Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых»

Регрессионный анализ в почвоведении

Учебное пособие

Рекомендовано Учебно-методическим Советом по почвоведению при УМО по классическому университетскому образованию РФ в качестве учебного пособия для студентов высших учебных заведений, обучающихся по направлению высшего профессионального образования 021900 «Почвоведение»

Владимир 2016

УДК 519.237.5:631.4(075)
ББК 22.172:40.3я 7
Р32

Авторы:
Е. В. Шеин, М. А. Мазиров, А. А. Корчагин, А. Б. Умарова,
В. М. Гончаров, С. И. Зинченко, А. В. Дембовецкий

Рецензенты:

Доктор биологических наук зав. лабораторией экологии почв Института водных и экологических проблем Дальневосточного отделения Российской академии наук (г. Хабаровск)
Г. В. Харитонова

Кандидат биологических наук доцент кафедры биологии и экологии Владимирского государственного университета имени Александра Григорьевича и Николая Григорьевича Столетовых
И. В. Князьков

Печатается по решению редакционно-издательского совета ВлГУ

Регрессионный анализ в почвоведении : учеб. пособие / Р32 Е. В. Шеин [и др.] ; Владим. гос. ун-т им. А. Г. и Н. Г. Столетовых. – Владимир : Изд-во ВлГУ, 2016. – 88 с.
ISBN 978-5-9984-0712-3

Содержит материал учебной программы курса «Статистические методы исследования в почвоведении» для студентов 4-го курса очной формы обучения направления 06.03.02 – Почвоведение.

Может быть использовано на практических занятиях и при самостоятельном изучении дисциплины «Статистические методы исследования в почвоведении».

Рекомендовано для формирования профессиональных компетенций в соответствии с ФГОС ВО.

Табл. 19. Ил. 30. Библиогр.: 20 назв.

УДК 519.237.5:631.4(075)
ББК 22.172:40.3я 7

ISBN 978-5-9984-0712-3

ВВЕДЕНИЕ

Цель данного пособия – познакомить читателей с назначением и применением в почвоведении регрессионного анализа, его возможностями и ограничениями. Регрессионный анализ способен давать прогнозные результаты, описывать зависимости между свойствами и многое другое. Но как статистический метод он не поможет дать однозначный ответ о связи двух явлений. Этот анализ способен дать ответ такого рода: с такой-то степенью вероятности мы можем утверждать, что между двумя явлениями (свойствами и процессами) существует связь, которая выражается определенным типом, устойчивостью и ограничениями.

Статистика – удивительная наука. На любой вопрос при наличии даже малого количества информации она всегда готова дать ответ, правда, в специфической форме. Она способна указать, с какой вероятностью произойдет то ли иное событие. Следует иметь в виду еще одну особенность: если изменится, пусть на немного, исходная информация, то и ответ будет иным. Как правило, пользуясь аппаратом статистики, говорят о вероятности того или иного события, когда оно возникает под действием одного или множества факторов. Будем постоянно иметь в виду, что если изменится исходная информация, то могут измениться итог и регрессионная зависимость. В статистике какова используемая информация, таков и прогнозный вывод. Это всегда надо иметь в виду. И если что-то не получается, то виноват не статистический анализ, а, скорее всего, исходной экспериментальной информации недостаточно для получения статистически достоверных выводов. Это важное правило использования статистического аппарата.

Регрессионный анализ – один из основных методов статистического анализа, который может быть использован для исследования связей между свойствами почвы и факторами окружающей среды на основе как единовременных, так и динамических наблюдений. Регрессионная модель строится для описания взаимосвязи отдельного почвенного свойства с воздействующим фактором. В таком типе анализа влияющий фактор (аргумент) принимается *независимой переменной,* или предиктором, а результат – зависимой *переменной, или функцией отклика* (переменной отклика). Наша задача – найти количественную взаимосвязь между предиктором и откликом. Термин

«переменная отклика» (response variable), или функция отклика, вытекает из идеи, что между характеристиками почвы и переменными факторами среды существует причинно-следственная связь. Однако механизм этой связи не может быть выведен из регрессионного анализа. Задача регрессионного анализа скромнее, а именно описать переменную отклика как функцию одного предиктора (аргумента) или бóльшего количества. Эта функция, названная функцией отклика (связи), обычно не может быть выбрана таким образом, чтобы прогнозировать зависимые переменные без ошибок. Регрессионный анализ даёт возможность минимизировать эти ошибки и свести их к нулю. Значение, прогнозируемое функцией отклика, в таком случае является средним (ожидаемым) откликом – значением с близкой к нулю средней ошибкой.

Сущность регрессионного анализа хорошо отражает термин Уиттекера «прямой градиентный анализ» (Whittaker, 1967). Подразумевается, что этот анализ исследует градиент (изменение) функции с ростом одного из аргументов (её предикторов). В почвоведении регрессионный анализ используется, главным образом:

– для оценки взаимосвязи некоторых свойств (например, влажности и плотности, содержания гумуса и средней температуры вегетационного периода, фильтрации, гранулометрического состава и др.), а также некоторых характеристических переменных и свойств почвы: оптимума температуры, влажности и урожая и пр.;

– определения того фактора-предиктора среды, который вносит наибольший вклад в реакцию отклика, и тех факторов, которые, по всей вероятности, не имеют большего значения. Например, на характеристику водоудерживания почвы оказывают влияние гранулометрический состав (содержание физической глины), содержание органического вещества, плотность почвы и другие факторы. С помощью регрессионного анализа мы сможем ответить на вопрос: какой фактор оказывает определяющее влияние на водоудерживание почвы, а каким фактом можно пренебречь? В этом случае мы можем прогнозировать отклик системы на воздействие одного фактора-предиктора среды или более.

Слово «статистика» происходит от латинского *status* – состояние дел. Выросла эта наука из подсчета государственных дел: ресурсов, народонаселения, денег и пр. В большинстве статистических методов сначала выдвигается гипотеза о том, соответствуют ли или не

соответствуют выборка или отдельные статистические параметры (называемые иногда просто «статистиками») выдвигаемой гипотезе, а затем эта гипотеза проверяется с помощью ряда критериев (Фишера, Стьюдента, Пирсона и др.). С этими методами вы уже достаточно хорошо знакомы по общему курсу статистики, их мы обязательно будем использовать при регрессионном анализе.

Однако, используя методы статистики, мы должны помнить два незыблемых правила:

1. Вся статистика основана на правиле «Как было, так и будет!». Получили мы с вами некоторый фактический материал, использовали его для статистического анализа, и он нам дал некоторый ответ. Но этот ответ справедлив только для той области фактического материала, на котором он был получен. Если мы используем полученный ответ для другой области данных, то он будет ненадежным, а может быть, и неверным. Помним правило № 1: «Как было, так и и будет».

