-критерий: 184.54
Число итераций приближений Фишера: 5
```

Прежде всего отметим, что оценка параметров модели `Mpl` выполняется универсальным методом максимального правдоподобия (MLE, maximum likelihood estimation). При этом в качестве "наиболее правдоподобного" значения параметров $\boldsymbol{\Theta}$ ищутся такие значения, при которых максимально вероятно получить при n опытах исходную выборку $X = (x_1, \ldots, x_n)$. Вид оптимизируемых функций правдоподобия $L(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{y})$, как правило, достаточно сложен и здесь рассматриваться не будет. С вычислительной точки зрения удобнее искать максимум логарифмической функции правдоподобия $\ln L$ (*log-likelihood*), т.е.

$$LL = \log\{L(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{y})\} \to \max.$$

Для формальной оценки адекватности моделей GLM, а также для сравнения качества нескольких построенных моделей, обычно используется такая статистика, как *девианс* (deviance), непосредственно вытекающая из оценок максимального правдоподобия. Девианс G^2 для анализируемой модели M определен как

$$G^2 = -2\,(LL_M - LL_S\,),$$

где LL_S – максимум логарифма функции правдоподобия для полной, или "насыщенной" (saturated), модели S, которая содержит так много параметров θ_S, чтобы по возможности точно восстановить значения y выборочных данных. Величина девианса G^2 по своему смыслу идентична сумме квадратов остатков классической модели регрессии.

Информационный критерий Акаике `AIC` (Akaike information criterion) в явном виде включает “штраф” за увеличение сложности модели

$$\mathrm{AIC} = -2\log\{L(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{y})\} + 2k,$$

где k – число оцениваемых параметров.

На примере вышеприведенных расчетов получена однофакторная линейная пробит-модель вида с обоими значащими коэффициентами

$$probit(P) = -0.84 + 2.38\ln(D)$$

Протокол результатов, выводимый функцией `summary()` включает:

- `Null deviance` – девианс "пустой" модели, не включающей ни одного параметра, кроме β_0;
- `Residual deviance` – остаточный девианс, который косвенно соответствует дисперсии в данных, оставшейся необъясненной после включения в модель предиктора `доза`.
- значение информационного критерия `AIC` и число итераций, выполненных перед схождением алгоритма, по которому вычисляются параметры модели (их число не должно быть слишком велико).

Зададимся вопросом, статистически значима ли в целом построенная модель. Проведем тест по критерию χ^2 на отличие от нуля разности девианса полученной пробит-модели от девианса нуль-модели без предикторов:

```
anova(Mpl,glm(dead ~ 1, family = binomial("probit"),
          data = df), test = "Chisq")
```

```
Анализ девианс-таблицы
Модель 1: dead ~ log(доза)
Модель 2: dead ~ 1
   Остат.Df  Остат.Дев Df  Девианс P(>|Chi|)
1       123     105.93
2       124     335.48 -1  -229.55 < 2.2e-16 ***
```

Разность между девиансами моделей 1 и 2 весьма велика, что определяет вывод о высокой статистической значимости полученной регрессионной зависимости.

Другой вопрос состоит в том, насколько оправданно мы провели логарифмирование доз, а также, не лучше ли было использовать логит в качестве функции связи. Сравним между собой четыре возможных модели с комбинациями этих условий и результат оценим по величине AIC-критерия (лучшая модель соответствует его минимуму). Попутно

рассмотрим, как меняется при этом такой показатель токсикометрии, как среднеэффективная доза LD_{50} , для чего используем функцию `dose.p()` из пакета `MASS`. Выведем график зависимости "доза-эффект" для двух лучших моделей.

```
library(MASS)
# Формируем список из четырех моделей пробита и логита
ModList <-list(Mpn <- glm(dead ~ доза,
          family = binomial(link = "probit"), data = df),
Mpl, Mln <- glm(dead ~ доза,
          family = binomial(link = "logit"), data = df),
Mll <- glm(dead ~ log(доза),
          family = binomial(link = "logit"), data = df))
# Для каждой модели рассчитываем AIC и LD50
AICmod <- sapply(ModList, function(i) AIC(i))
names(AICmod)<- c("Проб/доз","Проб/lnд","Лог/доз","Лог/lnд")
DL50 <- sapply(ModList, function(i) dose.p(i))
print (c(cat("AIC-критерий и значения LD50\n"), AICmod))
DL50l <- DL50[2]
DL50[c(2,4)]=exp(DL50[c(2,4)]) ; DL50
# Подготовка данных для графиков
df.plot = data.frame(df, эффект, доза_ln= log(df$доза),
      p1 = Mpl$fit, p2 = Mll$fit,
      пробит = predict(Mpl, type = "link"))
df.plot = df.plot[order(df.plot$доза_ln),]
# Прорисовка компонент графика
plot(df.plot$доза_ln, df.plot$p1, type="l", lwd = 2,
        xlab="log(Доза)", ylab="Доля умерших мышей", col=3)
points(df.plot$доза_ln, df.plot$эффект, cex = 0.7,
     pch = 23 + as.numeric(df.plot$пол),
     bg = 1+as.numeric(df.plot$пол))
lines(df.plot$доза_ln, df.plot$p2, lwd = 2, lty=2)
segments(min(df.plot$доза_ln),0.5, DL50l,0.5, col=4)
segments(DL50l,0.5, DL50l,0, lty=3)
text(0.55,0.52,"LD50")
legend("bottomright", c("Самки","Самцы","Пробит","Логит"),
     pch = c(24,25,NA,NA), pt.bg = 2:3,
     lwd = 2, col=c(NA,NA, 4,3))
```

```
AIC-критерий и значения LD50
        Пробит/   Пробит/   Логит/   Логит/
        Доза      ln(доза)  Доза     ln(доза)
AIC     190.6090  184.5372  189.4327 185.3461
LD50    1.550653  1.426395  1.530903 1.430855
```

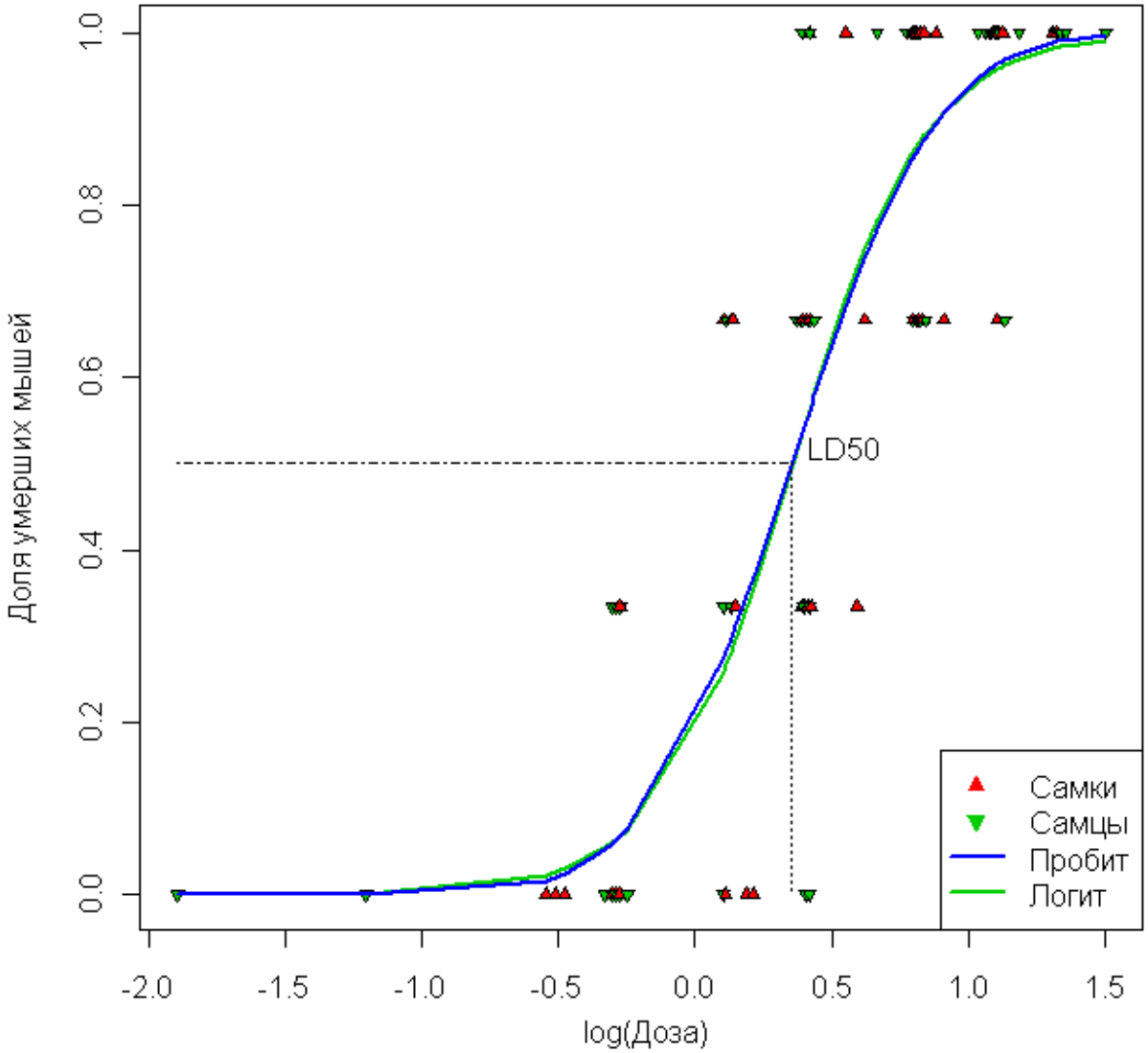


Рис. 2.2. Пробит- и логит-зависимости для яда гадюк

Оценка адекватности по AIC-критерию показывает, что переход к логарифмам дозы привел к объективно лучшим моделям, а использование логит-функции оказалось несколько хуже, чем пробит-трансформация.

Если проанализировать на рис. 2.2 распределение экспериментальных точек относительно модельной кривой, то визуально ошибка аппроксимации достаточно велика. На изменчивость эффекта может оказывать влияние как неоднородность тест-объектов (половая принадлежность мышей), так и вариация состава ядовитого секрета, обусловленная региональными популяциями гадюк.

Усложним модель и включим в качестве дополнительной переменной фактор "пол":

```
Mplfg <- glm(dead ~ log(доза)* пол,
             family = binomial(link = "probit"), data = df)
summary(Mplfg)
```

```
Коэффициенты:
                Оценка Ст.ошибка z-крит. Pr(>|z|)
(Св. член)     -0.88035    0.17693  -4.976 6.50e-07 ***
log(доза)       2.42306    0.29541   8.203 2.35e-16 ***
полМ            0.07049    0.24615   0.286    0.775
log(доза):полМ -0.08776    0.40726  -0.215    0.829
---
Коды значимости:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    Null-девианс: 335.48  для 124  степеней свободы
Девианс остатков: 105.85  для 121  степеней свободы
AIC: 188.45
```

В пробит-модели `Mplfg` свободный член `(Intercept)` и коэффициент при переменной `доза` корректируются в зависимости от пола животных, т.е. к ним добавляются значения 0.07 и -0.088 соответственно, но только при том условии, что выборка мышей состоит из самцов. Однако эти приращения столь малы, что статистически значимо не отличаются от 0 (*p*-значения, оцененные по *z*-критерию для этой нулевой гипотезы, равны 0.775 и 0.829 соответственно).

Построим теперь пробит-модель, учитывающую региональную изменчивость токсичности яда гадюк. Формулу модели `Mplfr` представим в несколько своеобразном виде, позволяющем получить набор коэффициентов индивидуальных регрессий отдельно для каждого региона `местооб`. Можно легко убедиться, что эта модель идентична модели, построенной по формуле `dead ~ log(доза)* местооб`.

```
Mplfr <- glm(dead ~ местооб/log(доза) - 1,
      family = binomial(link = "probit"), data = df)
summary(Mplfr)
anova(Mpl,Mplfr, test = "Chisq")
```

```
Коэффициенты:
                        Оценка Ст.ошибка z-крит. Pr(>|z|)
местообМоск             -1.5090    0.9029  -1.671 0.094655 .
местообНжНов            -1.7471    0.6243  -2.798 0.005135 **
местообНовг             -1.1055    0.3606  -3.066 0.002169 **
местообПерм             -1.1235    0.3711  -3.028 0.002464 **
местообСамар            -0.3826    0.2836  -1.349 0.177247
местообТатар            -2.8470    0.7995  -3.561 0.000369 ***
местообХвал             -0.3055    0.2323  -1.315 0.188488
местообМоск:log(доза)    5.4624    2.2958   2.379 0.017347 *
местообНжНов:log(доза)   3.6015    0.9345   3.854 0.000116 ***
местообНовг:log(доза)    2.7135    0.6434   4.218 2.47e-05 ***
местообПерм:log(доза)    2.1310    0.4802   4.438 9.09e-06 ***
местообСамар:log(доза)   2.8637    0.6844   4.184 2.86e-05 ***
местообТатар:log(доза)   4.3679    1.0704   4.080 4.49e-05 ***
местообХвал:log(доза)    1.6060    0.5166   3.109 0.001878 **
---
Коды значимости:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Null-девианс: 344.18  при 125  степенях свободы
Девианс остатков: 67.31 при 111  степенях
AIC: 169.92
```

Анализ девианс-таблицы

```
Модель 1: dead ~ log(доза)
Модель 2: dead ~ местооб/log(доза) - 1
   Остат.Df  Остат.Дев Df  Девианс P(>|Chi|)
1       123     105.93
2       111      67.31 12   38.622 0.0001215 ***
---
Коды значимости:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Таким способом в рамках регрессионного анализа одной и той же модели можно получить сразу семь зависимостей доза-эффект для семи популяций гадюки:

$probit(P) = -1.5 + 5.46 \ln(D)$ – Модель для московской популяции

$probit(P) = -1.74 + 3.6 \ln(D)$ – Модель для нижегородской популяции

…

$probit(P) = -0.31 + 1.6 \ln(D)$ – Модель для хвалынской популяции

Статистическое сравнение девиансов моделей `Mpl` и `Mplfr` с использованием функции `anova()` показывает, что включение регионального фактора `местооб` в аппроксимируемую зависимость доза-эффект является высоко значимым в смысле уменьшения ошибки модели по критерию χ^2 (p-значение равно 0.00012). Отметим также, что величина AIC-критерия после включения предиктора `местооб` уменьшилась с 184.5 для модели `Mpl` до 170.

Следовательно, можно сделать вывод: *яд гадюк, отловленных в разных регионах, имеет разную токсичность*. При этом форма представления коэффициентов модели `Mplfr` позволяет отдельно для каждого региона построить на графике (см. рис. 2.3) линии пробит-регрессии и рассчитать значения LD_{50}:

```
# Линии регрессии в координатах доза-пробит:
lf = length(levels(df$местооб))
plot(df.plot$доза_ln, df.plot$пробит, type = "l", lwd = 2,
xlab = "ln(Доза)", ylab = "Пробит доли умерших")
for (i in 1:lf) abline(coef(Mplfr)[i],
                coef(Mplfr)[lf+i],col = i+1)
segments(-2,0, 1,0, lty=6,lvd=2, col="grey")
text(1.2,0,"LD50")
legend("bottomright", c("Все", levels(df$местооб)),
      col = 1:8, lwd = c(2,rep(1,7)))
# Рассчитаем региональные величины DL50 для яда гадюки:
df.dl=data.frame(region =levels(df$местооб),
lnLD50 = rep(0,lf), LD50 = rep(0,lf), SE = rep(0,lf))
for (i in 1:lf) {
```

```
      a = dose.p(Mplfr, c(i,lf+i))
      df.dl[i,2] = a[1] ; df.dl[i,3] = exp(a[1])
      df.dl[i,4] = as.numeric(attr(a, "SE"))
          }
  df.dl
```

```
  region    lnLD50     LD50         SE
1   Моск 0.2762561 1.318185 0.07484909
2  НжНов 0.4850925 1.624325 0.08070838
3   Новг 0.4074207 1.502936 0.08605231
4   Перм 0.5271959 1.694175 0.10195969
5  Самар 0.1336115 1.142949 0.08761477
6  Татар 0.6518040 1.919000 0.06279256
7   Хвал 0.1901999 1.209491 0.12262101
```

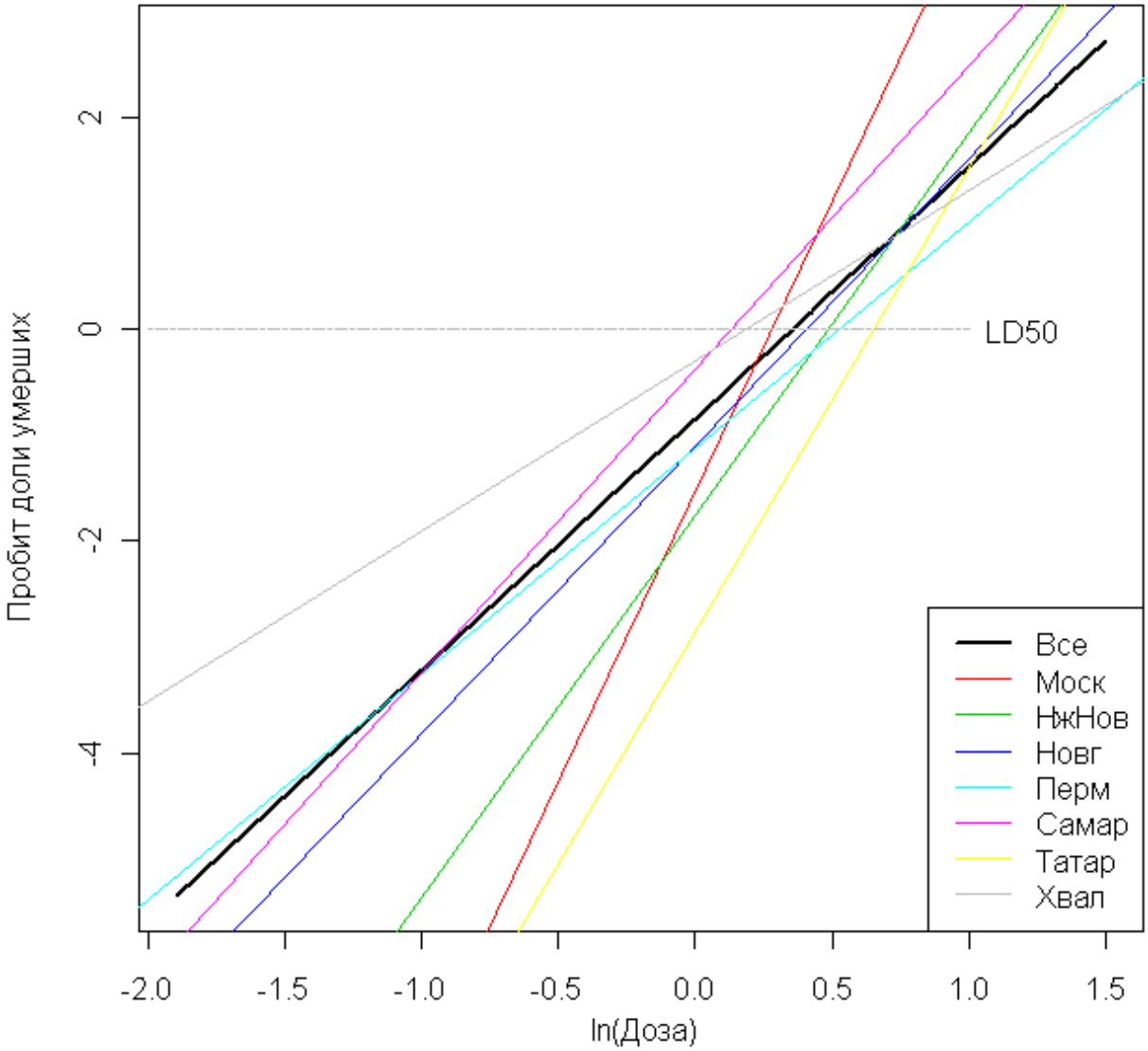


Рис. 2.3. Линейные зависимости пробит-эффекта гибели мышей от логарифма дозы для яда гадюк из различных регионов

Заметим, что LD_{50} – это точечный показатель, не дающий представления обо всем процессе развития эффекта при изменении дозы. Например, регионы Москва и Хвалынск имеют близкие LD_{50}, но они являются антиподами при сравнении зависимостей "доза-эффект" в целом. В частности, величина коэффициента угла наклона $\hat{\beta}_1$ (`местообМоск:доза` = 5.5 против `местообХвал:доза` = 1.6) определяет скорость нарастания эффекта, т.е. насколько круто S-образная кривая взмывает вверх в области ее перегиба.

2.4. Характеристики диагностического теста и ROC-анализ

По значению какого-нибудь важного количественного признака или их совокупности можно выполнить диагностику, а именно, оценить к какому из двух возможных классов следует отнести изучаемый объект (условно назовем эти классы "норма" или "патология"). Например, повышение температуры тела до 37.5^{o} чаще всего свидетельствует о заболевании, хотя не всегда болезнь может сопровождаться высокой температурой. Поскольку группы точек "патология/норма", в заданном пространстве, как правило, статистически неразделимы гиперплоскостью, то при тестировании вероятны ошибочные ситуации, такие как пропуск положительного (патологического) заключения или его гипердиагностика.

Результаты теста на некоторой контрольной выборке можно представить таблицей сопряженности:

Результаты теста	Состояние тест-объектов	
	Патология	Норма
Положительные	**a** (истинно-положительные)	**b** (ложно-положительные)
Отрицательные	**c** (ложно-отрицательные)	**d** (истинно-отрицательные)

Тогда объективная ценность рассматриваемого бинарного классификатора определяется следующими показателями:

- *Чувствительность* (sensitivity) Se = **a** / (**a** + **b**) , определяющая насколько хорош тест для выявления патологических экземпляров;
- *Специфичность* (specificity) Sp = **d** / (**c** + **d**) , показывающая эффективность теста для правильного исключения нормального состояния;
- *Точность* = (**a** + **d**) / (**a** + **b** +**c** + **d**) , определяющая общую вероятность теста давать правильные результаты.

Перспективным графоаналитическим методом оценки качества теста и интерпретации перечисленных показателей является *ROC-анализ* (Receiver Operator Characteristic – функциональная характеристика приемника), название которого взято из методологии оценки качества сигнала при радиолокации.

ROC-кривая получается следующим образом (Goddard, Hinberg, 1989). Пусть мы имеем выборку значений независимого количественного показателя, который варьирует от x_{min} до x_{max}, и сопряженного с ним альтернативного отклика (1 – патология, 0 – норма). Любое произвольное значение x на этом диапазоне может считаться классификационным *порогом*, или точкой отсечения {cutt-off value}, делящим вектор Y на два подмножества, и для этого разбиения можно рассчитать значения чувствительности *Se* и специфичности *Sp*. Если выполнить сканирование $x_{max} \geq x \geq x_{min}$, то можно построить график зависимости, где по оси Y откладывается *Se*, по оси X – (1 - *Sp*) – см. рис. 2.4.

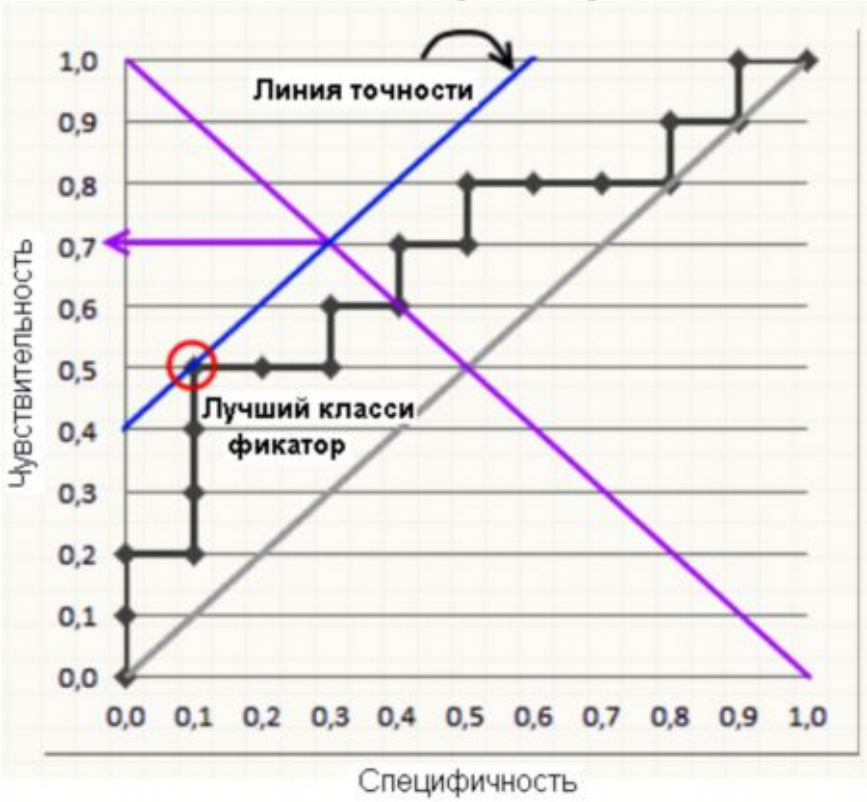


Рис. 2.4. ROC-кривая, линия максимальной точности и точка, соответствующая порогу лучшей классификации

В случае идеального классификатора график ROC-кривой проходит вблизи верхнего левого угла, где доля истинно-положительных случаев равна 1, а доля ложно-положительных примеров равна нулю. Поэтому, чем ближе кривая к верхнему левому углу, тем выше предсказательная способность модели. Наоборот, главная диагональная линия соответствует "бесполезному" классификатору, т. е. полной неразличимости двух классов:

близость ROC-кривой к диагонали говорит о низкой эффективности построенной модели. Для нахождения оптимального порога, соответствующего наиболее безошибочному классификатору, через крайнюю точку ROC-кривой проводят линию максимальной точности, параллельную главной диагонали.

Своеобразным методом сравнения ROC-кривых является численная оценка площади под кривыми AUC (Area Under Curve). Практически она изменяется от 0,5 ("бесполезный" классификатор) до 1,0 ("идеальная" модель). Показатель AUC предназначен исключительно для сравнительного анализа нескольких моделей, поэтому связывать его величину с прогностической силой можно только с большими допущениями.

Рассмотрим пример ROC-анализа с использованием экспериментальных данных по биотестированию техногенного загрязнения почвы на основе численности микрофауны: различных видов бактерий и микроскопических грибов (Шитиков и др., 2015). Образцы почвы были взяты с 15 площадок в районе урансодержащих отвалов вблизи оз. Иссык-Куль и с 7 условно-чистых площадок (альтернативный признак `Class`). Водные вытяжки образцов добавлялись в питательную среду при выращивании микроорганизмов. Получены данные для численностей трех групп микрофауны: Actinomycetes, Aspergilus и бактерий, выращенных на MPA-среде. Необходимо оценить, могут ли перечисленные группы являться эффективными биотестерами.

```
# Загружаем данные из файла (можно скачать с сайта пособия)
TTM <- read.table(file="bac_ROC.txt", sep="\t",
                  header=TRUE,row.names=1)
Class <-as.factor(TTM$Class)
head(TTM)
```

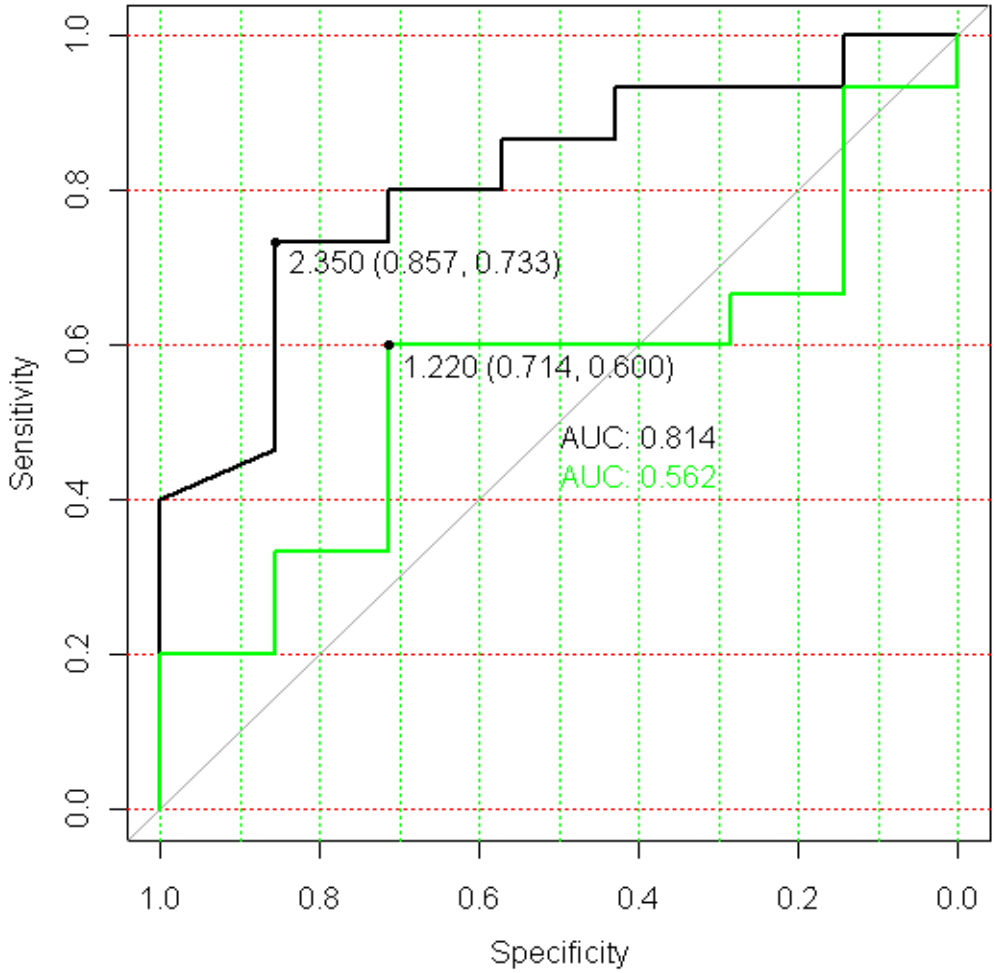
	MPA	Actinomycetes	Aspergilus	Class
0	0.87	1.28	2.10	1
1_1	3.76	1.36	0.00	1
1_2	0.57	2.87	2.10	1
2_1	0.83	0.46	0.00	1
3_1	2.37	1.80	0.00	1
3_2	0.72	2.08	0.00	1

Построим классификационные модели с использованием функции `roc(…)` пакета `pROC` на основе каждого из перечисленных показателей и сопоставим ROC-кривые для Actinomycetes и Aspergilus (кривая бактерий MPA проходит посередине между ними и пока интереса не вызывает):

```
library(pROC)
m1.roc <- roc(Class, TTM[,1])
m2.roc <- roc(Class, TTM[,2])
m3.roc <- roc(Class, TTM[,3])
```

```
roc.test(m2.roc,m3.roc)
plot(m3.roc, grid.col=c("green", "red"), grid=c(0.1, 0.2),
      print.auc=TRUE,print.thres=TRUE)
plot(m2.roc , add = T, col="green", print.auc=T,
      print.auc.y=0.45,print.thres=TRUE)
```

```
Бутстреп-тест на корреляцию двух ROC-кривых
данные:  m2.roc и m3.roc
D = -1.9027, boot.n = 2000, boot.stratified = 1,
p-значение = 0.05708
альтернативная гипотеза:
      действительная разность между AUC не равна 0
Выборочные оценки:
AUC по roc1 AUC по roc2
  0.5619048   0.8142857
```



2.5. ROC-кривые для Actinomycetes (зеленый цвет) и Aspergilus

Нетрудно заметить, что численность Aspergilus обладает существенно лучшими способностями к биотестированию: при пороговом значении T = 2.35 тыс. экз/100 г почвы чувствительность классификатора

составляет *Se* = 73 %, а специфичность *Sp* = 85.7. Бутстреп-тест на коррелированность обоих ROC-кривых (*p*-значение = 0.058) находится на пороге статистической значимости, а различия в площадях под кривой AUC на рис. 2.5 выглядят вполне внушительными.

Возникает закономерное предположение, а не будет ли классификатор более эффективным, если в прогнозировании использовать численности всех трех групп микроорганизмов в совокупности. Для этого построим модель логистической регрессии и постараемся оценить оптимальный состав ее предикторов.

```
#### логистическая регрессия для "полной" модели
(m.log <- glm(Class ~ . , data= TTM, # формула модели
            family=binomial(logit))) # вид модели
#### оптимизируем по AIC
mo.log <- step(m.log, trace = 0)
summary(mo.log)
#### сравниваем модели на предмет существенных отличий
anova(m.log, mo.log,test="Chi")
```

Полная модель

Коэффициенты:

```
    (Св. член)          MPA  Actinomycetes     Aspergilus
       4.8370       -0.7852       -0.1163        -0.6758
Null-девианс:       27.52
Девианс остатков: 15.14        AIC: 23.14
```

Оптимальная модель по AIC

```
Call:
glm(formula = Class ~ MPA + Aspergilus,
    family = binomial(logit), data = TTM)
```

Коэффициенты:

```
               Оценка  Ст.ошибка z value Pr(>|z|)
  (Св. член)    4.6778     1.8729   2.498   0.0125 *
MPA            -0.7892     0.4684  -1.685   0.0920 .
Aspergilus     -0.6629     0.3169  -2.092   0.0364 *
---
    Null-девианс: 27.522  on 21  degrees of freedom
Девианс остатков: 15.146  on 19  degrees of freedom
AIC: 21.146
```

Анализ девианс-таблицы

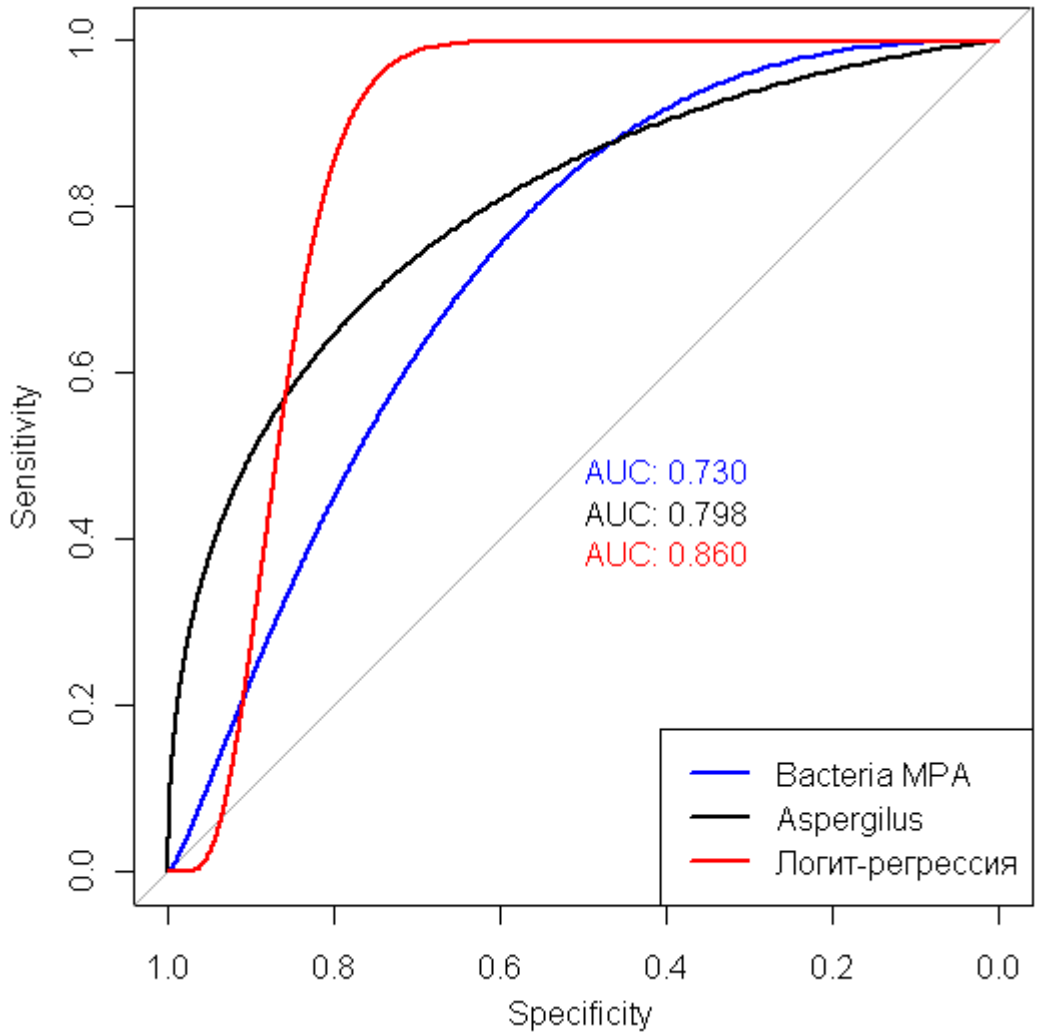
```
Модель 1: Class ~ MPA + Actinomycetes + Aspergilus
Модель 2: Class ~ MPA + Aspergilus
    Остат.Df  Остат.Дев Df  Девианс P(>|Chi|)
1         18     15.138
2         19     15.146 -1 -0.0088153    0.9252
```

Расчеты показывают, что исключение из модели переменной `Actinomycetes` статистически незначимо увеличивает ошибку модели и улучшает значение AIC-критерия. Используем предикторные значения оптимальной модели `mo.log` для классификации, построим ROC-кривую и сравним ее с кривыми, полученными для индивидуальных показателей – см. рис. 2.6.

В нижеприведенном скрипте подключение функции `smooth(…)` позволяет получить сглаженные кривые, а на основе функции `coord(…)` можно вывести полную информацию о точке оптимального классификатора. Использование функций `ci.auc()`, `ci.se()`, `ci.sp()`, `ci.thresholds()` даст нам доверительные интервалы для AUC, чувствительности, специфичности и классификационных порогов соответственно (мы показываем только одну из таких возможностей).

```
fit <- fitted(mo.log)
ms.roc <- roc(Class, fit)
coords(ms.roc, "best", ret=c("threshold", "specificity",
                "sensitivity", "accuracy"))
coords(m1.roc, "best", ret=c("threshold", "specificity",
                 "sensitivity", "accuracy"))
coords(m3.roc, "best", ret=c("threshold", "specificity",
                 "sensitivity", "accuracy"))
# Доверительные интервалы для параметров ROC-анализа:
ci.auc(m1.roc)
ci.thresholds(m1.roc)
plot(smooth(m1.roc), col="blue", print.auc=T)
plot(smooth(m3.roc) , add = T, col="black", print.auc=T,
         print.auc.y=0.45)
plot(smooth(ms.roc) , add = T, col="red", print.auc=T,
         print.auc.y=0.40)
legend("bottomright", c("Bacteria MPA","Aspergilus",
   "Логит-регрессия"),lwd=2, col=c("blue","black","red"))
```

```
Лучшие характеристики классификаторов
Модель       порог    специфичн.   Чувствит.    точность
ms.roc   0.5668445   0.8571429   0.9333333   0.8952381
m1.roc   1.3600000   0.8571429   0.6666667   0.7619048
m3.roc   2.3500000   0.8571429   0.7333333   0.7952381
Доверительные интервалы AUC для m1.roc
     95% CI: 0.4909-0.9948 (DeLong)
Интервальные оценки специфичности и чувствительности
     Порог  sp.low sp.median sp.high se.low se.median se.high
      6.530 0.0000    0.1429  0.4286 1.0000    1.0000  1.0000
      3.780 0.0000    0.4286  0.8571 0.8000    0.9333  1.0000
      3.000 0.1429    0.5714  0.8571 0.6667    0.8667  1.0000
      1.360 0.5714    0.8571  1.0000 0.4000    0.6667  0.8667
      0.425 1.0000    1.0000  1.0000 0.0000    0.1333  0.3333
```



2.6. ROC-кривые для логистической регрессии и двух ее независимых предикторов

2.5. Отклик, выраженный в номинальной и порядковой шкале

Логистическая регрессия обычно используется для моделирования выборок с альтернативной переменной отклика. Однако можно обобщить этот метод на номинальные или порядковые переменные с более чем двумя категориями. В обоих случаях для моделирования отклика Y мы сталкиваемся с многомерным биномиальным распределением.

Модель логита для номинальной переменной отклика (*multinomial logit*)

Пусть J обозначает число категорий для случайной величины Y. Многомерное распределение вероятности исхода отклика для каждой его категории может быть представлено вектором $\{\pi_1, ..., \pi_J\}$, $\sum_j \pi_j = 1$. Тогда для каждого из n независимых наблюдений предикторная оценка π_j будет соответствовать вероятности того, насколько предпочтительно присвоение произвольному тестируемому объекту категории j.

Отметим, что для оценки всех вероятностей достаточно построить J - 1 моделей, поскольку недостающее значение легко получить из условий нормировки (сумма π_j всегда равна 1). Поэтому одна любая из категорий принимается за базовую (примем для определенности, что это альтернатива с номером J) и тогда совокупность логит-моделей (Baseline-Category Logit) будет иметь вид:

$$\log(\pi_j / \pi_J) = \alpha_j + \beta_j \mathbf{x}, \quad j = 1,..., J - 1.$$

Кэффициенты системы логит-уравнений оцениваются одновременно и совместно путем максимизации общего критерия правдоподобия, т.е. другая форма анализа, основанная на трех последовательно построенных моделях регрессии локально для каждой альтернативы, может дать совсем иные результаты.

Рассмотрим пример, представленный в книге А. Агрести (Agresti, 2007), по пищевым предпочтениям флоридских крокодилов. В группе из 59 молодых особей, содержащихся на ферме, выделили три категории `choice` их основной диеты: рыбы (`F`), беспозвоночные (`I` – улитки, ракообразные, водные насекомые) и прочие (`O` – амфибии, млекопитающие, растения и другие крокодилы). В эксперименте измерялась длина особей `length`, составляющая от 1.24 до 3.89 м:

```
# Загружаем данные из файла (можно скачать с сайта пособия)
Alligator <- read.table(file = "Alligator.txt",
                        header = TRUE)
summary(Alligator)
# Устанавливаем категорию F в качестве базовой
Alligator$y <- relevel(Alligator$choice, ref = "F")
```

```
      length              choice
  Минимум:    1.240     F:31
  1st Qu.:    1.575     I:20
  Медиана :   1.850     O: 8
  Среднее:    2.130
  3rd Qu.:    2.450
  Максимум:   3.890
```

Обратим внимание, что в рассматриваемом примере традиционный смысл зависимости "доза-эффект" некоторым образом перевернут: мы оцениваем параметры модели `choice ~ length`, тогда как, наоборот, длина крокодила является зависимой от способа питания. Однако, во первых, техника расчетов никак не зависит от этого щекотливого обстоятельства и, во-вторых, принцип относительности причины и следствия, характерный для экологии, не исключает того, что пищевые предпочтения также могут зависеть от длины тела.