2. Статистика всегда дает ответ в понятиях вероятности. Она укажет, что вероятность P статистической зависимости такая-то (например, 95 %) или, что то же самое, уровень значимости α (0.05). В некоторых программах, таких как Statistica, **уровень значимости α обозначается *p*-level**. В большинстве естественных наук уровень значимости α (*p*-level) принят 0.05 (это означает, что в 95 случаях из 100 оцениваемое нами явление произойдет). Ответ всегда указывает лишь на вероятность события. Это не ответ физика на вопрос: какой будет ток в цепи при известном перепаде потенциалов и сопротивлении цепи? Физик назовет конкретную величину тока в амперах, использовав для расчета закон Ома. Экспериментатор, владеющий статистикой, получив экспериментальный материал, скорее всего, ответит: «С некоторой вероятностью (конкретной, называется цифра) можно утверждать, что величина тока будет следующая…». Заметим, ответ статистический, указана вероятность события. Статистика всегда дает ответ с некоторой вероятностью. Запомним и это правило № 2 и будем использовать его в дальнейшем.

Итак, мы приступаем к изучению регрессионного анализа и его применению в конкретной области естествознания – почвоведении. Познакомимся со сложным, но очень полезным и красивым методом, который можно использовать в исследованиях, чтобы сделать результаты своей работы понятными и значимыми.

Глава 1. ПЕРЕМЕННАЯ ОТКЛИКА И АРГУМЕНТ-ПРЕДИКТОР В РЕГРЕССИОННОМ АНАЛИЗЕ

1.1. Модель и типы переменных отклика

Регрессионный анализ базируется на модели отклика, которая состоит из двух составляющих: детерминистической, описывающей зависимость среднего значения отклика от влияющих факторов (предикторов), и случайной составляющей, описывающей отклонения наблюдаемого отклика от этой зависимости.

Детерминистическая составляющая модели описывается уравнением регрессии, случайная составляющая – статистическим распределением ошибки. Например, в случае линейной зависимости между переменными модель отклика выглядит как (рис. 1):

$$y = b_1 + b_2 x + \varepsilon,$$

где y – переменная отклика (функция отклика);
x – фактор (предиктор);
ε – ошибка определения отклика;
b_1 и b_2 – коэффициенты (или параметры регрессии).

Среднее значение отклика E_y равно $b_0 + b_1 x$. Детерминистическая составляющая модели описывается линейным уравнением регрессии

$$E_y = b_1 + b_2 x.$$

Случайная составляющая описывается распределением ошибок ε – случайных отклонений наблюдаемого отклика от среднего. Цель регрессионного анализа теперь может быть сформулирована более строго как определение детерминистической зависимости по данным измерений с учетом ошибки как составной части модели. В случае линейной зависимости идентификация сводится к оценке параметров

b_1 и b_2. Будем учитывать, что в классической регрессии, в которой используется метод наименьших квадратов (LSR-анализ – *least squares regression*), распределение ошибок ε принимается нормальным, что важно учитывать при расчетах.

1.2. Линейная регрессия

Линейная функция имеет вид $y = b_1 + b_2 x$ и графически это прямая линия.

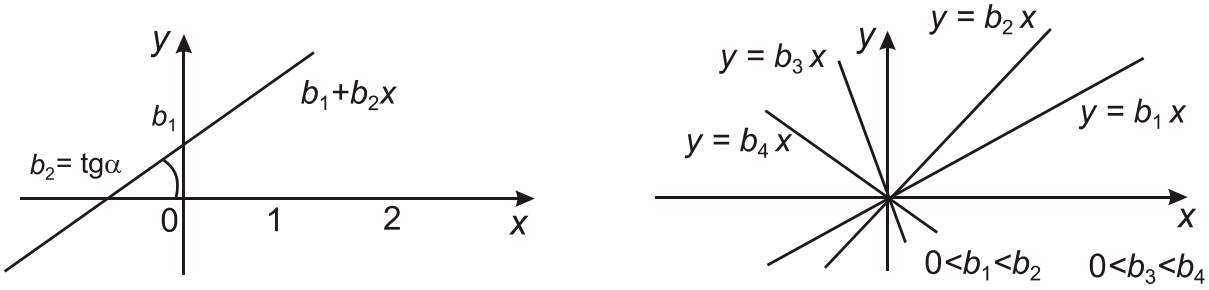


Рис. 1. Линейная регрессия $y = b_1 + b_2 x$

Коэффициент b_2 определяется как тангенс угла наклона, который образует прямая с положительным направлением оси абсцисс, свободный член b_1 – координатой точки пересечения прямой с осью ординат.

При $b_2 > 0$ прямая образует острый угол с осью абсцисс, при $b_2 < 0$ – тупой угол с осью абсцисс, при $b_2 = 0$ прямая параллельна оси абсцисс, а при $b_1 = 0$ прямая проходит через начало координат. Убывающая линейная функция получается из возрастающей при смене знака перед аргументом (запомним это правило!).

Линейная функция используется во всех областях науки для описания пропорциональной зависимости (градуировочные графики в аналитической химии). Линейная зависимость встречается во многих физических законах, технике (равноускоренное движение материальной точки, закон Ома, закон Дарси для процесса фильтрации воды в почве и др.), закон Гука для упругих деформаций, в биологии (скорость растворения вещества в крови). Однако для большинства почвенных процессов, таких как зависимость порозности от влажности (кривая усадки), функции влагопроводности, зависимости диапазона доступной влаги от гранулометрического состава, в частности, от содержания физической глины, зависимости температуропроводности

от влажности и множества других, такой вид функции непригоден, так как эти зависимости нелинейные. Для их описания требуются другие виды функций. Отметим, что вообще в науках об окружающей нас природе линейные функции используются нечасто. Наиболее употребительна эта форма регрессии при получении и использовании тарировочных кривых для какого-либо прибора при полевых исследованиях, определения вида взаимосвязи свойств в определенной области измеренных величин (где, судя по графику, зависимость близка к линейной). В основном взаимосвязи почвенных свойств и их зависимости от действующих факторов нелинейны. И все-таки для начала изучения регрессионного анализа подробнее познакомимся с линейной регрессией и её расчетом в программе STATISTICA на конкретных примерах.

Пример 1. Изучение зависимости плотности агрегата от влажности. Описание (аппроксимация) экспериментальных данных с помощью линейной регрессии.

Например, мы изучали плотность агрегата в зависимости от его влажности. Для этого существуют определенные методы, позволяющие измерять объем агрегата при различной влажности. Покрывают, например, агрегат при определенной влажности тонкой плёнкой и измеряют его объем, опуская в жидкость. Вес абсолютно сухого агрегата известен, объем изменяется при изменении влажности. Масса абсолютно сухого агрегата (величина постоянная) делится на изменяющийся объем, в результате получается величина плотности агрегата. Плотность агрегата увеличивается при уменьшении влажности – происходит его усадка. Хотя мы и знаем, что усадка – это нелинейный процесс, но в определенной области влажности вполне можно использовать и линейную зависимость плотности агрегата от влажности, что подтверждает построенный график зависимости

В табл. 1 приведены результаты нашего опыта.

Для получения регрессионной зависимости плотности агрегата от влажности мы используем статистический пакет STATISTICA. В открытом окне «Данные» представлены данные примера 1. Первым делом мы должны визуально оценить наши данные по графику. Для этого идем в меню «Графики», выбираем «Графики рассеяния», в высветившемся диалоговом окне указываем «Переменные»: «зависимая» – плотность агрегата, «независимая» (предиктор) – влажность. В результате получаем график (рис. 2), подтверждающий, что в исследованной области влажности мы можем использовать линейную регрессию.

Таблица 1

Изменение плотности агрегата от влажности (экспериментальные данные)

№ п/п	Плотность агрегата, г/см^3	Влажность, мас. %
1	1.38	33.4
2	1.40	32.2
3	1.46	28.6
4	1.51	27.1
5	1.56	25.4
6	1.58	23.7
7	1.60	22.1
8	1.64	21.7
9	1.67	20.9
10	1.69	20.5
11	1.69	20.1

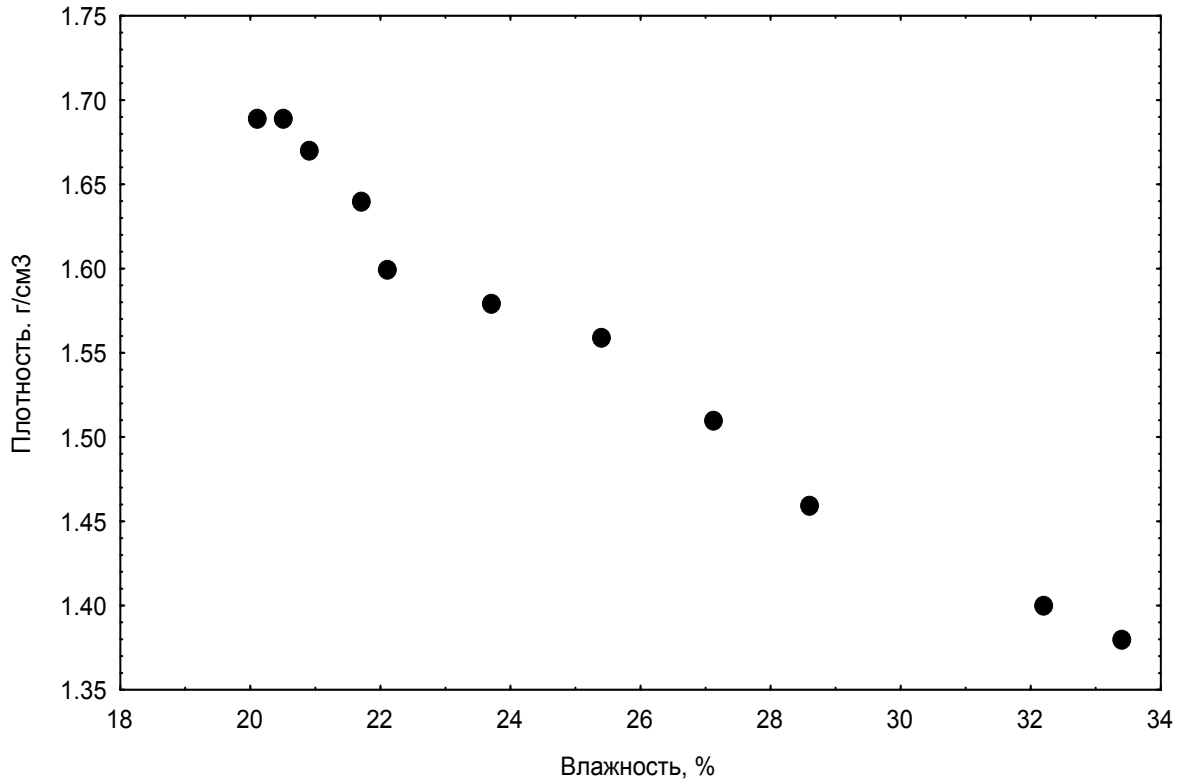


Рис. 2. Зависимость плотности почвенного агрегата (плотность, г/см^3) от влажности (мас. %)

Следующий шаг: получение регрессионного уравнения.

Используем программу STATISTICA. Идем в меню «Статистика». «Множественная регрессия», выбираем (для начала) «Продвинутый способ анализа (Advanced)». Попадаем в диалоговое окно с результатами анализа, где выбираем «Суммарные результаты» («Regression summary»). Перед нами итоговая таблица (табл. 2) проведенного регрессионного анализа со всеми необходимыми статистиками для анализа этой зависимости.

Таблица 2

Статистические результаты линейного регрессионного анализа

Regression Summary for Dependent Variable: Плотность, г/см^3 (пример 1) R = .98994247 R^2 = .97998609 Adjusted R = .97776232 $F(1,9)$ = 440.69 p < 0.00000 Std.Error of estimate: 0.01666						
Количество дат 11	Beta	Std.Err.	B	Std.Err.	t (9)	p-level
Intercept			2.150086	0.028469	75.5233	0.000000
Влажность, %	–0.989942	0.047157	–0.023471	0.001118	–20.9926	0.000000

Примечание. Число значащих цифр параметров модели определяется точностью экспериментальных измерений + один знак для расчетов.

Остановимся на этой итоговой таблице.

Линейный регрессионный анализ зависимости плотности агрегата от влажности дал нам результат зависимости плотности агрегата y от влажности x (см. в табл. 2 значения регрессионных коэффициентов в столбце, обозначенном B, и в строках «Intercept» и «Влажность», в которых приведены параметры регрессионного уравнения b_1 и b_2):

$y = 2{,}150 - 0{,}0235x$, где y – плотность агрегата (г/см^3), x – W, %.

Но мы помним о двух основных правилах использования статистических ответов (результатов). Напоминаем: 1. Как было, так и будет; 2. Статистика всегда дает ответ с некоторой вероятностью. Что означают эти правила, поясним на простом примере.

Мы получили экспериментальный материал с аналитическими ошибками (ошибки измерений сопровождают нас всегда) и с погрешностями, связанными с естественным варьированием свойств почв. Поэтому полученное уравнение применимо только для той области

влажности и соответственно плотности агрегата, для которой получено это уравнение, т. е. для области влажности от 33,4 до 22,1 %. Для более широкой области уравнение может быть уже не совсем точным. Кроме того, нужно учитывать, что мы получили уравнение по данным 11 измерений.

Здесь необходимо сделать еще одно очень существенное замечание и запомнить еще ***два очень важных правила*!** Понятно, что получили уравнение взаимосвязи плотности агрегатов и влажности. Мы проводили для этого специальный физический эксперимент, наша функция-отклик (плотность агрегатов) имеет физическую размерность – грамм на кубический сантиметр (г/см^3), предиктор (влажность) – прассовый процент (мас. %). И только при использовании этих размерностей приведенное уравнение справедливо. Изменим размерность фактора (предиктора) – уравнение будет другим, изменим размерность результирующей функции – функциональное выражение также изменится. Поэтому всегда следует неукоснительно сопровождать полученное регрессионное уравнение следующими справками:

1. Уравнение получено при указанных размерностях предиктора и функции-отклика. Обязательно отметить.
2. Уравнение получено в конкретном диапазоне варьирования предиктора при таком-то (указать) количестве повторностей. В итоге мы должны следующим образом написать регрессионную зависимость: $y = 2.1295 - 0.0229x$, где y – плотность агрегата (г/см^3), x – W, %, и эта зависимость справедлива для области влажности от 33.4 до 22.1 %. Сами же регрессионные коэффициенты безразмерны, так как уравнения регрессионного типа, как правило, не имеют физического смысла и указывают лишь на взаимосвязь явлений, ее достоверность и значимость.