Для анализа категориальных данных в R существуют мощные специализированные пакеты `VGAM` и `mlogit`. Мы же воспользуемся более простой в освоении функцией `multinom` из пакета `nnet`, которая

осуществляет оценку коэффициентов системы логит-моделей с использованием алгоритмов построения искусственных нейронных сетей (Шитиков и др., 2005). Поскольку функции, выполняющие подгонку моделей мультиноминального логита, обычно не проводят анализа статистической значимости коэффициентов, рассчитаем *p*-значения сами с использованием теста Вальда (хотя с пониманием относимся к общению в Интернете: «Попробуйте почитать наши книги. Там объясняется, почему тесты Вальда выполнять нельзя и что асимптотическая теория может здесь дико вводить в заблуждение» / Проф. B.D. Ripley/).

```
library(nnet)
m <- multinom(y ~ length, data = Alligator)
summary(m)
print (" p-значения с использованием теста Вальда")
z <- summary(m)$coefficients/summary(m)$standard.errors
(p <- (1 - pnorm(abs(z), 0, 1))*2)
```

```
Вызываемая функция:
        multinom(formula = y ~ length, data = Alligator)
Коэффициенты:
    (Intercept)     length
  I    4.079701 -2.3553303
  O   -1.617713  0.1101012
Ст. ошибки:
    (Intercept)    length
  I    1.468640 0.8032870
  O    1.307274 0.5170823
Девианс остатков: 98.34124
AIC: 106.3412
p-значения с использованием теста Вальда
    (Intercept)      length
  I 0.005471534 0.003366615
  O 0.215912393 0.831383210
```

Анализируя блок коэффициентов, можно заметить, что длина тела тела крокодила уменьшается при переходе с рыбной диеты F на беспозвоночных I, причем логарифм отношения шансов $\log(\pi_I / \pi_F)$ уменьшается на величину β_j = -2.35 при увеличении длины на 1 м. Если рассматривать животных с диетой O, то, согласно вычисленным *p*-значениям, значение $\log(\pi_O / \pi_F)$ статистически значимо не отличается от 0, т.е. шансы равны.

Рассмотрим, насколько удачно мы сможем выполнить прогноз пищевых предпочтений на основе измерений длины тела. Построим также график изменения вероятностей всех трех альтернатив от length – рис. 2.7.

```
y <- fitted(m)
```

```
# Записываем в таблицу эмпирические категории и их прогноз
Alligator$Pred=apply(y,1,function(x)
        colnames(y)[which(x==max(x))])
table(Alligator$Pred,Alligator$y)
# Формируем данные для графика
d.plot <- data.frame(y = rep(c("F", "I", "O"), each = 50),
          length = rep(seq(1.2, 4, len = 50), 3))
d.pplot <- cbind(d.plot, predict(m, newdata = d.plot,
             type = "probs", se = TRUE))
plot (1,1, xlim=c(1.2,4),ylim=c(0,1), type='n',
        xlab="Длина крокодила", ylab="Вероятность P")
lines(d.pplot[d.pplot$y=="F",c(2,3)],lwd=2, col="black")
lines(d.pplot[d.pplot$y=="I",c(2,4)],lwd=2, col="blue")
lines(d.pplot[d.pplot$y=="O",c(2,5)],lwd=2, col="green")
legend("topleft", c("Рыбы", "Беспозвоночные", "Прочие"),
          lwd=2, col=c(1,4,3))
```

```
        В эксперименте
            F  I  O
Прогноз  F 23  7  5
         I  8 13  3
```

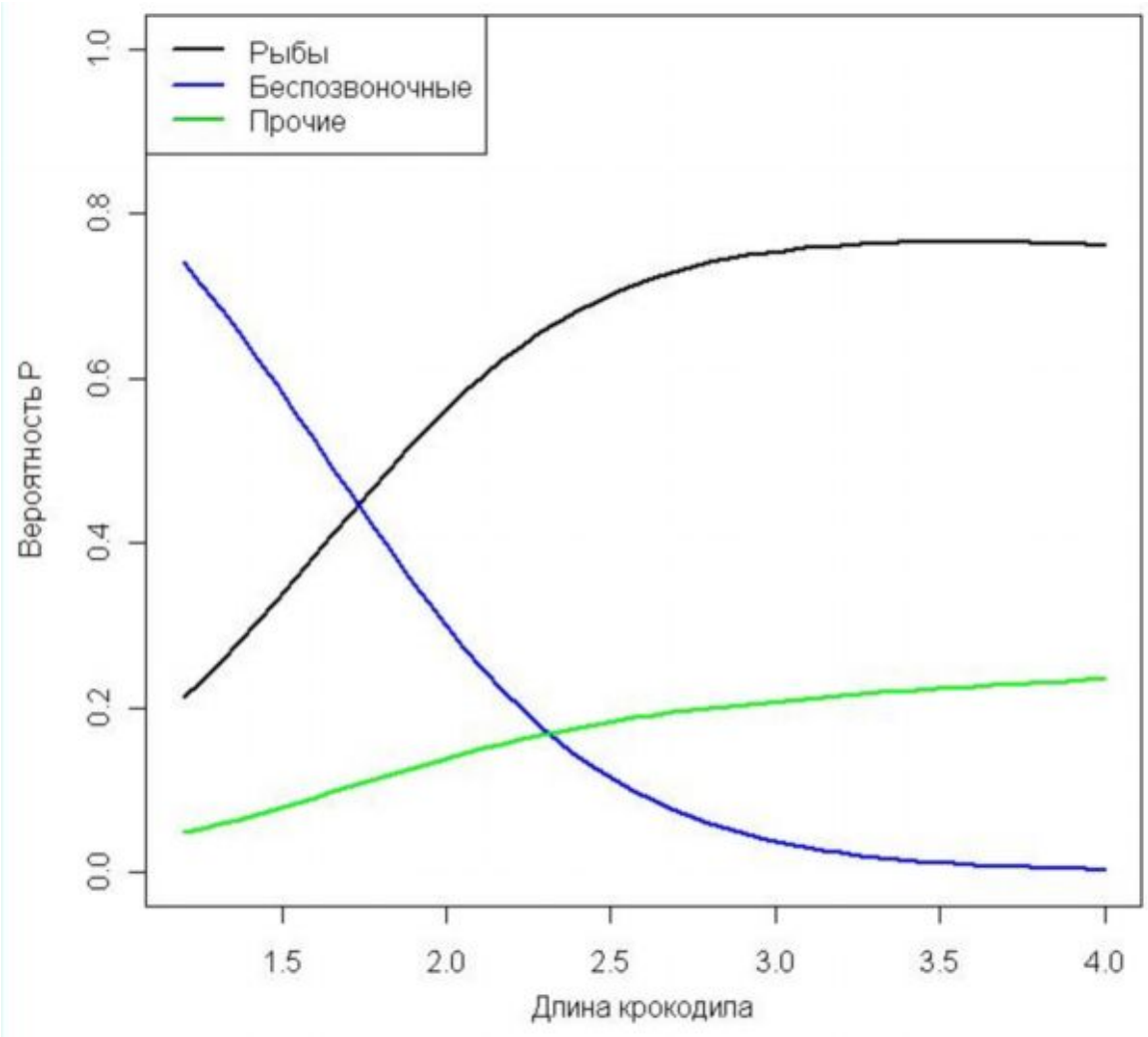


Рис. 2.7. Зависимость длины тела крокодила от пищевых предпочтений

На основе модели удалось правильно предсказать диету 36 крокодилов из 59. Отметим, что прогноз проигнорировал группу животных со смешанным питанием 0 ввиду небольшой априорной вероятности этой альтернативы и низкой статистической значимости этой модели.

Модель логита для порядковой переменной отклика (*Cumulative Logit*)

Если категории отклика являются упорядоченными, то можно использовать это обстоятельство и построить модель потенциально большей мощности и с более простой интерпретацией результатов, чем на основе номинальных переменных. Пусть для произвольной порядковой случайной величины *Y*, изменяющейся на интервале от 1 до *J*, справедливо неравенство $P(Y \le 1) \le P(Y \le 2) \le \ldots \le P(Y \le J) = 1$, определяющее процесс накопления вероятности:

$$P(Y \le 1) = \pi_1 + \ldots + \pi_j, \quad j = 1, \ldots, J.$$

Тогда модель для оценки логарифма отношения накопленных шансов или кумулятивного логита будет иметь вид:

$$\text{logit}[P(Y \le j)] = \log\left[\frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J}\right] = \alpha_j + \beta\mathbf{x}, \; j = 1, \ldots, J-1,$$

где параметр α_j определяет величину, на которую увеличивается логарифм отношения шансов при включении вероятности π_j, а коэффициенты β не зависят от j и объединяют эффекты от воздействия всей совокупности независимых переменных. На рис. 2.8 показан пример такой модели с одним и тем же эффектом от влияния независимой переменной x для каждой из трех функций накопленных вероятностей на основе отклика из четырех категорий: видно, что происходит пропорциональный сдвиг кривых вправо, определяемый величиной коэффициента α_i.

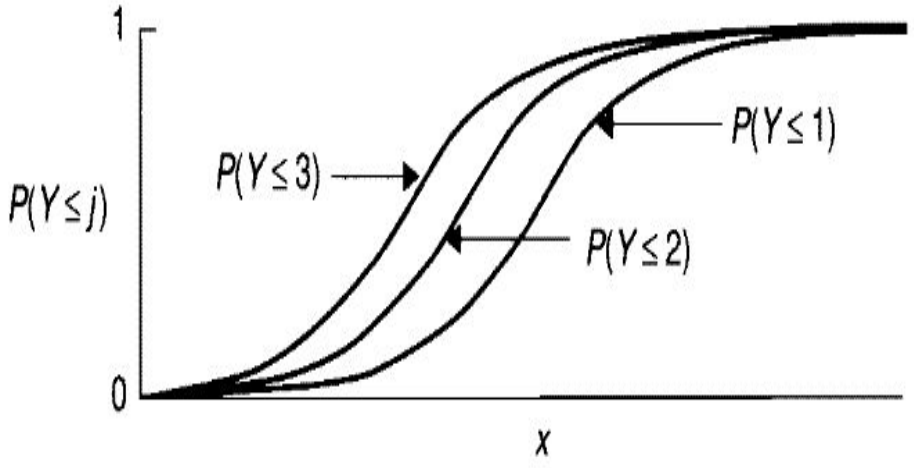


Рис. 2.8. Кривые накопленных вероятностей для модели пропорциональных шансов

Рассмотрим пример модели кумулятивного логита на основе данных, описанных в (Шитиков и др., 2005). Из 34 малых рек Самарской области, принадлежащих разным классам качества вод по ГОСТ 17.1.3.07–82, было взято 520 проб макрозообентоса, по каждой из которых оценивались различные гидробиологические показатели и биотические индексы, в том числе, число видов (`S.spe`), прологарифмированные суммарные численность (`N_ln` , экз/м^2) и биомасса (`B_ln` , мг/ м^2), доля хищных видов в общей биомассе (`H.spe`, %), индекс видового разнообразия Шеннона (`Ind.Shen`), биотический индекс Вудивисса (`Ind.Wud`), олигохетный индекс Пареле (`Ind.Par`) и хирономидный индекс Балушкиной (`Ind.Bal`). Как следует из корреляционной матрицы, представленной в колорированном виде на рис. 2.9, часть показателей (`S.spe`, `N_ln`, `B_ln`, `Ind.Shen`) образуют достаточно тесно связанный корреляционный комплекс.

```
# Загружаем данные из файла (можно скачать с сайта пособия)
Water_q <- read.table(file = "Water_q.txt", header = TRUE)
Water_q$Class <- as.factor(Water_q$Class)
M <- cor(Water_q[,-1])
library(corrplot)
corrplot(M, method="color", addCoef.col="green",
        addgrid.col = "gray33",tl.col = "black")
head(Water_q)
```

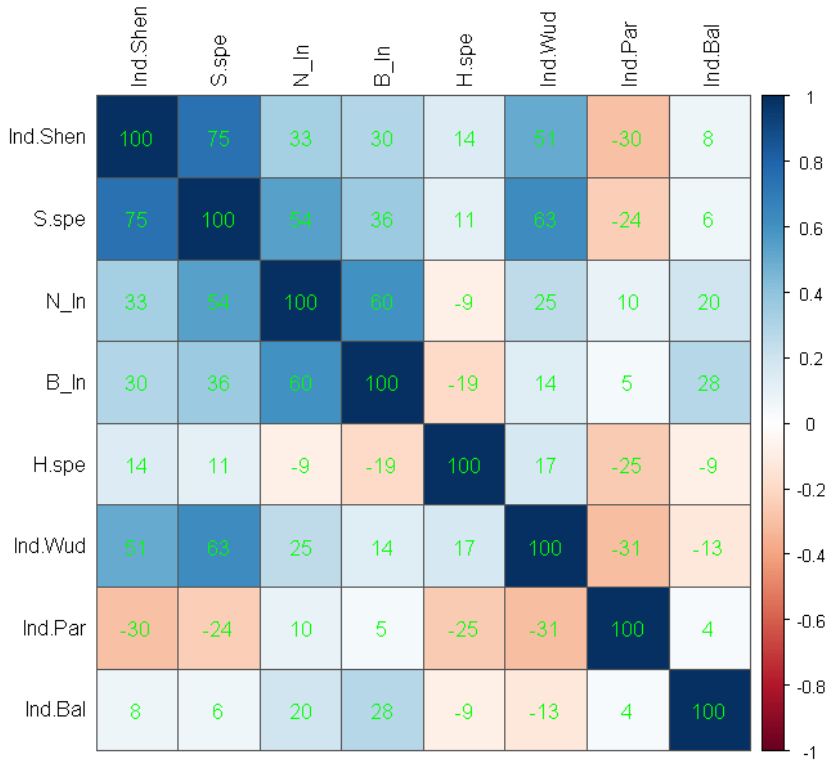


Рис. 2.9. Корреляционная матрица гидрбиологических показателей

```
  Class Ind.Shen S.spe     N_ln     B_ln     H.spe Ind.Wud Ind.Par   Ind.Bal
1     2 2.718601    13 6.620073 7.575585  4.615385     6.5   0.040 0.5326633
2     2 2.478081    19 7.981050 8.207947 53.678474     7.0   0.001 6.2852349
3     2 1.652091    10 9.442880 7.824046 15.600000     5.5   0.000 0.2397718
4     2 1.456787     9 7.731931 8.873468  4.901961     3.0   0.010 4.1315789
5     2 3.195423    18 7.177782 7.012115  8.108108     6.0   0.000 0.5254873
6     2 3.127793    17 7.106606 7.878534 59.469697     7.0   0.260 1.0294118
```

Отклик `Class` имеет 4 градации: 2 – "Чистые", 3 – "Умеренно загрязненные", 4 – "Загрязненные" и 6 – "Грязные и очень грязные" (класс 5 был объединен с 6-м). Построение модели кумулятивного логита будем осуществлять с использованием функции `polr` из пакета `MASS`.

```
library(MASS)
m.wq <- polr(Class ~ .,  data = Water_q)
summary(m.wq, digits = 3)
# Оценка доверительных интервалов для коэффициентов
confint.default(m.wq)
```

```
Вызываемая функция:
          polr(formula = Class ~ ., data = Water_q)
Коэффициенты:
              Оценка  Ст.ошибка  t-крит
   Ind.Shen  0.20889    0.14851   1.407
   S.spe    -0.05549    0.02353  -2.358
   N_ln      0.02564    0.07896   0.325
   B_ln      0.01432    0.04875   0.294
   H.spe    -0.00416    0.00368  -1.131
   Ind.Wud  -0.53012    0.06119  -8.663
   Ind.Par   1.69579    0.29010   5.846
   Ind.Bal   0.07218    0.02931   2.463
Свободные члены:
       Оценка  Ст.ошиб.  t-крит
   2|3 -4.050  0.557     -7.269
   3|4 -1.924  0.538     -3.574
   4|6  0.313  0.524      0.597
Девианс остатков: 1087.445
AIC: 1109.445
Доверительные интервалы коэффициентов
                  2.5 %       97.5 %
   Ind.Shen -0.08218760  0.499975309
   S.spe    -0.10160878 -0.009364040
   N_ln     -0.12912138  0.180392899
   B_ln     -0.08122762  0.109875742
   H.spe    -0.01137812  0.003052698
   Ind.Wud  -0.65005348 -0.410191283
   Ind.Par   1.12719951  2.264375686
   Ind.Bal   0.01473795  0.129613368
```

Заметим, что блок коэффициентов состоит из двух разделов: коэффициенты β для независимых переменных и оценки параметра α, корректирующие отношение шансов при каждом шаге объединения альтернатив. Расчет доверительных интервалов показал, что половина коэффициентов β статистически незначима, поскольку их доверительный интервал включает число 0 (отмечены курсивом). Это обычно является следствием высокой мультиколлинеарности данных, т.е. высокой степени взаимной зависимости гидробиологических показателей. Считается, что наиболее эффективный путь устранения мультиколлинеарности – исключение из регрессионной модели незначимых коэффициентов, или, выражаясь точнее, отбор информативного комплекса из q переменных ($q < m$). Пошаговый регрессионный анализ, выполняемый функцией `stepAIC(…)`, представляет собой последовательную процедуру включения и исключения отдельных предикторов в модель, пока не будет достигнута наилучшая регрессия по критерию Акаике.

```
ms.wq <- stepAIC (m.wq)
summary(ms.wq, digits = 3)
confint.default(ms.wq)
```

```
 Вызываемая функция:
polr (formula = Class ~ S.spe + Ind.Wud + Ind.Par + Ind.Bal)
 Коэффициенты:
             Оценка  Ст.ошибка  t-крит
    S.spe   -0.0320     0.0162   -1.97
    Ind.Wud -0.5297     0.0611   -8.67
    Ind.Par  1.7469     0.2771    6.30
    Ind.Bal  0.0808     0.0284    2.85
 Свободные члены:
          Оценка  Ст.ошиб.  t-крит
    2|3  -4.423   0.338    -13.100
    3|4  -2.313   0.285     -8.124
    4|6  -0.083   0.259     -0.320
Девианс остатков: 1091.131
AIC: 1105.131
                 2.5 %        97.5 %
    S.spe   -0.06373387 -0.0002284454
    Ind.Wud -0.64949497 -0.4099929381
    Ind.Par  1.20376745  2.2900959871
    Ind.Bal  0.02514575  0.1363918004
```

Число независимых переменных модели сократилось с 8 до 4, все коэффициенты которых оказались статистически значимыми. Проведем тест, попарно сравнивая отношения правдоподобия трех моделей: *а*) полной модели `m.wq` и оптимальной модели `ms.wq` , *б*) модели `ms.wq` и нуль-модели без параметров. В первом случае мы убеждаемся в том, что при исключении 4-х переменных не произошло статистически значимого

увеличения ошибки регрессии по сравнению с полной моделью, а во втором – что оптимальная модель адекватна в целом.

```
anova(m.wq, ms.wq)
m0 <- polr(Class ~ 1,  data = Water_q)
anova(m0, ms.wq)
```

```
Тест отношения лог. правдоподобия для порядковых данных
Отклик: Class
Модели:
1 S.spe + Ind.Wud + Ind.Par + Ind.Bal
2 Ind.Shen + S.spe + N_ln + B_ln + H.spe + Ind.Wud
                          + Ind.Par + Ind.Bal
Модель  Остат.df Остат.Dev   Тест    Df  LR стат.  Pr(Chi)
1          513   1091.131
2          509   1087.445  1 vs 2   4  3.685486 0.4502404

Отклик: Class
Модели:
1 -
2 S.spe + Ind.Wud + Ind.Par + Ind.Bal
Модель  Остат.df Остат.Dev   Тест    Df  LR стат.  Pr(Chi)
1         517  1357.317
2         513  1091.131   1 vs 2    4   266.1862       0
```

Аналогично предыдущему примеру, рассмотрим использование построенной модели для прогнозирования качества вод по значениям гидробиологических показателей. С помощью функции `fitted()` найдем значения вероятностей для каждого из 4 рассматриваемых классов, которые вычисляются на основе оценок коэффициентов α_j и β для оптимальной модели. Отнесение гидробиологического измерения к конкретному классу будем относить по максимальной вероятности. Функция `table()` поможет нам выделить несовпадение эмпирических и прогнозируемых значений класса качества вод.

```
y <- fitted(ms.wq)
head(y)
#  Построение таблицы сопряженности «Факт-Прогноз»
Pred=apply(y,1,function(x) colnames(y)[which(x==max(x))])
table(Water_q[,1], Pred)
```

```
            2          3          4          6
1 0.33687248 0.4705885 0.1675313 0.02500767
2 0.35046367 0.4661929 0.1597638 0.02357967
3 0.22980431 0.4814415 0.2469114 0.04184274
4 0.05227415 0.2606026 0.4960338 0.19108946
5 0.32920738 0.4728316 0.1720980 0.02586302
6 0.32983908 0.4726534 0.1717164 0.02579108
```

```
                Прогноз класса качества
                        2   3   4   6
Эмпирические   2   11  31   9   2
данные         3    6  59  49  14
класса         4    2  40 111  34
качества       6    0   6  46 100
```

Можно отметить, что модель достаточно хорошо справилась с оценкой грязных вод, но чистые воды часто принимала за загрязненные.

Многомерность параметров модели не позволяет показать графически изменение вероятности классов от всех значений предикторов одновременно. Однако это можно сделать, например, для индекса Вудивисса, если зафиксировать значения остальных независимых переменных на уровне их средних значений.

```
#   Подготовка данных для графика
d.plot <- data.frame( Ind.Wud = seq(min(Water_q$Ind.Wud),
       max(Water_q$Ind.Wud), len = 50),
       S.spe = rep(mean(Water_q$S.spe), each = 50),
       Ind.Par = rep(mean(Water_q$Ind.Par), each = 50),
       Ind.Bal = rep(mean(Water_q$Ind.Bal), each = 50))
d.pplot <- cbind(d.plot, predict(ms.wq, newdata = d.plot,
         type = "probs", se = TRUE))
#   Прорисовка компонент графика
plot (1,1, xlim=c(0,9),ylim=c(0,0.7), type='n',
     xlab="Индекс Вудивисса", ylab="Вероятность P")
lines(d.pplot[,c(2,5)],lwd=2, col="green")
lines(d.pplot[,c(2,6)],lwd=2, col="blue")
lines(d.pplot[,c(2,7)],lwd=2, col="gray")
lines(d.pplot[,c(2,8)],lwd=2, col="black")
legend("topright", c("2","3","4","6"), lwd=2,
       col=c(3,4,8,1))
```

Из графика на рис. 2.10 видно, крайние значения на шкале индекса Вудивисса соответствуют максимальным вероятностям отнесения качества вод к "Очень грязным" и "Чистым" категориям соответственно слева направо. Промежуточные значения индекса вероятнее всего приведут к классам качества 3 и 4.

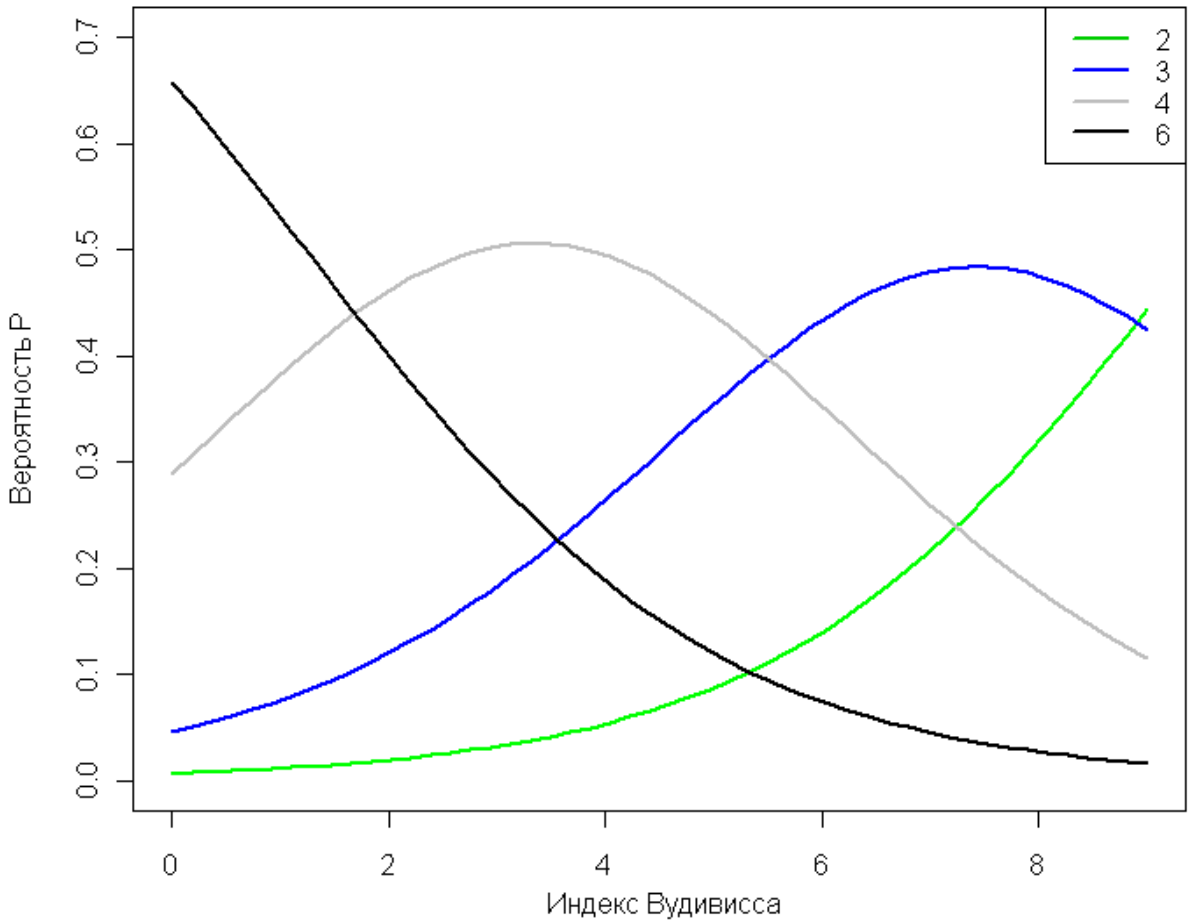


Рис. 2.10. Зависимость вероятности отнесения к классам качества вод от значений индекса Вудивисса

2.6. Процедуры сглаживания и обобщенные аддитивные модели

В общем случае, если отсутствуют явные теоретические предпосылки, вид функции регрессии от одной или нескольких независимых переменных заранее не определен. Исследователь волен в правой части уравнения использовать как непосредственно значения x-ов для построения линейной модели, так и любые их преобразования: логарифмы, экспоненты, дробные степени, фрагменты полиномов k-й степени, тригонометрические функции, преобразования Бокса-Кокса и др. Выбор конкретной функциональной формы модели сродни искусству, поскольку неверная спецификация может привести к серьезным искажениям при интерпретации результатов.

Полиномиальные и иные нелинейные модели регрессии являются фактически частными случаями общего подхода с использованием *базисных функций*. Идея здесь состоит в том, чтобы использовать семейство $b_1(x)$, $b_2(x)$, ..., $b_k(x)$ произвольных функций или преобразований, которые могут быть применены к переменной X. Тогда выполняется оценка параметров нелинейной модели расширенного вида:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \ldots + \beta_k b_k(x_i) + \varepsilon_i,$$

в результате чего даже на основе одной независимой переменной можно получить модель вполне внушительной сложности.

Но при этом возникает трудноразрешимая проблема: какое сочетание базисных функций даст нам оптимальную модель, наилучшим образом описывающую экспериментальные данные на всем диапазоне варьирования x? И тут же возникает другая идея: а нельзя ли для разных интервалов $\boldsymbol{x}$ строить не одну, а несколько самостоятельных моделей, или как-то динамически подстраивать коэффициенты полинома при движении слева направо по оси $\boldsymbol{x}$? Так оказалось, что прекрасной альтернативой классической регрессии являются полупараметрические модели *сглаживания* (*smoothing*). В настоящее время разработан целый набор гибких и эффективных средств аппроксимации неизвестных регрессионных зависимостей с использованием стандартных алгоритмов сглаживания: скользящей средней, экспоненциальной моделью, ядерной функцией, сплайнами, локальной полиномиальной регрессией и др. (Hastie at al., 2009; Шитиков, Розенберг, 2014; Мастицкий, Шитиков, 2015).

Метод *локальной регрессии* (или LOESS – акроним от *lo*cal reg*ress*ion) представляет собой процедуру скользящего усреднения параметров линейной или полиномиальной (квадратичной) модели таким образом, что коэффициенты $\beta_0(x_0)$, $\beta_1(x_0)$, $\beta_2(x_0)$ рассчитываются динамически для каждого текущего значения x_0. Технически это означает, что аппроксимирующая кривая последовательно подстраивается к данным по мере передвижения "окна" слева направо по шкале x, при этом каждый раз новая порция наблюдений включается в расчет локальных моделей взамен исключаемых старых. Подгонка такой модели методом наименьших квадратов осуществляется с учетом вектора весов $w(x_0, x_i, h)$, которые тем больше, чем ближе точки исходной выборки x_i располагаются по отношению к тестируемому объекту x_0.

Аппроксимация одномерной зависимости методом локальной регрессии реализована в R функцией `loess(…)`, где ключевым параметром является *степень сглаживания* `span`: чем выше его значение, тем меньше при аппроксимации появляется различных "горбов" (экстремумов), а модель становится более устойчивой и стационарной. Однако при этом одновременно ухудшается степень приближения к данным и ошибка регрессии.

Ядерная модель сглаживания отклика y на шкале независимой переменной x основывается практически на тех же исходных предпосылках, что и локальная регрессия. Основное отличие – в способе оценки значения $\hat{y}(x_0)$ в точке x_0. Для ядерного сглаживания применяется функция `ksmooth(…)`, параметр которой `bandwidth` ("ширина окна") имеет тот же смысл, что и `span`. Методы оптимизации параметров

сглаживания обоих моделей подробно рассматриваются, например, в (Hastie at al., 2009).

Рассмотрим пример, представленный в классической книге по пробит-анализу (Finney, 1971). Для последовательности из 7 доз инсектицида установлено число погибших насекомых. На перврм этапе построим традиционную модель пробита с использованием прологарифмированных значений доз:

```
library(drc); library (MASS)
data(finney71)
head(finney71)
## Подгонка glm-модели без свободного члена
fin.glm <- glm(cbind(affected,total-affected) ~ log(dose)-1,
      family=binomial(link = probit),
      data=finney71[finney71$dose != 0, ])
summary(fin.glm)
## Оценка изоэффективных доз
xp <- dose.p(fin.glm, p=c(0.50, 0.90, 0.95)) # функция MASS
xp.ci <- xp + attr(xp, "SE") %*%
       matrix(qnorm(1 - 0.05/2)*c(-1,1), nrow=1)
zp.est <- exp(cbind(xp, xp.ci[,1],xp.ci[,2]))
dimnames(zp.est)[[2]] <- c("Оценка","2.5% LCL","97.5 % UCL")
zp.est
```

```
Модель пробита
Коэффициенты:
            Оценка  Ст.ошибка  z-крит Pr(>|z|)
log(dose)  0.18816    0.04765   3.948 7.86e-05 ***
---
    Null-девианс: 98.505  при 5  степенях свободы
Девианс остатков: 82.904  при 4  степенях свободы
AIC: 104.12
Изоэффективные дозы и их доверительные интервалы:
             Оценка 2.5% LCL 97.5 % UCL
p = 0.50:  4.828918 4.363708   5.343724
p = 0.90:  9.802082 8.073495  11.900771
p = 0.95: 12.470382 9.748334  15.952512
```

Сравним пробит-регрессию, полученную на основе данных примера Финни, с моделями, построенными по идеологии сглаживания. Отметим, что мы, по сути (но с 70-летним лагом и на ином техническом уровне) воспроизводим противопоставление линейного пробит-метода Миллера-Тейнтера и алгоритма сглаживания накопленных частот Беренса-Шлоссер.

```
pval <- finney71$affected/finney71$total
# Модель сглаживания ядерной функцией
fin.ksm <- ksmooth(finney71$dose,pval, "normal",
          bandwidth=1.8, x.points=finney71$dose)
```

```
# Модель локальной регрессии
fin.loe <- loess(pval ~ finney71$dose, span=0.85)
# Оценка LD50
c(LOESS=approx(predict(fin.loe),finney71$dose, xout = 0.5)$y,
      KSMOOTH=approx(y=fin.ksm$x, x=fin.ksm$y, xout = 0.5)$y)
plot(finney71$dose, pval, xlab="Доза инсектицида",
          ylab="Доля погибших насекомых")
lines(finney71$dose,c(fin.glm$fit,0), lwd=2)
lines(fin.ksm, col="green", lwd=2)
lines(finney71$dose,predict(fin.loe), col="blue", lwd=2)
legend("bottomright",c("Пробит-модель","Локальная
       регрессия","Ядерная функция"),
          lwd = 2, col=c("black","blue","green"))
```

Среднеэффективные дозы для моделей сглаживания

```
     LOESS  KSMOOTH
  4.925957 5.121265
```

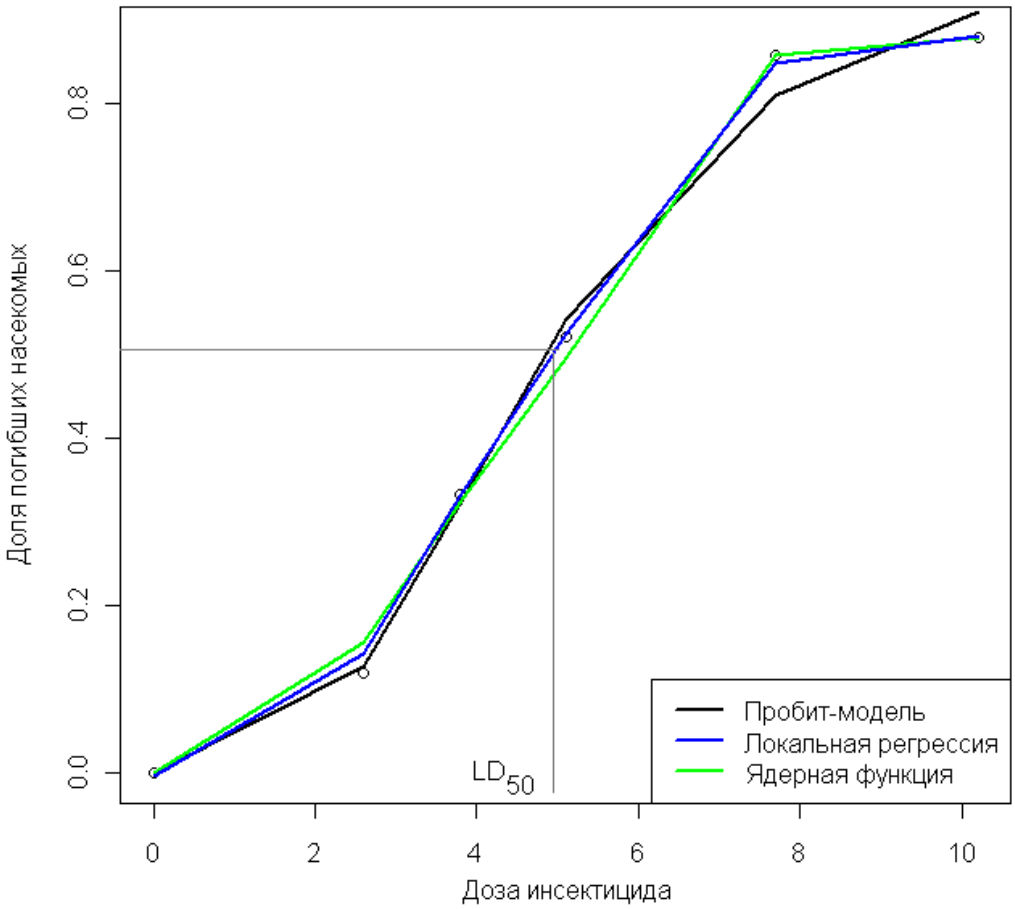


Рис. 2.11. Сопоставление моделей сглаживания и пробит- регрессии по данным примера Финни

Воспользовавшись процедурами сглаживания, мы существенно облегчаем себе задачу: уже нет необходимости задумываться над выбором базисной функции регрессии (в частности, нужно ли логарифмировать

дозы), либо остерегаться, что не подтвердятся те или иные исходные предположения анализа. Для всех трех моделей получены практически совпадающие значения LD_{50}, однако в случае ядерного сглаживания или локальной регрессии существуют проблемы оценки доверительных интервалов. Кроме того, при сглаживании совершенно неэффективна экстраполяция, в связи с чем мы не можем рассчитать LD_{90} или LD_{95}, поскольку максимальный эффект, достигнутый в эксперименте, составляет только 88 %.

Рассмотрим теперь смысл интерполяции сплайнами, сглаживание которыми можно представить как составление композиций из "кусочков" гладкой функции – как правило, кубических полиномов.

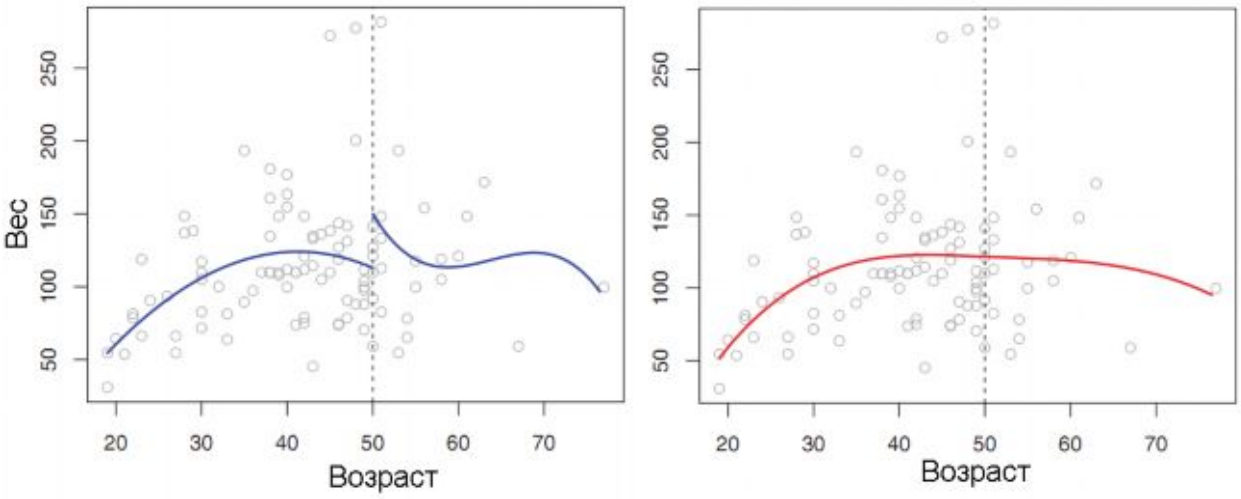


Рис. 2.12. Слева – несогласованные кубические полиномы, описывающие зависимость веса от возраста на двух интервалах, справа – кубический сплайн с одним узлом сопряжения

Сплайном, или кусочно-полиномиальной функцией с k сопряжениями в точках a_1, ..., a_k называется функция $\varphi_j(x)$, которая на каждом интервале $(a_j, a_j + 1)$, где $j = 0, ..., k$, описывается алгебраическим полиномом $P_j(x)$ степени m. Коэффициенты полиномов согласованы между собой так, чтобы в узлах сопряжений выполнялись условия непрерывности функции $\varphi_j(x)$ и ее $(m - 1)$ производных, вследствие чего отдельные участки состыковываются между собой для получения гладкой кривой.

Можно рассчитать кубический сплайн ($m = 3$) с узлами сопряжения в каждом уникальном значении x_i. Тогда ошибка аппроксимации будет близка к нулю, но сглаживающий сплайн будет иметь слишком много степеней свободы, полученная кривая состоять из резких локальных изменений, а модель – крайне низкую устойчивость при прогнозировании. Поэтому при сглаживании сплайнами основной проблемой является выбор оптимального (эффективного) числа степеней свободы k от 2 до n, для

регуляризации которого разработанные программные модули обычно используют методы ресэмплинга.

В функции `smooth.spline()`, которая реализует в R сглаживание сплайнами, степень сглаживания задается параметром `spar`, который обычно (но не обязательно) изменяется на интервале [0; 1]. Можно указать его конкретное значение, либо задать аргумент `cv = TRUE`, чтобы найти оптимальное число степеней свободы с использованием кросс-проверки.

Обобщенные аддитивные модели GAM (generalized additive model – Wood, 2006) объединили обе представленные идеи базисных функций и сглаживающих моделей под "оберткой" единого вида:

$$y = \beta_0 + \sum_{i=1}^{p} q_i(x_i) + \varepsilon,$$

где $q_i(x_i)$ – произвольные функции нелинейного преобразования независимых факторов, в качестве которых чаще всего используют вышеупомянутые модели сглаживания. В сущности, GAM можно представить как средство обобщения частных функций q_i в единое целое и с произвольным числом параметров, но без каких-либо предварительных предположений относительно формы регрессионной кривой.

В статистической среде R построение аддитивных моделей реализуется функцией, весьма напоминающей `glm()` и представленной в пакетах двух разработчиков `gam` (Trevor Hastie) и `mgcv` (Simon Wood):

```
gam(formula, data = data.frame,
      family = family.generator).
```

Здесь оператор `formula` в правой своей части задает вид сглаживающих функций для предикторов: локальной регрессией `lo(…)` или сплайнами `s(…)`. В качестве зависимой переменной могут использоваться произвольные данные – альтернативные, счетные или количественные, для чего параметр задается как `binomial`, `poisson` или `gaussian` соответственно.

Рассмотрим на примере построение моделей GAM для альтернативного показателя, сравнив две модели: традиционного пробита и GAM с использованием интерполяции сплайнами. Экологи, озабоченные идеей губительного влияния CO_2 на эволюцию лесного покрова, выдвинули предположение, что повышение углекислоты стимулирует способность сосен к размножению. Построим модели зависимости вероятности половой зрелости (по наличию шишек на дереве) `mature` от диаметра ствола сосны на уровне груди `DBH`.

```
library(boot)
library(mgcv)
# Загружаем данные из файла (можно скачать с сайта пособия)
cones <- read.table ( "pinecones.txt", header=T )
mature <- ifelse(cones$X2000 > 0, 1, 0)
```

```
DBH <-cones$dbh
summary (fit.log <- glm ( mature ~ DBH ,
         family=binomial(link=probit)))
summary (fit.gam <- gam( mature ~ s(DBH), family=binomial))
extractAIC(fit.gam)[2]
```

```
Модель пробита:     glm(formula = mature ~ DBH,
                          family = binomial(link = probit))
Коэффициенты:
              Оценка  Ст.ошибка  z-крит Pr(>|z|)
 (Св. член) -4.29105    0.95375  -4.499 6.82e-06 ***
DBH          0.20761    0.05163   4.021 5.80e-05 ***
---
    Null-девианс: 98.254  при 95  степенях свободы
Девианс остатков: 68.677  при 94  степенях свободы
AIC: 72.677
----------------------------------------
Модель  GAM:        mature ~ s(DBH)
Параметрические коэффициенты:
               Оценка  Ст.ошибка  z-крит Pr(>|z|)
 (Св. член)   -2.3094     0.5033  -4.589 4.46e-06 ***
---
Аппроксимирующая значимость сглаживающего члена:
       edf Ref.df Chi.sq  p-значен
s(DBH)   1      1  14.94 0.000111 ***
---
R-кв.(прив) = 0.271  Доля объяснения девианса = 29.7%
UBRE score = -0.23913  Scale est. = 1         n = 96
AIC: 73.043
```

По своим статистическим характеристикам обе модели практически идентичны: чуть большее значение AIC-критерия у модели GAM объясняется большим числом степеней свободы при использовании сплайнов (df = 2 против df = 1).

Построим графики зависимости с использованием обеих моделей:

```
# Формируем массивы предикторных значений
data.plot <- data.frame(DBH= seq(0, max(DBH),
             length.out = 100))
pred.log <- predict(fit.log,newdata=data.plot,"response")
pred.gam <- predict.gam(fit.gam,newdata=data.plot,
                se.fit=TRUE)
plot(mature ~ DBH,xlab="Диаметр ствола сосны", type="n",
     ylab="Вероятность наличия шишек")
# Исходные точки изобразим в виде насечек вверху и внизу
rug(DBH[mature==0],side=1,col="grey")
rug(DBH[mature!=0],side=3,col="grey")
X <- data.plot$DBH ; Y <- inv.logit(pred.gam$fit)
YUC <- inv.logit(pred.gam$fit+1.96*pred.gam$se.fit)
YLC <- inv.logit(pred.gam$fit-1.96*pred.gam$se.fit)
```

```
lines(X,pred.log,lty=4,lwd=2) #  Модель пробита
lines(X,Y,lwd=2)       #  Модель GAM
lines(X,YUC,lty=3)     # и ее доверительные интервалы
lines(X,YLC,lty=3)
legend("topleft", c("Сплайн GAM", "Пробит"),
        lty = c(1,4), lwd=2)
```

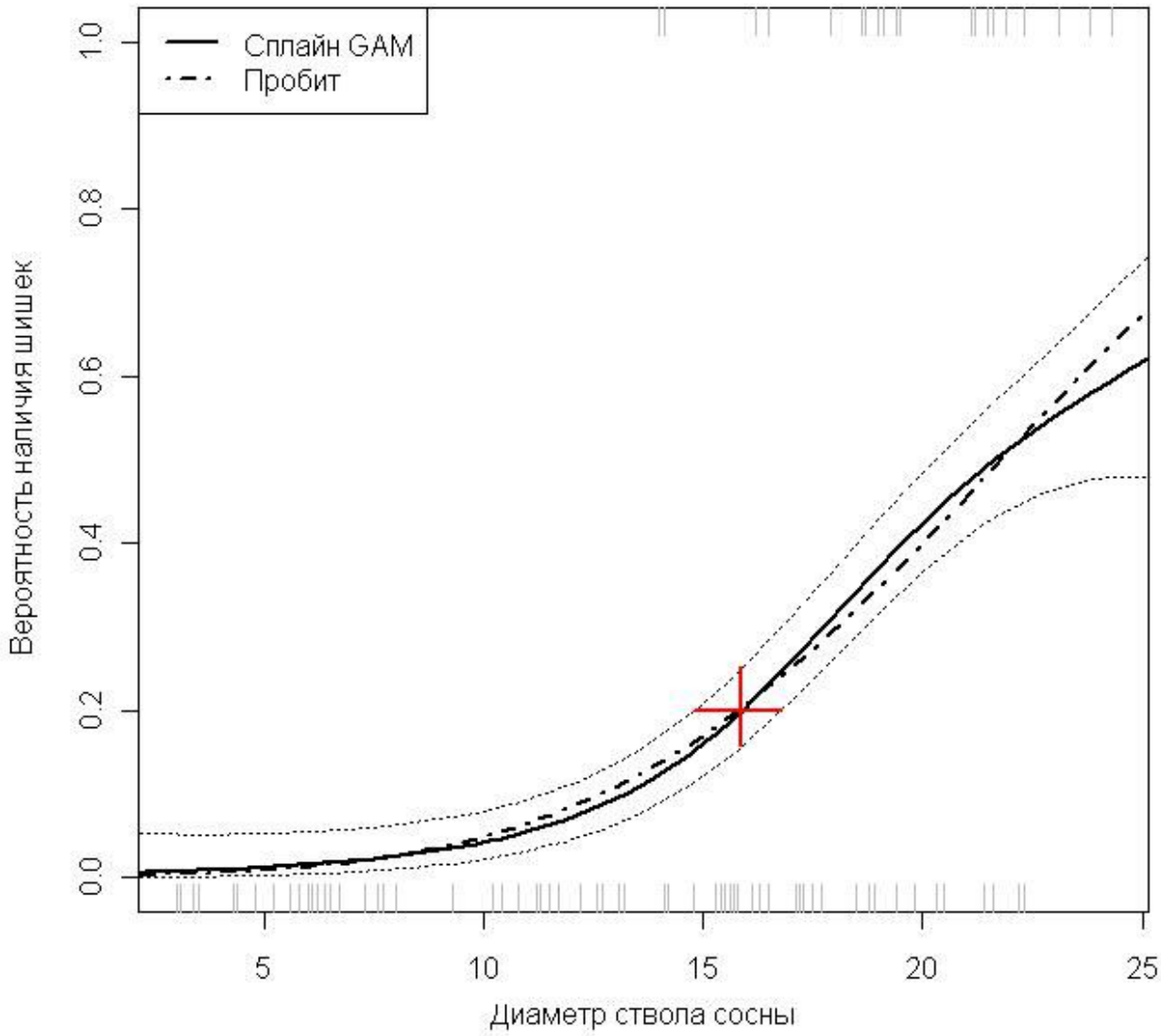


Рис. 2.13. Аппроксимация зависимости репродуктивной способности сосны от диаметра ствола моделями GAM и пробит; пунктиром показана доверительная полоса GAM-модели

Определенные трудности при интерполяции сплайнами возникают при оценке доверительных интервалов "изоэффективных доз" (под этим термином здесь выступает диаметр ствола сосен). Функция `predict.gam()` обеспечивает расчет стандартной ошибки для каждого прогнозируемого значения: на рис. 2.13 доверительные интервалы оцениваемой вероятности при p = 0.2 показаны вертикальной чертой красного цвета. Но доверительные вероятности независимой переменной

можно оценить только геометрически: на рис. 2.13 – им соответствует горизонтальная черта красного цвета. Попробуем выполнить расчеты:

```
# Для пробит-модели с использованием функции dose.p
library(MASS)
xp <- dose.p(fit.log, p=c(0.10, 0.20, 0.5))
xp.ci <- xp + attr(xp, "SE") %*% matrix(qnorm(1 - 0.05/2)*
          c(-1,1), nrow=1)
zp <- cbind(xp, attr(xp, "SE"), xp.ci[,1], xp.ci[,2])
dimnames(zp)[[2]] <- c("LD", "SE", "LCL","UCL")
zp
# Функция рассчитывает доверительные интервалы геометрически
CIdose.p <- function(p=0.5) {
     XE <- approx(Y, X, xout = p)$y
     X1 <- approx(YUC, X, xout = p)$y
     X2 <- approx(YLC, X, xout = p)$y
  return (c(XE, X1, X2))
}
print("Доверительные интервалы GAM для p= 0.2")
(S <- CIdose.p(p=0.2))
i <- findInterval(0.2, Y)+1
segments(S[2], 0.2, S[3], col="red", lwd=2)
segments(S[1], YUC[i], S[1], YLC[i], col="red", lwd=2)
```

```
Изоэффективные дозы и их доверительные интервалы
для модели пробита
                LD        SE      LCL      UCL
p = 0.1: 14.49613 1.2371528 12.07135 16.92090
p = 0.2: 16.61518 0.9203712 14.81128 18.41907
p = 0.5: 20.66909 1.0428476 18.62515 22.71303
для модели GAM при p= 0.2:
p = 0.2: 15.88262           14.78983 16.78565
```

При изучении графика на рис. 2.13 легко установить, что вычислить доверительные интервалы "изоэффективных доз" геометрическим способом часто не удается: при p < 0.1 они становятся неоправданно широки, а при p = 0.5 верхний доверительный интервал просто недостижим. В сравнении с этим, функция `dose.p()` для модели пробита исправно выдала полный комплект значений, внешне внушающий доверие.

3. МОДЕЛИРОВАНИЕ ВРЕМЕНИ НАСТУПЛЕНИЯ ЭФФЕКТА

3.1. Анализ выживаемости по методу Каплан-Майера

Анализ выживаемости (survival analysis) рассматривает эксперимент по установлению зависимости времени наступления неблагоприятного эффекта от продолжительности воздействия токсиканта. Как и в предыдущем разделе, описываемая техника расчетов может применяться не только для анализа смертности, но и любого другого альтернативного эффекта, поэтому термин "выживаемость" употребляется здесь только для краткости изложения. Анализ выживаемости – это в общем случае построение статистических моделей, в которых эффект y (отклик экосистемы) является функцией независимых переменных ($\mathbf{x}$; t), где $\mathbf{x}$ – уровень воздействия и/или иные экзогенные или эндогенные факторы, влияющих на время жизни объектов t.

Рассмотрим вначале алгоритм Каплан-Майера (Kaplan, Meier, 1958) по оценке показателей усредненной выживаемости в одной группе. Пусть мы имеем выборку из n объектов и переменная времени t отсчитывается от некоторой базовой величины (″точки старта″). В ходе последующего наблюдения каждая выборочная единица может находиться в одном из трех состояний:

- ″жив и находится под наблюдением″ (предположим, что в момент времени t в этом состоянии находится r_t объектов);
- ″наблюдение потеряно″ и таких объектов m_t;
- ″умер″ и тогда число объектов, для которых к моменту времени t имел место наблюдаемый эффект, равно $d_t = n - r_t - m_t$.

Текущую оценку вероятности выживаемости определим как $r_t /(r_t + d_t)$. Естественно, что при смерти одного объекта величина r_t уменьшается на единицу, а $(r_t + d_t)$ остается постоянной.

Данные для анализа выживаемости, как правило, являются неполноценными с точки зрения статистики: для многих объектов значение бинарного отклика (т.е. то, что нужно научиться предсказывать) неизвестно. Можно сформулировать лишь некоторые соображения относительно зависимой переменной: что предполагаемое время смерти превысит конечную точку наблюдений на какую-то неопределенную величину. Такие данные называются *цензурированными* (censored data). При анализе выживаемости встречается почти исключительно цензурирование справа, то есть знание, что время жизни объекта больше какой-то величины.

Формальной целью анализа Каплан-Майер является оценка функции распределения случайной величины τ (″времени жизни чего-либо″)

$$F(t) = P(\tau < t).$$

Вместо функции распределения F при анализе выживаемости обычно применяют функцию выживаемости (survival function)

$$S(t) = 1 - F(t) = P(\tau \geq t),$$

равную вероятности выжить к моменту t, и риск (hazard) равный плотности вероятности гибели в момент t при условии, что до этого момента дожили:

$$h(t) = - d[\ln S(t)]/dt.$$

Для множества из n объектов с разными распределениями времени жизни и, соответственно, функциями выживаемости $S_i(t)$ можно определить усредненную функцию выживаемости, равную математическому ожиданию доли объектов, выживших к моменту t:

$$S(t) = \frac{1}{n}\sum_i S_i(t).$$

Функция риска $h(t)$ при этом не является аддитивной.

Для того, чтобы оценить функцию распределения случайной величины τ по набору T_n из n ее независимых реализаций t_1; … ; t_n, ищется ее наилучшее приближение кусочно-постоянной ступенчатая функцией

$$\hat{S}(t) = \prod_{t_1 \le t} \frac{R_t - E_t}{R_t},$$

где R_t – число объектов, подвергающихся риску к моменту времени t (т.е. жив и находится под наблюдением); E_t – количество эффективных событий (летальных исходов) в момент t. Заметим, что можно перемножать значения только для тех моментов времени, когда произошёл хотя бы один исход, т.е. $E_t > 0$.

Оценку точности приближения кривой выживаемости дает стандартная ошибка выживаемости, которую можно рассчитать по формуле Гринвуда:

$$\sigma_{\hat{S}} = \hat{S}(t)\left[\sum_t \frac{E_t}{R_t(R_t - E_t)}\right]^{0.5}$$

В статистической среде R анализ выживаемости реализован в пакете `survival`, который и будет использоваться ниже. В другом пакете `KMsurv` представлены различные таблицы данных, рассматриваемые в книге (Klein, Moeschberger, 2003). Рассмотрим выборку `tongue` с данными о смертности больных от злокачественной анэуплоидной (`type` = 1) или диплоидной (`type` = 2) опухоли языка. Переменная-индикатор смерти `delta` = 1 соответствует точкам отсчета времени наблюдений `time` в неделях. Выполним расчеты для анэуплоидной (`type` = 1) опухоли:

```
library(survival)
# Инсталляция пакета нужна при первом запуске !!!
install.packages("KMsurv"); library(KMsurv)
data(tongue) ; attach(tongue)
head(tongue)
# Создаем объект Surv с данными [type==1], цензурированными справа
(tongue.surv <- Surv(time[type==1], delta[type==1]))
```

```
  type time delta
1    1    1     1
2    1    3     1
```

```
3     1     3      1
4     1     4      1
5     1    10      1
6     1    13      1
```

Объект с цензурированными данными

```
[1] 1 3 3 4 10 13 13 16 16 24 26 27 28 30
...
[43] 101+ 104+ 108+ 109+ 120+ 131+ 150+ 231+ 240+ 400+
```

Функция `Surv()`, выполняющая цензурирование данных справа, в качестве параметров имеет вектор времени и индикаторный вектор, обозначающий происходящие события (0 или 1). Объект, который возвращает эта функция, используется для конструирования формулы, задаваемой функции `survfit()` для подгонки кривой

```
(tongue.fit <- survfit(tongue.surv~1))
str(tongue.fit)     # Возвращает структуру итогового объекта
summary(tongue.fit) # Вывод расчетных данные для каждого t
```

```
Вызванная функция: survfit(formula = tongue.surv ~ 1)
Суммарные величины и расчеты выживаемости для отсчетов времени
records   n.max n.start  events  median 0.95LCL 0.95UCL
     52      52      52      31      93      67      NA
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     52       1    0.981  0.0190        0.944        1.000
    3     51       2    0.942  0.0323        0.881        1.000
    4     49       1    0.923  0.0370        0.853        0.998
   10     48       1    0.904  0.0409        0.827        0.988
   13     47       2    0.865  0.0473        0.777        0.963
…
   93     18       1    0.478  0.0728        0.355        0.644
…
  157      5       1    0.305  0.0918        0.169        0.550
  167      4       1    0.229  0.0954        0.101        0.518
```

Протокол обработки данных показывает, что к 167-й неделе наблюдений умер 31 пациент из 52. Прогноз гибели 50% объектов – 93 недели (доверительный интервал средне-смертельного времени начинается от 67 недель).

Точечные значения доверительного интервала, представленные выше, являются локальными, в то время как одновременная доверительная полоса, которую можно рассчитать с использованием фунции `confBands()`, относится одновременно ко всему диапазону значений времени и, например, будет накрывать истинную кривую выживания приблизительно в 95 случаях из 100. Покажем вид всех смоделированных кривых на рис. 3.1. Мы можем также сравнить статистики Каплан-Мейера для групп больных с обеими формами опухоли, используя переменную `type` в качестве регрессора для `Surv`-объекта.

```
# Найдем статистики выживания для разных форм опухоли type
type.fit <- survfit( Surv(time, delta) ~ type )
# Функция  confBands() находится в дополнительном пакете
install.packages("OIsurv"); library(OIsurv)
tongue.cb <- confBands(tongue.surv,
                 confLevel=0.95, type="hall")
plot(tongue.fit, xlab="Время, недели",
                 ylab="Функция выживания",lwd=2)
lines(tongue.cb$time, tongue.cb$lower, lty=3,col=4,type="s")
lines(tongue.cb$time, tongue.cb$upper, lty=3,col=4,type="s")
legend("topright", legend=c("K-M функция выживания",
    "точечные интервалы","доверительная полоса"), lty=1:3,
     col=c(1,1,4), lwd=c(2,2,1))
```

```
Вызванная функция: survfit(formula = Surv(time, delta) ~ type)
Суммарные оценки выживаемости для двух типов опухоли
       records n.max n.start events median 0.95LCL 0.95UCL
type=1      52    52      52     31     93      67      NA
type=2      28    28      28     22     42      23     112
```

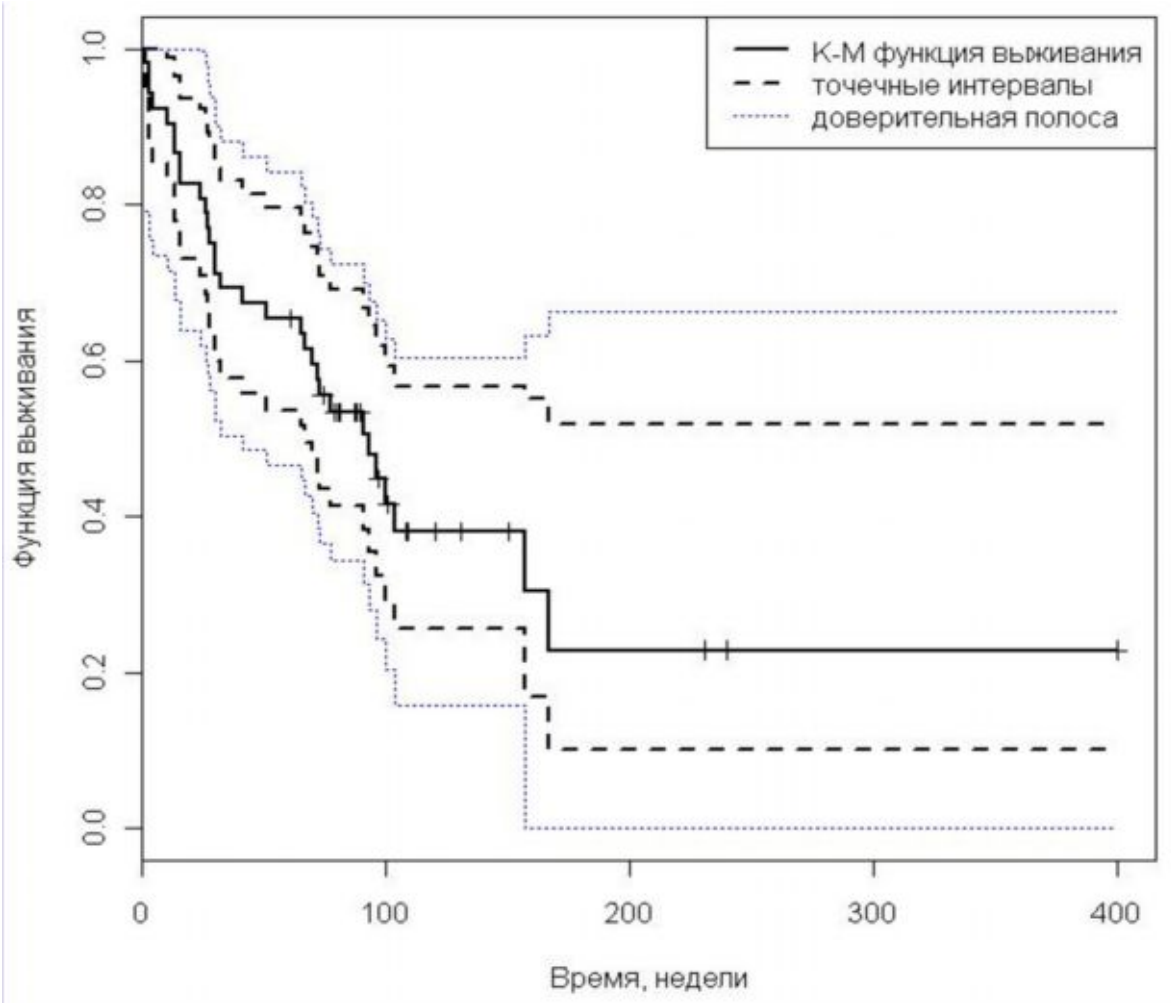


Рис. 3.1. Кривая функции выживания Каплан-Майера для больных раком языка и ее 95% доверительные интервалы

3.2. Сравнение времени жизни в двух и более группах.

После оценки средних характеристик времени жизни в двух или нескольких группах закономерна постановка вопроса о статистической значимости их отличий. При сравнении двух выборок с помощью того или иного критерия проверяется гипотеза о равенстве двух распределений времени жизни, т.е. нулевая гипотеза H_0: $F_1(t) = F_2(t)$.

Поскольку вид распределения времени жизни в сравниваемых выборках, как правило, не соответствует нормальному закону, наиболее корректным оказывается использование непараметрических критериев. Распределение в нескольких (3 и более) группах оценивается, как правило, по критерию χ^2 Пирсона, в двух группах – по другим непараметрическим критериям, семейство которых весьма обширно.

Наиболее часто используемый тест Гехана-Вилкоксона (Gehan's Wilcoxon test) рекомендуют применять в тех случаях, когда не выполняется модель пропорциональных рисков, то есть отношение интенсивностей ″исходов″ в сравниваемых группах не остается постоянным в течение всего периода наблюдения. Этот критерий наиболее эффективен, когда различия в кривых дожития наиболее выражены в начальный период наблюдения. *F*-тест позволяет осуществить проверку гипотезы о равенстве параметра интенсивности отказа в экспоненциально распределенных выборках. Критерий Кокса (Cox's) особенно чувствителен к различиям в кривых выживания, обнаруживающимся на концах распределений: это свойство может быть полезным при изучении отдаленных эффектов воздействия. Логранговый критерий (Log-rank test) рекомендуется применять, когда наблюдаемое число смертей мало. Мощность последних двух критериев максимальна при выполнении модели пропорциональных рисков.

Рассмотрим пример*, представленный в монографии М. Ньюмена (Newman, 2013, p. 178). 550 экземпляров живородящих лучепёрых рыб (*Gambusia Holbrooki*) помещали в аквариумы с соленой водой с разной концентрацией NaCl и выдерживали в течение 96 часов. Всего в опыте использовали 7 градаций солености от 0 до 20.1 г/л в 14 аквариумах с дублированием. Время смерти каждой особи фиксировалось, а экземпляры, прожившие 97 часов, считались выжившими. Таблица с исходными данными может быть загружена с сайта с указанным адресом или из файла `TOXICITY.csv`, предварительно размещенном в рабочем каталоге Вашего компьютера.

---

* Этот и многие другие примеры из книги Ньюмена, представлены Eduard Szöcs в блоге Интернет «Data in Environmental Science and Ecotoxicology» http://edild.github.io/tags/. Некоторые из них мы используем в нашем пособии.

```
# Данные можно загрузить как с ресурса github
# require(RCurl)
# url <- getURL("https://raw.github.com/EDiLD/
#     r-ed/master/quantitative_ecotoxicology/data/TOXICITY.csv",
#     ssl.verifypeer = FALSE, .opts=curlOptions(followlocation=TRUE))
# TOXICITY <- read.table(text = url, header = TRUE)
# Либо из файла, предварительно размещенного на компьютере
TOXICITY <- read.table(file = "TOXICITY.csv", header = TRUE)
# Вводим новую переменную  FLAG – индикатор гибели
TOXICITY$FLAG <- ifelse(TOXICITY$TTD > 96, 1, 2)
head(TOXICITY)
```

```
   TTD TANK  PPT WETWT STDLGTH FLAG
1    8    1 15.8 0.112     1.9    2
2    8    1 15.8 0.050     1.5    2
3    8    1 15.8 0.029     1.2    2
4    8    1 15.8 0.045     1.4    2
5    8    2 15.8 0.097     1.8    2
6    8    2 15.8 0.048     1.4    2
```

В столбцах таблицы помещены: TTD – время до смерти, TANK – номер цистерны, PPT - концентрация NaCl, г/л, WETWT и STDLGTH – вес и длина особи, FLAG = {1, 2} – особь выжила/умерла. Теперь мы можем представить данные на графике выживаемости Каплан-Майер – рис 3.2, где каждая линия соответствует аквариуму, а цвет линий определяет концентрацию NaCl.

```
mod <- survfit(Surv(TTD, FLAG) ~ PPT + strata(TANK),
               data = TOXICITY)
plot(mod, col  = rep(1:7, each=2), mark.time=FALSE,
  xlab = 'Время, час', ylab = '% выживших рыб',xlim=c(0,120))
legend('bottomright', legend = sort(unique(TOXICITY$PPT)),
       lwd=1, col = 1:7)
```

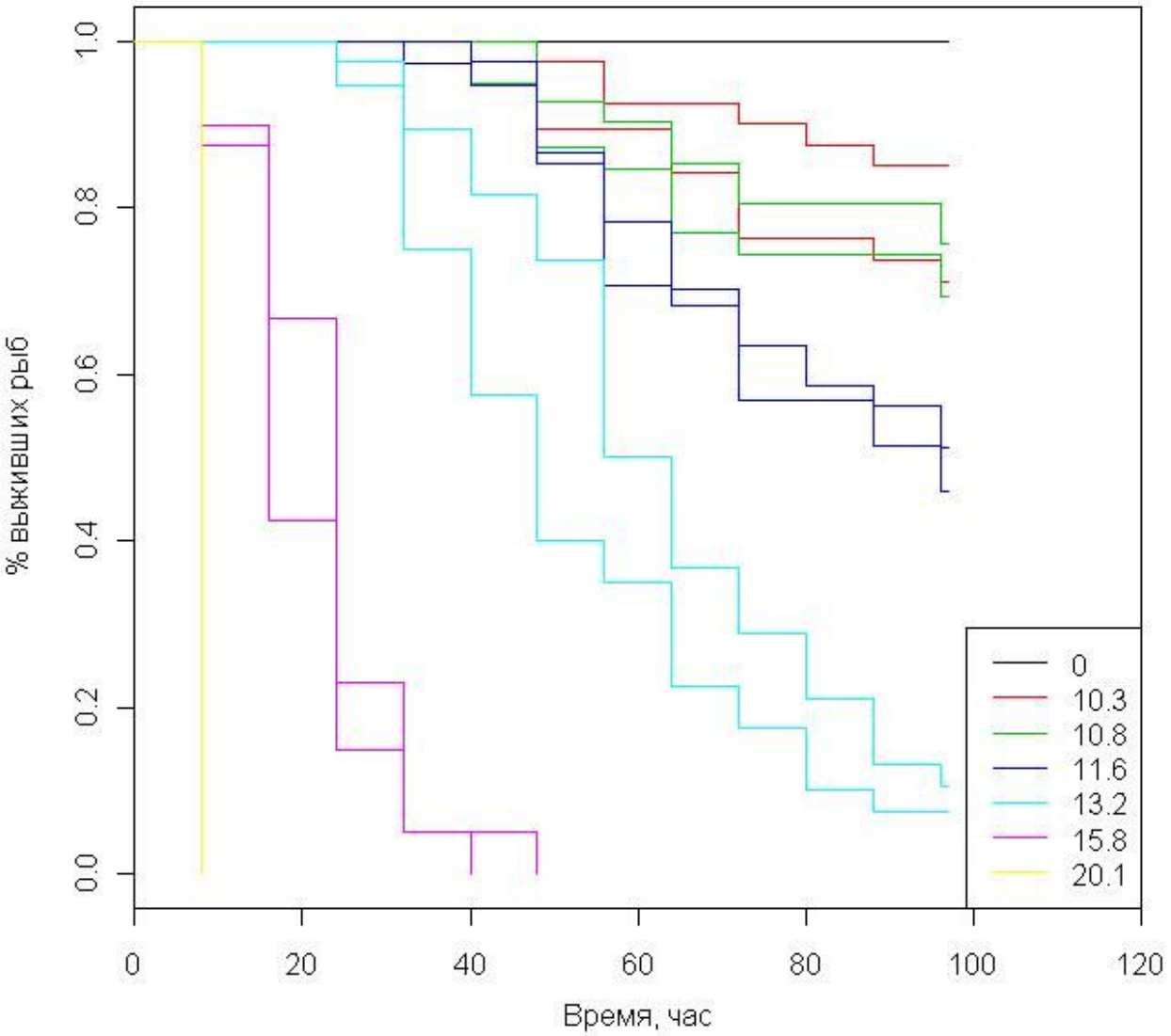


Рис. 3.2. Кривые выживаемости рыб в 14 аквариумах с 7 различными концентрациями NaCl

Мы видим на рис. 3.2 четкую зависимость кривых выживания от концентрации. Но в данном случае нас интересует, имеются ли различия в повторностях эксперимента: очевидно, что пара кривых с концентрацией 11.6 г/л достаточно близки, тогда как ситуация с дублем аквариумов 10.3 г/л не столь однозначна.

Используем проверку значимости различий с помощью методов, представленных в функции `survdiff()`, параметр `rho` которой специфицирует тип теста: при `rho = 0` применяется логранговый критерий, а при `rho = 1` – тест Гехана-Вилкоксона в модификации Пето (Peto).

```
survdiff(Surv(TTD, FLAG) ~ TANK, data =
                 TOXICITY[TOXICITY$PPT==13.2, ], rho = 0)
survdiff(Surv(TTD, FLAG) ~ TANK, data =
                 TOXICITY[TOXICITY$PPT==13.2, ], rho = 1)
# Сводные результаты по всем группам (только χ2 и p)
```

```
# Определим функцию, которая возвращает хи-квадрат
DiffTank <- function (Conc, rho = 0) {
  survdiff(Surv(TTD, FLAG) ~ TANK, data =
TOXICITY[TOXICITY$PPT==Conc, ], rho = rho)$chisq}
# Определим список концентраций
CList <- sort(unique(TOXICITY$PPT)[-c(2,7)])
print("Логранговый Тест  ChiQ/P-val" )
Chi2 <- sapply(CList, function(i) DiffTank(i))
names(Chi2) <- CList
Chi2 ; pchisq(Chi2,1,lower.tail=FALSE)
print("Тест Гехана-Вилкоксона ChiQ/P-val" )
Chi2 <- sapply(CList, function(i) DiffTank(i,1))
Chi2 ; pchisq(Chi2,1,lower.tail=FALSE)
```

```
Сравнение выживаемости в емкостях 3 и 4
Call: survdiff(Surv(TTD, FLAG) ~ TANK,… rho = 0)
        N Observed Expected (O-E)^2/E (O-E)^2/V
TANK=3 38       34     40.6      1.06       3.1
TANK=4 40       37     30.4      1.41       3.1
Хи-квадрат = 3.1  при 1 степени свободы, p= 0.0781
Call: survdiff(Surv(TTD, FLAG) ~ TANK,… rho = 1)
        N Observed Expected (O-E)^2/E (O-E)^2/V
TANK=3 38     17.8     23.5      1.37      5.02
TANK=4 40     24.9     19.2      1.68      5.02
Хи-квадрат = 5  on 1 степени свободы, p= 0.0251
Сводные результаты сравнений по всем парам емкостей
Логранговый Тест:
Конц      10.3      10.8      11.6      13.2      15.8
Хи-кв 2.2180477 0.5031704 0.1290130 3.1038589 3.0871953
P-зн  0.1364054 0.4781103 0.7194576 0.0781069 0.0789107
Тест Гехана-Вилкоксона:
Конц      10.3      10.8      11.6      13.2      15.8
Хи-кв 2.2824752 0.5955329 0.0524292 5.01862650 3.1482261
P-зн  0.1308429 0.4402874 0.8188888 0.02507605 0.0760095
```

Тест по логранговому критерию оказался статистически значимым (все *p*-значения меньше 0.05), в то время как тест Гехана-Вилкоксона для концентрации хлорида натрия 13.2 г/л отклонил нулевую гипотезу.

3.3. Модели влияния различных факторов на время жизни объектов.

Процедура Каплан-Майера оставляет открытым вопрос, какие факторы и в какой степени влияют на время жизни. Решение этой задачи возможно путем построения регрессионных моделей, имеющих для данных времени жизни свою специфику. Обычно на практике используются модель пропорциональных рисков Кокса (Cox proportional hazards model) или модели ускоренного времени AFT (Accelerated failure-time models), построенные,

исходя из некоторого предположения о теоретическом распределении времени жизни.

Основная идея модели Кокса состоит в том, что влияние экзогенных или эндогенных факторов **X** на процесс выживания соответствует умножению интенсивности смерти, существующей при стандартных условиях, на множитель, постоянный для всех t. Тогда функция риска для данных выживания с вектором ковариат **x** имеет вид:

$$h(t \mid \mathbf{x}) = h_0(t)e^{\boldsymbol{\beta}\mathbf{x}},$$

где $h_0(t)$ – функция риска смерти при стандартных условиях (baseline hazard), $\boldsymbol{\beta}$ – вектор оцениваемых параметров.

Для описанного выше примера с выживанием рыб при повышенной солености модель Кокса, учитывающую влияние двух факторов – концентрации NaCl `PPT` и массы тела рыбы `WETWT` – можно построить с использованием функции `coxph()`:

```
mod_cox <- coxph(Surv(TTD, FLAG) ~ PPT + WETWT,
           data = TOXICITY)
summary(mod_cox)
```

```
Вызванная функция:
coxph(formula = Surv(TTD, FLAG) ~ PPT + WETWT, data = TOXICITY)
  n= 480, число с наблюдаемым эффектом = 288
   (70 наблюдений удалено из-за пропусков)
Коэффициенты:
          coef exp(coef) se(coef)      z Pr(>|z|)
PPT    0.87021   2.38741  0.04451 19.550  < 2e-16 ***
WETWT -3.23134   0.03950  0.84568 -3.821 0.000133 ***
---
Коды значимости:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      exp(coef) exp(-coef) lower .95 upper .95
PPT      2.3874     0.4189   2.18795    2.6051
WETWT    0.0395    25.3135   0.00753    0.2072

Конкордантность = 0.937  (se = 0.02 )
R кВ = 0.834   (максимально возможно = 0.999 )
Тест отношения правдоподобия = 863.3 при 2 df,   p=0
Тест Вальда                  = 384.4 при 2 df,   p=0
Логранговый тест             = 388.3 при 2 df,   p=0
```

Обратим внимание на статистическую значимость коэффициентов β_1 и β_2 при обоих факторах, оцененную по z-критерию.

В последнее время удачной альтернативой модели Кокса являются модели AFT, основанные на предположении, что изменение объясняющих переменных сопряжено с изменением масштаба времени наблюдаемого состояния: ускорением или замедлением наступления момента смерти. Если обозначить как $S_0(t)$ опорную (baseline) функцию выживания, которая

описывает распределение времени жизни при отсутствии влияния внешних факторов ($\mathbf{x}$ = 0), то, при предположении, что регрессоры мультипликативно связаны с масштабом времени, будем иметь

$$S(t \mid \mathbf{x}) = S_0(te^{\boldsymbol{\beta}\mathbf{x}}).$$

Выбор экспоненты в качестве связующей функции обусловлен положительной областью значений и легкостью интерпретации результатов.

Дополнительно мы должны сделать предположение, что длительность жизни подчиняется некоторому закону распределения, чьи параметры, так или иначе, связаны с влияющими переменными. Например, если предположить, что время жизни подчиняется экспоненциальному закону с параметром $\lambda = e^{\boldsymbol{\beta}\mathbf{x}}$, то спецификация показательной регрессии в AFT-метрике будет иметь вид:

$$S(t \mid \mathbf{x}) = e^{(-te^{-\boldsymbol{\beta}\mathbf{x}})}.$$

Эта форма модели обосновывается предположением постоянства интенсивности смерти во времени (риск $h(t)$ = const).

Нормальное распределение времени жизни имеет возрастающую интенсивность смерти и является достаточно редким феноменом. Функция риска лог-нормального распределения сначала возрастает, а затем падает до нуля. Эта модель будет адекватна, в частности, для описания времени гибели экосистемы после катастрофического воздействия: если по истечении критического периода компоненты биосферы остаются живыми, то шансы последующей гибели с течением времени будут только убывать.

Если основная задача исследования состоит в изучении качественного влияния воздействующих факторов на время жизни, то выбор модели не имеет решающего значения. Тогда наиболее надежной и хорошо подгоняемой к данным часто оказывается регрессия, основанная на двухпараметрическом распределении Вейбулла, AFT-метрика которой может быть представлена в форме:

$$S(t \mid \mathbf{x}) = e^{-(te^{-\boldsymbol{\beta}\mathbf{x}})^P},$$

где параметр масштаба $\lambda = e^{-\boldsymbol{\beta}\mathbf{x}/p}$, а параметр формы p не связан с регрессорами.