Что же будет, если мы захотим «расширить области применения уравнения» и получить данные за пределами указанной области или же получим дополнительные данные в рассмотренной области влажности (33.4 – 22.1 %)? Рассмотрим эти случаи.

1. Итак, я решил продолжить мой эксперимент и получил дополнительные данные (даты) по плотности агрегата при меньшей влажности (табл. 3), данные № 12 и 13. Уже 13 дат, казалось бы, должно быть лучше, ведь чем больше данных, тем надежнее зависимость. Но давайте посчитаем.

Таблица 3

Зависимость плотности агрегата от влажности

№ п/п	Плотность агрегата, г/см^3	Влажность, мас. %
1	1.38	33.4
2	1.4	32.2
3	1.46	28.6
4	1.51	27.1
5	1.56	25.4
6	1.58	23.7
7	1.6	22.1
8	1.64	21.7
9	1.67	20.9
10	1.69	20.5
11	1.69	20.1
12	1.71	15.5
13	1.72	9.6

В результате расчетов по вышеприведенным правилам получаем табл. 4 и регрессионное уравнение $y = 1.975 - 0.017x$.

Таблица 4

Результирующая таблица

	Regression Summary for Dependent Variable: Плотность (Пример 2) R = .93122357 R^2= .86717733 Adjusted R = .85510255 $F(1,11)$ = 71.817 p < .00000 Std.Error of estimate: .04457					
	Beta	Std.Err.	B	Std.Err.	t (11)	p-level
Intercept			1.974551	0.047557	41.51985	0.000000
Влажность, %	–0.931224	0.109885	–0.016819	0.001985	–8.47450	0.000004

Отметим, что при использовании 11 дат было уравнение $y = 2.150 - 0.024\,x$ (табл. 5). Сейчас, добавив только две точки в «сухой области», получили уже другое уравнение! Это очень важно: опять работает правило «Как было, так и будет». Мы изменили набор

величин, более того, область аргумента, и само уравнение регрессии стало другим. Иначе говоря, изменили «как было», изменился и результат.

2. Я решил проверить свои данные и получил два добавочных результата в исходной области влажности. Теперь у меня 13 дат, а область влажности не изменилась, так и осталась от 33.4 до 22.1 %. Эти данные приведены в табл. 5 под номерами 6* и 6**. Посмотрим, что нам говорит применение регрессионного линейного анализа.

Мы получили регрессионное уравнение $y = 2.130 - 0.023x$. Первоначально для той же области влажности уравнение было $y = 2.150 - 0.024x$. Значит, даже добавив данные в той же самой исследованной области, мы изменили результат: регрессионное уравнение стало иным (немного, но все-таки другим) (табл. 6).

Таблица 5

Зависимость плотности агрегата от влажности

№ п/п	Плотность агрегата, г/см^3	Влажность, мас. %
1	1.38	33.4
2	1.4	32.2
3	1.46	28.6
4	1.51	27.1
5	1.56	25.4
6	1.58	23.7
6*	1.57	23.0
6**	1.58	22.8
7	1.6	22.1
8	1.64	21.7
9	1.67	20.9
10	1.69	20.5
11	1.69	20.1

Таблица 6

Результирующая таблица

	Regression Summary for Dependent Variable: Плотность* (Пример 2) R = 0.98053971 R^2 = 0.96145813 Adjusted R = 0.95795432 $F(1,11)$ = 274.40 p < .00000 Std.Error of estimate: 0.02094					
	Beta	Std.Err.	B	Std.Err.	t(11)	p-level
Intercept			2.129547	0.034640	61.4761	0.000000
Влажность*	-0.980540	0.059193	-0.022874	0.001381	-16.5651	0.000000

Опять работает правило «Как было, так и будет!». Важно всегда указывать число дат и область аргумента (независимой переменной), в которой они получены.

Теперь зададимся вопросом: «насколько надежны наши коэффициенты регрессии? или, как говорят статистики, достоверно ли наши

полученные коэффициенты отличаются от нуля? Здесь самое время вспомнить правило 2 (статистика всегда дает ответ с некоторой вероятностью) и задаться вопросом: с какой вероятностью мы может утверждать, что коэффициенты регрессии отличны от нуля (достоверно ли отличаются от нуля)? Перейдем к п. 1.3.

1.3. Статистическая оценка регрессионного уравнения

Для начала мы должны применить критерий Фишера. *F*-критерий, или критерий Фишера, обычно используется для общей оценки достоверности полученного регрессионного уравнения. Этот критерий рассматривает отношений дисперсий. В данном случае это отношение дисперсии наших измерений и дисперсии ошибок модели (отличий расчетных данных от реально полученных). Если это отношение (*F*-критерий) большое, т. е. ошибки использования модели несравненно малы по сравнению с экспериментальными ошибками, тогда с определенной вероятностью (правило 2!) мы можем утверждать, что наше уравнение достоверно.

В статистическом разделе аппроксимации (в верхней ячейке табл. 2) приведено значение *F*-критерия и его уровень значимости (*p*), указано «Regression Summary for Dependent Variable: Плотность г/см^3 (Пример 1).

$R = 0.98994247$, $R^2 = 0.97998609$ Adjusted $R = 0.97776232$.

$F(1,9) = 440.69$ p < 0.00000 Std.Error of estimate: 0.01666»

F (критерий Фишера) и уровень значимости *p* равны $F(1,9) = 440.69$ и $p < 0.00000$. Заметим, что в скобках при критерии Фишера стоят величины (1, 9). Это значения степеней свободы, которые зависят от количества переменных и общего количества дат.

Количество переменных (аргументов) у нас 1, а количество степеней свободы $f(или\ v) = n - m - 1$, где *n* – число измерений; *m* – число аргументов (11 дат минус 2). Далее мы должны были оценить достоверность уравнения по критерию Фишера, сравнив рассчитанный для этого уравнения критерий с табличным значением для соответствующей вероятности и числа степеней свободы. Вероятность мы используем 0.95 (или уровень значимости *p*-level 0.05), степеней свободы – 9. Но искать таблицу и проводить сравнение нет необходимости, это делает программа STATISTICA, указывая достоверность отличия полученного *F*-критерия от табличного. Если эти отличия достоверны (т. е. $p < 0.05$), мы можем утверждать, что наше уравнение справедли-

во. Рассчитанное значение F оказывается значительно больше критического, и с очень большой (больше 99.999 %) вероятностью принимается альтернативная гипотеза: дисперсия наших измерений больше, чем дисперсия ошибок модели, и в этом случае мы имеем право использовать модель при заданном уровне значимости.

1.4. Статистическая оценка полученных параметров аппроксимации и их достоверность

Самой распространенной статистикой для оценки значений параметров аппроксимации является прежде всего t-статистика – критерий Стьюдента, который рассчитывается по формуле (Е. А. Дмитриев, 2009)

$$t = \frac{|b|}{S_b},$$

где b – найденное среднее значение параметра; S_b – его стандартное отклонение.