В статистической среде R параметрические AFT-модели могут быть построены с использованием функции `survreg()`, аргумент `dist` которой специфицирует задаваемое распределение. Построим серию моделей с использованием четырех видов распределений: `exponential`, `weibull`, `lognorm` и `loglogistic`. Из каждой модели извлечем достигнутый логарифм правдоподобия. В расчете формируется два значения `OBJ$loglig`: без учета и с учетом ковариатов (нас интересует второе). Наилучшей модели

соответствует максимум оценки правдоподобия или минимум AIC-критерия, вычисляемого функцией `extractAIC()`:

```
mod_exp <- survreg(Surv(TTD, FLAG) ~ PPT + WETWT,
            data = TOXICITY, dist = 'exponential')
mod_wei <- survreg(Surv(TTD, FLAG) ~ PPT + WETWT,
            data = TOXICITY, dist = 'weibull')
mod_lnorm <- survreg(Surv(TTD, FLAG) ~ PPT + WETWT,
            data = TOXICITY, dist = 'lognorm')
mod_llog <- survreg(Surv(TTD, FLAG) ~ PPT + WETWT,
            data = TOXICITY, dist = 'loglogistic')
# извлечение loglik и заполнение таблицы
df <- data.frame(
     model = c('Cox','exp', 'weibull', 'lnorm', 'loglog'),
        logLik = c(mod_cox$loglik[2],
                   mod_exp$loglik[2],
                   mod_wei$loglik[2],
                   mod_lnorm$loglik[2],
                   mod_llog$loglik[2]))
# добавление AIC в столбец таблицы
df$AIC <- c(extractAIC(mod_cox)[2],
            extractAIC(mod_exp)[2],
            extractAIC(mod_wei)[2],
            extractAIC(mod_lnorm)[2],
            extractAIC(mod_llog)[2])
df
```

```
        Модель   logLik      AIC
     1     Cox -1234.790 2473.581
     2     exp -1309.719 2625.439
     3 weibull -1114.330 2236.659
     4   lnorm -1121.756 2251.511
     5  loglog -1118.090 2244.180
```

Модель Вейбулла имеет минимальный AIC-критерий из всех пяти протестированных моделей. Функцией `summary()` можно получить оценки коэффициентов этой модели и их стандартные ошибки. Мы можем также получить график, где модельные кривые функций выживания, полученные из распределения Вейбулла для каждого уровня концентраций, совмещены с кривыми Каплан-Майера – рис. 3.3.

```
summary(mod_ wei)
# График Каплан-Майера
km <- survfit(Surv(TTD, FLAG) ~ PPT, data = TOXICITY)
plot(km, lty = 1:7 , xlab = 'Время, час',
          ylab = '% выживших рыб', xlim=c(0,120))
# Подгоняем модель только для концентраций
mod_ppt <- survreg(Surv(TTD, FLAG) ~ PPT,
          data = TOXICITY, dist = 'weibull')
# Добавляем модельные кривые на график
PPT_u <- sort(unique(TOXICITY$PPT))
```

```
for(i in seq_along(PPT_u)){
  lines(predict(mod_ppt,
        newdata = list(PPT=PPT_u[i]),type = "quantile",
        p = 1:99/100), 99:1/100, col=i, lwd=2)
}
legend('bottomright', legend = PPT_u,  col = 1:7, lwd=2)
```

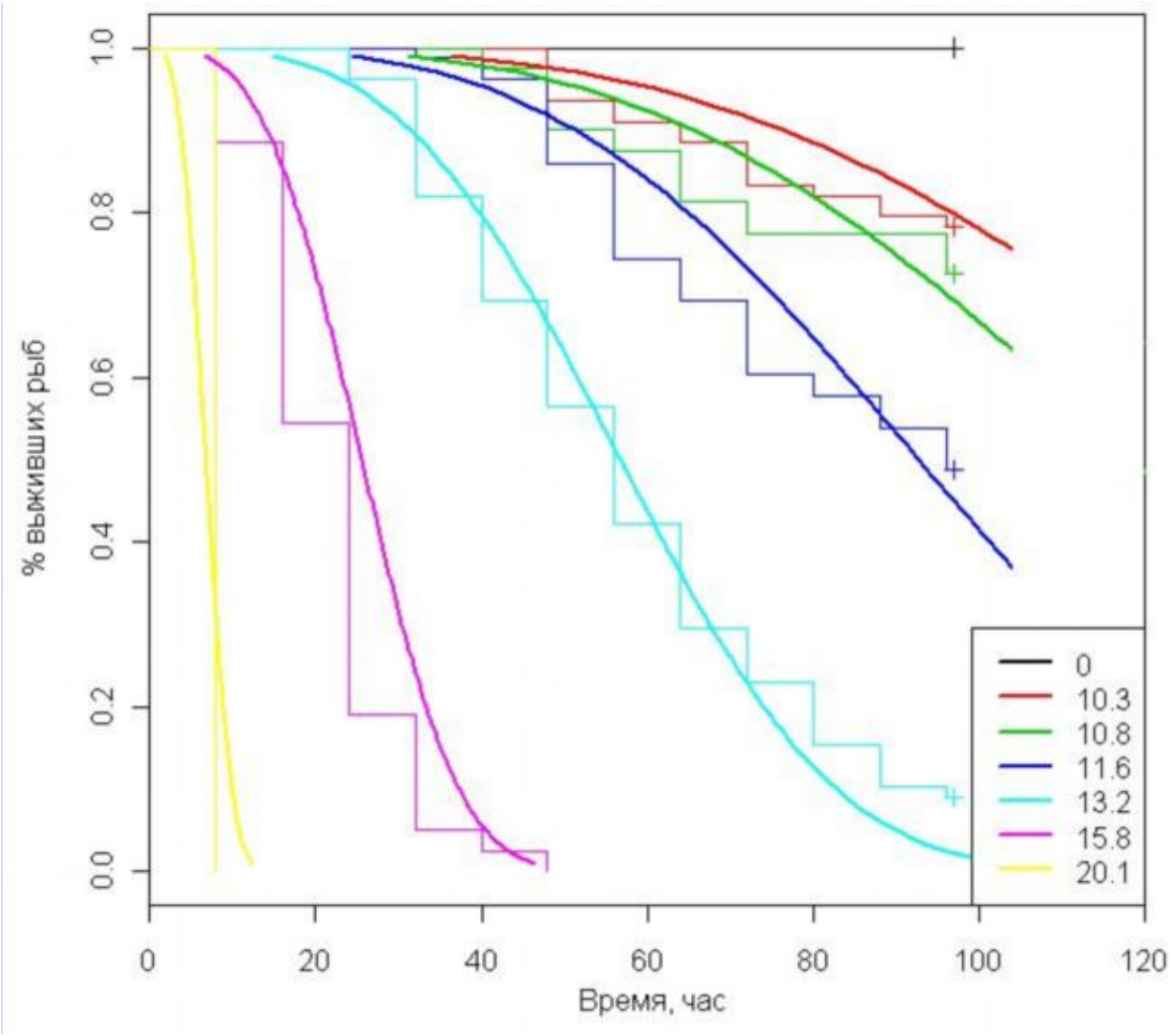


Рис. 3.3. Кривые выживаемости с различными концентрациями NaCl по методу Каплан-Майер и с использованием регрессии Вейбулла

```
Вызванная функция:  survreg(formula =
                Surv(TTD, FLAG) ~ PPT + WETWT,
                data = TOXICITY, dist = "weibull")
Коэффициенты модели Вейбулла:
              Value Std. Error      z         p
(Intercept)  7.849    0.08449  92.89  0.00e+00
PPT         -0.295    0.00512 -57.54  0.00e+00
WETWT        1.066    0.25604   4.16  3.15e-05
Log(scale)  -1.189    0.04455 -26.69 6.09e-157
```

```
Scale= 0.305
Распределение Вейбулла
Лог. функции правдоподобия (для модели)      = -1114.3
Лог. функции правдоподобия (только св.члена) = -1606.3
        Chisq= 983.9 при 2 степенях свободы, p= 0
Число итераций Ньютона-Рафсона: 9
n=480 (70 наблюдений удалены из-за пропусков)
```

По построенной модели можно выполнить прогноз совокупности среднемедианных значений времени жизни особей с различной массой тела. Для этого используется функция `expand.grid()`, генерирующая все возможные комбинации веса рыб от 0 до 1.5 г для каждого из 4-х уровней солености. На основе этих данных функция `predict()` выполняет расчет медиан, что обуславливается задаваемыми параметрами `type='quantile', p = 0.5`. Вывод графика на рис. 3.4 осуществляется с использованием графической системы `ggplot2`, которую большинство пользователей R считают весьма продвинутой и удобной в реализации.

```
# Новые данные: все комбинации признаков WETWT и PPT
newtox <- expand.grid(WETWT = seq(0, 1.5, length.out=100),
            PPT = seq(12.5, 20, by = 2.5))
# прогноз среднего всемени жизни
newtox$preds <- predict(mod_wei, newdata=newtox,
             type='quantile', p = 0.5)
# Вывод графика
require(ggplot2)
ggplot(newtox, aes(x = WETWT, y = preds,
        linetype = factor(PPT), group = PPT)) +
        geom_line(size = 2) +
        theme_bw() +
        labs(x = 'Вес (г)',
        y = 'Среднемедианное время жизни (час)') +
        coord_cartesian(ylim=c(0,100))
```

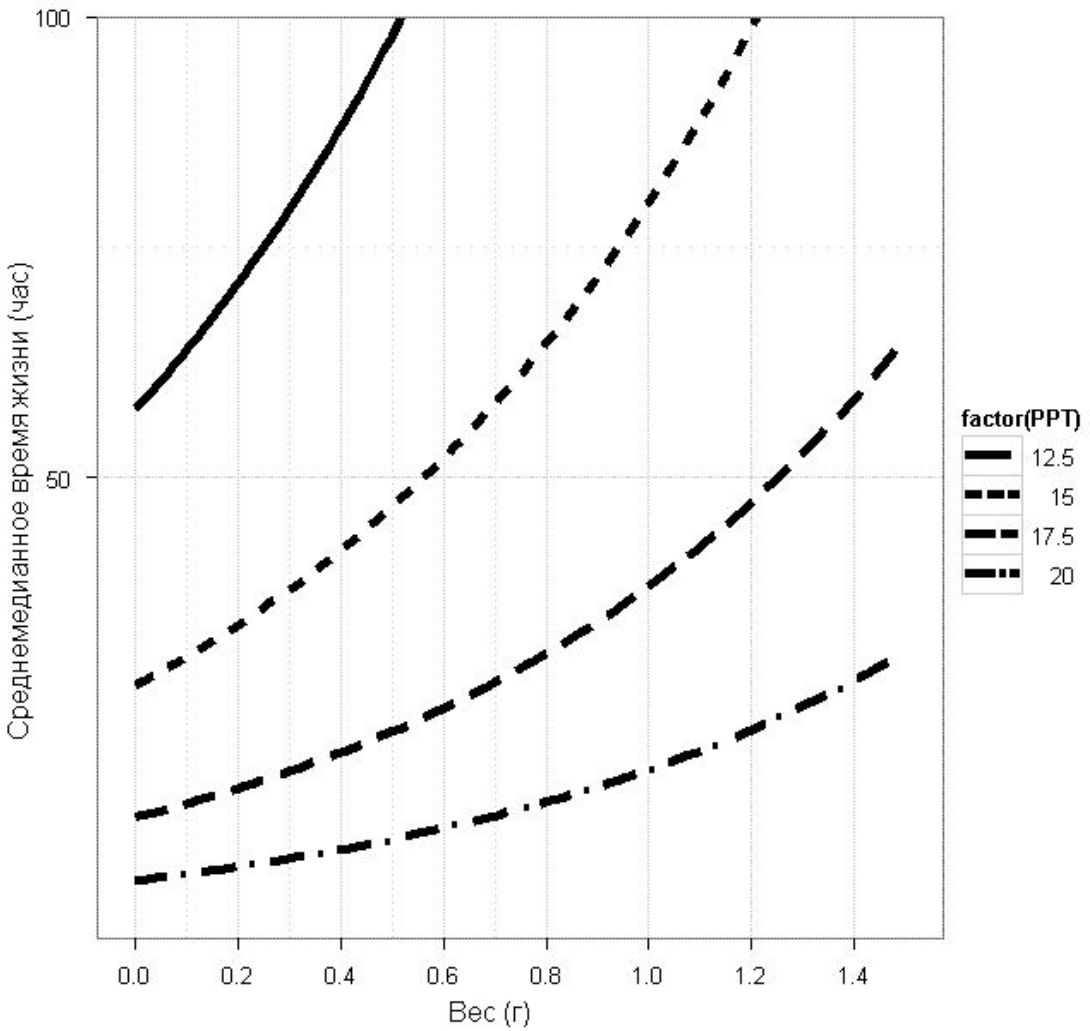


Рис. 3.4. Зависимость среднего времени жизни от массы тела рыб при различных концентрациях солености PPT

Общий итог статистического и графического анализа: увеличение солености воды отрицательно сказывается на времени жизни рыб *Gambusia Holbrooki* , причем этот эффект сильнее сказывается на рыбах с меньшей массой тела.

3.4. Биоаккумуляция

С позиций токсикокинетики организм представляет собой сложную гетерогенную систему, состоящую из большого числа компартментов (отделов): кровь, ткани, внеклеточная жидкость, отделенных друг от друга биологическими барьерами. Токсикокинетические модели описывают закономерности, а также устанавливают качественные и количественные характеристики процессов резорбции (т.е. проникновения вещества в организм через желудочно-кишечный тракт, дыхательные пути, кожный покров), распределения ксенобиотиков внутри и между компартментами и их элиминации.

Под элиминацией понимают процесс, приводящий к снижению концентрации веществ в крови, органах и тканях путем экскреции (выведения вещества из организма в окружающую среду) и биотрансформации (химических превращений молекул ксенобиотика в ходе метаболизма). Рассмотрим простейшую однокамерную модель элиминации, которая является основой для оценки процессов аккумуляции вещества в организме.

Пусть C_t – концентрация ксенобиотика в компартменте (масса, отнесенная к его объему) в момент времени t, k_u и k_e, – константы скорости поглощения и элиминации, мл/(г·час), и C_1 – концентрация (константа) вещества в окружающей среде. Если предположить кинетику первого порядка процесса аккумуляции при постоянной скорости поглощения k_u и элиминации k_e, то можно записать уравнение изменения концентрации во времени:

$$dC/dt = -k_eC\ ;\ \ dC/dt = k_uC_1 - k_eC,$$

или в интегральном виде:

$$C_t = \frac{k_u}{k_e}C_1(1-e^{-k_et})\ .$$

Рассмотрим пример (Newman, 2013, p. 108) биоаккумулирования инсектицида бромофоса из водной среды рыбками гуппи (*Poecilia reticulata*) по эксперименту De Bruijn и Hermens (1991). Сначала рыб поместили на 264 часа в аквариум, содержащий раствор бромофоса C_1 = 10.5 нг/л, а затем переместили в чистую водную среду. На всех фазах эксперимента измеряли концентрацию бромофоса BRPHOS в жировых тканях рыб (нанограмм/г извлекаемого жира).

Вначале оценим константы скорости по первой стадии эксперимента, где имеет место одновременно поглощение и элиминация. Используем функцию построения нелинейных моделей `nls()`:

```
ACCUM <- matrix(ncol=2, byrow = TRUE, c(
0.5,1900,1,3000,2,5200,4,6900,8,24000,24,50000,
72,200000,144,400000,240,500000,264,500000),
dimnames=list(1:10,c("HOUR","BRPHOS")))
(ACCUM <- as.data.frame(ACCUM))
mod_accum <- nls(BRPHOS ~ KU/KE*10.5 * (1 – exp(-KE*HOUR)),
       data = ACCUM, start = list(KU = 100, KE = 0.01))
summary(mod_accum)
AIC(mod_accum)
```

```
     HOUR BRPHOS
1     0.5   1900
2     1.0   3000
3     2.0   5200
4     4.0   6900
5     8.0  24000
6    24.0  50000
7    72.0 200000
```

```
   8  144.0 400000
   9  240.0 500000
   10 264.0 500000
```

```
Модель кинетики 1-го порядка поглощения-элиминации
Формула: BRPHOS ~ KU/KE * 10.5 * (1 - exp(-KE * HOUR))
Коэффициенты:
      Оценка  Ст.ошибка  t крит Pr(>|t|)
KU 3.448e+02  3.186e+01  10.824 4.69e-06 ***
KE 5.249e-03  1.032e-03   5.088 0.000944 ***
---
Коды значимости: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Ст. отклонение остатков: 18940 при 8 степенях свободы
AIC    229.1299
```

Получили коэффициенты скорости поглощения k_u = 344.8 ± 31.9 и элиминации k_e = 0.00523 ± 0.00103. Тогда скорость накопления бромофоса будет осуществляться по уравнению:

$$C_t = 334.8/0.00523\ C_1(1 - e^{-0.00523t}) = 672\cdot10^3\cdot(1 - e^{-0.00523t})\ \text{нг/г}$$

Оценим теперь константу скорости элиминации по второй фазе эксперимента.

```
# Данные по выведению гербицида
ELIMIN <- matrix(ncol=2, byrow = TRUE, c(
0,500000,12,450000,24,370000,48,290000,72,190000,96,150000,
144,70000,216,21000,319,5000),
dimnames=list(1:9,c("HOUR","BRPHOS")))
(ELIMIN <- as.data.frame(ELIMIN))
ELIMIN$LBROMO <- log(ELIMIN$BRPHOS)
# Строим линеаризованную модель в логарифмах
mod_elimin_lm <- lm(LBROMO ~ HOUR, data = ELIMIN)
summary(mod_elimin_lm)
# Коэффициент детерминации для обычной линейной модели
summary(lm(BRPHOS ~ HOUR, data = ELIMIN))$r.squared
# Вывод графика линеаризированной модели
plot( ELIMIN$HOUR, ELIMIN$LBROMO, type="p", xlab="Часы",
                ylab="Логарифм концентрации")
matplot(ELIMIN$HOUR, predict(mod_elimin_lm,
 interval="confidence"),type="l", lwd=c(2,1,1), lty=c(1,2,2),
 col = c(1,2,2), add = TRUE)
```

```
   HOUR BRPHOS    LBROMO
1     0 500000 13.122363
2    12 450000 13.017003
3    24 370000 12.821258
4    48 290000 12.577636
5    72 190000 12.154779
6    96 150000 11.918391
7   144  70000 11.156251
8   216  21000  9.952278
9   319   5000  8.517193
```

```
Линейная модель: отклик – логарифм концентрации
lm(formula = LBROMO ~ HOUR, data = ELIMIN)
Коэффициенты:
                Оценка  Ст.ошибка  t крит Pr(>|t|)
(Св. член)  13.2126182  0.0360403  366.61 2.97e-16 ***
HOUR        -0.0146900  0.0002503  -58.69 1.09e-10 ***
---
Коды значимости: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Ст. отклонение остатков: 0.07521 при 7 степенях свободы
Множеств. R-квадрат: 0.998,     Приведенный R-квадрат: 0.9977
F-статистика:  3444 при 1 и 7 DF,  p-значение: 1.095e-10
Линейная модель: концентрация без логарифмирования
R-квадрат: 0.7876746
```

Таким образом, найденная константа скорости элиминации бромофоса рыбками в чистой среде $k_e = 0.0147 \pm 0.0003$, что существенно выше ранее вычисленного значения для модели кумуляции. Легко убедиться с использованием графика на рис. 3.5 и путем сравнения значений коэффициентов детерминации, что на самом деле скорость элиминации постоянна относительно логарифма концентрации бромофоса в тканях рыб, а не ее натуральных значений.

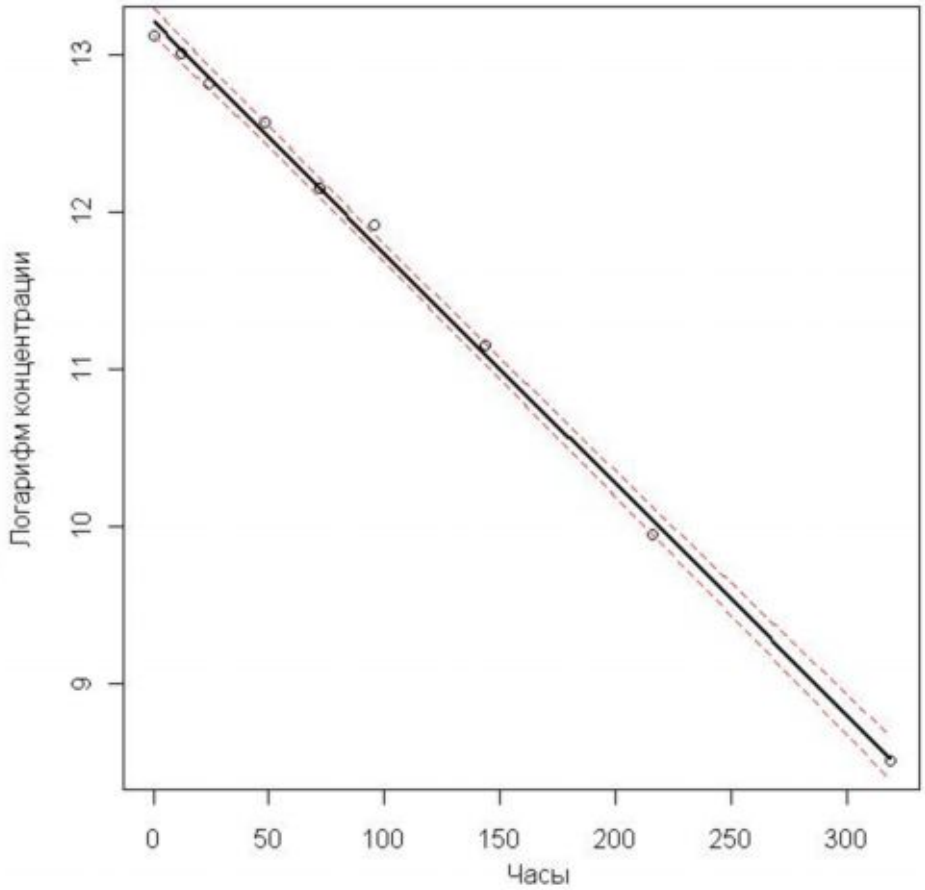


Рис. 3.5. Скорость выведения бромофоса из тела гуппи

Вычислим теперь коэффициенты модели аккумуляции с использованием найденной нами на втором этапе эксперимента константы скорости элиминации. Построим графики накопления бромофоса рыбками гуппи для обоих случаев.

```
mod_accum2 <- nls(BRPHOS ~ KU / -coef(mod_elimin_lm)[2]
      * 10.5 * (1 - exp(coef(mod_elimin_lm)[2] * HOUR)),
      data = ACCUM,start = list(KU = 100))
summary(mod_accum2)
AIC(mod_accum2)
# Выводим графики обеих моделей
HOUR_pred <- seq(min(ACCUM$HOUR), max(ACCUM$HOUR), by = 0.1)
plot(ACCUM, xlab="Часы", "Концентрация гербецида")
lines(HOUR_pred, predict(mod_accum, newdata =
       data.frame(HOUR = HOUR_pred)), lwd=2)
lines(HOUR_pred, predict(mod_accum2, newdata =
       data.frame(HOUR = HOUR_pred)),col=3, lwd=2)
legend("bottomright",c("Модель по 1-й фазе", "С подстановкой
          скорости элиминации"),col = c(1,3), lwd=2)
```

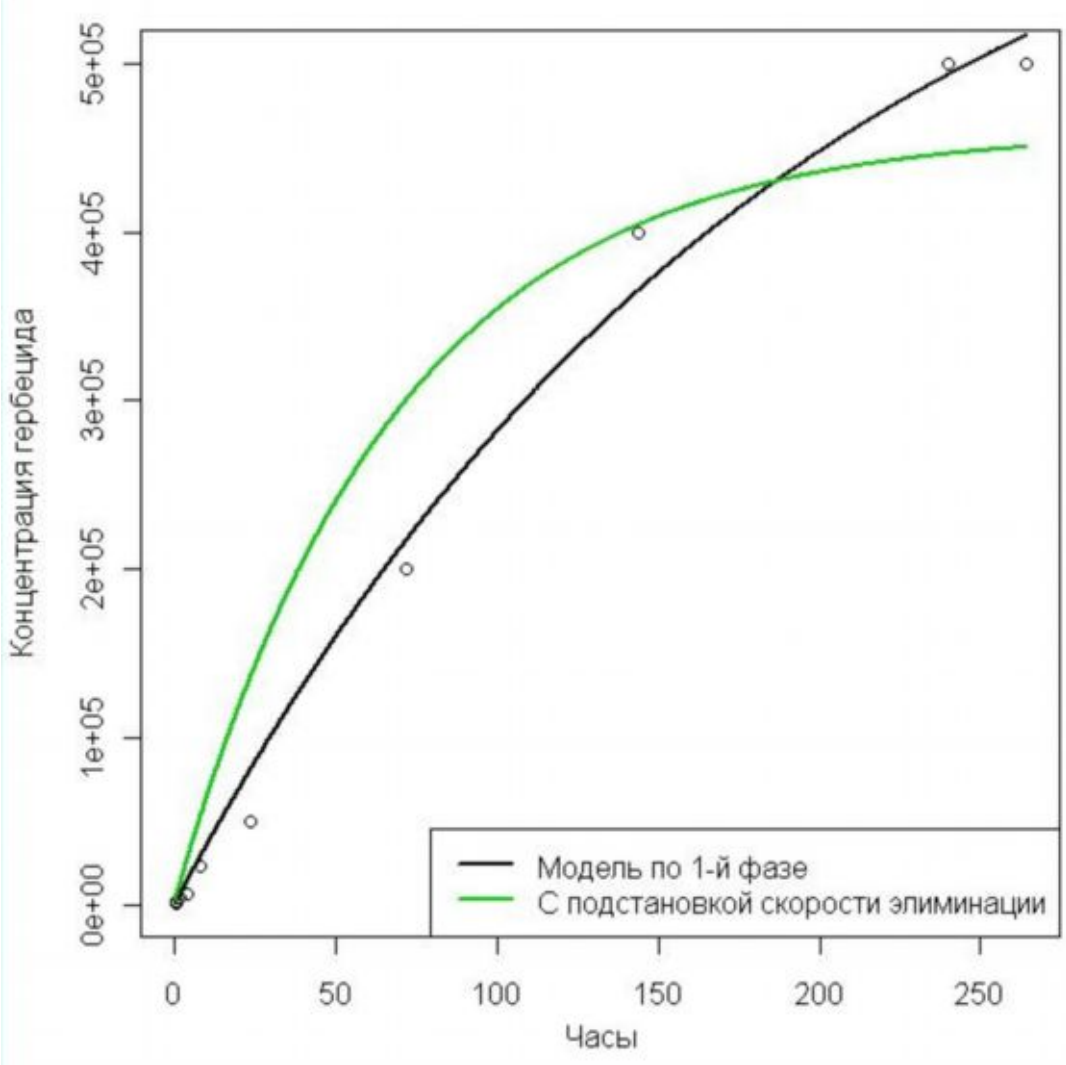


Рис. 3.6. Две модели аккумуляции бромофоса

```
Форула: BRPHOS ~ KU/-coef(mod_elimin_lm)[2] * 10.5 *
          (1 - exp(coef(mod_elimin_lm)[2] * HOUR))
Коэффициенты:
     Оценка  Ст.ошибка  t крит Pr(>|t|)
KU   643.95      40.43   15.93 6.69e-08 ***
---
Ст. отклонение остатков: 51730 при 9 степенях свободы
AIC    248.4029
```

Необходимо отметить, что по статистическим параметрам (стандартной ошибке регрессии и AIC-критерию) первая модель оказалась лучше второй. М.Ньюман подробно останавливается на смысле этих отличий и объясняет, почему 2-я модель, построенная на использовании "чистой" элиминации, все же является предпочтительней.

4. ОЦЕНКА ПАРАМЕТРОВ ТОКСИКОМЕТРИИ ДЛЯ ЭФФЕКТОВ В МЕТРИЧЕСКОЙ ФОРМЕ

4.1. Счетные данные и регрессия Пуассона

Особое место в статистике занимают счетные данные (англ. count data), которые часто подчиняются отрицательному биномиальному распределению или распределению Пуассона. Они широко распространены в экологии, например, при оценке численности популяций, когда практически всегда сталкиваются с неравномерностью распределения особей в пространстве. Если изучается популяция редкого вида, то в большинстве случаев можно не встретить ни одной особи, реже попадутся единичные особи, и лишь иногда – их большие скопления. Для анализа характера распределения данных такого рода не подходят обычные линейные модели, предполагающие нормальное распределение остатков.

Распределение Пуассона имеет случайная величина Y, равная количеству событий, произошедших за некоторый промежуток времени, если эти события независимы и происходят с постоянной интенсивностью λ, $\lambda \in (0, \infty)$. Предполагается, что среднее $E(Y) = \lambda$ и дисперсия $Var(Y) = \lambda$, а функция плотности вероятности $p(k) = e^{-\lambda} \lambda^k/k!$, $k = 1, 2, \ldots, \infty$, где $k!$ – факториал от числа событий.

Пуассоновская регрессия применяется, если отклик является счетной переменной (например, численность организмов некоторого вида в пробе) и подразумевается, что Y имеет пуассоновское распределение и что можно подогнать линейную модель вида

$$\ln(\lambda) = \beta_0 + \sum_p \beta_j x_j$$

где $x_1,\ldots, x_p$ – набор из p независимых переменных, β_0 – математическое ожидание Y при равенстве нулю всех предикторов x_j, β_j – коэффициенты при независимых переменных. Эта регрессионная модель, как и модель логита в разделе 2.2, может быть подогнана под выборочные данные путем нахождения максимума правдоподобия при помощи команды

```
glm(Y~X1+…, family=poisson(link="log"), data=mydata)
```

Рассмотрим в качестве примера таблицу `nitrofen`, представленную в пакете `boot` и состоящую из 50 строк и 5 столбцов. Нитрофен является эффективным гербицидом, относительно нетоксичным к взрослым млекопитающим, но оказывающим тератогенное и мутагенное действие. В эксперименте по выявлению репродуктивной токсичности в 50 емкостей было помещено по 10 экземпляров ветвистоусых рачков (*Ceriodaphnia dubia*) и внесены 5 градаций дозы нитрофена до концентраций `conc` от 0 до 310 мкг/л. Через 7-дневные периоды регистрировалась численность животных первой, второй и третей генерации `brood1`, `brood2` и `brood3` соответственно.

```
library(boot)
data(nitrofen)
attach(nitrofen)
head( nitrofen)
```

```
   conc brood1 brood2 brood3 total
1     0      3     14     10    27
2     0      5     12     15    32
3     0      6     11     17    34
…
48  310      4      0      0     4
49  310      6      0      0     6
50  310      5      0      0     5
```

Построим три модели Пуассона, используя в качестве отклика численность после каждой генерации `brood1`, `brood2`, `brood3`, а в качестве независимой переменной – концентрацию нитрофена `conc`.

```
summary(m1 <- glm(brood1 ~ conc, family="poisson"))
with(m1, cbind(res.deviance = deviance, df = df.residual,
      p = pchisq(deviance, df.residual, lower.tail=FALSE)))
summary(m2 <- glm(brood2 ~ conc, family="poisson"))
with(m2, cbind(res.deviance = deviance, df = df.residual,
      p = pchisq(deviance, df.residual, lower.tail=FALSE)))
summary(m3 <- glm(brood3 ~ conc, family="poisson"))
with(m3, cbind(res.deviance = deviance, df = df.residual,
     p = pchisq(deviance, df.residual, lower.tail=FALSE)))
```

```
   Модель численности через 7 дней
   glm(formula = brood1 ~ conc, family = "poisson")
Коэффициенты:
                 Оценка  Ст.ошибка  z крит Pr(>|z|)
 (Св. член)   1.6631592  0.1076030  15.456   <2e-16 ***
conc         -0.0000193  0.0005625  -0.034    0.973
Тест девианса остатков:  res.deviance df         p
                              27.25977 48 0.9930985
   Модель численности через 14 дней
   glm(formula = brood2 ~ conc, family = "poisson")
Коэффициенты:
                 Оценка  Ст.ошибка z крит Pr(>|z|)
 (Св. член)   2.7207223  0.0699717  38.883   <2e-16 ***
conc         -0.0047367  0.0004841  -9.784   <2e-16 ***
Тест девианса остатков:  res.deviance df           p
                              169.1091 48 2.11564e-15
   Модель численности через 21 день
   glm(formula = brood3 ~ conc, family = "poisson")
Коэффициенты:
                 Оценка  Ст.ошибка z крит Pr(>|z|)
 (Св. член)   2.8953833  0.0645940   44.82   <2e-16 ***
conc         -0.0051860  0.0004607  -11.26   <2e-16 ***
Тест девианса остатков:  res.deviance df           p
                              130.6729 48 1.39344e-09
```

Выше представлена информация о коэффициентах моделей, из которой следует, что в случае первой генерации влияние концентрации нитрофена статистически незначимо (коэффициент β_0 близок к нулю). Из двух остальных моделей `m2` и `m3` явствует, что при увеличении концентрации нитрофена на 100 мкг/л логарифм численности воспроизводства цериодафний уменьшается в среднем на 0.47 и 0.52 соответственно.

Однако для проверки, насколько хорошо была выполнена подгонка модели под данные, используем тест на отношение девианс G^2 (likelihood ratio test), который эквивалентен F-статистике для обычных регрессионных моделей. Как показано в разделе 2.3, остаточный девианс (Residual deviance) G^2 представляет собой разность суммарных значений логарифмов функции правдоподобия для текущей LL_M и идеальной модели LL_S, где предсказанные значения идентичны наблюдаемым. Величина G^2 распределена как χ^2 с числом степеней свободы df, равным разности количества параметров моделей LL_M и LL_S:

$$G^2 = -2\,(LL_M - LL_S\,) = -\,2(LL_M/LL_S) \sim \chi^2_{df}$$

Таким образом, если остаточный девианс будет достаточно мал, как это имеет место для модели `m1`, то пуассоновская регрессия хорошо описывает имеющиеся данные. В противном случае, если тест G^2 статистически значим (что имеет место для регрессий `m2` и `m3`), то это указывает на недостаточно хорошее соответствие данных выбранной модели. В этой ситуации можно попытаться оценить: *а*) нет ли важных переменных, пропущенных в анализе, *б*) верно ли предположение о линейности модели или *в*) существует ли проблема избыточной дисперсии.

Для проверки предположения *б*) построим графики зависимостей:

```
# Подготавливаем данные для прорисовки кривых: 100 пар
# значений на шкале концентрации и прогнозы по моделям
newxs <- seq(0, max(conc),length.out = 100)
pred_b1 <- predict(m1, data.frame(conc = newxs), type="response")
pred_b2 <- predict(m2, data.frame(conc = newxs), type="response")
pred_b3 <- predict(m3, data.frame(conc = newxs), type="response")
plot(conc, brood3, cex=0.7, pch=19, xlab="Концентрация
      нитрофена", ylab = "Численность новорожденных рачков")
lines (newxs,pred_b3, lwd=2)
points (conc, brood2, cex=0.7, pch=21, bg="grey")
lines (newxs,pred_b2, lty=4, lwd=2)
points (conc, brood1, cex=0.7)
lines (newxs,pred_b1, lty=3, lwd=2)
legend("topright",c("1 генерация", "2 генерация",
          "3 генерация"), lty=c(3,4,1), lwd=2)
```

На рис. 4.1 нетрудно заметить, что основная модель `m3` недостаточно адекватна относительно данных: почти все экспериментальные точки в средней части кривой находятся выше нее, что свидетельствует о существенной нелинейности искомой функции регрессии.

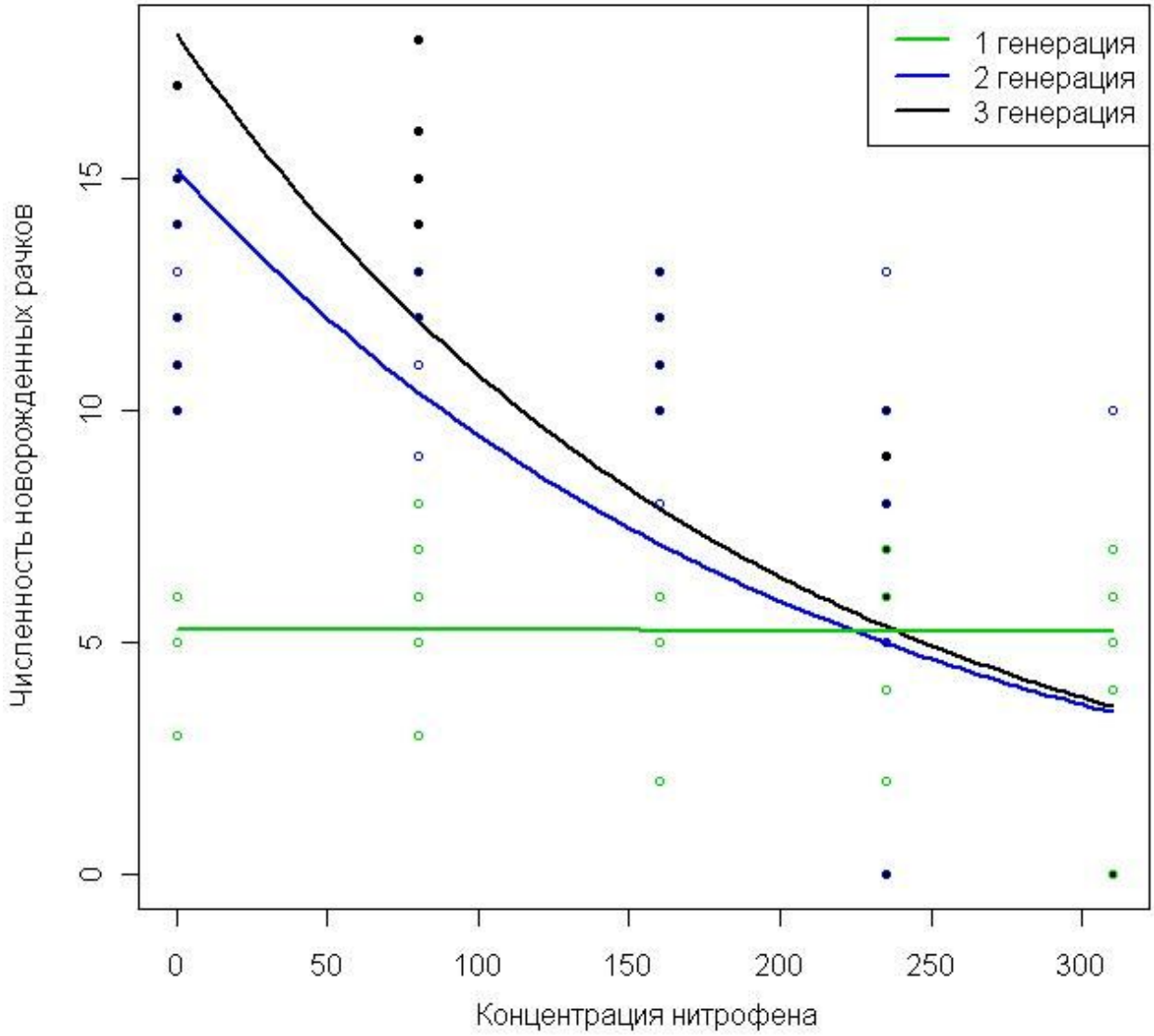


Рис. 4.1. Графики моделей пуассоновской регрессии для трех генераций цериодафнии

Выполним построение нелинейной пуассоновской регрессии, используя полином второй степени от концентрации нитрофена:

```
summary(m3p <- glm(brood3 ~ conc + I(conc^2),
                   family="poisson"))
with(m3p, cbind(res.deviance = deviance, df = df.residual,
  p = pchisq(deviance, df.residual, lower.tail=FALSE)))
```

```
Аппроксимация полиномом второй степени
glm(formula = brood3 ~ conc + I(conc^2), family = "poisson")
Коэффициенты:
                 Оценка   Ст.ошибка   z крит Pr(>|z|)
 (Св. член)    2.561e+00   8.438e-02   30.347  < 2e-16 ***
conc           7.140e-03   1.571e-03    4.546 5.47e-06 ***
I(conc^2)     -4.908e-05   6.196e-06   -7.921 2.35e-15 ***
    Null-девианс: 269.794 при 49  степенях свободы
Девианс остатков:  56.202 при 47  степенях свободы
AIC: 229.76
```

```
Тест девианса остатков:  res.deviance df         p
                              56.20224 47 0.1682243
```

Для этой модели остаточный девианс уже статистически значимо не отличается от 0. Предположим также, что средний эффект E_{50} от угнетения репродуктивной функции равен медиане предикторных значений числа особей. Доверительный интервал среднеэффективной величины рассчитаем на основе стандартной ошибки, возвращаемой функцией `predict()`. Значения среднеэффективной концентрации EC_{50} и ее доверительных интервалов рассчитаем интерполяцией с использованием функции `approx()`.

```
pred_b3p <- predict(m3p,newdata = data.frame(conc = newxs),
         type="response", se=TRUE)
E50l <- median(pred_b3p$fit)
std.err <- min(pred_b3p$se.fit [pred_b3p$fit>E50l])
E50 <- c(E50l, E50l - 1.96 * std.err,E50l + 1.96 * std.err)
names (E50) <- c("Оценка","2.5 %  LCL","97.5 %  UCL")
EC50<-approx(pred_b3p$fit, newxs, xout = E50)$y
rbind(E50,EC50)
```

```
Среднеэффект. концентрация с доверительными интервалами
        Оценка 2.5 % LCL 97.5 % UCL
E50   12.03905  10.45641  13.62169
EC50 155.00000 170.91872 123.03814
```

Хорошим вариантом выяснения, насколько корректно была выбрана нами аппроксимирующая функция, является сопоставление ее с моделью сглаживания. Выполним построение обобщенной аддитивной модели GAM, описанной в разделе 2.5, и терм `s(conc, df=3)` в функции `gam()` определит нам явно задаваемое число степеней свободы `df` (Wood, 2006).

```
library(gam)
summary(m3.gam <- gam(brood3~s(conc,df=3), poisson))
with(m3.gam, cbind(res.deviance = deviance, df = df.residual,
  p = pchisq(deviance, df.residual, lower.tail=FALSE)))
```

```
Обобщенная аддитивная модель
Вызываемая функция: gam(formula = brood3 ~ s(conc, df = 3),
                                        family = poisson)
DF членов и Хи-квадрат для непараметрических эффектов:
                 Df Npar Df Npar Chisq   P(Chi)
(Св. член)        1
s(conc, df = 3)  1       2     68.811 1.11e-15 ***
    Null-девианс: 269.7944 при 49 степенях свободы
Девианс остатков:  40.8619 при 46 степенях свободы
AIC: 216.4202
Тест на величину остаточного девианса
     res.deviance df         p
         40.86187 46 0.6867997
```

Обратим внимание, что все статистические показатели модели GAM (остаточный девианс, AIC-критерий, значимость по χ^2) существенно лучше, чем у обеих моделей Пуассона m3 и m3p. Смысл отличий легко уточнить, построив графики зависимостей ″доза-эффект″:

```
#  Подготавливаем данные для прорисовке графика
newxs <- seq(0, max(conc),length.out = 100)
pred_b3 <- predict(m3,newdata = data.frame(conc = newxs),
           type="response")
pred_b3p <- predict(m3p,newdata = data.frame(conc = newxs),
           type="response")
pred_b3gam <- predict(m3.gam,newdata = data.frame(conc =
             newxs) , type="response")
plot(conc, brood3, cex=0.7, pch=19, xlab="Концентрация
      нитрофена",ylab = "Численность новорожденных рачков")
lines (newxs,pred_b3, col=3, lwd=2)
lines (newxs,pred_b3p, col=4, lwd=2)
lines (newxs,pred_b3gam, col=2, lwd=2)
legend("bottomleft",c("Линейная Пуассона", "Квадратичная
   Пуассона", "Сплайн-интерполяция"), col=c(3,4,2), lwd=2)
```

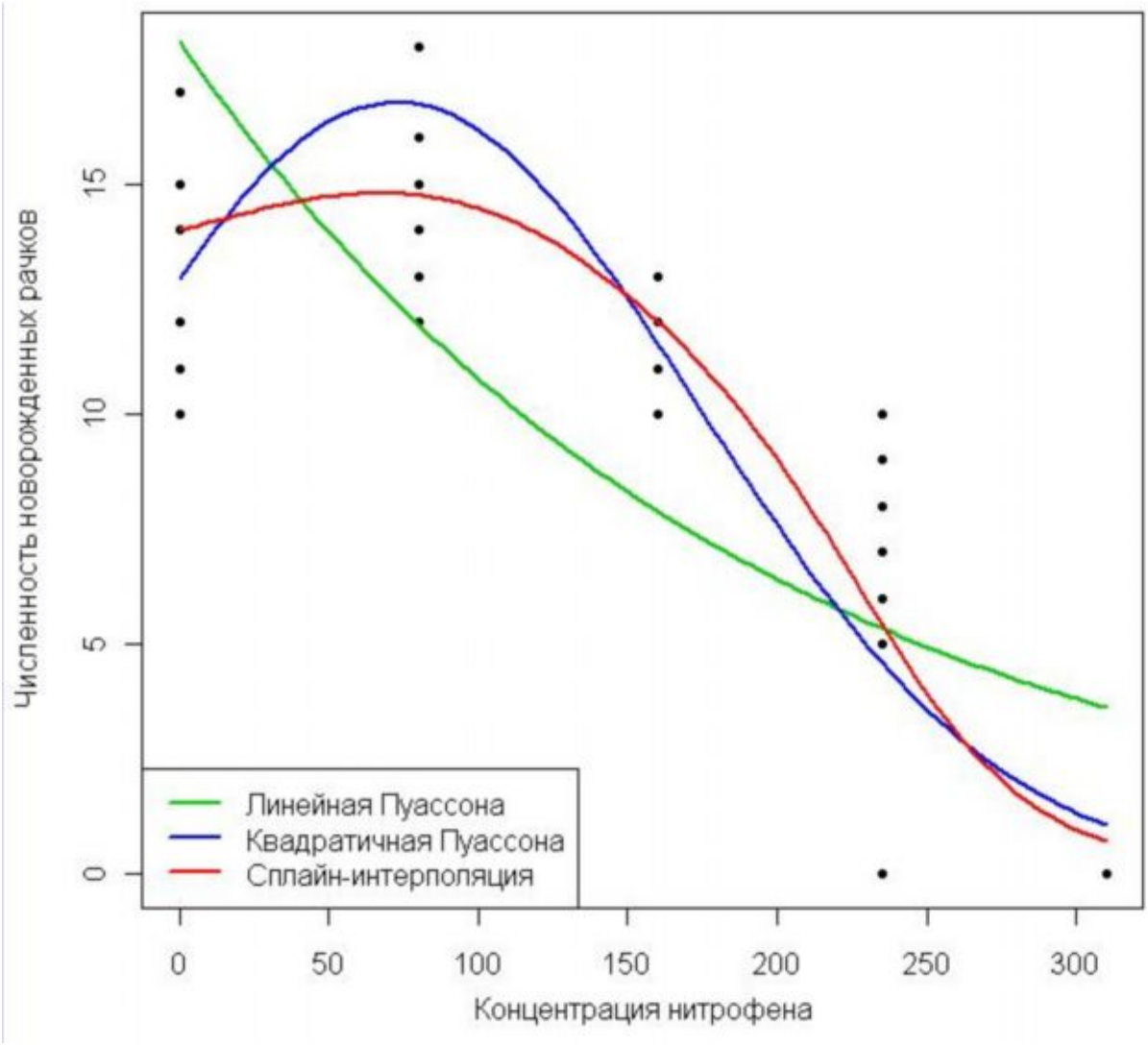


Рис. 4.2. Графики моделей пуассоновской регрессии и сглаживания для третей генерации цериодафний

Нетрудно заметить, что квадратичная функция пуассоновской регрессии недостаточно реалистична в области малых концентраций нитрофена, что приводит к парадоксальному выводу о его благотворном влиянии на репродукцию. Значение среднеэффективной концентрации и ее доверительных интервалов рассчитаем с использованием бутстреп-метода: этот пример приведен в классической монографии (Davison, Hinkley, p.383) и воспроизводится нами без комментариев:

```
nitro <- rbind(nitrofen,nitrofen,nitrofen,nitrofen,nitrofen)
nitro <- rbind(nitro,nitro,nitro,nitro,nitro)
nitro$conc <- seq(0,310,length=nrow(nitro))
nitro.fun <- function(data,i, nitro)
   {  assign ("d" ,data[i,] )
     d.fit <- gam(brood3~s(conc,df=3), poisson, data=d)
     f <- predict(d.fit,nitro,"response")
     f.gam <- max(nitro$conc[f>0.5*f [1]])
     f.gam }
nitro.boot <- boot(nitrofen, nitro.fun, R=499,
              strata=rep(1:5,rep(10,5)), nitro=nitro)
c("Среднеэффективная концентрация:  ", median(nitro.boot$t))
boot.ci(nitro.boot,type=c("norm","basic","bca"))
```

```
Среднеэффективная концентрация:  219.904
Доверительные интервалы, вычисленные бутстрепом
Основаны на 499 бутстреп-повторностях
Уровень  Нормальные          Основные          BCa
95%   (207.3, 234.2 )  (208.6, 234.4 )  (205.5, 231.4 )
```

По сравнению с моделью Пуассона `m3p` значение среднеэффективной концентрации несколько сместилось по шкале вправо, а доверительный интервал несколько сузился.

4.2. Модели "доза-эффект" для метрической шкалы отклика

Ранее в разделе 1.4 отмечалось, что вид функциональной зависимости "доза-эффект" далеко не всегда может быть постулирован, исходя из эколого-медицинских соображений. С другой стороны, используя современные технические средства и компьютерные программы, достаточно несложно построить целый набор самых разнообразных моделей, почти в равной мере претендующих на "оптимальность". Поэтому понятно, что подбор функций регрессии, наиболее правдоподобно объясняющих экспериментальные данные – важная и сложная область токсикометрии.

Особое значение селекция оптимальной модели имеет при использовании в качестве отклика метрических показателей. Статистическая (т.е. основанная на средней тенденции) зависимость величины эффекта от уровня воздействия x в этом случае описывается нелинейными моделями, которые в общем виде могут быть представлены как

$$\varphi\,(x;\, b,\, c,\, d,\, e,\ldots) = c + (d - c)\,\psi\,(x;\, b,\, e,\ldots),$$

где параметры c и d являются нижним и верхним пределами отклика, а ψ – некоторая задаваемая нелинейная функция с параметрами b и e. Список некоторых основных моделей приведен в табл. 1, где использованы наименования и формульная нотация из статьи (Ritz , 2010), претендующей на определенное тематическое обобщение.

Таблица 1. Математические формулы функций нелинейных моделей регрессии, используемых для аппроксимации зависимостей доза-эффект; код соответствует спецификации моделей в пакете *drc* среды R

Наименование модели	Вид функции регрессии	Код
Логнормальная модель (пробит) с четырьмя параметрами (b, c, d, e)	$\varphi(x) = c + (d - c)\Phi\{b(\log(x) - \log(e))\}$, где Ф - кумулятивная функция плотности для стандартного нормального распределения	LN.4
Log-логистическая модель с четырьмя параметрами (b, c, d, e)	$\varphi(x) = c + \dfrac{d - c}{1 + \exp\{b(\log(x) - \log(e))\}}$	LL.4
Log-логистическая модель с тремя параметрами (b, d, e)	$\varphi(x) = \dfrac{d}{1 + \exp\{b(\log(x) - \log(e))\}}$	LL.3
Log-логистическая модель с двумя параметрами (b, e)	$\varphi(x) = \dfrac{1}{1 + \exp\{b(\log(x) - \log(e))\}}$	LL.2
Модель Weibull - 1	$\varphi(x) = c + (d - c)\{\exp[\exp(b(\log(x) - \log(e))]\}$	W1.4
Модель Weibull - 2	$\varphi(x) = c + (d - c)\{1 - \exp[\exp(b(\log(x) - \log(e))]\}$	W2.4
Логистическая модель с четырьмя параметрами (b, c, d, e)	$\varphi(x) = c + \dfrac{d - c}{1 + \exp\{b(x - e)\}}$	L.4
Модель Gompertz с четырьмя параметрами (α, b, d, e)	$\varphi(x) = c + (d - c)\{\exp[\exp(b(x - e))]\}$	G.4
Экспоненциальная модель с тремя параметрами (c, d, e)	$\varphi(x) = c + (d - c)(1 - \exp(-x/e))$ $\varphi(x) = c + (d - c)(\exp(-x/e))$	AR.3 EXP.3
Экспоненциальная модель с двумя параметрами (d, e)	$\varphi(x) = d(1 - \exp(-x/e))$ $\varphi(x) = d(\exp(-x/e))$	AR.2 EXP.2
Модель Michaelis-Menten с тремя параметрами (c, d, e)	$\varphi(x) = c + \dfrac{d - c}{1 + (e/x)}$	MM.3
Модель Brain-Cousens (1989) с пятью параметрами (b, c, d, e, f)	$\varphi(x) = c + \dfrac{d - c + fx}{1 + \exp\{b(\log(x) - \log(e))\}}$	BC.5
Модель Cedergreen-Ritz-Streibig (2005) с четырьмя параметрами (α, b, d, e)	$\varphi(x) = \dfrac{d + f\exp(-1/x^{\alpha})}{1 + \exp\{b(\log(x) - \log(e))\}}$	CRSc.4

Два параметра c и d для большинства моделей из табл. 1, определяют нижнюю и верхнюю горизонтальные асимптоты сигмоидной кривой. Разность (d - c) параметров модели "доза-эффект" определяет, собственно, *полный эффект* или величину максимального превышения количественного отклика над уровнем нормы. Тогда для всей области определения модели легко найти любую дозу ED_y, которая приводит к (100 - y) % доле от (d - c), т.е. выполнить, как и в случае альтернативного показателя, построение шкалы изоэффективных доз. Например, если в результате воздействия происходит угнетение жизнедеятельности биологического показателя на y = 10 % от общего эффекта, то можно полагать, что эффективная доза ED_{10} представляет собой критический экологический риск.

Логнормальная (LN.4) и лог-логистическая модели (LL.4) в таб. 1 представляют собой градированный аналог описанным в разделе 2 двухпараметрическим моделям пробита и логита с прологарифмированными значениями x_i. При этом параметры моделей b и e для альтернативных и метрических показателей имеют идентичный физический смысл: b – коэффициент угла наклона в области переходного состояния, а e определяет положение точки перегиба. Для этих моделей изоэффективные дозы выражаются непосредственно через параметры b и e , как например, для модели LL.4:

$$ED_y = e\,(y / (100 - y))^{1/b}.$$

Поскольку все параметры модели оцениваются с известными статистическими ошибками регрессии, то, тем самым, можно легко обосновать доверительные интервалы показателя ED_y и другие характеристики его надежности.

Во многих случаях бывает целесообразно полагать, что минимальное значение эффекта равно нулю, т.е. приравнять нижнюю асимптоту c = 0, и тогда функция регрессии в общем случае будет иметь вид:

$$\varphi\,(x;\, b,\, d,\, e,...) = d\,\psi\,(x;\, b,\, e,...).$$

В табл. 1 таким примером является спецификация трехпараметрической модели LL.3.

При использовании исходных данных с откликом, состоящим из квантованных значений 0 и 1, задается параметр d = 1 и общий вид функции регрессии сводится к традиционным моделям логита и пробита (см. модель LL.2 в табл. 1). Здесь ED_y = (d - c) = (1 – 0) имеет уже вероятностный смысл, т.е. дозы, вызывающей регистрируемый эффект у y % особей, взятых в эксперимент.

При использовании логнормальной и лог-логистической моделей исходят из предположения, что сигмоидная кривая является симметричной относительно ED_{50}, т.е. дисперсия, обусловленная индивидуальной чувствительностью, изменяется по одному и тому же закону как в направлении низких доз, так и при достижении экстремальных воздействий. Смысл практического использования моделей Вейбулла W1.4 и W2.4 с четырьмя параметрами связан с учетом асимметрии сигмоиды: кривая Weibull-1 медленнее убывает в области верхнего предела, но быстрее

приближается к более низкому пределу, что позволяет точнее отследить зависимость в области низких уровней воздействия. С кривой Weibull-2 происходит обратный результат. Можно отметить, что в большинстве случаев, когда асимметрия проявляется не слишком явно, модели Вейбулла дают весьма близкие результаты по сравнению с LL.4.

Существует общее (недостаточно обоснованное) мнение, что значения доз должны быть обязательно прологарифмированы перед статистическим анализом моделей. Считается, что log-преобразования часто помогают получать симметричные распределения, близкие к нормальному, хотя практически это утверждение доказать трудно. Тут важнее другое: использование трансформации ″логарифм-экспонента″ является прекрасным средством построения моделей, которые надежно определены в неотрицательных координатах и имеют биологически приемлемые асимптоты в области очень низких и очень больших доз. Однако по определенным соображениям иногда вполне можно воздержаться от логарифмирования независимой переменной и оценить необходимое число параметров для обычной логистической модели (L.4) или модели Гомперца (G.4), которая отличается от моделей Вейбулла формой экспоненты (Seber, Wild, 1989).

Если нет оснований предполагать зависимость сигмоидного типа, то наиболее часто используемым ″штатным″ вариантом является построение моделей экспоненциальной регрессии с тремя EXP.3 или двумя EXP.2 параметрами. Следует также упомянуть обобщенную кривую Михаэлиса-Ментен – универсальную гиперболическую зависимость снижения скорости роста, широко применяемую в биохимии и микробиологии.

Наличие промежуточных экстремумов в медицинской токсикологии считается сомнительным явлением и получило название "парадоксальная токсичность". Степень парадоксальности, которая может трактоваться как "сомнительная", "достоверная" или "абсолютная", принято оценивать, с одной стороны, по критерию монотонности убывания функции "доза-эффект" на достаточно представительном участке и, во-вторых, по отсутствию статистически значимых различий величины эффекта в критических точках (Криштопенко и др., 2008).

В экологической токсикологии базовой зависимостью доза-эффект принято считать колоколообразную кривую гауссового типа, теоретическое обоснование которой восходит к законам толерантности Шелфорда. Действительно, в последние десятилетия многочисленными исследованиями было показано, что при умеренных нагрузках обилие многих популяций и разнообразие биоценозов может возрастать при росте химического или радиационного воздействия (это явление принято называть гормезиз-эффект, от греч. hórmēsis). Иными словами, сигмоидная зависимость рассматривается в экотоксикологии как частный случай гауссианы (например, ее правая ветвь) и не исключается, что при проведении дальнейших исследований будет доказано, что вредное действие могут оказывать как большие, так и супермалые дозы токсиканта. Яркими примерами этому является содержание

йода в питьевой воде или любые зависимости от воздействия pH, где оптимум располагается в районе pH = 7.

Стимулирующее действие умеренных доз токсикантов может быть учтено при использовании моделей Brain-Cousens (BC.5) и Cedergreen-Ritz-Streibig (CRSc.4), которые приводят к кривым с верхним или нижним промежуточным оптимумом (Cedergreen et al., 2005).

4.3. Использование пакета `drc` статистической среды R

Представленная в табл. 1 иерархическая структура моделей оказалась очень удобной для их общей программной реализации, поскольку каждый частный случай наследует, как правило, основные особенности родительской модели. Этот подход был заложен в основу пакета `drc` (dose-response curves) статистической среды R (Ritz C., Streibig, 2005; Knezevic et al., 2007). Текущая версия пакета обеспечивает практически все необходимые потребности пользователя по построению и исследованию зависимостей "доза-эффект". Здесь можно найти разнообразный сервис: удобную спецификацию и настройку встроенных функций для получения разнообразных моделей, развитые процедуры оценки адекватности и селекции наилучших регрессий, универсальные средства оценки изоэффективных доз и их доверительных интервалов, необходимый графический инструментарий и т.д.

Функция `drm()` пакета `drc` является основным генератором моделей, оценивающим их параметры и возвращающим объекты класса "`drc`":

```
drm(formula, curveid, pmodels, weights, data, subset,
 fct, type, bcVal, bcAdd, start, na.action, robust,
 logDose, control, lowerl, upperl, separate, pshifts)
```

Перечислим ее основные параметры:

`formula` – символическое описание структуры переменных модели (в простейшем случае `'response ~ dose'`);

`curveid` – вектор, определяющий возможную группировку данных;

`pmodels` – таблица, содержащая столько столбцов, сколько параметров модели (ее смысл объясняется далее на примере);

`weights` – вектор весов, на которые умножается квадраты ошибок при оценке коэффициентов модели;

`data` – таблица, содержащая исходные данные для построения модели;

`subset` – вектор, задающий необходимое подмножество таблицы наблюдений;

`fct` – спецификация подгоняемой модели, представленная столбцом *Код* табл. 1, и списком дополнительных опций (наименования оцениваемых коэффициентов нелинейной функции, вектор, определяющий стартовые параметры и т.д.);

`type` – параметр, определяющий тип обрабатываемых данных, необходимый для построения функции максимального правдоподобия

("continuous", "binomial", "Poisson", "quantal", "event").

Так же, как и при использовании иных функций построения статистических моделей, таких как lm(), glm(), gam(), к объекту класса drc применимы все методы извлечения необходимой информации:

- anova: сравнение критериев качества подгонки для двух моделей;
- coef: оценки параметров (коэффициентов) модели;
- fitted: значения, прогнозируемые по модели;
- logLik: логарифм значения правдоподобия;
- plot: вывод графика подогнанной кривой;
- residuals: значения остатков модели;
- summary: сводные итоги подгонки модели;
- vcov: матрица оценок дисперсий-ковариаций.

Построим модель ингибирования роста корневой системы плевела многолетнего (*Lolium perenne*) под действием ароматических кислот. Таблица исходных данных ryegrass, доступная в пакете drc, содержит 24 строки наблюдений над двумя следующими показателями: rootl – длина корневища растения (см) и conc – концентрация феруловой кислоты (мМоль).

```
library(drc)
## Посмотрим набор исходных данных
ryegrass
## Подгонка 4-параметрической лог-логистической модели
## с наименованиями оцениваемых параметров
ryegrass.m0 <- drm(rootl ~ conc, data = ryegrass,
      fct = LL.4(names = c("Св.член", "Нижний предел",
      "Верхний предел", "ED50")))
summary(ryegrass.m0)
## Сравнение с моделью без параметров
noEffect(ryegrass.m0)
```

```
Оценки параметров:
                          Оценка    Ст.ошибка  t-крит  p-знач.
Св.член:(Intercept)        2.98222    0.46506  6.41251  0.0000
Нижний предел:(Intercept)  0.48141    0.21219  2.26876  0.0345
Верхний предел:(Intercept) 7.79296    0.18857 41.32722  0.0000
ED50:(Intercept)           3.05795    0.18573 16.46440  0.0000
Стандартное отклонение для остатков:
 0.5196256 (20 степеней свободы)
Сравнение с моделью без параметров
Тест Хи-квадрат              Df          p-value
       91.87776          3.00000          0.00000
```

Обратим внимание на функцию noEffect(), которая сравнивает по критерию χ^2 остатки построенной модели с остатками линейной модели с одним свободным членом (т.е. горизонтальной линией регрессии, для которой имеет место отсутствие эффекта). Поскольку нулевая гипотеза об

идентичности ошибок отвергается с $p \approx 0$, то модель лог-логистической регрессии с высоким уровнем значимости адекватна исходным данным.

Построим для сравнения модели Вейбулла 1 и 2, асимметричные относительно ED_{50}, и сопоставим кривые зависимости на одном графике – рис. 4.3. Решение вопроса, какая из возможных моделей является наилучшей со статистической точки зрения, можно осуществить с использованием функции `mselect()`. В список проверяемых претендентов включим модели Вейбулла и лог-логистической регрессии с разным числом параметров.

```
## Подгонка 4-параметрической модели Вейбулла (тип 2 и 1)
ryegrass.m1 <- drm(rootl ~ conc, data = ryegrass,
                   fct = W1.4())
ryegrass.m2 <- drm(rootl ~ conc, data = ryegrass,
                   fct = W2.4())
summary(ryegrass.m2)
## Сравнение лог-логистической регрессии и моделей Вейбулла
mselect(ryegrass.m0, list(LL.3(), LL.5(), W1.3(),
                    W1.4(), W2.4()), linreg = TRUE)
plot(ryegrass.m0, broken=TRUE, xlab="Доза (mM)",
      ylab="Длина корня(см)", lwd=2,
      cex=1.2, cex.axis=1.2, cex.lab=1.2)
plot(ryegrass.m1, add=TRUE, broken=TRUE, col=4, lwd=2)
plot(ryegrass.m2, add=TRUE, broken=TRUE, col=3, lwd=2)
arrows(3, 7.5, 1.4, 7.5, 0.15, lwd=2)
text(3,7.5, "Weibull-2", pos=4, cex=1.2)
arrows(2.5, 0.9, 5.7, 0.9, 0.15, lwd=2)
text(3,0.9, "Weibull-1", pos=2, cex=1.2)
```

```
Подбираемая модель: Weibull (тип 2) (4 параметра)
Оценки параметров:
                  Оценка   Ст.ошибка    t-крит   p-знач.
b:(Intercept) -1.96791    0.29070 -6.76957  0.0000
c:(Intercept)  0.32459    0.24902  1.30346  0.2072
d:(Intercept)  7.72630    0.17339 44.55945  0.0000
e:(Intercept)  2.48765    0.14781 16.83039  0.0000
Стандартное отклонение для остатков:
 0.5144203 (20 степеней свободы)
```

```
Сравнение характеристик протестированных моделей
         logLik        IC Lack of fit   Res var
W2.4   -15.91352  41.82703  0.94507131 0.2646283
LL.4   -16.15514  42.31029  0.86648304 0.2700107
LL.5   -15.87828  43.75656  0.85384758 0.2777393
W1.4   -17.46720  44.93439  0.45056762 0.3012075
LL.3   -18.60413  45.20827  0.35316787 0.3153724
W1.3   -22.22047  52.44094  0.04379149 0.4262881
Cubic  -25.53428  61.06856          NA 0.5899609
Quad   -35.11558  78.23116          NA 1.2485122
Lin    -50.47554 106.95109          NA 4.2863247
```

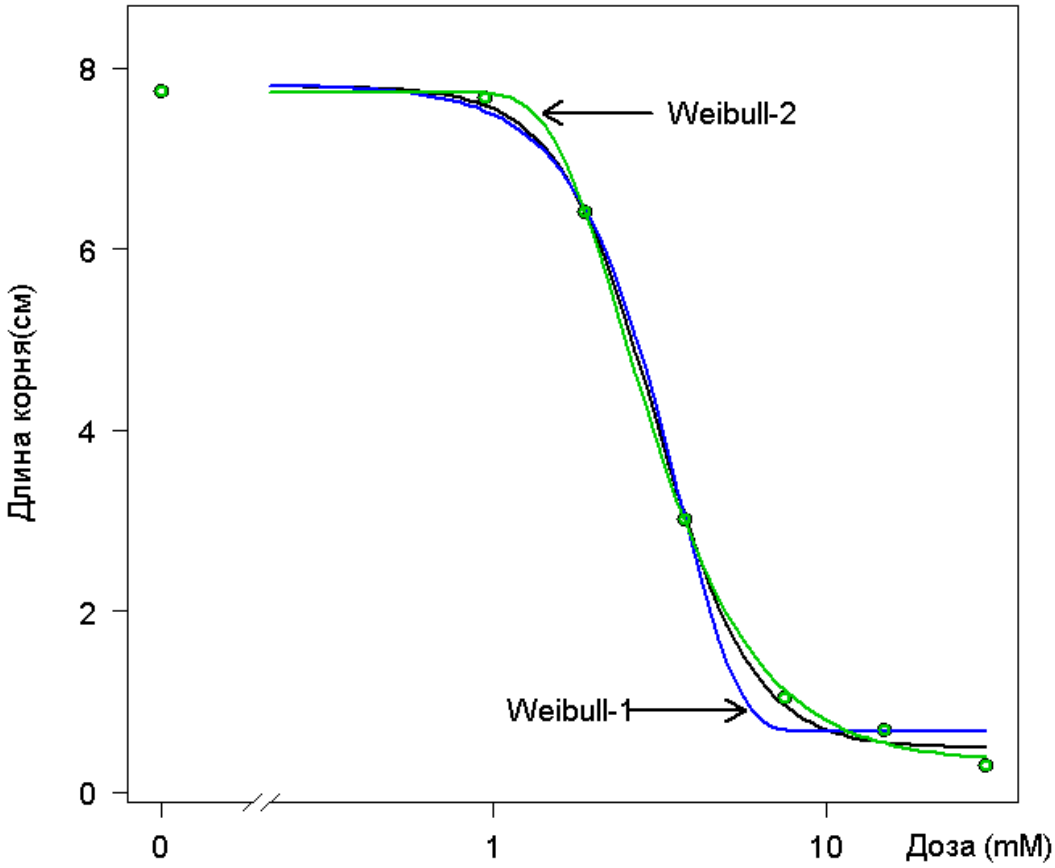


Рис. 4.3. Модели зависимости длины корня райграса от концентрации феруловой кислоты

В протокол проверки, формируемый функцией `mselect()`, для каждой сравниваемой модели включены логарифм максимального значения функции правдоподобия (`logLik`), информационный критерий Акаике (`IC`), величина *p*-значения для *F*-критерия в тесте на потерю соответствия (`lack of fit` – Bates, Watts, 2007; Ritz, Martinussen, 2011) и остаточная дисперсия (`Res var`). Если использовать параметр `linreg = TRUE`, то в список протестированных моделей добавляются кубическая (`Cubic`), квадратичная (`Quad`) и линейная (`Lin`) регрессии. Представленная таблица отсортирована по возрастанию AIC, но все остальные критерии также свидетельствуют об оптимальности модели Вейбулла `W2.4`.

Оценка коэффициента $e = 3.06$ модели `LL.4` соответствует показателю токсикометрии EC_{50}. С использованием функции `ED()` можно рассчитать произвольный комплект изоэффективных доз для большинства моделей с использованием различных методов оценки доверительных интервалов (рекомендуется использовать дельта-метод).

```
print("Модель LL.4")
ED(ryegrass.m0,c(5,50,95), interval="delta")
print("Модель W2.4")
ED(ryegrass.m2,c(5,50,95), interval="delta")
```

```
Оценки изоэффективных доз
(доверительные интервалы основаны на дельта-методе(s))
Модель LL.4
      Оценка   Ст.ошибка  Нижний Верхний
1:5   1.13930    0.18575 0.75184  1.5268
1:50  3.05795    0.18573 2.67053  3.4454
1:95  8.20774    1.37857 5.33209 11.0834
Модель W2.4"
      Оценка   Ст.ошибка  Нижний Верхний
1:5   1.42447    0.14402  1.12404  1.7249
1:50  2.99691    0.19692  2.58614  3.4077
1:95 11.25317    2.60250  5.82445 16.6819
```

Поскольку доверительные интервалы изоэффективных доз пересекаются, то можно сделать вывод, что в смысле показателей токсикометрии модели `LL.4` и `W2.4` являются идентичными.

Пакет `drc` позволяет строить зависимости "доза-эффект" для отклика в альтернативной шкале, идентичные моделям логита или пробита, полученным с помощью функции `glm()`. Для этого достаточно определить параметры `weights =` и `type = "binomial"`. Однако в ряде случаев со специфическим планом эксперимента гибкость задания спецификации моделей, характерная для `drc`, может оказаться решающей.

Рассмотрим пример с численностью земляных червей, которые свободно перемещаются между двумя смежными участками: с землей, загрязненной токсичным веществом, и с чистой землей. Столбцы таблицы `earthworms` с исходными данными содержат дозу токсиканта `dose`, общее `total` и обнаруженное на загрязненном участке количество червей `number`. Напомним, что обычная логистическая регрессия LL.2 подразумевает верхний предел для нулевой дозы *d*, равный 1. В нашем примере при нулевой дозе земляные черви распределятся равномерно и приблизительно половина их числа ожидается быть обнаруженной на каждом из двух соседних участков ($E_0 = 0.5$).

```
head(earthworms)
## Подгонка обычной лог-логистической регрессионной модели
earthworms.m1 <- drm(number/total~dose, weights = total,
       data = earthworms, fct = LL.2(), type = "binomial")
summary(earthworms.m1)
##  Подгонка лог-логистической регрессионной модели,
## где оценивается верхний предел
earthworms.m2 <- drm(number/total~dose, weights = total,
       data = earthworms, fct = LL.3(), type = "binomial")
summary(earthworms.m2)
## Сравнение моделей 1 и 2
mselect(earthworms.m1, list(LL.3()))
anova(earthworms.m1, earthworms.m2)
```

```
Две лог-логистические модели:
LL.2: Нижний предел = 0 и верхний предел = 1  (2 параметра)
Оценки параметров:
                  Оценка   Ст.ошибка    t-крит   p-знач.
b:(Intercept) 1.260321   0.246707 5.108564   0e+00
e:(Intercept) 0.145140   0.036797 3.944389   1e-04

LL.3: Нижний предел = 0 (3 parms)
Оценки параметров:
                  Оценка   Ст.ошибка    t-крит   p-знач.
b:(Intercept) 1.505679   0.338992 4.441641   0e+00
d:(Intercept) 0.604929   0.085800 7.050498   0e+00
e:(Intercept) 0.292428   0.083895 3.485636   5e-04
Сравнение характеристик протестированных моделей
         logLik        IC Lack of fit
LL.3  -36.15518  78.31036  0.09049411
LL.2 -347.55013 699.10026  0.16311466

ANOVA-подобная таблица
1 модель  fct:      LL.2()
2 модель  fct:      LL.3()
          ModelDf  Loglik Df LR знач. p знач.
1 модель        2 -347.55
2 модель        3  -36.16  1   622.79       0
```

Заметим, что мы использовали функцию `anova()` для сравнения девианса двух тестируемых моделей `LL.2()` и `LL.3()`, поскольку одна из них является родительской по отношению к другой. Задав трехпараметрическую модель `LL.3()` с вычисляемым верхним пределом и биномиальным законом распределения отклика, мы получили значительно более корректную зависимость.

Преимущества модели с вычисляемым верхним пределом легко оценить на графике рис. 4.4. Использование параметра `type="confidence"` дает нам возможность вывести 95%-ную доверительную полосу относительно кривой анализируемой зависимости.

```
plot(earthworms.m1, lty=4, lwd=2,
         xlab="Доза", ylab="Доля обнаруженных червей" )
plot(earthworms.m2, lwd=2, add=TRUE, type="confidence")
legend("topright",c("LL.2","LL.3"), lwd=2, lty=c(4,1))
```

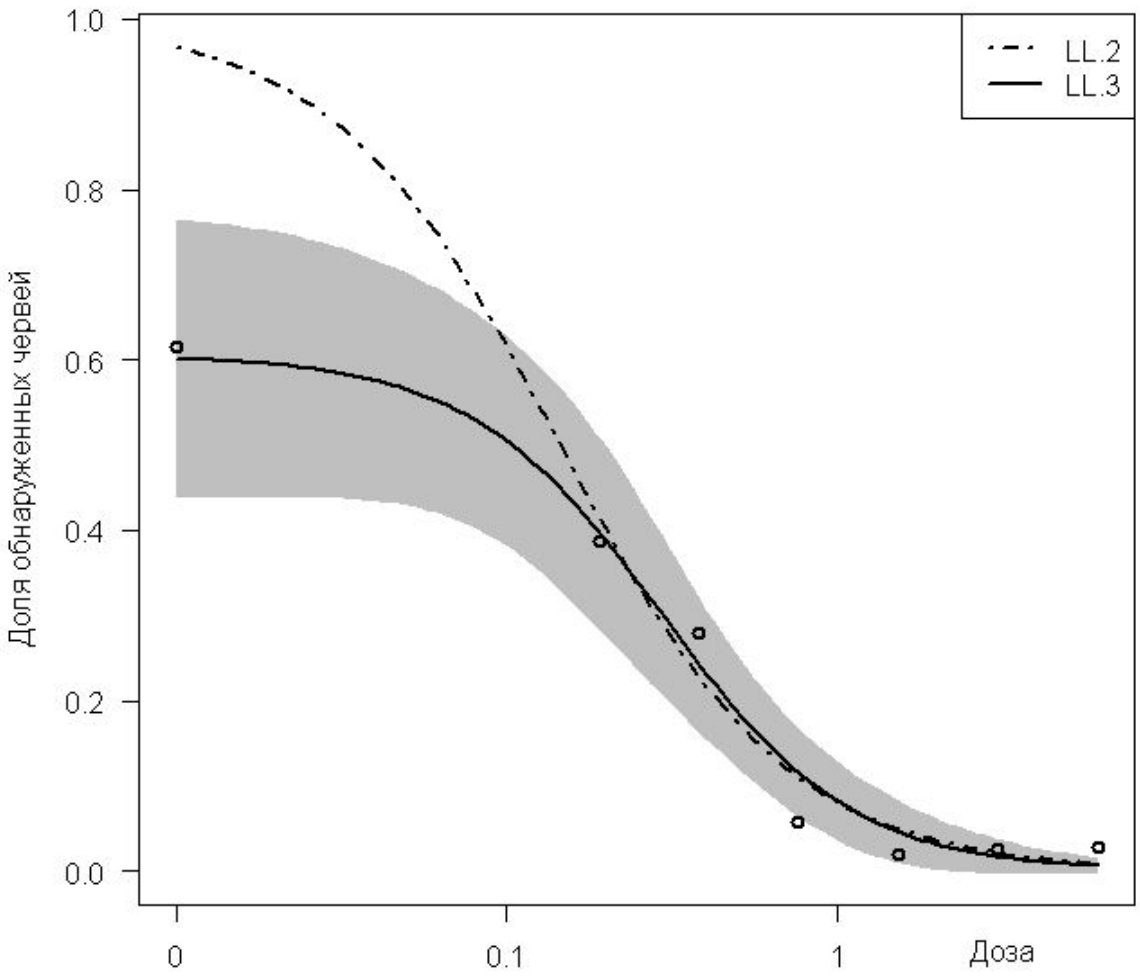


Рис. 4.4. Модели зависимости численности червей от концентрации

4.4. Сравнение параметров кривых отклика

Рассмотрим пример, связанный со сравнительной оценкой эффективности двух гербицидов глифосата (`Glyphosate`) и бентазона (`Bentazone`). Набор данных `S.alba` включает данные по кривой отклика – сухому весу растений белой горчицы `DryMatter` (г/пот) в зависимости от внесенной дозы `Dose` (г/га) одного из двух типов гербицида `Herbicide` (фактор с двумя уровнями).

```
Library(drc)
head( S.alba)
S.alba.m1<-drm(DryMatter~Dose, Herbicide, fct=LL.4(),
                  data=S.alba)
summary(S.alba.m1)
modelFit(S.alba.m1)
```

```
Подобранная модель: лог-логистическая
                  (ED50 как параметр) (4 параметра).
Оценки параметров:
                  Оценка   Ст.ошибка    t-крит   p-знач.
b:Glyphosate   2.715409   0.748279  3.628873   6e-04
b:Bentazone    5.134810   1.130949  4.540266   0e+00
```

```
c:Glyphosate  0.891238   0.194703  4.577429   0e+00
c:Bentazone   0.681845   0.095111  7.168925   0e+00
d:Glyphosate  3.875759   0.107463 36.066087   0e+00
d:Bentazone   3.805791   0.110341 34.491101   0e+00
e:Glyphosate 62.087606   6.611444  9.390929   0e+00
e:Bentazone  29.268444   2.237090 13.083268   0e+00
Стандартное отклонение для остатков:
 0.3730047 (60 степеней свободы)
```

Тест на потерю соответствия

```
          ModelDf    RSS  Df F крит p знач.
ANOVA          53 8.0800
Модель DRC     60 8.3479  7  0.2511  0.9696
```

Была получена модель с 8 параметрами отдельно для каждого вида гербицида. Тест на потерю соответствия (Lack-of-fit test) показывает, что полученная модель столь же хорошо описывает наблюдаемые данные, что и обычный дисперсионный анализ.

Поставим задачу оценить сравнительную эффективность (токсическую силу – Relative potency) обоих гербицидов. Для этого нужно уменьшить параметричность модели, сделав предположение, что оба токсиканта отличаются между собой в средней части кривых, а параметры нижнего и верхнего пределов имеют одинаковое значение. Чтобы задать это условие в модели, необходимо определить параметр

`Pmodels = data.frame(Herbicide,1,1,Herbicide)`,

который указывает, какие коэффициенты модели должны быть одинаковыми для каждой кривой. Последовательность параметров модели определяется в алфавитном порядке, т.е. мы объявили условие, что b и e вычисляются в зависимости от значения фактора `Herbicide`, а c и d принимаются одинаковыми для всех кривых.

```
S.alba.m2<-drm(DryMatter~Dose, Herbicide, fct=LL.4(),
   data=S.alba, pmodels=data.frame(Herbicide,1,1,Herbicide))
summary(S.alba.m2)
anova(S.alba.m1,S.alba.m2)
ED(S.alba.m1, c(10, 90), interval="delta")
ED(S.alba.m2, c(10, 90), interval="delta")
plot(S.alba.m1, col="red",lty=c(3,4), xlab="Доза",
                ylab="Сухая биомасса")
plot(S.alba.m2, lwd=2, col=3:4, add=TRUE)
```

Оценки параметров:

```
                  Оценка   Ст.ошибка    t-крит   p-знач.
b:Bentazone     5.046141   1.040135  4.851430         0
b:Glyphosate    2.390218   0.495959  4.819387         0
c:(Intercept)   0.716559   0.089245  8.029117         0
d:(Intercept)   3.854861   0.076255 50.551925         0
e:Bentazone    28.632355   2.038098 14.048566         0
e:Glyphosate   66.890545   5.968819 11.206663         0
```

```
Стандартное отклонение для остатков:
 0.3705151 (62 степеней свободы)

Таблица дисперсионного анализа ANOVA
1 модель:  fct LL.4()  pmodels: Herbicide, 1, 1, Herbicide
2 модель:  fct LL.4() pmodels: Herbicide (для всех параметров)
          ModelDf    RSS Df  F крит  p знач
2 модель        62 8.5114
1 модель        60 8.3479  2  0.5876  0.5588

Оценки изоэффективных доз  Модель S.alba.m1
                 Оценка   Ст.ошибка  Нижний Верхний
Bentazone:10   19.0793     2.5219  14.0348  24.124
Bentazone:90   44.8991     4.9139  35.0698  54.728
Glyphosate:10  27.6431     6.1664  15.3086  39.978
Glyphosate:90 139.4513    37.5119  64.4164 214.486

Оценки изоэффективных доз Модель S.alba.m2
                 Оценка   Ст.ошибка  Нижний Верхний
Bentazone:10   18.5248     2.1667  14.1936  22.856
Bentazone:90   44.2548     4.9605  34.3389  54.171
Glyphosate:10  26.6770     5.3496  15.9833  37.371
Glyphosate:90 167.7231    36.9282  93.9048 241.542
```

Тест ANOVA показывает, что сумма квадратов отклонений для обеих моделей статистически незначимо отличается между собой, поэтому потеря информации от объединения коэффициентов c и d статистически мала. Однако значения максимальных изоэффективных концентраций ED_{90} серьезно отличаются между собой, что ясно следует из графика на рис. 4.5.

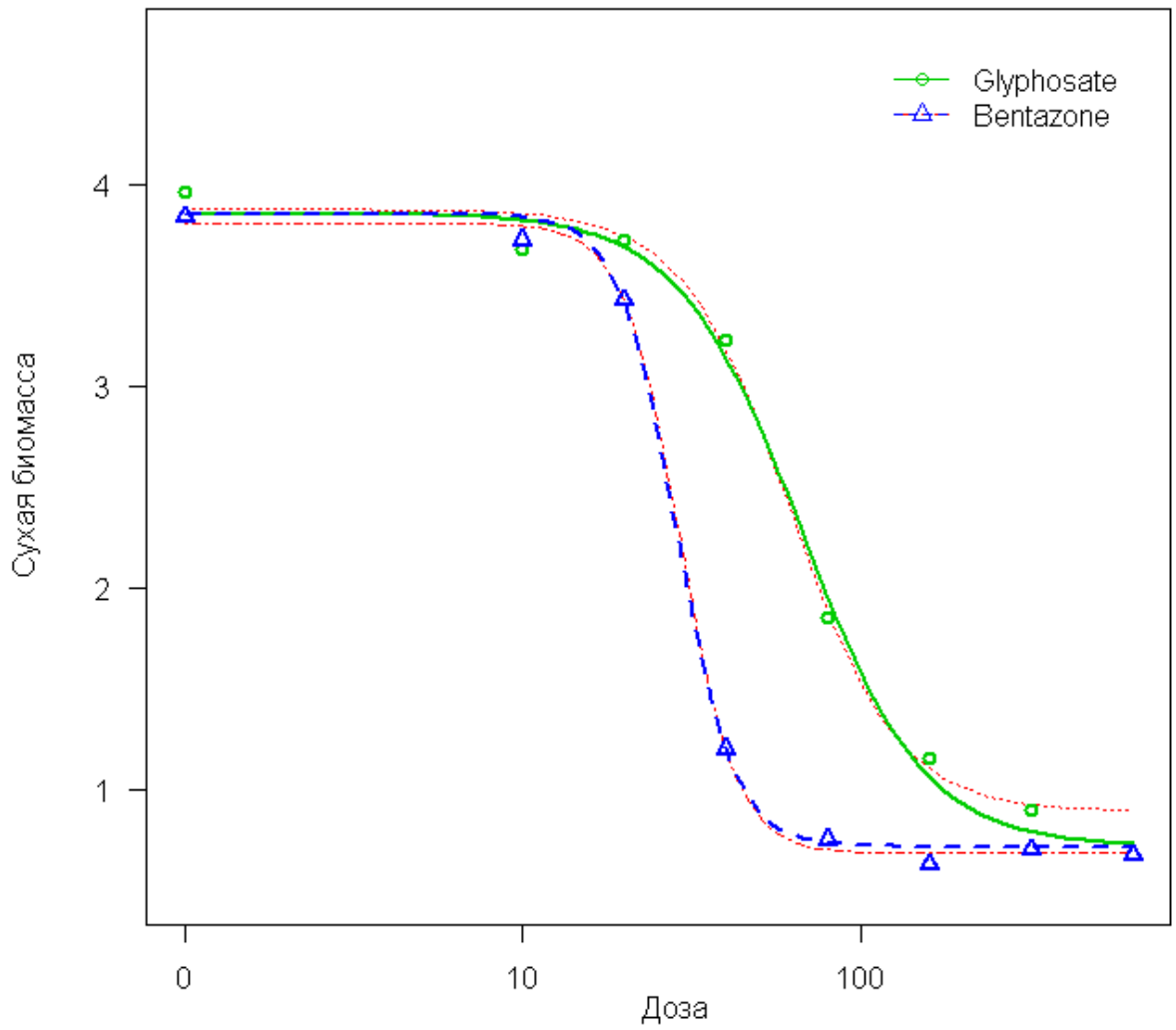


Рис. 4.5. Сравнение эффективности двух гербицидов (красным пунктиром показана модель с разными пороговыми значениями

Пакет `drc` включает несколько функций сравнения параметров построенных моделей (Ritz et al., 2006). Оценить статистически различия коэффициентов *e* (ED_{50}) и *b* (угол наклона сигмоиды) можно, сравнив их разность с 0 или их отношение с 1 с использованием функции `compParm()`.

```
## Сравнение ED50 по их отношениям и разностям
compParm(S.alba.m2, "e", "/")
compParm(S.alba.m2, "e", "-")
## Сравнение угла наклона по их отношениям и разностям
compParm(S.alba.m2, "b", "/")
compParm(S.alba.m2, "b", "-")
```

```
Сравнение параметра 'e'  по отношению
                     Оценка   Ст.ошибка     t-крит   p-знач.
Bentazone/Glyphosate 0.428048   0.043915 -13.024074       0
Сравнение параметра 'e'  по разности
                     Оценка    Ст.ошибка   t-крит  p-знач.
Bentazone-Glyphosate -38.2582     5.9148  -6.4682       0
```

```
Сравнение параметра 'b' по отношению
                        Оценка   Ст.ошибка t-крит  p-знач.
Bentazone/Glyphosate  2.11116    0.55928 1.98678  0.0514
Сравнение параметра 'b' по разности
                        Оценка   Ст.ошибка t-крит  p-знач.
Bentazone-Glyphosate   2.6559     1.0689  2.4847  0.0157
```

Выполненные расчеты показывают, что имеется существенное превышение среднеэффективной дозы ED_{50} `Glyphosate` над `Bentazone`, однако отличия в угле наклона средней части сигмоиды не являются статистически значимыми.

С использованием функции `EDcomp()` можно рассчитать относительную эффективность каждого гербицида по отношению к другому (Ritz et al., 2006). Если угол наклона обеих сигмоид одинаков на всем интервале доз, то этот индекс постоянен и равен, например ED_{50}^{1}/ED_{50}^{2} = 0.428. В нашем случае это не так и отчасти усугубляется принятым предположением о равенстве пороговых коэффициентов. Изменение соотношения эффективностей для различных уровней отклика можно проследить на графике рис. 4.6, который можно получить с использованием функции `relpot()`.

```
## Вычисление коэффициента относительной эффективности
EDcomp(S.alba.m2, c(50,50), interval = "delta")
EDcomp(S.alba.m2, c(10,90), interval = "delta")
EDcomp(S.alba.m2, c(90,10), interval = "delta")
EDcomp(S.alba.m2, c(50,50),interval = "delta",reverse=TRUE)
relpot(S.alba.m2, interval = "delta", lwd=2)
```

```
Оценка отношения изоэффективных доз
(доверительные интервалы основаны на дельта-методе (s))
                                 Оценка Верхний Нижний
Bentazone/Glyphosate:50/50  0.42805 0.34026 0.5158
                                 Оценка Верхний Нижний
Bentazone/Glyphosate:10/90 0.110449 0.053985 0.1669
                                 Оценка Верхний Нижний
Bentazone/Glyphosate:90/10  1.65891 0.86845 2.4494
                                 Оценка Верхний Нижний
Glyphosate/Bentazone:50/50   2.3362 1.8571 2.8153
```

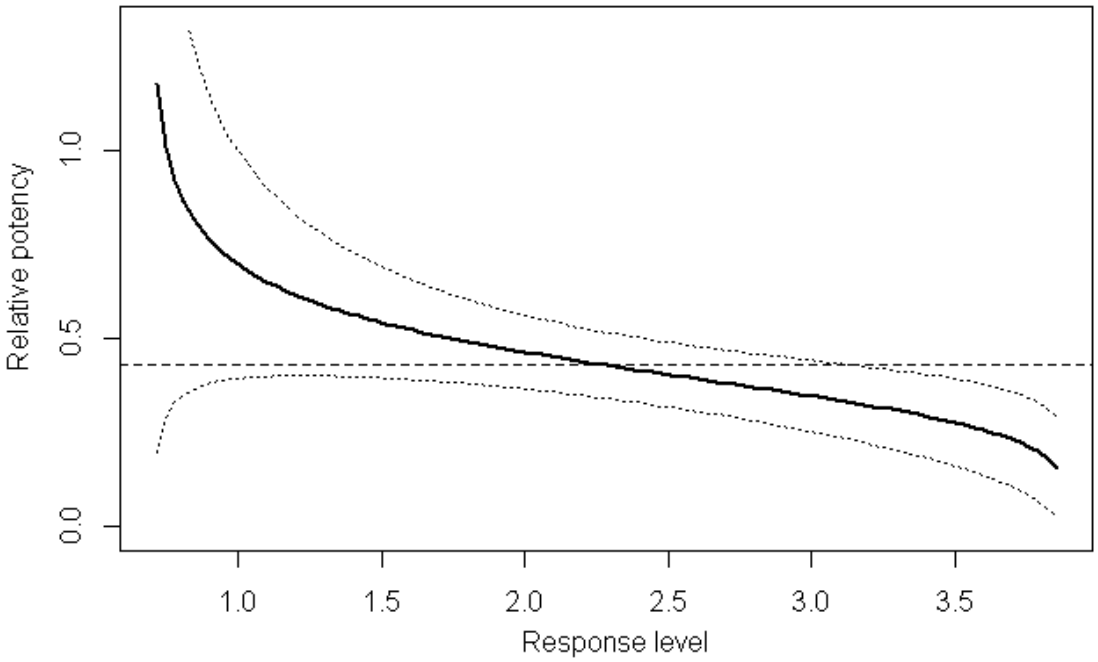


Рис. 4.6. Изменение относительной эффективности (relative potency) двух гербицидов в зависимости от уровня отклика (response level); пунктиром показаны доверительные интервалы индекса

4.5. Модели экспоненциального роста, Михаэлиса-Ментен и гормезиса

Асимптотические процессы кинетики роста или угнетения в биологии принято описывать экспоненциальными моделями. В пакете `drc` представлено два комплекта функций, генерирующих такие модели с одним (`AR.2, EXD.2`) или двумя (`AR.3, EXD.3`) предельными параметрами.