Рассчитанное значение t-критерия сравнивается с его табличным значением при заданном уровне значимости и числе степеней свободы $f = n - m - 1$. Если рассчитанный критерий Стьюдента выше, чем табличный, то нулевая гипотеза отвергается в пользу альтернативной и мы можем утверждать, что параметр с заданным уровнем значимости отличен от нуля. Сравнение t-рассчитанного с t-табличным производится в программе, и в последнем столбце для обоих параметров указан уровень значимости p-level, который значительно меньше 0.05. Параметры b_1 и b_2 значимы с вероятностью больше 99.9999 %.

Мы доказали статистическую значимость (отличие от нуля) обоих параметров. На этом краткий статистический анализ модели может быть закончен. Итог использования линейной регрессии:

1. Получено уравнение связи – линейная регрессия между плотностью агрегата и влажностью почвы в виде $y = 2.150 - 0.024x$ для области влажности от 33.4 до 22.1 %, где y – плотность, г/см^3; x – влажность (%, мас.).

2. Согласно критерию Фишера полученное уравнение высокодостоверно ($P > 99.999$ %), и мы имеем право использовать эту модель.

3. Анализ значимости параметров уравнения регрессии с помощью t-критерия показал, что параметры b_1 (свободный член) и b_2 (коэффициент перед значением влажности) значимы с вероятностью боль-

ше 99.9999 %. Мы имеем право использовать рассчитанное уравнение с рассчитанными коэффициентами при данном уровне значимости.

Пример 2. Аппроксимация экспериментальных данных линейной функцией (линейная регрессия).

Определение числовых значений параметров.

К примеру, мы имеем экспериментально полученную выборку данных относительной транспирации (T / T_0 – функция отклика) и матричного давления влаги в почве (pF – фактор, предиктор). Теоретически (Е. В. Шеин, 2005) мы знаем, что в «засушливой» области относительная транспирация уменьшается практически линейно при снижении pF (т. е. при увеличении почвенной засухи). Поэтому, опираясь на известные литературные данные, мы используем линейную регрессию. Ниже в таблице приведены экспериментальные данные.

T/T_0	1.0	0.94	0.83	0.91	0.86	0.78	0.76
pF	2.5	2.56	2.71	2.78	2.87	2.92	3.2

Первоначально надо нарисовать график экспериментальных значений. Он действительно указывает, что линейная функция выбрана правильно.

Используя программу STATISTICA, раздел «Линейная регрессия»), можно провести линейную аппроксимацию данных уравнением вида $y = b_1 - b_2 x$ (рис. 3).

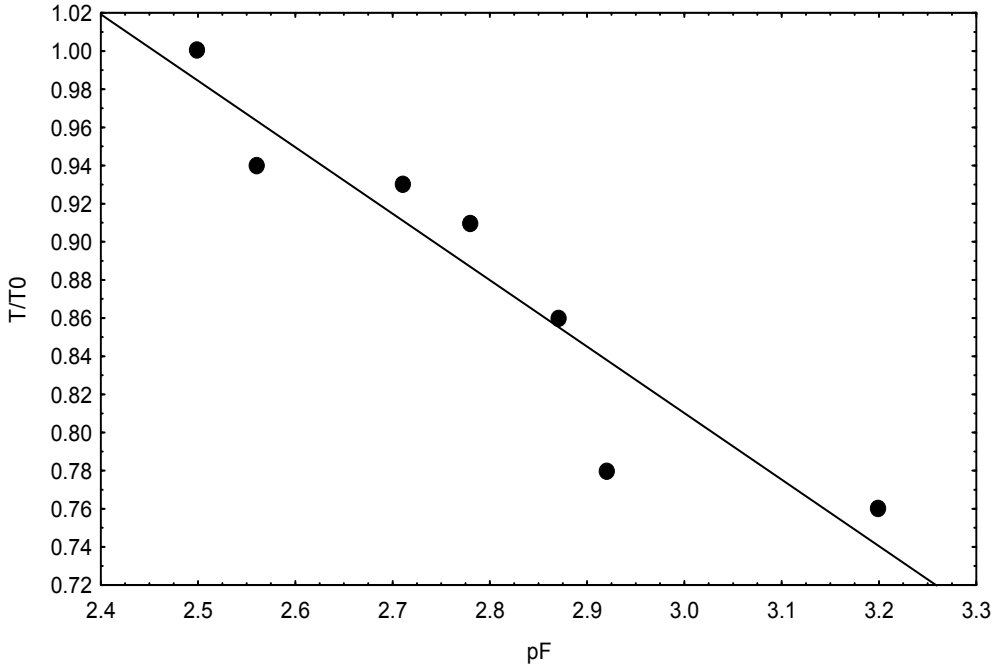


Рис. 3. Линейная зависимость T/T_0 от pF

Далее ее проводим и получаем конкретный вид функции для наших экспериментальных данных и статистические характеристики из программы STATISTICA (табл. 7).

Таблица 7

Пример линейной регрессии Model is: $v_1 = b_1 - b_2 v_2$ ().
Dep. Var.: T / T_0. Level of confidence 95.0 % (α = 0.050)

Параметр	Estimate	Standard	*t*-value	*p*-level
b_1	1.855844	0.159826	11.61169	0.000083
b_2	0.348562	0.057081	6.10647	0.001706

Уравнение для наших данных приобретает конкретный вид $T / T_0 = 1.856 - 0.349 \cdot \mathrm{pF}$. Итак, мы получили конкретный вид уравнения и можем перейти к решению второй задачи «Статистическая оценка полученного уравнения».

Как указано выше, для начала используем подход с применением *F*-критерия, который оценивает в целом достоверность нашего уравнения.

В статистическом разделе аппроксимации приведены значение *F*-критерия и его уровень значимости *p*-level, 2518.8 и 0.00008. Рассчитанное значение *F* значительно больше критического, т. е. с очень большой (P > 99.999 %) вероятностью принимается альтернативная гипотеза, что дисперсия наших измерений больше, чем дисперсия ошибок модели. В этом случае мы имеем право использовать модель при данном уровне значимости, что свидетельствует о достоверности нашей линейной модели. Можно перейти к следующему пункту: «Оценка достоверности полученных параметров аппроксимации».

Для оценки значений параметров аппроксимации наиболее часто используется критерий Стьюдента *(t*-критерий). В современных программах расчет *t*-критерия производится автоматически. Мы разберем это на нашем примере.

В табл. 7 для параметров b_1 и b_2 указан уровень значимости 0.000083 и 0.001706 соответственно. Сравнение *t*-рассчитанного с *t*-табличным производится в программе автоматически и выдается в последнем столбце в виде уровня значимости p = level. Параметры значимы с вероятностью больше 99 %.

Мы доказали значимость обоих параметров и теперь можем перейти к п. 1.5 «Анализ полученных ошибок моделирования (погрешностей аппроксимации)».

1.5. Анализ полученных ошибок моделирования (погрешностей аппроксимации)

Погрешности аппроксимации: **абсолютная и относительная, случайная и систематическая**

Итак, мы имеем определенную экспериментальную выборку. Эту выборку мы получили, задавая поочередно некоторые значения аргумента и оценивая всякий раз значения отклика. Заметим, что мы получили наши данные экспериментально, что означает присутствие некоторой экспериментальной ошибки. К этим данным мы должны теперь подобрать функцию (модель), которая наилучшим образом их опишет. При этом опять возникнут погрешности, на сей раз погрешности аппроксимации, у которых тоже будет некоторый разброс.

Характеристика качества найденной модели, т. е. то, насколько модель достоверна, оценивается путем сравнения реальных и предсказанных значений. Найденная аппроксимационная кривая никогда не может пройти точно по всем экспериментальным точкам, ведь задачей аппроксимации является именно «сглаживание» реальных данных и получение модели (уравнения) для их описания в возможно наиболее широкой области аргумента. Несоответствие отклика найденной зависимости и значения, полученного в эксперименте, называется погрешностью моделирования Δ (в литературе употребляется определение «ошибка», хотя строго говоря, оно не является термином) (рис. 4).

В зависимости от целей работы погрешность можно представить как абсолютную и относительную. Так, ***абсолютная погрешность*** есть отклонение предсказанного по модели значения от реально наблюдаемого: $\Delta_{\text{абс}} = y_{\text{эксп}} - y_{\text{расч}}$. Абсолютная погрешность удобна при сравнении отклонений предсказанных значений в конкретной аппроксимации или моделей сходных явлений. В целях сравнения моделей разного рода явлений можно использовать **относительную погрешность**, кото-

рая является отношением к абсолютной погрешности экспериментальному значения: $\Delta_{\text{отн}} = \dfrac{y_{\text{эксп}} - y_{\text{расч}}}{y_{\text{эксп}}} 100\ \%$.

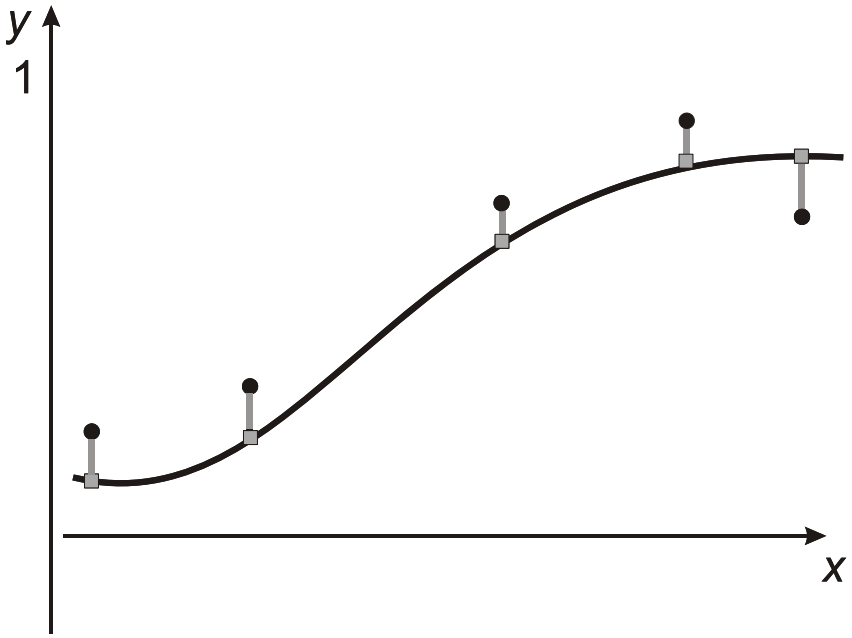


Рис. 4. Ошибки (погрешности моделирования).
Расчетная (модельная) кривая – сплошная линия;
точки – реальные экспериментальные значения.
Столбиками указаны погрешности: $\Delta_{\text{абс}} = y_{\text{эксп}} - y_{\text{расч}}$

Экспериментальные точки ($Y_{\text{эксп1}}$, $Y_{\text{эксп2}}$, $Y_{\text{эксп3}}$, $Y_{\text{эксп4}}$, $Y_{\text{эксп5}}$)
Расчетные значения по $y = \varphi(x)$ ($Y_{\text{расч1}}$, $Y_{\text{расч2}}$, $Y_{\text{расч3}}$, $Y_{\text{расч4}}$, $Y_{\text{расч5}}$)
Погрешности Δ ($Y_{\text{эксп1}}$,-$Y_{\text{расч1}}$, $Y_{\text{эксп2}}$ -$Y_{\text{расч2}}$, $Y_{\text{эксп3}}$ -$Y_{\text{расч3}}$, $Y_{\text{эксп4}}$ -$Y_{\text{расч4}}$, $Y_{\text{эксп5}}$ -$Y_{\text{расч5}}$)

Следует понимать, что если погрешность рассматривается не по модулю, то погрешности аппроксимации могут компенсировать друг друга, и в сумме погрешность всего прогноза будем минимальной. Поэтому для расчетов средней **абсолютной ошибки прогноза MAE** (*the mean absolute error*) используют только величины ошибок без учета знака: $\text{MAE} = \dfrac{1}{n} \sum_{j=1}^{n} \left| \Delta_{\text{абс}} \right|$, где j – число экспериментальных значений, n – число наблюдений.

Тот же принцип используется и для расчета **средней относительной ошибки прогноза MAPE** (*the mean absolute percentage*

error): $\mathrm{MAPE} = \frac{1}{n}\sum_{j=1}^{n}\left|\Delta_{\text{отн}}\right|$, где j – число экспериментальных значений, n – в этом случае число наблюдений.

Используют также не только величину абсолютной погрешности модели, но и логарифм этой погрешности, предполагая, что погрешности модели могут быть распределены логнормально. Однако в этом случае усложняется анализ этих погрешностей и оценки адекватности выбранной функции. Поэтому лучше всего пользоваться указанными выше абсолютными или относительными погрешностями моделирования.

Наиболее часто используют величину среднеквадратической погрешности (root mean square error) – *RMSE*

$$S_r = \left(\frac{1}{N}\sum_{j=1} n_j \Delta_j^2\right)^{\frac{1}{2}},$$

где S_r – среднеквадратическая ошибка; N – общее число измерений с учетом повторностей; n_j – количество повторностей измерений в j-м варианте; Δ – абсолютная ошибка для j-го варианта.

Кроме деления по величине и способу расчета, погрешности разделяют и по их происхождению на случайные и систематические. **Случайные погрешности** не имеют преимущественного направления (в сторону плюса или минуса) и уравновешивают друг друга. Они возникают в результате случайных отклонений значений изучаемого показателя в исследуемой выборке (ошибки экспериментатора, единичных случаев изменения внешних условий и др.). Эти отклонения объясняет теория вероятностей.

В отличие от случайных отклонений **систематические погрешности** направлены в сторону только преувеличения или только преуменьшения в результате действия на изучаемую систему постоянного неучтенного фактора (рис. 5). Таким фактором, как правило, бывает методическая неточность (смещение нуля шкалы прибора, шкала линейки нанесена неравномерно, капилляр термометра в разных участках имеет разное сечение и др.), которая нередко приводит к заметным ошибкам и, более того, к неверной интерпретации процессов.

Поэтому анализ появления случайных ошибок – чрезвычайно важный этап регрессионного анализа.