В таблице `O.mykiss` представлены исходные данные по снижению веса `weight` радужной форели *Oncorhynchus mykiss* после 28-дневной экспозиции в присутствии токсиканта с концентрацией `conc`. Выполним аппроксимацию данных ниспадающей экспоненциальной моделью `EXD.2`:

```
library(drc)
head(O.mykiss)
## Подгонка ниспадающей экспоненциальной зависимости EXD.2
O.mykiss.m1 <- drm(weight ~ conc, data = O.mykiss,
        fct = EXD.2(), na.action = na.omit)
summary(O.mykiss.m1)
## Вывод графика
plot(O.mykiss.m1, type = "all",  xlab = "Концентрация
        (мг/л)", ylab = "Вес (г)", broken = TRUE
       , lwd=2, xlim = c(0, 500), ylim = c(0,4))
```

```
Подбираемая модель: Экспоненциальный спад
                    с нижним пределом, равным 0 (2 параметра)
Оценки параметров:
                    Оценка    Ст.ошибка     t-крит   p-знач.
d:(Intercept)    2.846794    0.092526  30.767352  0.0000
e:(Intercept) 111.738614   33.196876   3.365938  0.0013
```

```
Стандартное отклонение для остатков:
0.5598508 (59 степеней свободы)
```

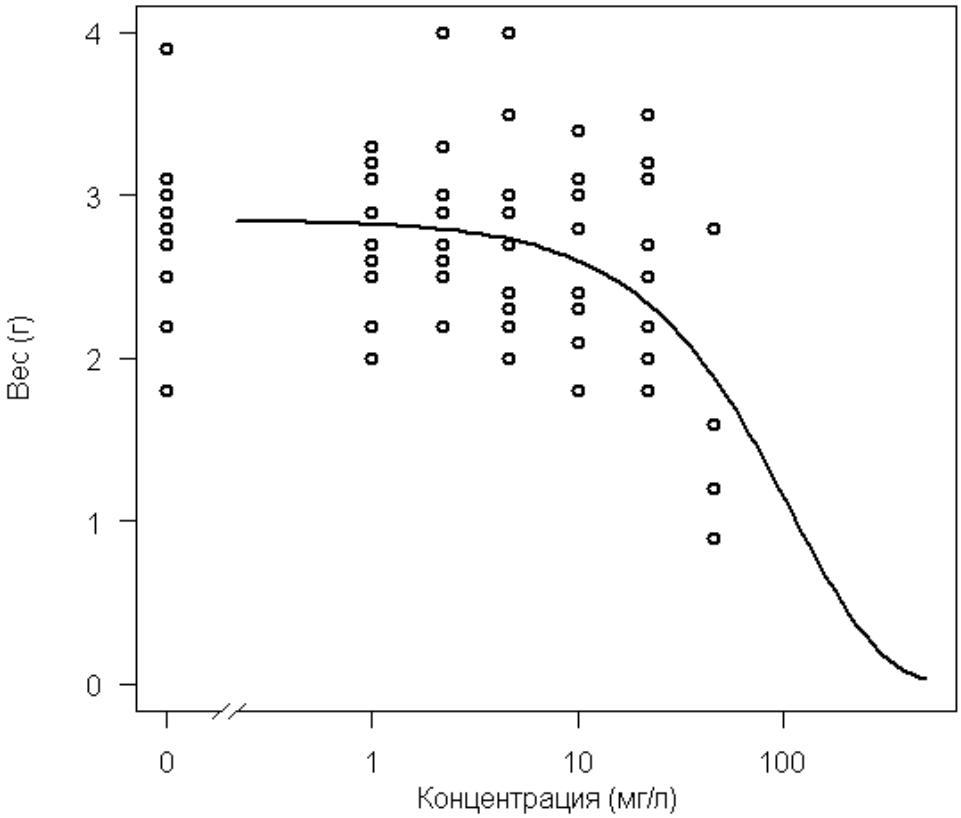


Рис. 4.7. Экспонента уменьшения веса форели под действием токсиканта

Модели кинетики роста могут быть описаны не только экспоненциальными моделями асимптотического роста AR.2, AR.3, но и моделями Михаэлиса-Ментен MM.2, MM.3. Данные таблицы methionine учитывают среднее прибавление в весе цыплят gain, в пищу которым добавляли один из двух препаратов (DLM или HMTBA – фактор product), содержащих метионин, в количестве dose.

```
head(methionine)
## Подгонка асимптотической экспоненты   AR.3
met.as.m1<-drm(gain ~ dose, product, data = methionine,
     fct = AR.3(),pmodels = list(~1, ~factor(product),
      ~factor(product)))
summary(met.as.m1)
## Подгонка модели Михаэлиса-Ментен    MM.3
met.mm.m1 <- drm(gain~dose, product, data=methionine,
     fct=MM.3(),pmodels = list(~1, ~factor(product),
    ~factor(product)))
summary(met.mm.m1)
# Сравнение моделей
mselect(met.as.m1, list(MM.3()))
plot(met.mm.m1, log = "", col=c(3,4), lty=1,
     ylim = c(1450, 1800), lwd=2, xlab="Доза",
     ylab="Привес цыплят",legendPos=c(0.2,1550))
```

```
plot(met.as.m1, log = "", col=c(3,4), lty=2, add=TRUE,
        legend=FALSE, lwd=2)
```

```
Подбираемая модель: регрессия асимптотического роста
                                          (3 параметра)
Оценки параметров:
                   Оценка    Ст.ошибка      t-крит   p-знач.
c:(Intercept) 1.4536e+03 1.0764e+01 1.3504e+02  0.0000
d:DLM         1.6892e+03 9.8280e+00 1.7188e+02  0.0000
d:MHA         1.7541e+03 2.1369e+01 8.2086e+01  0.0000
e:DLM         4.5386e-02 7.4128e-03 6.1226e+00  0.0036
e:MHA         9.2668e-02 1.6516e-02 5.6109e+00  0.0050
Стандартное отклонение для остатков:
 11.20328 (4 степеней свободы)

Подбираемая модель: возрастающая Михаэлиса-Ментен
                                          (3 параметра)
Оценки параметров:
                   Оценка    Ст.ошибка      t-крит   p-знач.
c:(Intercept) 1.4520e+03 1.0886e+01 1.3338e+02  0.0000
d:DLM         1.7361e+03 1.8922e+01 9.1754e+01  0.0000
d:MHA         1.8685e+03 4.3930e+01 4.2534e+01  0.0000
e:DLM         3.8946e-02 1.0184e-02 3.8241e+00  0.0187
e:MHA         1.1104e-01 2.8484e-02 3.8984e+00  0.0176
Стандартное отклонение для остатков:
 11.14285 (4 степеней свободы)
Сравнение характеристик протестированных моделей
        logLik       IC Lack of fit  Res var
MM.3 -30.81844 73.63688          NA 124.1630
AR.3 -30.86712 73.73424          NA 125.5135
```

Обратим внимание, что задав параметр `pmodels`, мы свели в одну точку, соответствующую контрольному результату с нулевой дозой, начало обеих кривых зависимостей.

При сравнении статистических показателей двух протестированных моделей можно обнаружить некоторое, хотя и микроскопическое преимущество у модели Михаэлиса-Ментен (см. график на рис 4.8). Используем эту модель для расчета изоэффективных доз и сравнительной оценки привеса цыплят при действии каждого из препаратов метионина. В простейшем случае можно просто сравнить между собой коэффициенты `e:MHA` и `e:DLM`, т.е. ED_{50}^{2}/ED_{50}^{1} = 0.111/0.0389 = 2.851, но функция `SI()` дополнит расчет доверительными интервалами этого отношения.

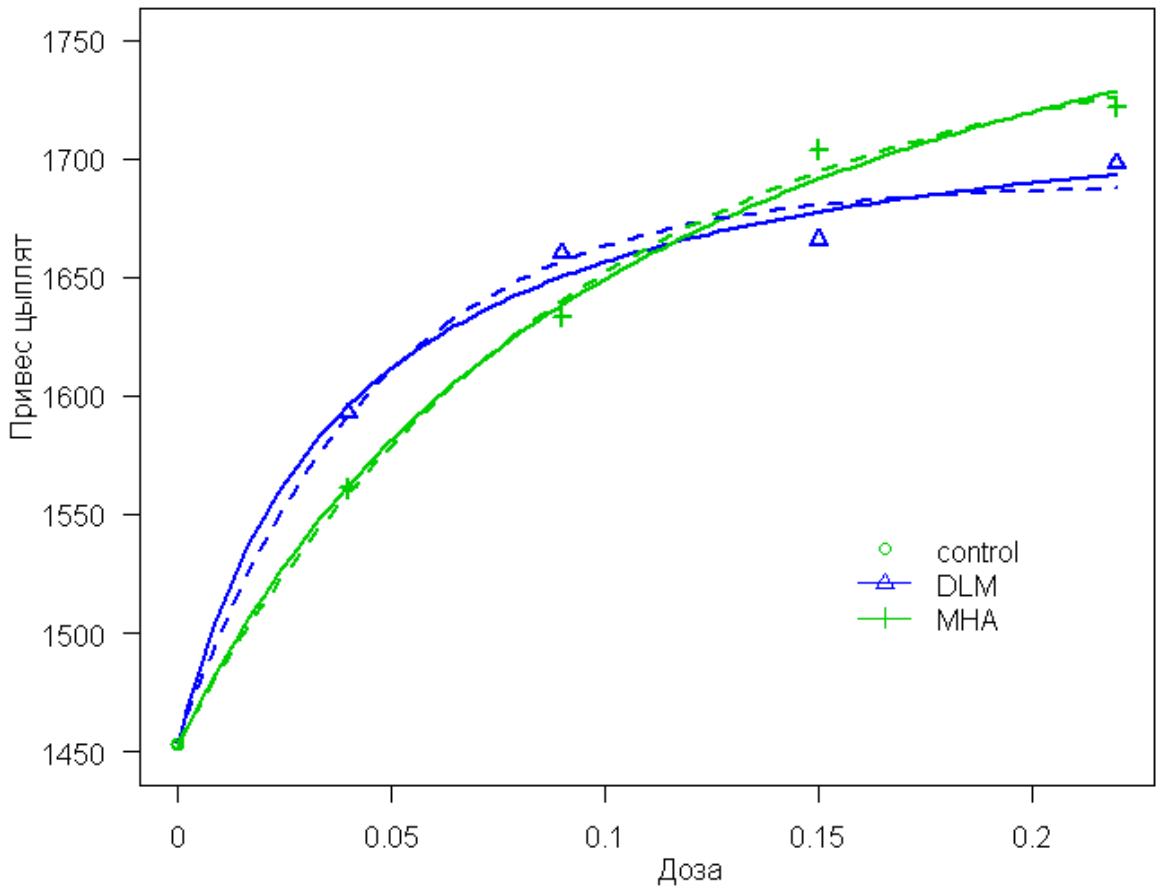


Рис. 4.8. Кривые привеса под действием двух препаратов метионина для модели Михаэлиса-Ментен и асимптотической регрессии (штриховая линия)

```
ED(met.mm.m1, c(10, 50), interval="delta")
SI(met.mm.m1, c(50,50), interval="delta")
```

```
Оценки изоэффективных доз
(доверительные интервалы основаны на дельта-методе(s))
                 Оценка   Ст.ошибка  Нижний Верхний
DLM:10        0.0043274  0.0011316 0.0011855 0.0075
DLM:50        0.0389462  0.0101843 0.0106699 0.0672
MHA:10        0.0123379  0.0031649 0.0035508 0.0211
MHA:50        0.1110410  0.0284837 0.0319576 0.1901
Оценка отношения изоэффективных доз
(доверительные интервалы основаны на дельта-методе (s))
                      Оценка Верхний Нижний
MHA/DLM:50/50        2.85114 0.18494 5.5173
```

Выше мы рассматривали исключительно нелинейные модели, где отклик монотонно изменяется при увеличении дозы. Однако, как правило, это далеко не так: достаточно вспомнить изречение Парацельса (1494— 1541) «Все вещества являются одновременно ядом и лекарством и различие лишь в величине дозы». Поэтому мы не должны упускать из виду модели с

промежуточным оптимумом, когда при низких концентрациях ксенобиотика наблюдается обратная тенденция снижения вредного действия.

Таблица `lettuce` содержит данные эксперимента с выращиванием салат (*Lactuca sativa*), в питательную среду которого включался изобутанол с концентрацией `conc` (мг/л). Через 21 день салат был срезан и взвешивалась его биомасса `weight`. Сравним результаты моделирования данных лог-логистической регрессией `LL.3` и моделью Brain-Cousens `BC.4`.

```
head(lettuce)
## Монотонная модель доза-эффект
lettuce.m1 <- drm(weight~conc, data=lettuce, fct=LL.3())
## Модель, предложенная van Ewijk and Hoekstra (1994)
lettuce.m2 <- drm(weight~conc, data=lettuce, fct=BC.4())
summary(lettuce.m2)
## Гормезис-эффект высоко значим
anova(lettuce.m1, lettuce.m2)
plot(lettuce.m2, broken = TRUE, lwd=2,
       xlab="Концентрация", ylab="Биомасса")
plot(lettuce.m1, add = TRUE, broken = TRUE, type = "none",
          lwd=2, col = 3)
legend("topright",c("BC.4","LL.3"), lwd=2, col=c(1,3))
```

```
Подбираемая модель: Brain-Cousens (гормезиз)
   с нижним пределом, зафиксированным в 0 (4 параметра)
Оценки параметров:
                    Оценка    Ст.ошибка      t-крит   р-знач.
b:(Intercept)  1.282812   0.049346 25.996404  0.0000
d:(Intercept)  0.967302   0.077123 12.542350  0.0000
e:(Intercept)  0.847633   0.436093  1.943698  0.0806
f:(Intercept)  1.620703   0.979711  1.654266  0.1291
Стандартное отклонение для остатков:
 0.1117922 (10 степеней свободы)

Таблица дисперсионного анализа ANOVA
1 модель    fct:      LL.3()
2 модель     fct:     BC.4()
          ModelDf     RSS Df F крит р знач.
1 модель       11 0.24222
2 модель       10 0.12498  1  9.3817  0.0120
```

Дисперсионный анализ ANOVA остатков модели показал, что адекватность модели с промежуточным оптимумом `BC.4` статистически значимо выше, чем у монотонной модели `LL.3`. В этом легко также убедиться, рассмотрев графики кривых на рис. 4.9.

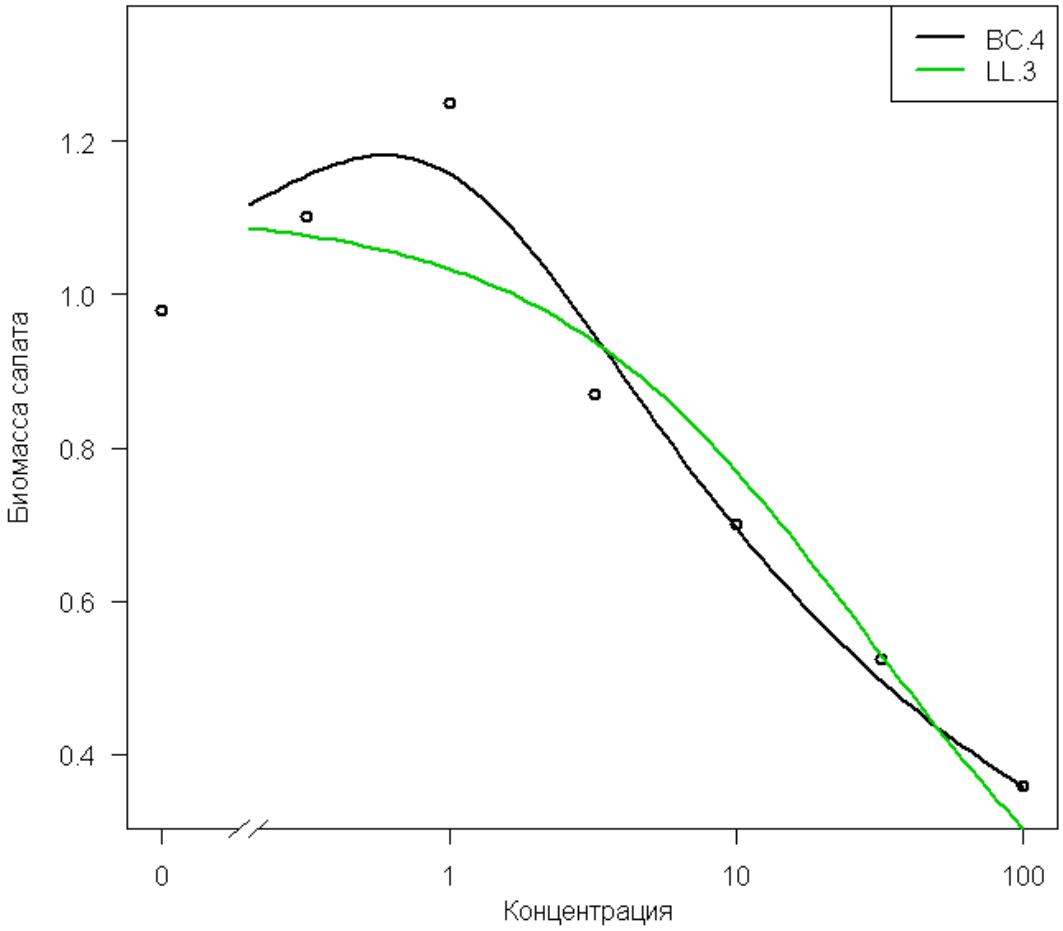


Рис. 4.9. Изменение биомассы салата при внесении раствора изобутанола, описанное монотонной `LL.3` и немонотонной `BC.4` моделями

4.6. Тест на аддитивность воздействия смеси токсикантов

Как обсуждалось выше в разделе 1.4, антагонизм и синергизм токсического действия смеси веществ являются отклонениями от модели суммирования концентраций CA (Concentration Addition). Рассмотрим процедуру оценки статистической значимости наличия комбинаторного эффекта с использованием функции `mixture()` пакета `drc`.

Набор данных `acidiq` содержит результаты эксперимента (Soerensen et al, 2007) по оценке снижения относительных темпов роста `rgr` макрофитовых водорослей *Lemna minor* под воздействием дозы `dose` смеси двух гербицидов ацифлуорфена и диквата в различных соотношениях `pct`. Построим вначале серию лог-логистических моделей «доза-эффект» для каждого соотношения компонентов.

```
library(drc)
head(acidiq)
## Построение модели со свободной вариацией значений ED50
acidiq.free <- drm(rgr ~ dose, pct, data = acidiq,
    fct = LL.4(),
   pmodels = list(~factor(pct), ~1, ~1, ~factor(pct) - 1))
summary(acidiq.free)
```

```
plot(acidiq.free,xlab="Концентрация", ylab="Относительный
        прирост",col=c(1,4,1,1,1,1,1,3), lty=c(NA,1:6,1))
```

```
Подобранная модель: лог-логистическая
                    (ED50 как параметр) (4 параметра).
Оценки параметров:
                  Оценка  Ст.ошибка    t-крит   р-знач.
b:100          1.3589e+00 1.1035e-01 1.2315e+01       0
b:83           1.7675e+00 1.5803e-01 1.1185e+01       0
b:67           2.1577e+00 2.0216e-01 1.0673e+01       0
b:50           2.2777e+00 2.2913e-01 9.9407e+00       0
b:33           2.2302e+00 2.5416e-01 8.7746e+00       0
b:17           2.5058e+00 2.6607e-01 9.4176e+00       0
b:0            2.3076e+00 2.5911e-01 8.9060e+00       0
c:(Intercept) 2.9700e-02 3.0952e-03 9.5953e+00       0
d:(Intercept) 3.0209e-01 2.5854e-03 1.1684e+02       0
e:100          3.0844e+02 2.1265e+01 1.4504e+01       0
e:83           3.7660e+02 2.2280e+01 1.6903e+01       0
e:67           4.8746e+02 2.6072e+01 1.8697e+01       0
e:50           5.1669e+02 2.6541e+01 1.9468e+01       0
e:33           5.2288e+02 2.8379e+01 1.8425e+01       0
e:17           3.7891e+02 1.8619e+01 2.0352e+01       0
e:0            3.4766e+02 1.7712e+01 1.9628e+01       0
Стандартное отклонение для остатков:
 0.01681793 (164 степеней свободы)
```

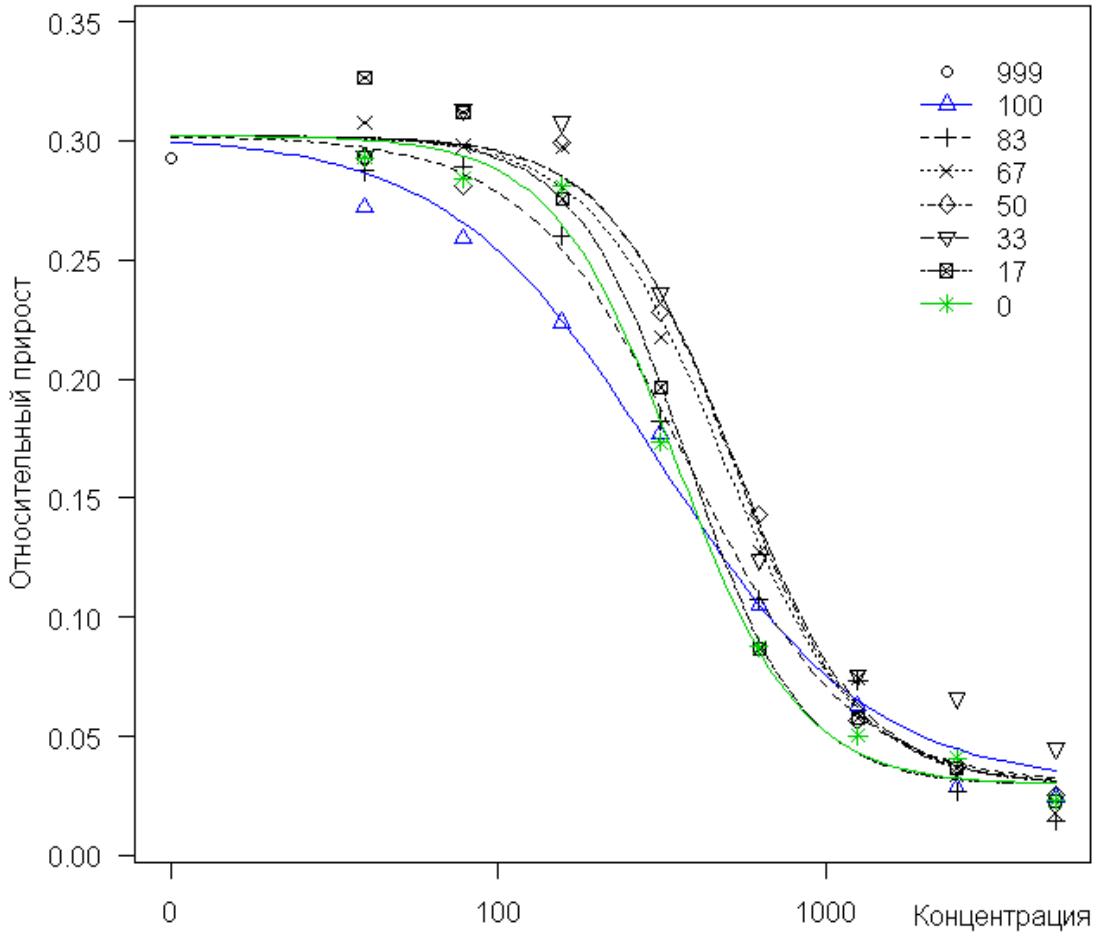


Рис. 4.10. Изменение темпов роста водорослей при обработке смесями двух гербицидов с различным соотношением компонентов

Каждая из 7 кривых на рис. 4.10 соответствует смеси с различным содержанием ацифлуорфена по соотношению к диквату (от 0 до 100%) и определяется своим комплектом коэффициентов *b* угла наклона и *e* (ED_{50}). С использованием параметра `pmodels` мы установили для всех моделей единые значения коэффициентов *d* и *c*, определяющих верхний и нижний пороги сигмоиды.

На графике отчетливо видно, что токсичность смеси ниже, чем каждого из составляющих компонентов в отдельности, т.е. имеет место антагонизм. Однако проверим это предположение статистическими методами. Построим с использованием функции `mixture()` аддитивную модель CA и проверим с использованием теста ANOVA, насколько значимо ухудшилась ошибка регрессии:

```
## Построение  модели аддитивности концентраций
acidiq.ca <- mixture(acidiq.free, model = "CA")
plot(acidiq.ca)
summary(acidiq.ca)
## Сравним с моделью свободной вариации e-параметра
anova(acidiq.ca, acidiq.free)
```

```
Подобранная модель: Аддитивности в смеси (5 параметров)
Оценки параметров:
                        Оценка    Ст.ошибка   t-крит   p-знач.
b:100                 1.3844e+00 1.4407e-01 9.6096e+00       0
b:83                  1.7642e+00 1.9210e-01 9.1841e+00       0
b:67                  2.1972e+00 2.6153e-01 8.4014e+00       0
b:50                  2.2111e+00 2.7359e-01 8.0815e+00       0
b:33                  2.4103e+00 3.4444e-01 6.9977e+00       0
b:17                  2.4239e+00 3.1130e-01 7.7865e+00       0
b:0                   2.1510e+00 2.7739e-01 7.7544e+00       0
c:(Intercept)         2.8821e-02 3.8833e-03 7.4219e+00       0
d:(Intercept)         3.0000e-01 3.1660e-03 9.4756e+01       0
e:I(1/(pct/100))      4.2566e+02 2.2587e+01 1.8845e+01       0
f:I(1/(1 - pct/100)) 4.2800e+02 1.9844e+01 2.1568e+01       0
Стандартное отклонение для остатков:
 0.02080481 (169 степеней свободы)

Таблица дисперсионного анализа ANOVA
1 модель      fct:      CA модель
2 модель      fct:      LL.4()
          ModelDf      RSS  Df  F крит p знач.
1 модель       169 0.073150
2 модель       164 0.046386  5  18.925   0.000
```

В построенной модели все кривые проходят через центр тяжести, соответствующий единой для всех смесей среднеэффективной дозе $ED_{50} = 426$. Однако ошибка регрессии такой модели слишком велика, т.е. сделанные нами предположения об аддитивности следует отклонить.

Построим теперь модель Хьюлита:

```
## Построение модели Хьюлита
acidiq.hew <- mixture(acidiq.free, model = "Hewlett")
summary(acidiq.hew)
anova(acidiq.free, acidiq.hew)
## Построение графика изоболы на основе модели Хьюлита
isobole(acidiq.free, acidiq.hew, xlim = c(0, 420),
    ylim = c(0, 450), xlab="Ацифлуорфен", ylab="Дикват")
segments(426,0,0,426, col=4, lwd=2)
segments(308,0,0,348, col=3, lwd=2)
```

```
Подобранная модель: Хьюлита для смеси (6 параметров)
Оценки параметров:
                        Оценка    Ст.ошибка   t-крит   p-знач.
b:100                 1.3704e+00 1.1184e-01 1.2253e+01   0e+00
b:83                  1.7757e+00 1.5964e-01 1.1123e+01   0e+00
b:67                  2.1808e+00 2.0685e-01 1.0543e+01   0e+00
b:50                  2.2925e+00 2.3345e-01 9.8198e+00   0e+00
b:33                  2.3154e+00 2.6237e-01 8.8252e+00   0e+00
b:17                  2.4666e+00 2.5919e-01 9.5167e+00   0e+00
b:0                   2.3347e+00 2.6714e-01 8.7397e+00   0e+00
c:(Intercept)         3.0042e-02 3.0711e-03 9.7820e+00   0e+00
d:(Intercept)         3.0176e-01 2.5825e-03 1.1685e+02   0e+00
e:I(1/(pct/100))      3.1683e+02 1.3191e+01 2.4020e+01   0e+00
f:I(1/(1 - pct/100)) 3.3710e+02 1.1814e+01 2.8534e+01   0e+00
g:(Intercept)         2.8063e-01 6.9489e-02 4.0385e+00   1e-04
Стандартное отклонение для остатков:
 0.01692075 (168 степеней свободы)

Таблица дисперсионного анализа ANOVA
1 модель     fct:     модель Хьюлита
2 модель     fct:     LL.4()
          ModelDf     RSS  Df  F крит p знач.
1 модель      168 0.048100
2 модель      164 0.046386  4  1.5151  0.2001
```

В модель включен дополнительный параметр – показатель взаимодействия или λ Хьюлита g = 0.28. Принимая во внимание стандартную ошибку этого коэффициента s = 0.07, можно утверждать, что λ статистически значимо меньше 1 и антагонизм токсического действия не противоречит результатам эксперимента. Кроме того, тест ANOVA показал, что значимых искажений основной модели "доза-эффект" не произошло, поскольку остаточные суммы квадратов обеих регрессий не отличаются между собой по F-критерию.

Наконец, комбинаторный характер взаимодействия токсикантов можно графически представить изоболами на рис. 4.11, что выполняется функцией `isobole()`. Точками представлены использованные в эксперименте комбинации смесей, концентрация которых соответствует ED_{50} (вместе с их стандартными ошибками). Показаны изоболическая кривая Хьюлита при λ = 0.28, прямая, соединяющая точки изоэффективных доз индивидуальных

токсикантов (зеленым цветом) и изоболическая прямая, построенная при предположении об аддитивности воздействия (синим цветом).

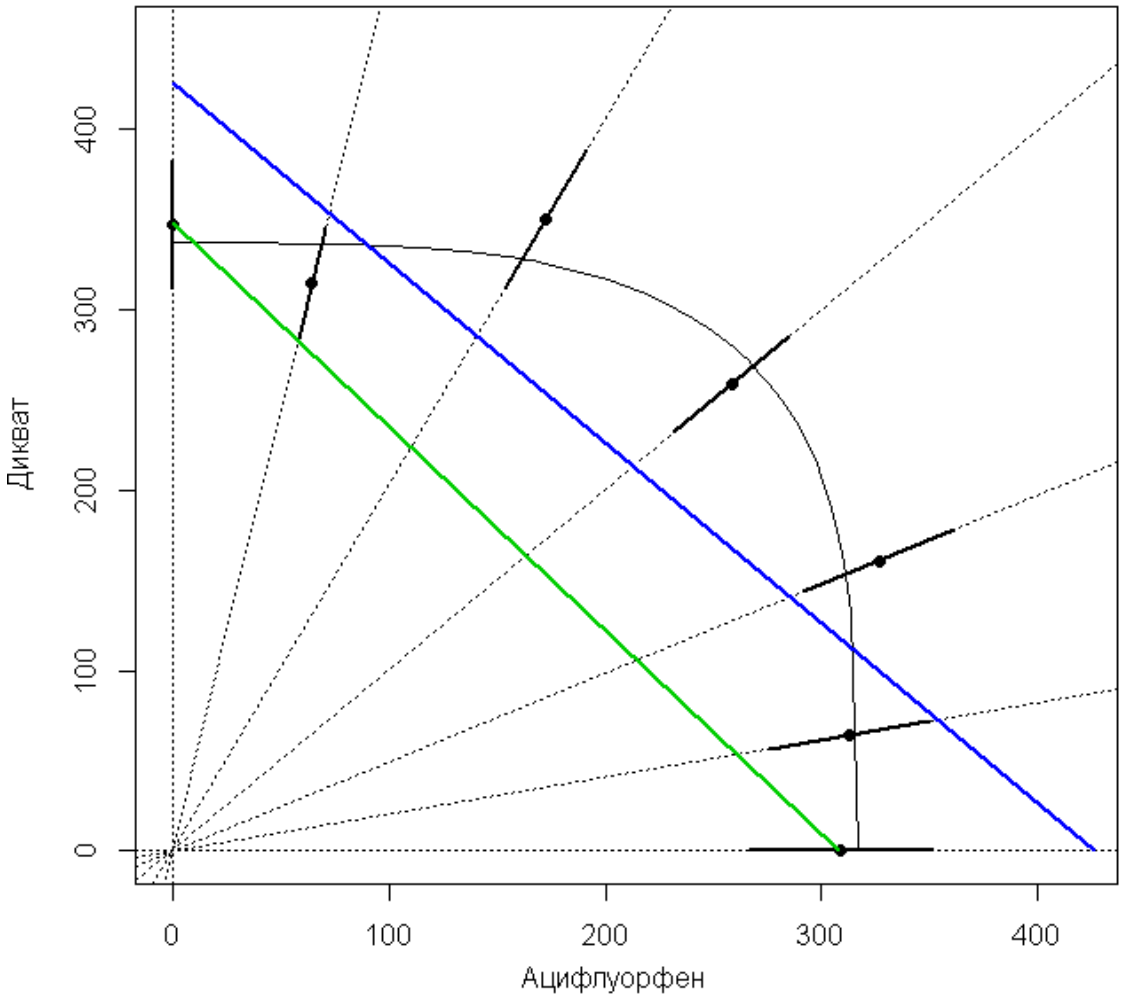


Рис. 4.11. Изоболы аддитивной модели и модели Хьюлита

5. РАСПРЕДЕЛЕНИЕ ЧУВСТВИТЕЛЬНОСТИ ВИДОВ И ОЦЕНКА РИСКА

5.1. Оценка безопасных уровней воздействия для биоценозов

Основным предметом исследований экотоксикологии являются биологические системы надорганизменного уровня, подверженные техногенному загрязнению. Естественно, что общей теоретической основой являются принципы устойчивости и стабильности природных систем, которые находятся в процессе активного осознания современной теоретической экологией. Это определяет своеобразие применяемых концептуальных подходов и методик экотоксикологии, оценивающих степень нарушения популяционных и биоценотических механизмов.

Живые организмы составляют огромное разнообразие по таксономическим признакам, биологическим циклам, физиологии, морфологии, особенностям поведения и географическому местоположению. Столь же разнообразен и отклик разных организмов на токсическое воздействие. Эти различия могут объясняться, например, на уровне организменных механизмов: кинетикой накопления-поглощения-выделения, внутренними факторами изолирования и биотрансформации, природой биохимических рецепторов и скоростью их восстановления, эффективности механизмов регенерации и т.д. При переносе изучаемого объекта в естественную среду на характер биотического отклика начинают оказывать значимое влияние абиотические факторы, питание, внутрипопуляционные и биоценотические связи.

В 80-х годах прошлого века экотоксикологи пришли к выводу, что анализ межвидовой изменчивости к ксенобиотикам – не только их приоритетная задача, но и основа для поиска решений. «Легко увидеть, что чувствительность различных видов к воздействию большинства токсикантов (оцениваемая по LC_{50} или LD_{50}) имеет выраженный характер вариационного ряда, распределенного по логнормальному закону. Таким образом, тест на токсичность каждого вида не является представительной оценкой для любых других видов, но служит одной из составляющих общего критерия чувствительности всего биоценоза» (Mount, 1982).

Но что нужно сделать, если в ходе длительных и планомерных исследований нам оказалось доступным некоторое множество значений *NOEC* или LC_{50}? Проанализировав значения LC_{50} 14 химических пестицидов в морской воде для различных видов, включая бактерии, морские водоросли, ракообразных, насекомых, рыбу, и амфибии, С. Кооймaн (Kooijman, 1987) обосновал понятие опасной концентрации для чувствительных видов HCS (hazardous concentration for sensitive species). Алгоритм ее оценки для локального сообщества, включающего *m* видов (например, выборка беспозвоночных, живущих в определенном водоеме), представлен на рис. 5.1 (Posthuma, 2001).

- предполагается, что значения показателей токсикометрии log(LC_{50}) для всех возможных водных беспозвоночных могут быть аппроксимированы симметричным колоколообразным логлогит-распределением;

- из этой генеральной совокупности извлекается выборка из n видов, имеющих наименьшие значения log*(LC_{50}); тогда эталонная $\hat{HCS}$ определена как концентрация, вероятность превышения которой нормативами самых чувствительных видов из n равняется произвольному небольшому значению p;

- если по результатам наблюдений (например, гидробиологической съемки) в водоеме обнаружено m видов, то можно построить аналогичное распределение опасных концентраций и оценить его параметры x_m и s_m; небольшое число δ задает вероятность того, что истинная безопасная концентрация HCS окажется больше, чем оцененная по параметрам

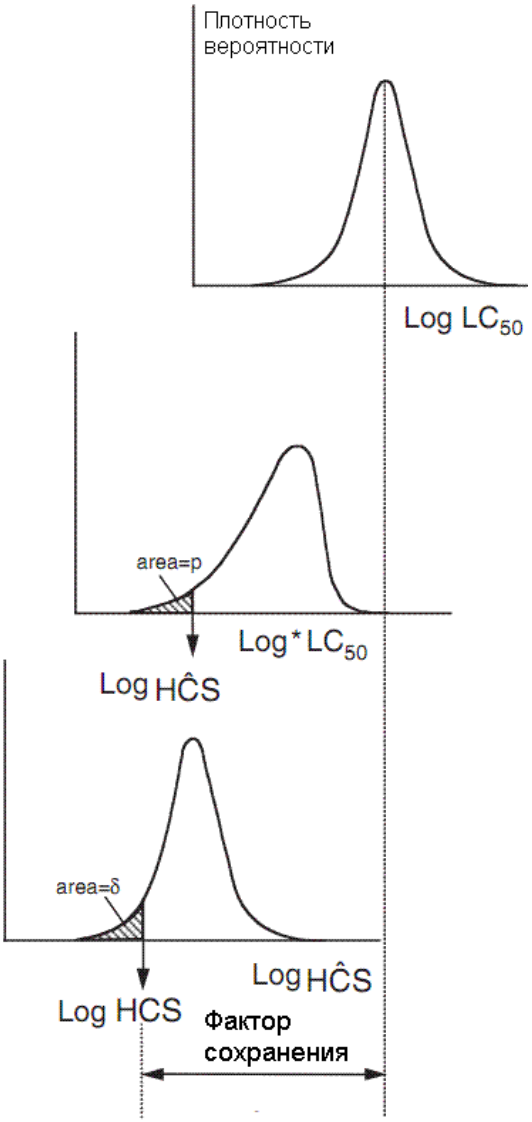


Рис. 5.1. Графики, иллюстрирующие алгоритм вычисления безопасных концентраций для чувствительных видов HCS

В итоге алгоритм, разработанный Кооймaном, может быть записан в виде следующего уравнения:

$$HCS = \exp\,[x_m - f\,(p, \delta, m, n)\, s_m]$$

где x_m и s_m – среднее и стандартное отклонение располагаемых m значений HCS после логарифмирования, а f является функцией, выражение для которой выбирается в зависимости от остальных параметров (Kooijman, 1987). Разность членов в квадратных скобках соответствует "фактору сохранения" (safety factor), или, с иной точки зрения, диапазону опасных концентраций для группы наиболее чувствительных видов.

В представленной статистической модели учитывалось, что при увеличении числа видов в сообществе n безопасная концентрация уменьшается (достигая 0 при $n \to \infty$). Это аргументировалось тем, что с увеличением n появляется больше редких видов, нуждающихся в охране.

Отсюда утверждение, что океан с его громадным обилием видов нуждается в более жестких нормативах загрязнения, чем ручей (этот довод может показаться сомнительным).

Поскольку выборка из m значений составляет лишь небольшую часть генеральной совокупности, то оценки x_m и s_m обычно находятся с некоторой погрешностью. Степень потери доверия за счет этой неопределенности задается другим небольшим числом δ. Кроме того значение HCS уменьшается с уменьшением объема m выборки и становится очень низким при $m < 6$. Такой встроенный штраф за потерю доверия к небольшой обучающей выборке также не лишен оснований.

Хотя понятие HCS, предложенное Коойманом, уже включало основные элементы вероятностной методологии оценки риска, однако нормативы, вычисленные по этой модели, на практике никогда не применялись в качестве реальных критериев качества окружающей среды. Главная причина была в том, что оценки HCS неизменно оказывались очень низкими (обычно ниже любого предела чувствительности или концентрации фона). Алгоритм стал подвергаться не слишком принципиальным усовершенствованиям (Van Straalen, Denneman, 1989; Aldenberg, Slob, 1993), пока, наконец, основная его идея не оформилась в целое научное направление с аббревиатурой SSD (Species Sensitivity Distribution), т.е. распределение чувствительности видов.

5.2. Общие принципы моделирования чувствительности видов

Вариационный ряд показателей токсикометрии (LC_{50} или $NOEC$) может быть использован для описания чувствительности видов некоторой параметрической функцией статистического распределения, такого как треугольное, нормальное, или логистическое. Поскольку истинные параметры модели неизвестны, то доступные экотоксикологические данные интерпретируются как независимая выборка из этого распределения и используются, чтобы оценить выборочные характеристики (например, среднее и дисперсию).

Графически SSD обычно представляется как кумулятивная кривая распределения CDF (Cumulative Distribution Function), соответствующая интегралу функции плотности вероятности – рис. 5.2 (Posthuma, 2001). Как и в случае зависимости "доза-эффект" (рис. 1.1), стрелки на графике указывают, что кривая чувствительности может использоваться как "прямым" так и "обратным" способом. При обратном использовании устанавливаются критерии качества окружающей среды (environmental quality criterion, EQC), т.е. на оси Y задают вероятность p, обеспечивающую защиту от воздействия (1 - p) % видов, и в результате оценивается требуемая безопасная концентрация HC_p (hazardous concentration for p% of the species).

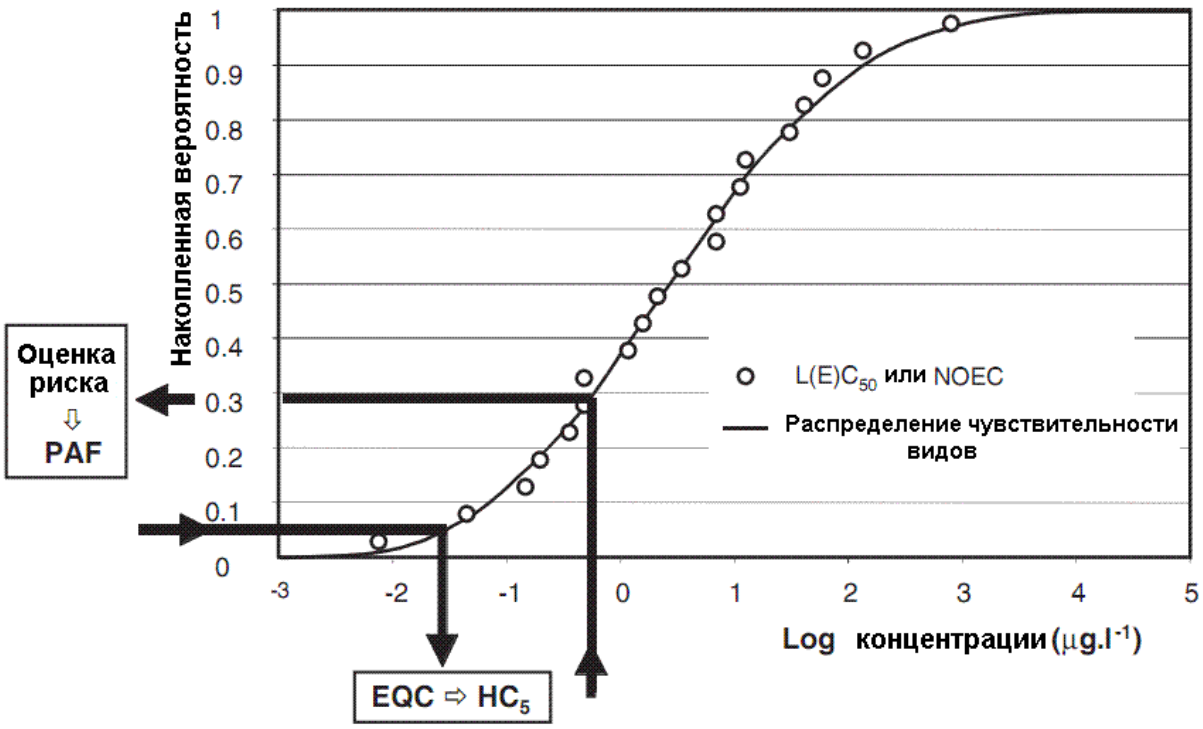


Рис. 5.2. Оценка качества окружающей среды (EQC), безопасных концентраций (HC_p) и потенциально затронутой фракции видов (PAF) с использованием распределения чувствительности видов (SSD)

Прямое применение кривой чувствительности связано с оценками экологического риска на загрязненных участках в соответствии со значениями концентраций, выбранными на шкале X (по результатам натурных наблюдений или из иных соображений). Тогда с использованием SSD может быть установлены "потенциально затронутые фракции" (PAF - potentially affected fraction) видов для каждой из этих концентраций. Если доля *p* % числа видов от их общего количества, которые могут исчезнуть из изучаемого сообщества при наблюдаемой концентрации токсиканта *EC*, превысит порог допустимого риска (например, 5%), то это может представлять существенную угрозу. Характер этой опасности соответствует смыслу самих экотоксикологических показателей, использованных для построения SSD.

Как обычно в подобных случаях, реализация метода требует трех шагов: *а*) формирование выборки токсикометрических данных, *б*) статистический анализ этих данных и *в*) интерпретацию полученных результатов.

Набор данных по токсикометрии видов обычно формируется по результатам лабораторного биотестирования спектра видов изучаемого сообщества для каждого загрязняющего вещества, входящего в программу исследований (или их смеси). Эта выборка должна быть *экологически представительной* для изучаемой композиции видов и *статистически*

репрезентативной. Для качественной аппроксимации функций распределения и несмещенной оценки пороговых концентраций общее число анализируемых видов должно превышать 30 (по другим оценкам – от 15 до 55 – Newman et al., 2000).

Основное требование статистики – набор таксономических групп, по которым моделируется SSD, должен быть случайной выборкой из регионального фонда видов, которые нуждаются в природоохранных мероприятиях. В некоторых случаях оценка безопасных концентраций требует минимального таксономического разнообразия до нескольких родов или семейств. С другой стороны, виды для токсикометрического теста обычно отбираются совсем по иным критериям: насколько они подходят для лабораторного эксперимента, какова их чувствительность к токсичным веществам, насколько вообще уместно их охранять и т.д. Если взять истинную случайную выборку из всего видового богатства биосферы (исключая бактерии), то она, по крайней мере, на 50 % состояла бы из насекомых. Поэтому на практике исходный набор данных обычно определен доступной информационной базой параметров токсикометрии.

Для поиска необходимой информации могут быть использованы такие базы экотоксикологических данных как ECOTOX (Американское агентство охраны окружающей среды US EPA), ECHA Европейского химического агентства или ETOX. Предпочтительнее было бы использовать результаты хронического теста, но обычно для оценки экологического риска используют параметры острой токсичности, которые легкодоступны, хорошо стандартизованы и часто достаточно уместны по условиям экспозиции.

Попытка при помощи SSD и подобных методов статистической экстраполяции распространить содержательные выводы, полученные на ограниченной выборке данных о токсикометрии видов, на весь остальной реальный мир основана на не вполне корректных предположениях и вызывает многочисленные критические замечания. Во-первых, условия химического воздействия на животных при лабораторных испытаниях могут сильно отличаться от полевых условий по самым важнейшим факторам: транспорту загрязняющих веществ к рецепторам организма, условиям экспозиции, характеру биоаккумуляции и т.д.

Во-вторых, представленный метод оценки экологических нормативов, по сути, никак не использует информацию об экологии сообществ (межвидовых взаимодействиях, трофических связях, условиях среды обитания или относительной значимости ключевых видов и функциональных групп). Например, в реальных условиях комбинации совместно встречающихся видов часто составляются, в значительной мере, случайно и вероятность появления каждого вида определяется как факторами среды, так и плотностью распределения относительного обилия организмов для каждого сообщества.

Наконец, еще один аргумент связан с тем, что принцип охраны возможно большего числа видов более консервативен, чем защита функций экосистемы. К сожалению, экология пока не может дать точного ответа, как видовое разнообразие связано с функциями системы: по этой проблеме высказываются, как минимум, три различных гипотезы (Lawton, 1994). Наибольшей поддержкой пользуется идея функциональной избыточности: потеря одного из ключевых видов обычно компенсируется ростом обилия других видов в той же функциональной группе. Экспериментально установлено (Klepper et al., 1999), что основные функции экосистемы обычно деградируют только при высоких уровнях загрязнения, особенно в сообществах с высоким биологическим разнообразием. Несмотря на приведенные частные возражения, основную концепцию, направленную на охрану возможно большего числа видов, можно считать вполне справедливой.

Многие специалисты по охране окружающей среды высказывают возражения против идей HC_p на том основании, что *p* % видов в сообществе все же остаются "незащищенными". Охрана 95% видов в сообществе не всегда означает охрану самого сообщества. Например, Хопкин (Hopkin, 1993), обсуждая экологическое нормирование металлов в почве, обратил внимание, что, если в исчезнувшие 5 % войдут виды земляных червей, то это приведет к полной деградации поступления питательных веществ. Эта аргументация не имеет прямого смысла, поскольку назначение критических уровней *p* – дело не науки, а природоохранной политики. В любом случае, попытка SSD использовать для оценки экологического риска лабораторные данные по токсичности отдельных видов предоставляет специалистам полезную информацию об охраняемых сообществах

5.3. Статистические аспекты построения SSD с использованием R

Выше отмечалось, что распространение на изучаемые экологические сообщества выводов, полученных из распределения чувствительности видов (SSD), связано с несколькими плохо проверяемыми предположениями: *а*) выборочные данные соответствуют некоторому теоретическому распределению, *б*) объем выборки достаточен для корректного предсказания, и *в*) виды, включенные в набор данных, являются представительными индикаторами экологического состояния.

Впрочем, рассматривая чувствительность видов как статистически случайную величину, справедливо указывается (Suter, 1998), что не вполне правильно рассматривать ее распределение как вероятностное. Каковы тут формы изменчивости, обуславливающие распределение вероятности? С одной стороны, измерение токсикометрических показателей всегда связано со случайной ошибкой, обусловленной погрешностью

эксперимента. Другая составляющая изменчивости – «чувствительность каждого вида оценена без ошибки, но одни виды существенно более чувствительны к токсиканту, чем другие». Вторую часть нельзя считать вероятностной, поскольку обуславливающий ее механизм является полностью детерминированным. Если не принимать во внимание токсикометрическую ошибку, то зависимость SSD представляет собой последовательное нарастание эффекта вредного действия, а не накопление вероятности. Поэтому, как и в общем случае при аппроксимации кривых "доза-эффект", зависимость чувствительности видов может быть "подогнана" под данные не только кумулятивной кривой распределения вероятности, но и любой иной подходящей статистической моделью, включая, например, сглаживание ядерными функциями.

Вариационный ряд токсикометрических показателей представляет собой выборку дискретной случайной величины, интегральная функция распределения которой определена как

$$F(x) = P(X < x) = \sum_{x_i < x} P(X = x_i),$$

где неравенство $x_i < x$ под знаком суммы указывает, что суммирование распространяется на все те значения x_i , которые меньше x. Эмпирическая функция распределения – всегда разрывная ступенчатая функция, скачки которой происходят в точках, соответствующих возможным случайным значениям величины, и равны вероятностям этих значений. На графике Хазена (Hazen – рис. 5.3) при числе видов n = 7 выделяется 7 ступенек высотой 100/n каждая, причем каждое значение $p_i = 100(i - 0.5)/n$.

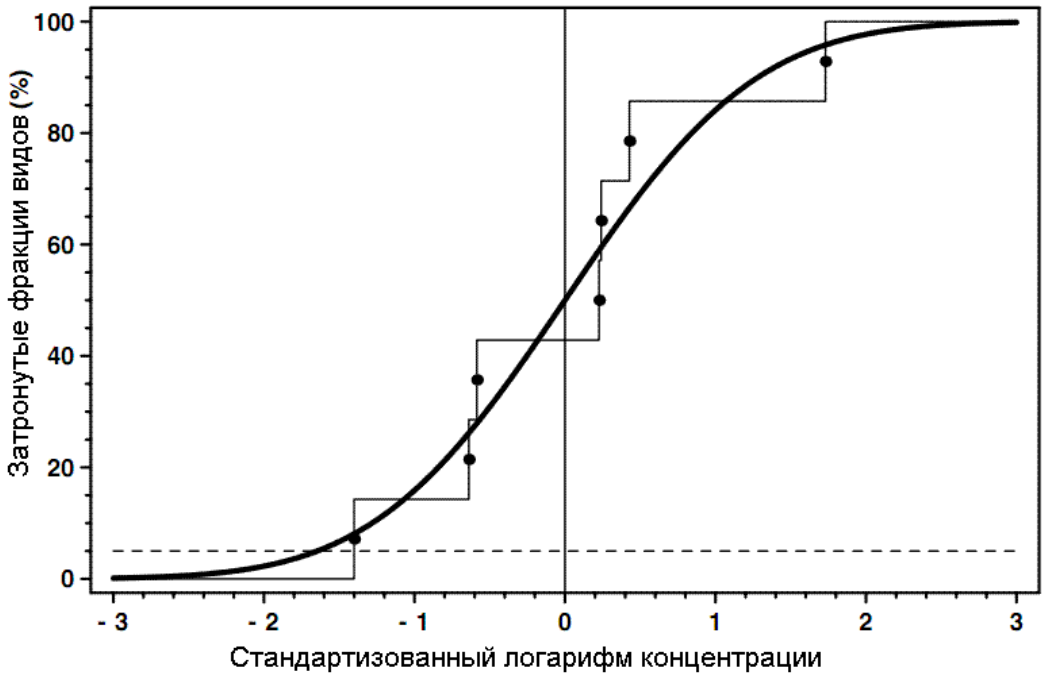


Рис. 5.3. Построение интегральной функции распределения на графике Хазена.

Наилучшая аппроксимация ступенчатой кривой тем или иным теоретическим распределением осуществляется с использованием различных алгоритмов: оценкой максимального правдоподобия, методом моментов или квантилей, а также на основе критериев качества подгонки, таких как Крамера-фон Мизеса, Колмогорова-Смирнова, Андерсона-Дарлинга и др.

Рассмотрим процесс моделирования распределения чувствительности видов (SSD) на примере, представленном в блоге Интернет «Data in Environmental Science and Ecotoxicology» (http://edild.github.io/ssd/ , автор Eduard Szöcs). Файл с исходными данными `ssd_data.csv` содержит данные о летальной концентрации LC_{50} (`val`) распространенного пестицида хлорпирифоса для 36 видов водных беспозвоночных (`species`).

```
# Данные можно загрузить как с ресурса github
# require(RCurl)
# url <- getURL("https://raw.githubusercontent.com/EDiLD/r-
#     ed/master/post_ssd/ssd_data.csv",
#     ssl.verifypeer = FALSE)
#df <- read.table(text = url, header = TRUE, sep = ',',
#  stringsAsFactors = FALSE)
# Либо из файла, предварительно размещенного на компьютере
df <- read.table(file = "ssd_data.csv", header = TRUE,
    sep = ',', stringsAsFactors = FALSE)
# Сортировка по возрастанию val
df <- df[order(df$val), ]
df$frac <- ppoints(df$val, 0.5)
head(df)
```

После загрузки и сортировки данных по величине токсикометрического показателя функция `ppoints()` генерирует последовательность эмпирических вероятностей для каждого вида (`frac`). Первые шесть строк сформированной таблицы данных имеют вид:

```
                  species       val n       frac
                  -------       --- -       ----
17        Deleatidium sp. 0.0500000 1 0.01428571
28          Procloeon sp. 0.0810000 1 0.04285714
10    Chironomus riparius 0.1749286 2 0.07142857
1    Aedes taeniorhynchus 0.2491987 2 0.10000000
5  Atalophlebia australis 0.2526544 3 0.12857143
34      Simulium vittatum 0.2800000 1 0.15714286
```

Для аппроксимации данных теоретическим распределением, характер которого нам неизвестен, воспользуемся функцией `fitdist()` пакета `fitdistplus`. В анализ включим пять возможных распределений: нормальное, логнормальное, логистическое, Коши и Вейбулла. Функция `gofstat()` даст нам возможность сопоставить между собой статистики и

критерии качества подгонки, а функция `cdfcomp()` вывести графики эмпирического и теоретического распределений – см. рис. 5.4.

```
library(fitdistrplus)
no = fitdist(df$val,"norm")
lo = fitdist(df$val,"lnorm")
lg = fitdist(df$val,"logis")
ca = fitdist(df$val,"cauchy")
we = fitdist(df$val, "weibull")
fitlist <- list(no, lo, lg, ca, we)
legendtext=c("Нормальное", "Логнормальное",
    "Логистическое","Коши", "Вейбулла")
cdfcomp(fitlist,xlab="Концентрация хлорпирифоса (лог)",
    xlogscale = TRUE,legendtext=legendtext, lwd=2 )
sapply(fitlist, function(i) i$loglik),
gofstat(fitlist, fitnames=legendtext)
summary((fit<-lo))  # Лучшая модель распределения
```

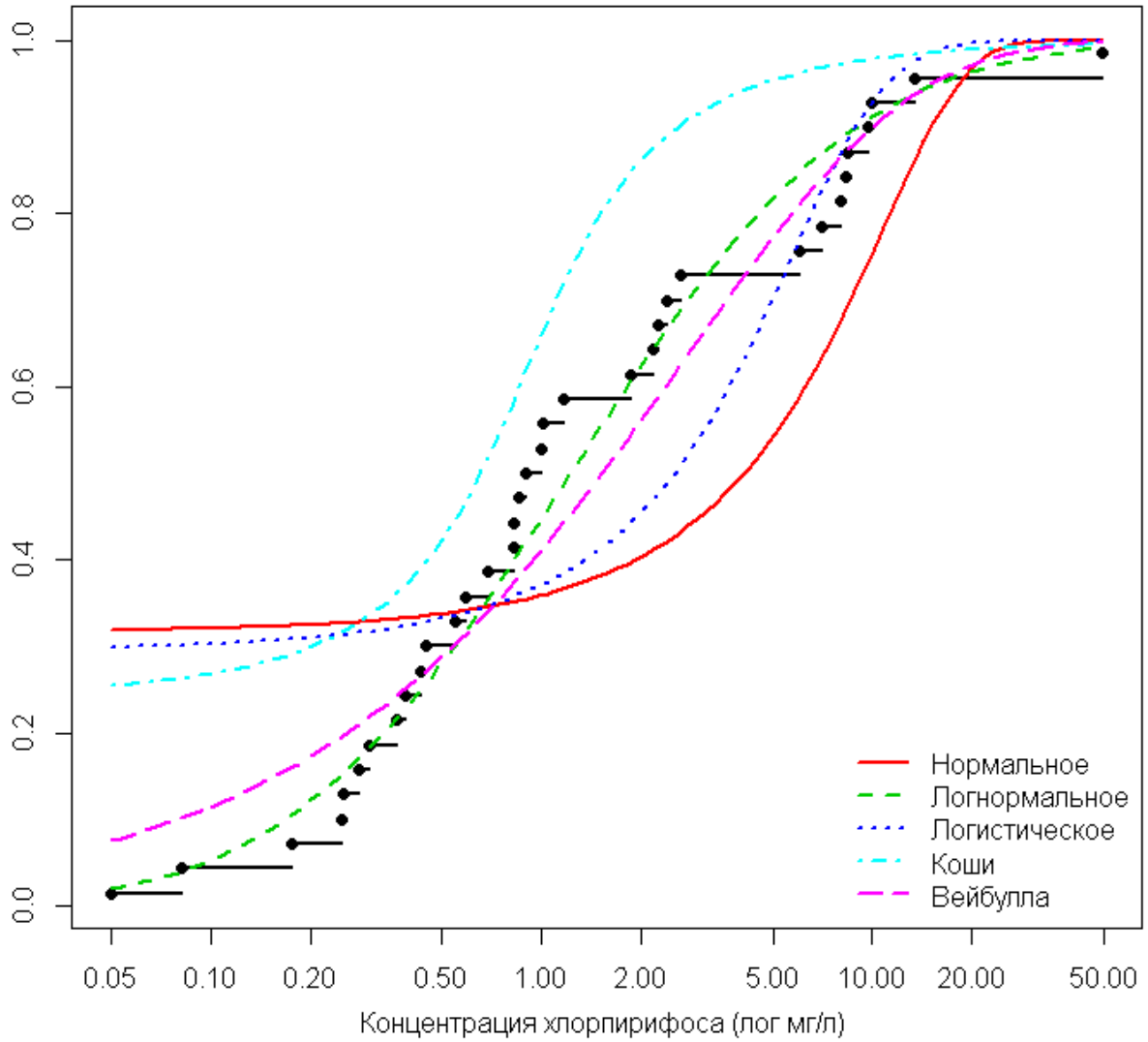


Рис. 5.4. Подбор распределений для аппроксимации эмпирических данных

```
Статистики качества подгонки (Goodness-of-fit)
                      Нормальное Логнормальное Логистическое   Коши  Вейбулла
Колмогорова-Смирнова  0.3187578    0.12007747     0.2991735 0.2549100 0.1575468
Крамера-фон Мизеса    1.0916456    0.07969651     0.6914383 0.8802006 0.1690479
Андерсона-Дарлинга    5.8681905    0.47021561     3.9085268 6.6391518 0.9820729
Критерии качества подгонки (Goodness-of-fit)
Функция правдоподобия -124.88069    -72.09035    -111.55586 -88.21673 -75.49560
Акаике                 253.7614     148.1807      227.1117  180.4335  154.9912
Байесовский            256.8721     151.2914      230.2224  183.5442  158.1019
```

Наилучшая аппроксимация соответствует логнормальному распределению:

```
Параметры подогнанного распределения 'lnorm'
по критерию максимума правдоподобия:
         Оценка    Ст. ошибка
meanlog 0.2023528  0.2620444
sdlog   1.5502758  0.1852930
```

Хотя SSD нельзя считать истинным распределением плотности вероятности, но в оценке его параметров есть элемент неопределенности. Значения параметров μ и σ – неизвестные константы, которые оцениваются по случайной выборке, взятой из генеральной совокупности (в частности, базы токсикометрических показателей). Так как сам метод отбора предполагает некоторый статистический разброс, то в оценке риска возникает вероятностная составляющая. Точнее, сам по себе экологический риск остается детерминированной мерой (то есть, долей видов, затронутых токсическим действием), но из-за выборочной ошибки оценивается на некотором доверительном интервале.

Оценка доверительных интервалов SSD может быть осуществлена с использованием методов генерации повторных выборок, к которым относится бутстреп (bootstrap). *Параметрический бутстреп* использует предположение, что исходные выборочные данные представляют собой случайные реализации вероятностного процесса, определяемого заданным теоретическим распределением (Davison, Hinkley, 2006; Шитиков и др., 2014). Он основан на следущей процедуре:

1. По выборочным данным $\{x_1, x_2, \ldots, , x_n\}$ осуществляется построение заданной модели и оцениваются ее параметры $\hat{\boldsymbol{\theta}}$ (μ и σ в нашем случае).

2. Случайным образом из подобранного распределения с параметрами $\hat{\boldsymbol{\theta}}$ генерируются n элементов $\{x^*_1, x^*_2, \ldots, , x^*_n\}$. Бутстреп-повторность, полученная имитацией, используется для оценки параметров $\hat{\boldsymbol{\theta}}^*$ и других анализируемых статистик.

3. Шаг 2 повторяется большое число раз (например, 1000) и в семействе сымитированных показателей или кривых распределения выделяется доверительная область с заданной доверительной вероятностью.

Определим предварительно скрипт R, содержащий функции, выполняющие генерацию псевдо-выборок.

```
# 1. Функция для нахождения p-квантили случайной выборки из
#  логнормального распределения с параметрами  fit
myboot <- function(fit, p){
  # генерация случайной выборки из заданного распределения
  xr <- rlnorm(fit$n, meanlog = fit$estimate[1],
                      sdlog = fit$estimate[2])
  # подгонка параметров распределения под новые данные
  fitr <- fitdist(xr, 'lnorm')
    hc5r <- qlnorm(p, meanlog = fitr$estimate[1],
                    sdlog = fitr$estimate[2])
  # возвращает значение выборки, соответствующее p-квантили
  return(hc5r)
}
# 2. Функция, возвращающая значения вероятностей для случайной
#  выборки из логнормального распределения с параметрами  fit
myboot2 <- function(fit, newxs){
  # генерация случайной выборки из заданного распределения
  xr <- rlnorm(fit$n, meanlog = fit$estimate[1],
                      sdlog = fit$estimate[2])
  # подгонка параметров распределения под новые данные
  fitr <- fitdist(xr, 'lnorm')
  # прогноз вероятностей под новые данные
  pyr <- plnorm(newxs, meanlog = fitr$estimate[1],
                       sdlog = fitr$estimate[2])
  return(pyr)
}
#---------------------------------------------
set.seed(1234) # Установка генератора случайных чисел
#  Концентрация, соответствующая 5%-му эффекту
(hc5 <- qlnorm(0.05, meanlog = fit$estimate[1],
                     sdlog = fit$estimate[2]))
#  Концентрация с доверительными интервалами
hc5_boot <- replicate(1000, myboot(fit, p = 0.05))
quantile(hc5_boot, probs = c(0.025, 0.5, 0.975))
```

```
Концентрация, соответствующая 5%-му эффекту
0.09559604
Ее доверительные интервалы, найденные бутстрепом
     2.5%       50%     97.5%
0.0460268 0.1024267 0.2144115
```

Функция `myboot()` используется для оценки доверительных интервалов концентрации токсиканта, приводящей к p%-му эффекту. Если принять риск p = 0.05, то, основываясь только на оценках параметров логнормального распределения, мы получим опасную концентрацию HC_5 = 0.096 мг/л хлорпирифоса. Но если рассчитать 1000 значений HC_5 для различных случайных выборок из этого распределения, то можно установить, что с доверительной вероятностью 95% эта концентрация будет находиться на интервале от 0.046 до 0.214 мг/л.

Функция myboot2(), наоборот, для заданного вектора значений концентраций возвращает вектор вероятностей и может быть использована для построения кривой SSD с доверительными интервалами.

```
# новые данные для построения плавной кривой
newxs <- 10^(seq(log10(0.01), log10(max(df$val)),
                                length.out = 1000))
# получение матрицы для построения 1000 кривых
boots <- replicate(1000, myboot2(fit, newxs))
require(reshape2)
bootdat <- data.frame(boots)
bootdat$newxs <- newxs
bootdat <- melt(bootdat, id = 'newxs')
# извлечение доверительных интервалов
cis <- apply(boots, 1, quantile, c(0.025, 0.975))
rownames(cis) <- c('lwr', 'upr')
# добавление в итоговую таблицу подогнанных значений
pdat <- data.frame(newxs, py = plnorm(newxs,
       meanlog = fit$estimate[1], sdlog = fit$estimate[2]))
# добавление доверительных интервалов
pdat <- cbind(pdat, t(cis))
# координаты x для названия видов
df$fit <- 10^(log10(qlnorm(df$frac,
   meanlog = fit$estimate[1], sdlog = fit$estimate[2])) -0.4)
# Вывод полноценного графика с использованием пакета ggplot2
library(ggplot2)
ggplot()+
  geom_line(data = bootdat, aes(x = newxs, y = value, group = variable),
col = 'steelblue', alpha = 0.05) +
  geom_point(data = df, aes(x = val, y = frac)) +
  geom_line(data = pdat, aes(x = newxs, y = py), col = 'red') +
  geom_line(data = pdat, aes(x = newxs, y = lwr), linetype = 'dashed') +
  geom_line(data = pdat, aes(x = newxs, y = upr), linetype = 'dashed') +
  geom_text(data = df, aes(x = fit, y = frac, label = species),
              hjust = 1, size = 4) +
  theme_bw() +
  scale_x_log10(breaks = c(0.1, 1, 10, 100, 1000),
              limits = c(0.003, max(df$val))) +
  labs(x = 'Концентрация хлорпирифоса, мг/л ',
           y = 'Доля видов, затронутых эффектом')
```

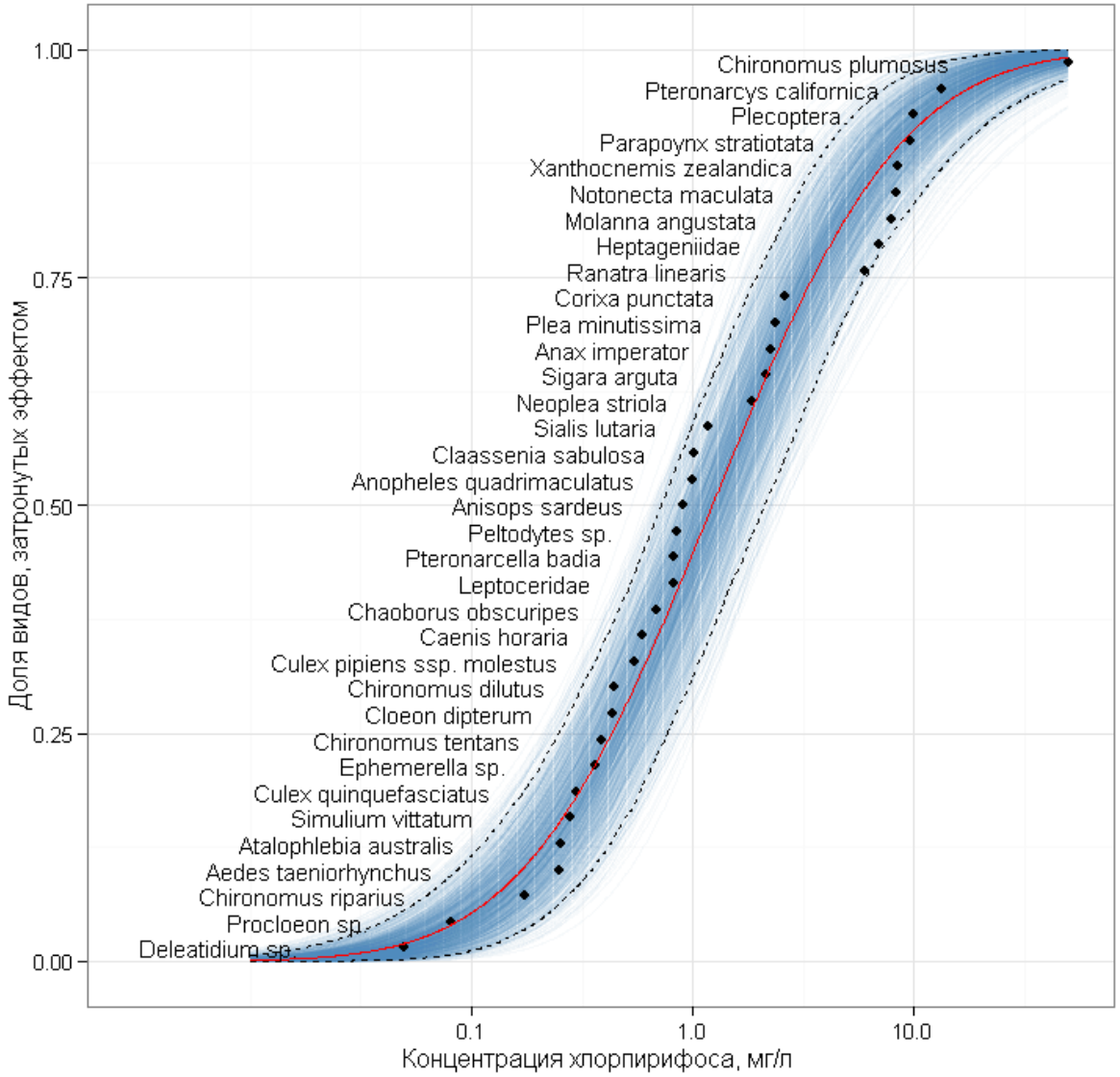


Рис. 5.5. График распределения чувствительности видов с доверительными интервалами; точки – эмпирическое распределение эффективных концентраций, красная линия – подогнанное под данные логнормальное распределение, штриховые линии определяют 95% доверительную полосу, а голубые – кривые бутстреп-распределения,

При непараметрическом бутстрепе параметры заданного теоретического распределения не используются, а генерация повторной выборки осуществляется непосредственно на основе эмпирического набора данных. Для этого из исходной выборки $x_1, x_2, \ldots, x_{n-1}, x_n$ на каждом шаге из n последовательных итераций с помощью датчика случайных чисел извлекается произвольный элемент x_k, который снова возвращается в исходную выборку (т.е. может быть извлечен повторно).

Использование собственных функций для реализации бутстрепа делает процесс вычислений более осознанным, но проще воспользоваться функцией `bootdist()` пакета `fitdistplus`.

```
# Создание 1000 выборок непараметрического бутстрепа
fit_boot <- bootdist(fit, bootmethod='nonparam', niter=1000)
#  Концентрация при p=5% с доверительными интервалами
quantile(fit_boot, probs=0.05)
```

```
Медиана бутстреп-оценки при p=0.05
Значение   0.09934126
Двусторонние 95 % доверительные интервалы для  p=0.05
2.5 %      0.05251031
97.5 %     0.18848304
```

Отметим, что при использовании параметра `bootmethod='param'` функции `bootdist()`, реализуется описанная нами выше процедура параметрического бутстрепа.

5.4. Обоснование экологического риска с использованием SSD

Непосредственное определение риска – это *вероятность возникновения нежелательного события* (т.е. число между 0 и 1, иногда умноженное на 100 для перевода в проценты). Для оценки фактического риска эта вероятность интерпретируется как относительная частота, т.е. отношение числа фактических нежелательных событий к общему количеству возможных событий.

Нежелательные события могут происходить на уровне экосистемы, сообщества, отдельных видов, популяций или особей, поэтому спецификация риска требует ответа на вопрос: что конкретно мы собираемся охранять в окружающей среде (Suter, 1993)? Распределение чувствительности видов (SSD) связано только с одним, узко определенным сегментом оценки экологического риска. Здесь под вероятностью нежелательного события для произвольного вида, случайно выбранного из большого сообщества, понимается статистическая значимость превышения эффекта вредного действия текущей концентрации ксенобиотика над аналогичным эффектом, оказываемым NOEC (Verdonck et al., 2003). Формальная условность понятия "вид, случайно выбранный из большого сообщества" обсуждалась выше (редкие виды имеют равный вес, что и виды с высокой численностью особей; позвоночные животные рассматриваются наравне с беспозвоночными; насекомые, вероятно, будут доминировать во взятой выборке и т.д.).

Предположим, мы пришли к мнению, что некоторое сообщество является прекрасным индикатором токсической опасности изучаемого ксенобиотика. Предварительно в ходе лабораторных испытаний или по литературным данным для отдельных таксонов тестируемого сообщества определяются токсикометрические показатели и строится зависимость SSD. Однако, сформировав по результатам полевых наблюдений над природной экосистемой массив данных, включающий реально

действующие концентрации токсикантов EC (exposure concentration), мы сталкиваемся еще с одним источником изменчивости.

Естественно, мера токсического воздействия (т.е. то, что понимается в качестве дозы) не может характеризоваться некоторыми средними уровнями содержания ксенобиотиков в абиотической среде. В качестве такой меры следует рассматривать спектр концентраций, или некоторую функцию ECD (Exposure Concentration Distribution), описывающую плотность статистического распределения содержания токсических элементов во времени или пространстве. Тогда количественной характеристикой риска воздействия токсичного вещества может быть результат наложения двух статистических распределений SSD и ECD.

Естественным условием корректного наложения распределений должны быть смысловое соответствие обоих множеств значений в отношении оцениваемого токсического эффекта (Suter, 1998). Например, проблематично без предварительных пересчетов сравнивать результаты 96-часовых токсикологических тестов с концентрациями, ежечасно регистрируемыми в точке сброса сточных вод, или с профилями содержания вещества в почве на основе географической информационной системы. Необходимо осуществлять определенные коррекции сравниваемых показателей с учетом продолжительности экспозиции и способности к биоаккумуляции. Будем считать, однако, что вся предварительная подготовка данных была успешно выполнена, и мы имеем два сопоставимых распределения случайных величин.

Задача статистической оценки экологического риска заключается в нахождении вероятности того, что наблюдаемые концентрации $\log(X_{\mathrm{ESD}})$ превысят критические $\log(X_{\mathrm{SSD}})$, которые задаются кривой чувствительности видов. В формальной записи это выглядит так:

$$P[\log(X_{SSD}) < \log(X_{ECD})] = \int_{-\infty}^{\infty} F_{SSD}(x)dF_{ECD}(x)\,,$$

где F_{SSD}– интегральная (CDF, Cumulative Distribution Function) и dF_{ECD} – дифференциальная (PDF, Probability Density Function) функции распределения обеих случайных величин по шкале концентрации токсиканта x.

Рассмотрим практическую процедуру оценки экологического риска в зависимости от концентрации хлорпирифоса с использованием данных для построения SSD из примера предыдущего раздела.

```
# Построение распределения чувствительности видов SSD
df <- read.table(file = "ssd_data.csv", header = TRUE,
      sep = ',', stringsAsFactors = FALSE)
df <- df[order(df$val), ]
df$frac <- ppoints(df$val, 0.5)
fit = fitdist(df$val,"lnorm")
# новые данные для построения плавных кривых
```

```
newxs <- seq(0.01, max(df$val), length.out = 1000)
# Извлекаем доверительные интервалы из бутстреп-выборок
fit_boot <- bootdist(fit, bootmethod='nonparam', niter=1000)
pp <- apply(fit_boot$estim, 1, function (x) plnorm(newxs,
           x[1], x[2]))
cis <- apply(pp, 1, quantile, c(0.025,0.975))
rownames(cis) <- c('lwr', 'upr')
# Формирование итоговой таблицы
pdat <- data.frame(newxs, py = plnorm(newxs,
       meanlog = fit$estimate[1], sdlog = fit$estimate[2]))
pdat <- cbind(pdat, t(cis))
```

Предположим теперь, что в разных частях изучаемого водоема (или в разное время) была 100 раз измерена концентрация хлорпирифоса. Необходимо оценить вероятность экологического риска.

Поскольку у нас реальных данных нет, извлечем случайную выборку из логнормального распределения с несколько измененными параметрами, а сглаживание кривой плотности вероятности выполним ядерной функцией `density()`.

```
# Массив наблюдаемых концентраций - случайная выборка n = 100
xr <- rlnorm(100, meanlog = fit$estimate[1]/10,
    sdlog = fit$estimate[2]/2)
dc <- density(xr, n = 1000, from=0.01, to=max(df$val))
# Вывод графика
plot(x = df$val, y = df$frac, type='n', xlab="Концентрация
хлорпирифоса, мг/л", ylab = "Распределение видов по фракциям",
    xlim = c(0.01, max(df$val)),log="x" )
lines(pdat,  lwd=2)
lines(dc, col = "red", lwd = 2)
#  распределение вероятностей экологического риска
er <- dc$y*pdat[,2]
lines(pdat[,1], er, col = "blue", lwd = 2)
#  Заливка области под кривой и интегрирование
polygon(newxs,er,col="gray70",border=NA)
f1 <- approxfun(pdat[,1], er)
f2 <- function(x) abs(f1(x))
integrate(f2, 0.01, 1)
#  Оценка нижнего и верхнего интервалов риска
lines(pdat[,1], pdat[,3], lwd=2, lty=3)
lines(pdat[,1], pdat[,4], lwd=2, lty=3)
f1 <- approxfun(pdat[,1], dc$y*pdat[,3])
integrate(f2, 0.01, 1)
f1 <- approxfun(pdat[,1], dc$y*pdat[,4])
integrate(f2, 0.01, 1)
lines(pdat[,1], pdat[,4], lwd=2, lty=3)
```

```
Риск            0.1496993 абсолютная ошибка < 5.1e-05
Нижн.интервал   0.0977022 абсолютная ошибка < 9.9e-05
Верх.интервал   0.2077391 абсолютная ошибка < 0.00012
```

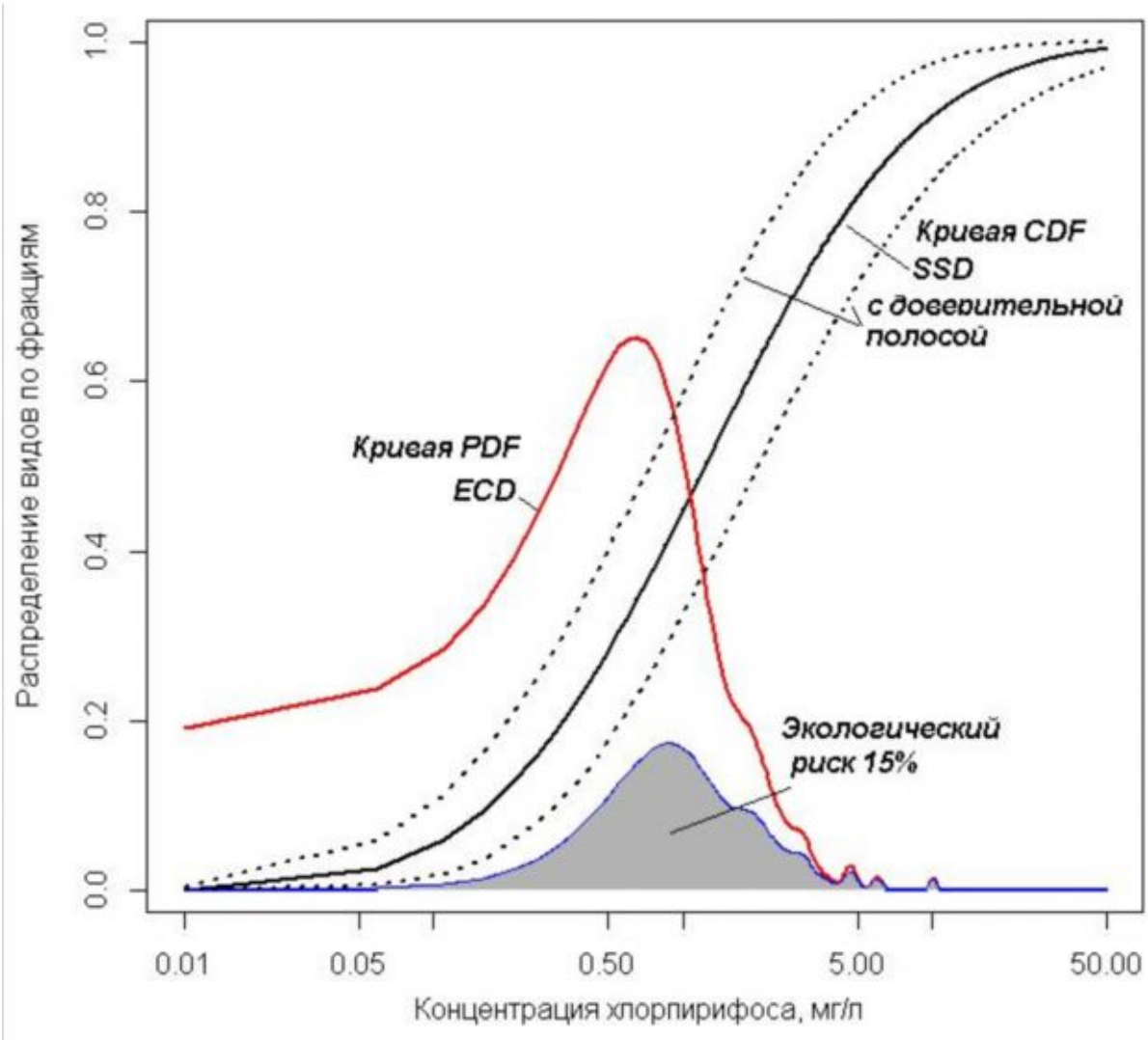


Рис. 5.6. Интерпретация расчета экологического риска; PDF ECD – функция плотности распределения наблюдаемых концентраций, CDF SSD – кумулятивная функция распределения чувствительности видов с доверительными интервалами, построенными непараметрическим бутстрепом

Как показано на рисунке 5.6, кривая распределения экологического риска получена произведением значений плотности распределения ECD на накопленные вероятности SSD. Средний экологический риск равен площади под кривой распределения и составляет 15%.

Оцененная вероятность риска связана с некоторой ее неопределенностью, зависящей, например, от того, насколько хорошо использованная выборка видов представляет тестируемое нами сообщество. Если вместо SSD использовать ее 95%-ные доверительные огибающие, то можно рассчитать интервальную оценку экологического риска, которая составляет от 9.8 до 20.8%.

Технология построения распределений чувствительности видов и обоснования экологического риска постоянно совершенствуется:

- предложен алгоритм, в котором на основании распределений плотности вероятности показателей токсикометрии отдельных видов, доступных на основании литературных или экспериментальных данных, строится функция PDF для всей совокупности таксонов, которая затем переводится в традиционную интегральную форму SSD (Gottschalk, Nowack, 2013);
- на базе пакета `fitdistplus` разработан метод построения SSD с использованием оптимальной процедуры цензурирования данных, который реализован в форме общедоступной web-платформы MOSAIC_SSD (Kon Kam King et al., 2014);
- рассмотрена процедура, учитывающая оценку вероятности максимального эффекта воздействия на все виды и неопределенности при обосновании риска на основе ограниченного числа токсикометрических тестов, использующих только две группы наблюдений (Scott-Fordsmand, Damgaar, 2006) и т.д.