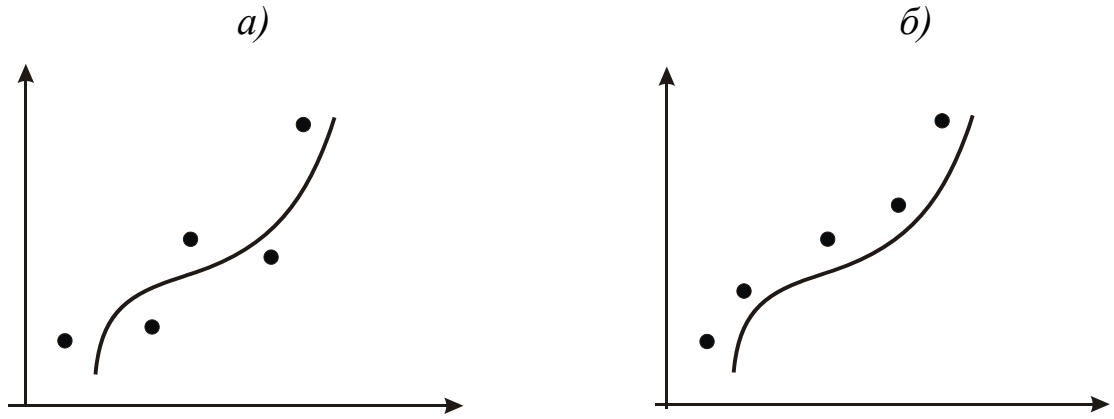


Рис. 5. Графическое изображение ошибок: *а* – случайных; *б* – систематических

Если величина систематической ошибки известна, то необходимо внести поправки в экспериментальные данные или ввести соответствующий коэффициент в модель.

К сожалению, наиболее часто в качестве критериев точности модели используют коэффициенты корреляции R либо детерминации R^2. Однако это может привести к существенным ошибкам. Действительно, коэффициент корреляции указывает не на близость рассчитанной и экспериментальной величин, а на их выстраивание в линейный вид. Этот критерий не указывает на систематическую ошибку, что чрезвычайно важно! Настоятельно рекомендуем ни в коем случае не ограничиваться расчетом коэффициента корреляции, так как наличие систематических ошибок, их значимость и диапазон эта величина не показывает. Необходимо рассчитывать другие показатели (критерии) совпадения расчетных и экспериментальных величин. Рассмотрим основы анализа ошибок моделирования на приведенном выше примере расчета относительной транспирации от величины pF.

Пример 3. Продолжение анализа зависимости $T / T_0 = f(\mathrm{pF})$ на наличие возможных систематических ошибок моделирования.

Сначала надо проанализировать зависимость экспериментальных значений от расчетных (рис. 6). В идеале это должна быть прямая – биссектриса угла начала координат. Очень похоже, что у нас все так и

получается, лишь при малых значениях наблюдаются некоторые отличия. Однако окончательный вывод о наличии/отсутствии систематических ошибок и области их распространения сделать трудно. Необходимо построить зависимость ошибок аппроксимации от расчетной (или реальной) величины функции отклика. Эта зависимость приведена на рис. 7. Из рисунка определенно следует, что ошибки имеют случайный разброс, а наибольшие наблюдаются в точках 7 (около 0.02) и 6 (–0.058). Видимо, функция отклика (T / T_0) при высоких значениях нестабильна, возникают ошибки разного знака и значительные по величине: это нестабильная область определения T / T_0. Лучше, однако, провести аналитическую проверку значимости зависимости ошибок от реальной величины, которая предлагалась для аналитического анализа ошибок, провести проверку значимости зависимости ошибок от реальной величины $y(T / T_0)$, а именно $\Delta_{ij} = a_i + b_i y_{ij}$.