5.5. Индексы SPEAR, основанные на чувствительности видов

Деградация экологических сообществ, которая выражается в снижении видового разнообразия, низкой продуктивности, исчезновении ключевых таксономических групп и т.д., может происходить в результате совместного влияния обеих групп факторов: антропогенного воздействия и изменения комплекса естественных факторов внешней среды. В частности, видовая структура биоценозов водных беспозвоночных определяется не только концентрацией в воде нефтепродуктов, пестицидов, СПАВ, тяжелых металлов или других ингредиентов, но и морфологией донных отложений, скоростью течения и глубиной реки, наличием зон экологических флуктуаций.

В настоящее время используются два основных подхода, позволяющие отделить эффекты антропогенного воздействия от эффектов естественных экологических факторов. Первый сводится к применению методик статистического анализа, которые оценивают доли вариации, объясняемые каждым из факторов, в основе которых положены различные многомерные алгоритмы ординации. Другой подход сводится к нахождению такого подмножества показателей, которые не зависят от естественных экологических факторов (например, устойчивы вдоль градиента речного континуума), но чувствительны к экстремальной изменчивости токсикантов техногенного происхождения. Важным инструментом для оценки экологического состояния пресноводных экосистем являются, например, биотические индексы, основанные на таксономических свойствах сообществ беспозвоночных: видовое богатство, доля групп Ephemeroptera, Plecoptera и Trichoptera (% EPT) или отношение числа наблюдаемых и ожидаемых таксонов (O/E).

Перспективной альтернативой таксономическим индексам является использование биологических характеристик видов (traits), таких как размер тела, особенности поведения, сменяемость поколений, многие из которых относительно постоянны в рамках больших экологических регионов. Такой системой биологической оценки является SPEAR (Species At Risk) – подход, основанный на свойствах анализируемого сообщества, которые связываются с определенным экологическим стрессором.

Первым по хронологии из этого семейства индексов был разработан $SPEAR_{pesticides}$ (Liess, von der Ohe, 2005), который вычисляется как относительное обилие таксонов, чувствительных к воздействию пестицидов:

$$\mathrm{SPEAR}_{\mathrm{pesticides}} = \frac{\sum_{i=1}^{n} \log(x_i + 1) y_i}{\sum_{i=1}^{n} \log(x_i + 1)} 100,$$

где n – число таксонов на различных таксономических уровнях, x_i – обилие таксона i, $y_i = 1$, если таксон i классифицирован как "вид с риском", и $y_i = 0$ в противном случае.

Выделение списка видов, испытывающих стресс в присутствии пестицидов, осуществляется следующим образом. Для каждого i-го анализируемого вида рассматриваются значения LC_{50} в отношении различных токсикантов и рассчитывается максимальное значение чувствительности (sensitivity) по сравнению с LC_{50} для *Daphnia_magna*:

$$S_i = \log_{10} (LC_{50\ D.magna} / LC_{50\ i}).$$

Классификация видов (т.е. присвоение $y_i = 1$) выполняется по сумме следующих условий:

- значение физиологической чувствительности S_i превышает -0.36;
- период смены поколений превышает 0.5 года;
- стадии развития (яйца, личинки, куколки) проходят в воде во время периодов интенсивного использования пестицидов;
- вид имеет низкую способность к миграции.

Оценка индекса $SPEAR_{pesticides}$ по значениям обилия видов, найденных в гидробиологических пробах, может быть осуществлена интерактивно с помощью программного продукта SPEAR Calculator (http://www.systemecology.eu/spear/spear-calculator/), который содержит базу данных с вышеперечисленными характеристиками 2400 таксонов водных беспозвоночных и выполняет все необходимые расчеты и графическую интерпретацию.

Пакет `rspear` является своеобразной R-надстройкой к этому web-приложению. Функция `get_traits()` пакета осуществляет загрузку всех необходимых данных с сайта SPEAR и помещает их в локальный файл traits.csv в рабочей директории. Поле `name` этого файла содержит базовые наименования таксонов.

```
require(rspear)
# Загрузка характеристик видов из базы данных
get_traits()
traits_data <- read.csv("traits.csv", header = TRUE)
head(traits_data)
```

```
                  name   region exposed generationTime sensitivity migration
1                Acari  Eurasia       1           0.5    -1.11100         0
2              Acentria  Eurasia       1           1.0    -0.06000         0
3  Acentria ephemerella  Eurasia       1           1.0    -0.06000         0
4               Acilius  Eurasia       1           1.0    -0.80728         0
5 Acilius canaliculatus  Eurasia       1           1.0    -0.80728         0
6           Acilius sp.  Eurasia       1           1.0    -0.80728         0
```

Используем учебный набор данных `spear_example` из пакета `rspear`:

```
data(spear_example)
head(spear_example)
# Сверка наименований таксонов
spear_example$Taxon[is.na(match(spear_example$Taxon,
                      traits_data$name))]
```

```
            Taxon Abundance Year           Site
1          Baetis         1 2007 Sample Point A
2 Baetis rhodani         1 2007 Sample Point A
3  Baetis rodani         1 2007 Sample Point A
4       xxxxxxxxx        1 2007 Sample Point A
5      Baetis sp.        1 2007 Sample Point A
6    Athericidae         2 2007 Sample Point A
Названия таксонов, не совпадающих с базой
"Baetis rodani" "xxxxxxxxx"
```

Важным моментом в расчетах является установление соответствия названий таксонов в тест-примере и базе. В нашем случае отсутствует совпадение в наименованиях двух видов.

Функция `spear` осуществляет формирование сводной таблицы характеристик и расчет индексов для заданного набора обилия видов.

```
# Формирование таблицы характеристик и расчет индексов
sp <- spear(spear_example ,taxa = "Taxon",
       abundance = "Abundance",group = c("Year", "Site"),
       check = FALSE)
head(sp$traits)
sp$spear
spear_df <- sp$spear
plot(SPEAR ~ factor(Year), data = spear_df, xlab = "Year")
```

Таблица свойств

```
     taxa_data     taxa_matched match_val    region exposed generationTime sensitivity migration SPEAR
     xxxxxxxxx            <NA>        NA      <NA>      NA             NA          NA        NA     0
 Baetis rodani   Baetis rhodani       0.1   Eurasia       0        0.50000     0.02159         0     0
       Baetis           Baetis      -1.0   Eurasia       1        0.64564     0.02159         0     1
Baetis rhodani   Baetis rhodani      -1.0   Eurasia       0        0.50000     0.02159         0     0
    Baetis sp.       Baetis sp.      -1.0   Eurasia       1        0.50000     0.02159         0     1
```

```
Таблица индексов
  Year            Site    SPEAR
1 2007 Sample Point A 35.00612
2 2007 Sample Point B 63.24266
3 2007 Sample Point C 34.98550
4 2007 Sample Point D 58.64163
5 2008 Sample Point A 42.31371
6 2008 Sample Point B 19.38471
7 2008 Sample Point C 28.15862
8 2008 Sample Point D 30.64599
```

При выполнении расчетов нет необходимости предварительно загружать данные с сайта SPEAR функцией `get_traits()` – это может быть сделано в ходе самой процедуры. Обратим внимание на "интеллектуальный стиль" сверки наименований: соответствие *Baetis rodani* и *Baetis rhodani* найдено автоматически. Если совпадение не найдено, то принимается $y_i = 0$ – см. "xxxxxxxxx".

Значение индекса `SPEAR` получено для всех комбинаций станций (`Site`) и лет наблюдений (`Year`). Эти данные могут быть использованы в дальнейшем с применением любых средств статистической обработки среды R. Например, на рис. 5.7 можно увидеть существенное увеличение опасности пестицидного загрязнения в 2008 г. по сравнению с предыдущим годом.

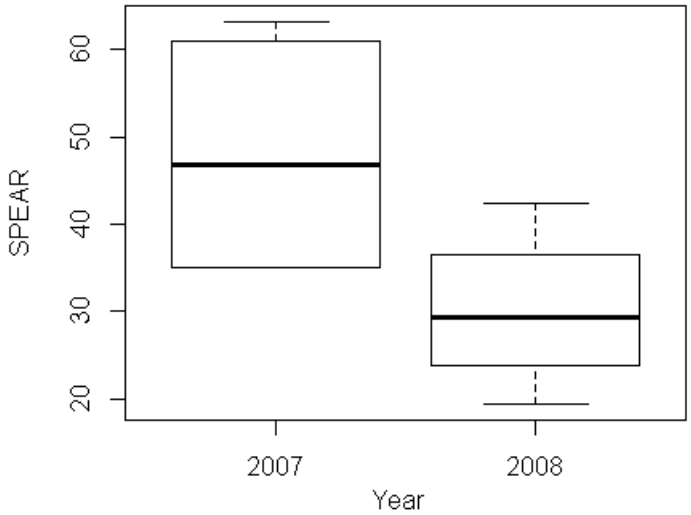


Рис. 5.7. Изменение индекса SPEAR для тестового примера

В дальнейшем исследователи начали обосновывать аналогичные индексы, учитывающие физиологическую чувствительность к другим стрессорам: органическим веществам $SPEAR_{organic}$ (нефтепродукты, СПАВ – Beketov, Liess, 2008) и минерализации $SPEAR_{salinity}$ (Schafer et al. 2011). Расчет перечисленных индексов возможен только с применением SPEAR Calculator и использование пакета `rspear` для этого не предусматривается.

5.6. Построение главных кривых многомерного отклика

Анализ избыточности (RDA, redundancy analysis) рассматривается как каноническая версия метода главных компонент, и, одновременно, как распространение множественной регрессии на данные с многомерным откликом (P. Legendre, L. Legendre, 2012). RDA оценивает параметры зависимости матрицы отклика $\mathbf{Y}(n \times p)$ от матрицы независимых переменных $\mathbf{X}(n \times m)$, где n – число выполненных наблюдений. Поскольку анализ избыточности часто применяется для ординации экологических сообществ, предположим, что **Y** составляют данные обилия p видов, а **X** – m факторов окружающей среды.

Как и в методе главных компонент, RDA выполняет подбор линейных комбинаций, позволяющих сформировать редуцированное описание матрицы видов. Коэффициенты канонической корреляции между **Y** и **X** рассчитываются при этом также с учетом линейных комбинаций факторов среды, полученных для каждого местообитания. В качестве критерия связи каждого фактора среды с характеристиками видов используется минимальная остаточная сумма квадратов для всех возможных линейных комбинаций (Джонгман и др. 1999, с.144).

Построение главной кривой отклика PRC (Principal response curves – Van den Brink, ter Braak, 1999) является частным случаем RDA с одним воздействующим фактором и одной переменной, представляющей временной ряд повторных наблюдений. При этом моделируется характер зависимости многомерного отклика: например, как изменяется соотношение численностей p видов под воздействием различных уровней токсиканта в течение некоторого периода времени.

В этом типе анализа интерес представляют, с одной стороны, графики изменения значений весовых коэффициентов относительно первой оси RDA, вычисленных для каждого уровня воздействия и отражающих контраст наблюдаемого эффекта по сравнению с контролем во времени. С другой стороны, целесообразно проверить статистическую значимость канонических корреляций с первой осью, если планом эксперимента предусмотрены повторности. Это – омнибусный тест: нулевая гипотеза H_0 соответствует утверждению "нет никакой зависимости эффекта от уровня воздействия", а H_1 включает все возможные формы проявления эффекта вне зависимости от принадлежности к различным группам наблюдений. Если анализ показывает, что нулевая гипотеза может быть отклонена (кривые временных рядов на графике отличаются), то можно начинать поиск тех уровней воздействия, которые сильнее отличаются от других.

Рассмотрим пример, представленный набором данных `pyrifos` пакета `vegan` и содержащий результаты эксперимента по оценке воздействия хлорпирофоса на сообщество водных беспозвоночных.

Исследования проводились в 12 водоемах (экспериментальных каналах), в 8 из которых вносился инсектицид с номинальными уровнями дозы 0.1, 0.9, 6, и 44 г/л в двух повторностях, а 4 канала использовались в качестве контрольного мезокосма. Гидробиологические пробы беспозвоночных в каждом объекте брались 11 раз на протяжении 28 недель: два раза до внесения токсиканта и 9 раз – после воздействия. Таблица **Y** включает прологарифмированные значения численностей p = 178 видов по результатам $n = 11 \times 12 = 132$ выполненных наблюдений.

Таблица `pyrifos` в качестве наименований столбцов содержит аббревиатуру видов, а наименования строк включают ссылки на номера каналов и градаций концентраций. Это позволяет сформировать три вектора `week`, `dose` и `ditch`, связанные со строками `pyrifos` и определяющие время, дозу и канал соответственно.

```
require(vegan)
data(pyrifos)
head(pyrifos[, c(1:10)])
# Задаем шкалу времени взятия проб в неделях
week <- gl(11, 12, labels = c(-4, -1, 0.1, 1, 2, 4, 8, 12,
           15, 19, 24))
# Определяем как факторы дозы инсектицида и номера водоемов
dose <- factor(rep(c(0.1, 0, 0, 0.9, 0, 44, 6, 0.1, 44,
           0.9, 0, 6), 11))
ditch <- gl(12, 1, length = 132)
```

	Simve	Daplo	Cerpu	Alogu	Aloco	Alore	Aloaf	Copsp	Ostsp	Slyla
w.4.c1	3.951	0	0	0	0	0	0	2.773	0.000	1.386
w.4.c2	2.303	0	0	0	0	0	0	2.079	0.000	0.000
w.4.c3	4.595	0	0	0	0	0	0	3.761	0.000	0.693
w.4.c4	2.398	0	0	0	0	0	0	3.296	0.693	0.000
w.4.c5	4.025	0	0	0	0	0	0	3.466	0.000	0.000
w.4.c6	2.303	0	0	0	0	0	0	2.197	0.000	0.000

Выведем ординационную диаграмму, показывающую траекторию изменения отклика в пространстве двух главных координат PC1-PC2:

```
mod <- rda(pyrifos)
plot(mod, type = "n")
ordisegments(mod, ditch, label = TRUE,
       show.groups = c("2", "3", "5", "11"))
ordisegments(mod, ditch, label = TRUE, show = c("6", "9"),
      col = 2)
legend("topright", c("Контроль", "44 г/л"), lty = 1,
      col = c(1,2))
```

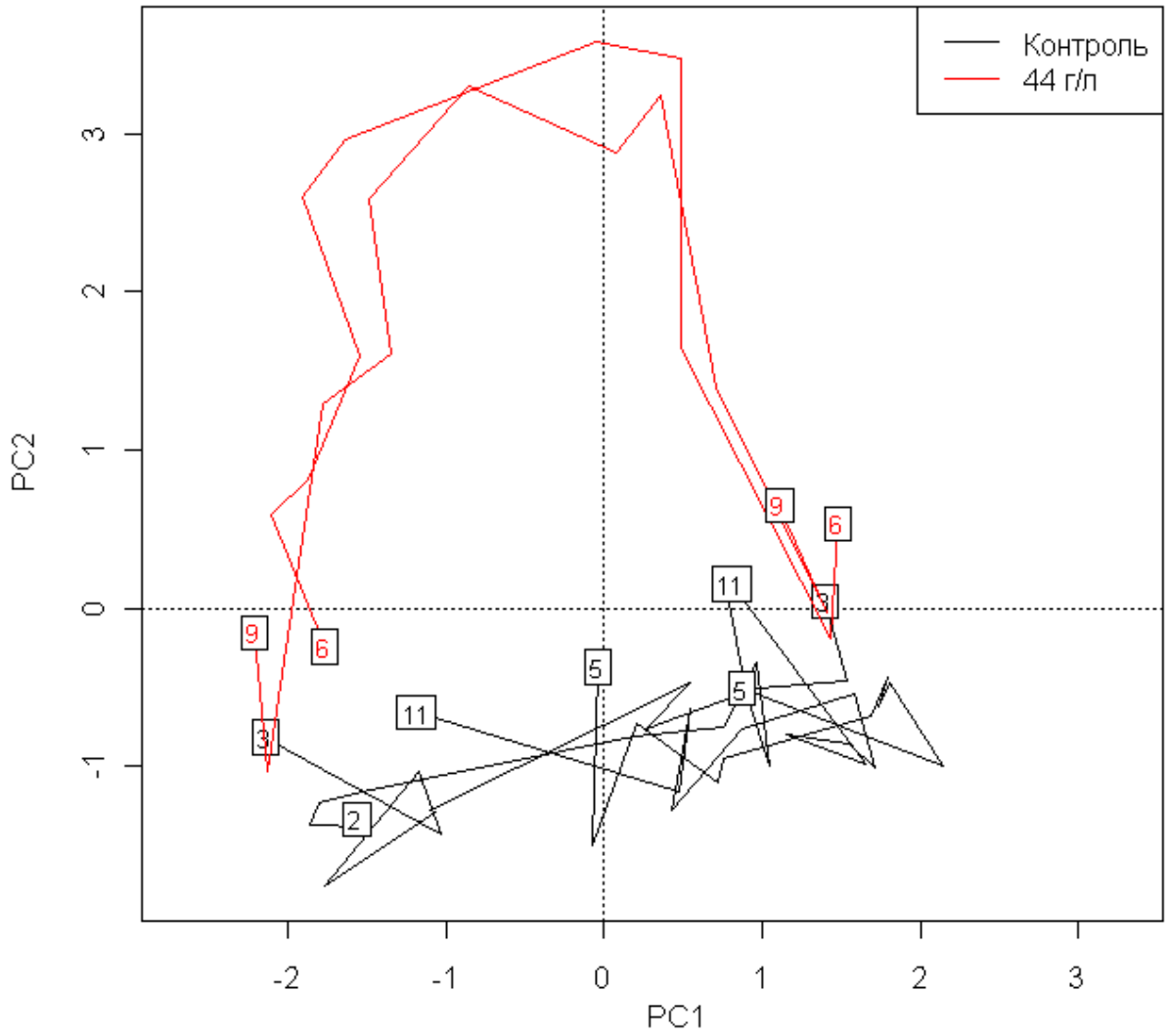


Рис. 5.8. Ординационная диаграмма изменения видовой структуры гидробионтов в мезокосмах при воздействии максимальной концентрации хлорпирофоса (№ 6 и 9) и в условиях контроля (№ 2, 3, 5 и 11)

На рис. 5.8 видно, что во всех экспериментальных водоемах видовая структура гидробионтов за время эксперимента претерпевает закономерный дрейф, вызванный, например, сезонными изменениями. Если сравнить, например, взаимное расположение точек 3 (контроль) и 6 (опыт) в начальном состоянии за 4 недели до воздействия, то они находятся справа на диаграмме в относительной близости. Однако, начиная с момента внесения инсектицида и далее на протяжении 8-х недель, сообщество в точке 6 находится в условиях токсического стресса и траектория развития сравниваемых биоценозов начинает резко расходиться. По истечение 8 недель после воздействия экологическая ситуация в экспериментальном канале начинает стабилизироваться, в сообществе проходят самовосстановительные процессы и, в конечном итоге, точки 3 и 6 справа на диаграмме снова оказываются в относительной близости.

Построение главной кривой отклика PRC, которое можно осуществить с использованием функции `prc()` из пакета `vegan`, позволяет конкретизировать траекторию изменения сообществ во времени:

```
(pyr_prc <- prc(response = pyrifos, treatment = dose,
       time = week))
#  Доля объясненной вариации
head (pyr_prc$CCA$eig/sum(pyr_prc$CCA$eig))
sum_prc <- summary(pyr_prc)
# Весовые коэффициенты видов:
sum_prc$sp[abs(sum_prc$sp) > 0.5]
# Построение кривых главного отклика
plot(pyr_prc, select = abs(sum_prc$sp) > 0.5, scaling = 1,
         lwd=2,xlab="недели", ylab="Эффект", ylim=c(-4,2))
```

```
Разложение изменчивости в данных
                          Inertia      Доля Ранг
Общая                    288.9920    1.0000
Условная во времени       63.3493    0.2192   10
Связанная с загрязнением  96.6837    0.3346   44
Несвязанная              128.9589    0.4462   77
Inertia - изменчивость

Веса ординационных осей
      RDA1       RDA2       RDA3       RDA4       RDA5       RDA6
0.26149460 0.08581472 0.06251539 0.04929725 0.04290436 0.03989127

Весовые коэффициенты для видов
     Simve      Daplo      Copsp      Ostsp      Copdi      NauLa
-1.4619340 -0.7965104 -0.6359661 -1.2574923 -0.7770883 -2.6360998
     Strvi      amosp      Lepsp   olchaeta    sphidae   armicris
-1.6694744  0.7383708 -0.5429219  0.6336741 -0.7960152 -0.9136809
  binitent   popyanti   hycarina   gammpule   aselaqua   caenhora
 1.0607879 -0.6919040 -0.5678026 -0.8301662 -0.8586062 -3.1368667
  caenluct   cloedipt   cloesimi   conagrae   hytuinae   hytuvers
-1.2923002 -2.5746251 -0.6755795 -0.8867945 -1.2596637 -0.9639006
  sialluta   ablaphmo   chironsp   cepogoae   chaoobsc   mystloni
-0.6033200 -1.6275900 -1.0278389 -1.3897671 -1.3282610 -1.6307253
```

Каждый из 178 обнаруженных видов со значениями прологарифмированных численностей участвует в формировании главных осей RDA-модели со своими весовыми коэффициентами (Species scores). На основании последних можно оценить значения "коллективного" отклика (в относительных единицах по контрасту с контролем) для каждого из 11 состояний относительно времени внесения токсиканта и для всех 12 экспериментальных водоемов. После сглаживания данных можно получить графики изменения многомерного эффекта во времени для различных уровней воздействия – см. рис. 5.9.

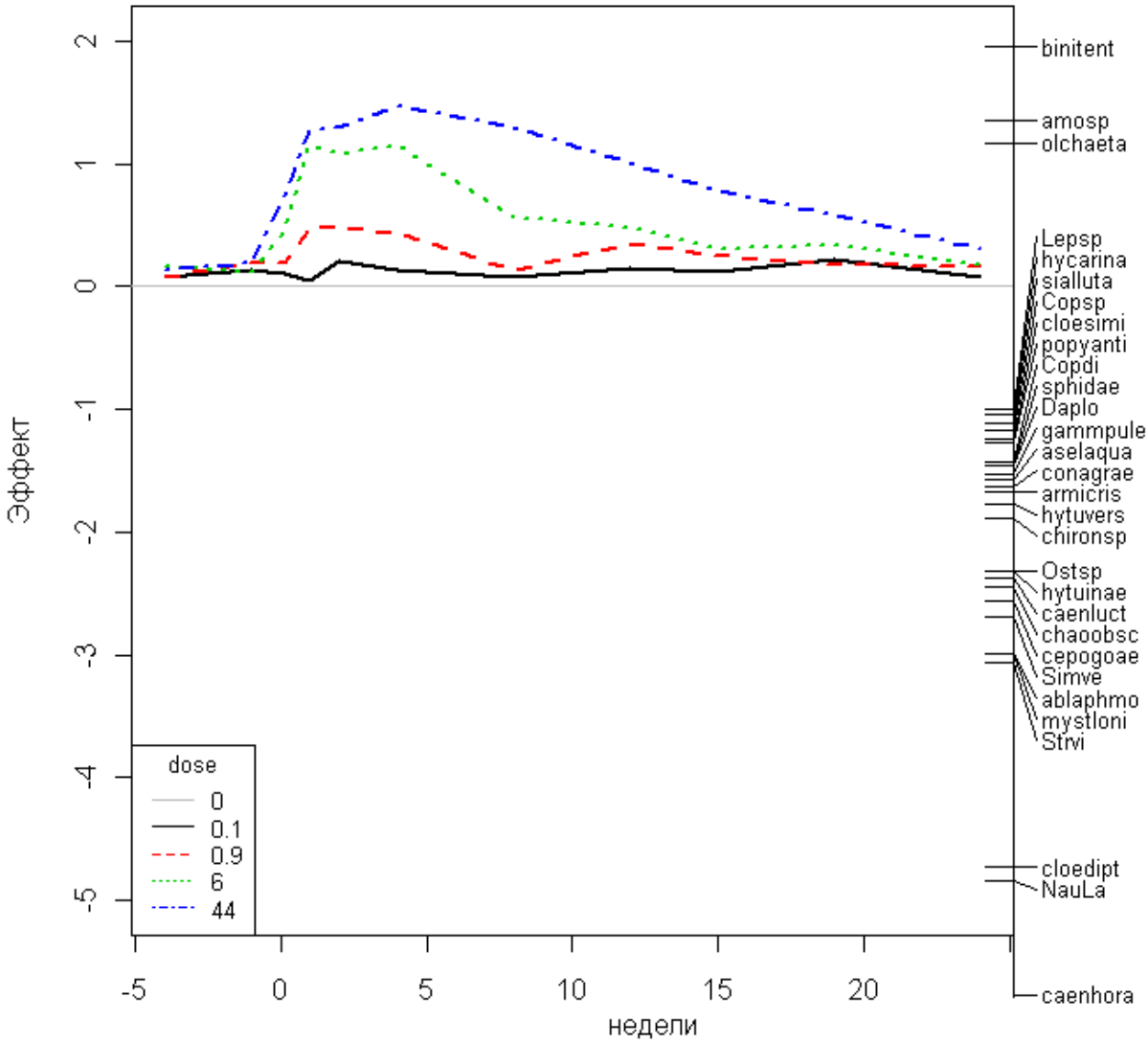


Рис. 5.0. Кривые главного отклика (PRC) сообщества беспозвоночных гидробионтов в результате залпового внесения различных доз (dose) хлорпирофоса

Справа на диаграмме показаны весовые коэффициенты, с которыми виды участвуют в оценке отклика. Для оптимизации графика из всего списка видов отобрано 36, у которых величина коэффициента по абсолютной величине превысила 0.5. Например, `ceanhora` (=*Caenis horaria*) имеет самый низкий вес (~-3), таким образом обилие этого вида в наибольшей степени уменьшается в результате воздействия, тогда как у `binitent` (=*Bithynia tentaculata*) положительный весовой коэффициент и его численность немного увеличивается.

Выполним статистический анализ полученных результатов. Из вышеприведенного протокола расчетов следует, что изменчивость во времени `Conditional` объясняет 22 % вариации данных, а эффект обработки токсикантом `Constrained` – 33 % дисперсии. При этом первая

ось RDA аккумулирует 26% и вторая – 8.6% общего статистического разброса вектора обилия видов.

Проверим две статистические гипотезы: *а*) о том, что отсутствуют изменения в структуре рассматриваемых биоценозов в течение времени проведения эксперимента и *б*) о том, что разные дозы внесенного токсиканта приводят к одинаковому эффекту относительно контроля. В первом случае выполняем цикл для всей продолжительности эксперимента, и на сформированных выборках для каждого временного отрезка выполняем перестановочный тест (permutation test) с использованием функции `anova.CCA()` из пакета `vegan`. Предварительно создаем вектор прологарифмированных доз:

```
ln_dose <- log(20 * as.numeric(as.character(dose)) + 1)
out <- NULL
for (i in levels(week)) {
    take_spec <- pyrifos[week == i, ]
    take_dose <- ln_dose[week == i]
    out[[i]] <- anova(rda(take_spec ~ take_dose),
         by = "terms", step = 1000)
}
sapply(out, function(x) x[1, 5])  # Извлекаем p-значения
```

```
P-значения для каждого отсчета времени
   -4    -1   0.1     1     2     4     8    12    15    19    24
0.430 0.895 0.009 0.001 0.001 0.002 0.008 0.003 0.030 0.021 0.161
```

Очевидно, что в течение 19 недель после внесения токсиканта следует отклонить гипотезу, что биоценозы в 12 каналах были извлечены из одной генеральной совокупности с конкретной видовой структурой.

Тестирование второй гипотезы не может быть выполнено подобным образом, поскольку для 4 уровней доз нельзя получить достаточное количество уникальных перестановок. Поэтому будем использовать тест множественных сравнений на основе контрастов Даннета (Dunnett-Test) или Вильямса (Williams, 1972) применительно к счетам PCA, относящимся к первой главной компоненте. Тем самым для каждой даты наблюдений мы осуществляем сравнение эффекта от действия каждой экспериментальной дозы токсиканта с нулевой дозой контроля. Для выполнения расчетов используем функции `aov()` и `glht()` из пакета `multcomp`.

```
require(multcomp)
df <- data.frame(dose = dose, week = week)
out_willi <- NULL     # Создаем пустой объект
# Цикл по времени, реализующий тест Даннета
for (i in levels(week)) {
    take_spec <- pyrifos[week == i, ]
    pca <- rda(take_spec)  # Расчет матриц PCA
```

```
    # выделение счетов первой клавной компоненты
    pca_scores <- scores(pca, display = "sites", choices = 1)
    # множественные сравнения с использованием контрастов
    out_willi[[i]] <- summary(glht(aov(pca_scores ~ dose,
        data = df[week == i, ]), alternative = "t", linfct =
        mcp(dose = "Dunnett")))
}
# извлечение p-значений
result <- lapply(out_willi, function(x) data.frame(
       comp = levels(df$dose)[-1], pval = x$test$pvalues,
       sig = x$test$pvalues < 0.05))
# результаты теста Даннета для PCA-счетов после первой недели
result[["1"]]
```

```
Значимость отличий видового состава
Каналы с концентрацией токсиканта против контроля
  конц        pval   sig
1  0.1 0.999856959 FALSE
2  0.9 0.061024497 FALSE
3    6 0.002465169  TRUE
4   44 0.001233040  TRUE
```

Таким образом, можно полагать, что концентрация, не вызывающая токсического эффекта в сообществе водных беспозвоночных, равна NOEC = 0.9 мг/л.

СПИСОК ЛИТЕРАТУРЫ

Бабич П.Н., Чубенко А.В., Лапач С.Н. Применение пробит-анализа в токсикологии и фармакологии с использованием программы Microsoft Excel для оценки фармакологической активности при альтернативной форме учета реакций // Современные проблемы токсикологии. 2003. №4. С. 80-88.

Безель В.С. Экологическая токсикология: популяционный и биоценотический аспекты. Екатеринбург: Гощицкий, 2006. 279 с.

Безель В.С., Большаков В.Н., Воробейчик Е.Л. Популяционная экотоксикология. М.: Наука, 1994. 80 с.

Беленький М.Л. Элементы количественной оценки фармакологического эффекта. Л.: Медгиз. 1969. 143с.

Венэбльз У.Н., Смит Д.М. Введение в R: Заметки по R: среда программирования для анализа данных и графики : пер. с англ. Вер. 3.1.0. Москва, 2014. 109 с.

Воробейчик Е.Л., Садыков О.Ф., Фарафонтов М.Г. Экологическое нормирование техногенных загрязнений наземных экосистем (локальный уровень). Екатеринбург: Наука, 1994. 280 с.

Гелашвили Д.Б. Популяционная экотоксикология и экологические риски // Теоретические проблемы экологии и эволюции (Шестые Любищевские чтения). Тольятти : Кассандра, 2015. С. 89-93.

Гелашвили Д.Б., Безель В.С., Романова Е.Б., Безруков М.Е., Силкин А.А., Нижегородцев А.А. Принципы и методы экологической токсикологии: В 2-х т. Нижний Новгород: Нижегородский госуниверситет – 2014 – Т.1….с. Т.2….с.

Голубев А.А., Люблина Е.И., Толоконцев Н.А., Филов В.А. Количественная токсикология. Л., 1973. 287 с.

Джонгман Р.Г.Г., тер Браак С.Дж.Ф., ван Тонгерен О.Ф.Р. Анализ данных в экологии сообществ и ландшафтов. М.: РАСХН. 1999. 306 с.

Зарядов И.С. Введение в статистический пакет R: типы переменных, структуры данных, чтение и запись информации, графика. Москва: Изд-во РУДНБ, 2010а. 207 с.

Зарядов И.С. Статистический пакет R: теория вероятностей и математическая статистика. Москва: Изд-во РУДНБ, 2010б. 141 с.

Зеленин К.Н. Что такое химическая экотоксикология // Соросовский образовательный журнал. 2000. Т. 6. С. . 32-36.

Кабаков Р.И. R в действии: Анализ и визуализация данных в программе. М.: ДМК Пресс, 2014. 580 с.

Каплин В.Г. Основы экотоксикологии. Учебное пособие. М.: КолосС, 2006. 232 с.

Криштопенко С.В., Тихов М.С. Токсикометрия эффективных доз. Нижний Новгород: ННГУ, 1997. 156 с.

Криштопенко С.В., Тихов М.С., Попова Е.Б. Доза-эффект. М.: Медицина, 2008. 288 с.

Курляндский Б.А., Филов В.А. (ред). Общая токсикология. М.: Медицина, 2002. 608 с.

Куценко С.А. Основы токсикологии. М.: Фолиант, 2004. 570 с.

Лазарев Н.В. Общие основы промышленной токсикологии. М. Л., 1938. 388 с.

Мастицкий С.Э., Шитиков В.К. Статистический анализ и визуализация данных с помощью R . Москва: ДМК Пресс, 2015. 496 с. (PDF, данные и скрипты доступны на сайте авторов ievbras.ru/ecostat)

Рашевский Н. Некоторые медицинские аспекты математической биологии. М.:Медицина, 1966. 242 с.

Справочник по прикладной статистике. В 2-х т. Пер. с англ. / Под ред. Э. Ллойда, У. Ледермана, Ю. Н. Тюрина. М.: Финансы и статистика. Т. 1. 1989. 510 с., Т. 2. 1990. 526 с.

Толоконцев Н.А., Филов В.А. Основы общей промышленной токсикологии Л: Медицина, 1976. 304 с.

Трахтенберг И.М., Сова Р.Е., Шефтель В.О., Оникиенко Ф.А. Проблема нормы в токсикологии (современные представления и методические подходы, основные параметры и константы). М.: Медицина, 1991. 208 с.

Шипунов А.Б. и др. Наглядная статистика. Используем R! Москва: ДМК Пресс, 2014. 298 с.

Шитиков В.К., Розенберг Г.С. Рандомизация и бутстреп: статистический анализ в биологии и экологии с использованием R. Тольятти : Кассандра, 2014. 314 с. (PDF, данные и скрипты доступны на сайте авторов ievbras.ru/ecostat).

Шитиков В.К., Розенберг Г.С., Зинченко Т.Д. Количественная гидроэкология: методы, критерии, решения. М.: Наука, 2005. Кн. 1. 281 с.; Кн. 2. 337 с.

Шитиков В. К., Терехова В. А., Узбеков Б. А., Кыдралиева К. А., Худайбергенова Б. М. Модели “доза-эффект” для оценки экологического риска при техногенном загрязнении почвы // Принципы экологии. 2015. № 3. С. 73–88.

Agresti A. An introduction to categorical data analysis. Wiley. 2007. 372 p.

Aldenberg T., Slob. W. Confidence limits for hazardous concentrations based on logistically distributed NOEC toxicity data. // Ecotoxicology and Environmental Safety. 1993. V. 25. P. 48–63.

Altenburger R., Backhaus T., Boedeker W., Faust M., Scholze M. Simplifying complexity: Mixture toxicity assessment in the last 20 years. // Environ. Toxicol. Chem. 2013. V. 32, No 8. P. 1685 – 1687.

Bates DM., Watts DG. Nonlinear Regression Analysis and Its Applications. New-York: John Wiley, 2007. 392 p.

Beketov MA, Liess M. An indicator for effects of organic toxicants on lotic invertebrate communities: independence of confounding environmental factors over an extensive river continuum. // Environmental Pollution. 2008. V. 156. P. 980-987.

Berenbaum MC. Criteria for analyzing interactions between biologically active agents. // J Am Stat Assoc. 1981. V. 78. P. 90–98.

Bruin de A. Biochemical toxicology of environmental agents. Elsevier Worth, Holland: Biomedical PGS, 1976. 1544 p.

Cedergreen N., Ritz C., Streibig JC. Improved empirical models describing hormesis // Environmental Toxicology and Chemistry. 2005. V. 24. P. 3166–3172.

Crane M., Newman M.C. What level of effect is a no observed effect? // Environmental Toxicology and Chemistry. 2000. V. 19, N 2. P. 516-519.

Davison A.C., Hinkley D.V. Bootstrap methods and their application. Cambridge: Cambridge University Press, 2006. 592 p.

Fan A.M., Alexeeff G., Khan E. Toxicology and Risk Assessment. CRC Press, Taylor & Francis Group, 2015. 1368 p.

Finney D.J. Probit Analysis. Cambridge: Cambridge University Press. 1971. 333 p.

Fraser TR. An experimental research on the antagonism between the actions of physostigma and atropia. // Proc R Soc Edinb. 1870. V. 7. P. 506-511.

Goddard M.J., Hinberg I. Receiver operator characteristic (ROC) curves and non-normal data: An empirical study. // Statistics in Medicine. 1989. Vol. 9, No 3. P. 325–337.

Gottschalk F., Nowack B. A probabilistic method for species sensitivity distributions taking into account the inherent uncertainty and variability of effects to estimate environmental risk // Integr. Environ. Assess. Manag. 2013. V. 9. P. 79-86

Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction. N.Y.: Springer-Verlag, 2009. 763 p

Hewlett PS. Measurement of the potency of drug mixtures. // Biometrics. 1969. V. 25. P. 477–487

Hopkin S.P. Ecological implications of “95% protection levels” for metals in soil. // Oikos. 1993. V. 66. P. 137–141.

Kaplan E.L,. Meier P. Nonparametric estimation from incomplete observations // J. Am. Stat. Assoc. 1958. V. 53. P. 457–481.

Klein J.P., Moeschberger M.L. Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer, 2003. 536 p.

Klepper O., Traas T.P., Schouten A., Korthals G.W., de Zwart D. Estimating the effect on soil organisms of exceeding no-observed effect concentrations (NOECs) of persistent toxicants. // Ecotoxicology. 1999. V. 8. P. 9–21.

Knezevic SZ, Streibig JC., Ritz C. Utilizing R software package for dose-response studies: The concept and data analysis // Weed technology. 2007. V. 21, N 3. P. 840-848.

Kon Kam King G.,·Veber P., Charles S.,·Delignette-Muller M-L.·MOSAIC_SSD: a new web-tool for the Species Sensitivity Distribution, allowing to include censored data by maximum likelihood // Environmental Toxicology and Chemistry. 2014. V. 33, N 9.

Kooijman S.A.L.M. A safety factor for LC50 values allowing for differences in sensitivity among species. // Water Research. 1987. V. 21. P. 269–276.

Lawton J.H. What do species do in ecosystems? // Oikos. 1994. V. 71. P. 367–374.

Legendre P., Legendre L. Numerical Ecology. Amsterdam: Elsevier Sci. BV, 2012. 990 p.

Liess M, von der Ohe PC. Analyzing effects of pesticides on invertebrate communities in streams. // Environmental Toxicology and Chemistry. 2005. V. 24, No 4. P. 954-965.

McCullagh P., Nelder J.A. Generalized Linear Models. London: Chapman & Hall, 1989. 511 p.

Mount D.I. 1982. Aquatic surrogates. In Surrogate Species Workshop Report, TR-507-36B. U.S. Environmental Protection Agency, Washington, D.C., USA: pp. A6-2–A6-4.

Newman M.C. Quantitative Ecotoxicology. CRC Press, 2012. 556 p.

Newman M.C., Ownby D.R., Mezin, L.C.A. et al. Applying species-sensitivity distributions in ecological risk assessment: assumptions of distribution type and sufficient numbers of species // Environ. Toxicol. Chem. 2000. V. 19, No 2. P. 508–515.

Posthuma L., Suter G.W.H., Traas T.P. Species Sensitivity Distributions in Ecotoxicology. CRC Press, 2001. 616 p.

Ritz C. Towards a unified approach to dose-response modeling in ecotoxicology // Environ. Toxicol. Chem. 2010. V. 29. P. 220–229.

Ritz C., Cedergreen N., Jensen JE., Streibig JC. Relative potency in nonsimilar dose-response curves. // Weed Science. 2006. V. 54. P. 407–412.

Ritz C., Martinussen T. Lack-of-fit tests for assessing mean structures for continuous dose-response data. // Environmental and Ecological Statistics. 2011. V. 18. P. 349–366.

Ritz C., Streibig J.C. Bioassay analysis using R // J. Stat. Soft. 2005. V. 12. P. 1-22.

Ritz C., Streibig JC. From additivity to synergism - A modelling perspective. // Synergy. 2014. V. 1. P. 22–29.

Schafer R.B., Kefford B., Metzeling L. et al. A trait database of stream invertebrates for the ecological risk assessment of single and combined effects of salinity and pesticides in South-East Australia. // Science of the Total Environment. 2011. V. 406. P. 484-490.

Scholze M., Boedeker W., Faust M., Backhaus T., Altenburger R., Grimme LH. A general best-fit method for concentration-response curves and the estimation of low-effect concentrations. // Environ. Toxicol. Chem. 2001, V.20, N 2. P. 448-457.

Scott-Fordsmand J., Damgaar C. Uncertainty analysis of single-concentration exposure data for risk assessment introducing the species effect distribution approach. // Environ. Toxicol. Chem., 2006. V. 25, N 11. P. 3078-3081

Seber G.A., Wild C.J. Nonlinear Regression. New-York: John Wiley, 1989. 768 p.

Soerensen H., Cedergreen N., Skovgaard I.M., Streibig J.C. An isobole-based statistical model and test for synergism/antagonism in binary mixture toxicity experiments // Environmental and Ecological Statistics. 2007. V. 14. P. 383–397.

Solomon K.R., Brock T.C.M., De Zwart D. et al., eds. Extrapolation Practice for ecotoxicological effect characterization of chemicals. SETAC Press & CRC Press, Boca Raton, FL, USA, 2008.

Suter G.W. (Ed.). Ecological Risk Assessment. Boca Raton: Lewis Publishers, 1993. 538 p.

Suter G.W. Comments on the interpretation of distributions in “Overview of recent developments in ecological risk assessment.” // Risk Analysis. 1998. V. 18. P. 3–4.

Timbrell J. Principles of Biochemical Toxicology. Taylor & Francis, 2008. 464 p.

Van den Brink PJ., ter Braak CJF. Principal response curves: Analysis of time-dependent multivariate responses of biological community to stress. // Environmental Toxicology and Chemistry. 1999. V. 18. P. 138–148.

Van Straalen N.M., Denneman C.A.J. Ecotoxicological evaluation of soil quality criteria. // Ecotoxicology and Environmental Safety. 1989. V. 18. P. 241–251.

Verdonck F.A.M., Aldenberg T., Jaworska .J, Vanrolleghem P.A. Limitations of current risk characterization methods in probabilistic environmental risk assessment. // Environ. Toxicol. Chem. 2003. V. 22. P. 2209–2213.

Williams DA. The comparison of several dose levels with a zero dose control. // Biometrics. 1972. V. 28. P. 519–531.

Wilson A.G.E. (Ed.) New Horizons in Predictive Toxicology: Current Status and Application. Royal Society of Chemistry, 2012. 702 p.

Wood S.N. Generalized Additive Models: An Introduction with R. Chapman, Hall/CRC, 2006. 410 p.

Woolley A. A guide to practical toxicology: evaluation, prediction, and risk. USA: Informa Healthcare, 2008. 465 p.