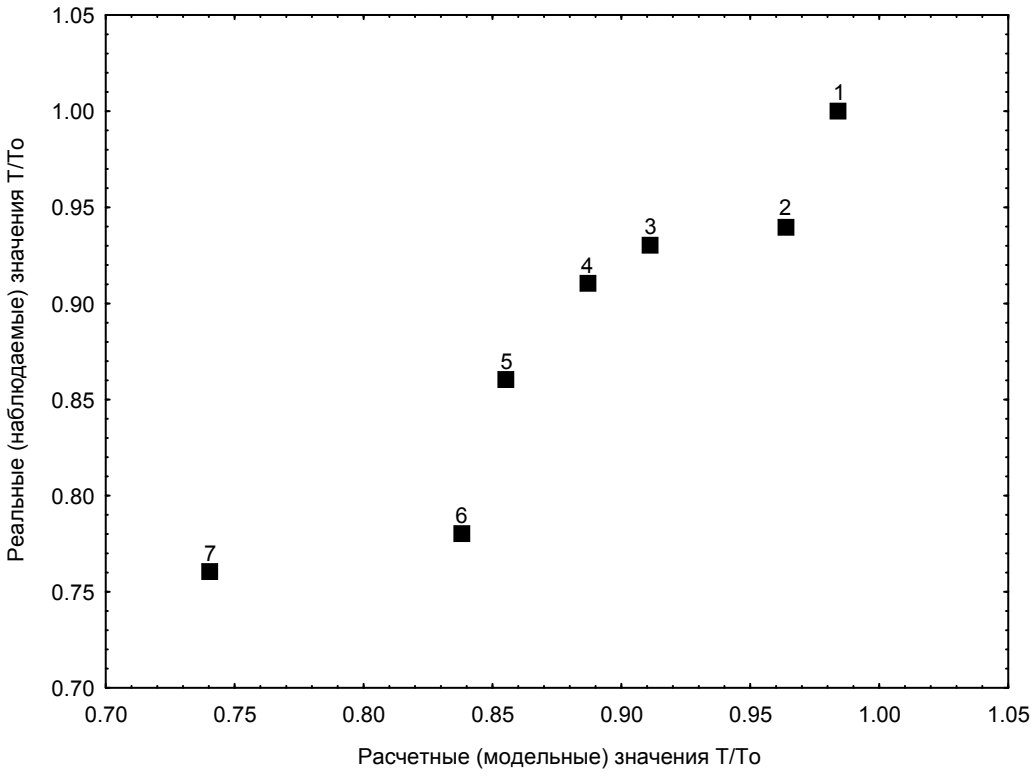


Рис. 6. Зависимость реальных (наблюдаемых) значений от расчетных величин

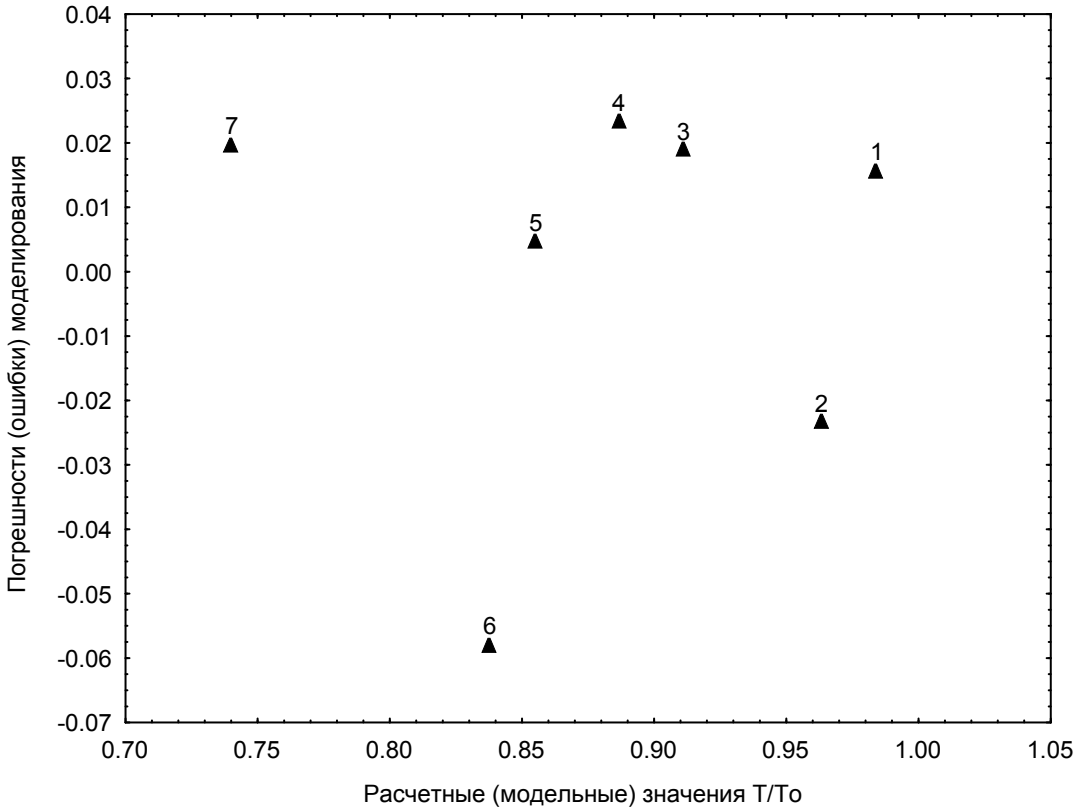


Рис. 7. График расположения значений ошибок (погрешностей) от расчетных (модельных) значений T / T_0

Попробуем получить регрессионное уравнение и проанализировать его на значимость по *F*-критерию, а параметры *a* и *b* уравнения – на значимое отличие от нуля с помощью *t*-критерия. Во-первых, данное уравнение зависимости ошибок от реальной величины T / T_0 оказалось при проверке по *F*-критерию незначимым. *F*-критерий составляет всего лишь 0.6704 при его достоверности 0.45, так же, как и оба параметра *a* и *b* по *t*-критерию. Это определенно показывает, что у нас нет зависимости ошибок от реальной величины, что говорит об отсутствии систематических ошибок. Только после этих проверок и четырех этапов статистических проверок самого уравнения, его параметров, анализа ошибок аппроксимации мы можем определенно сказать: «Полученное линейное уравнение вида $T / T_0 = 1.856 - 0.349\text{pF}$ значимо и не дает систематических ошибок при его использовании».

Глава 2. МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

2.1. Использование метода наименьших квадратов для расчета множественной регрессии

В предыдущей главе переменная отклика выражалась различными способами как функция одного фактора среды. Однако, как это всегда бывает в природных объектах, свойство (характеристика) почвы может зависеть от большего числа факторов. В подобной ситуации используется множественный регрессионный анализ.

Линейное уравнение в случае двух независимых переменных описывает плоскость

$$E_y = b_1 + b_2 x_1 + b_3 x_2 + \varepsilon,$$

где E_y – функция отклика; x_1 и x_2 – факторы (предикторы); b_1, b_2 и b_3 – параметры, или коэффициенты регрессии; ε – ошибка определения отклика.

Средний отклик E_y равен b_1 при $x_1 = 0$ и $x_2 = 0$. Скорость изменения среднего отклика вдоль осей x_1 и x_2, b_2 и b_3 соответственно. Таким образом, b_2 есть изменение E_y в зависимости от x_1 при фиксированном значении x_2, b_3 – изменение E_y в зависимости от x_2 при фиксированном значении x_1.

Параметры оцениваются методом наименьших квадратов (МНК), т. е. путём минимизации суммы квадратов разностей между значениями наблюдаемой и ожидаемой переменными. В геометрической интерпретации это означает следующее: плоскость регрессии выбрана так, что сумма квадратов вертикальных расстояний между наблюдаемыми значениями и плоскостью минимальна. Аналитически метод наименьших квадратов означает, что **надо найти минимальное значение** суммы квадратов отклонений некоторой функции от искомых переменных. Этот метод – один из основных в регрессионном анализе, когда оценивается неизвестный отклик по выборочным данным предиктора на основе регрессионной модели.

Сущность МНК (обычного, классического) заключается в том, чтобы найти такие параметры b, при которых сумма квадратов отклонений (ошибок для регрессионных моделей, которых часто называют остатками регрессии) ε будет минимальной; т. е. в регрессионном анализе используются вероятностные модели зависимости между переменными

$$y_t = f(x_t, b) + \varepsilon_t,$$

где ε_t – так называемые *случайные ошибки* модели.

Методы нахождения минимальных значений случайных ошибок для выборки значений предиктор – отклик могут быть различными, но в основе лежит метод наименьших квадратов. Находят такое значение связи между откликом и предиктором, при котором ошибки (остатки или погрешности регрессионной модели) будут минимальны. Разработка этого метода принадлежит таким великим математикам, как Гаусс (1795 г.), который первым применил метод, Лежандру (1805 г.), независимо открывшему и опубликовавшему его под современным названием, а также Бесселью, Ганзену и другим, которые общими усилиями распространили и усовершенствовали метод. Отметим, что в начале XX в. работы А. А. Маркова позволили включить метод наименьших квадратов в аппарат математической статистики, которым мы сейчас пользуемся.

Множественный регрессионный анализ, выполненный с помощью компьютера, даёт не только оценки коэффициентов b_1, b_2 и b_3, но также и стандартные ошибки оценок и соответствующие значения t (см. пример 4). Значения t можно применять для проверки, равен ли один из параметров нулю, или, что точнее, достоверно ли отличается от нуля этот регрессионный параметр.

Пример 4. Использование множественной регрессии

Рассмотрим пример множественной регрессии, используя данные эксперимента по изучению зависимости относительной транспирации T/T_0 от давления влаги pF, однако в этом эксперименте контролировалась еще и относительная влажность воздуха $W_{\text{отн}}$. Получили следующие данные (табл. 8).

Таблица 8

Изучение зависимости относительной транспирации T/T_0 от давления влаги pF и относительной влажности воздуха $W_{отн}$

№ п/п	T/T_0	pF	$W_{отн}$
1	1	2.5	72
2	0.94	2.56	70
3	0.83	2.71	62
4	0.91	2.78	58
5	0.86	2.87	48
6	0.78	2.92	36
7	0.76	3.2	30

Так как действуют уже два фактора (давление почвенной влаги pF и влажность воздуха $W_{отн}$), а функция отклика одна (T/T_0), то график образует трехмерную поверхность (рис. 8). Конечно, понять этот график, а тем более его интерпретировать весьма затруднительно. Попробуем выразить зависимость T/T_0 от pF и от $W_{отн}$ в виде множественной регрессии.

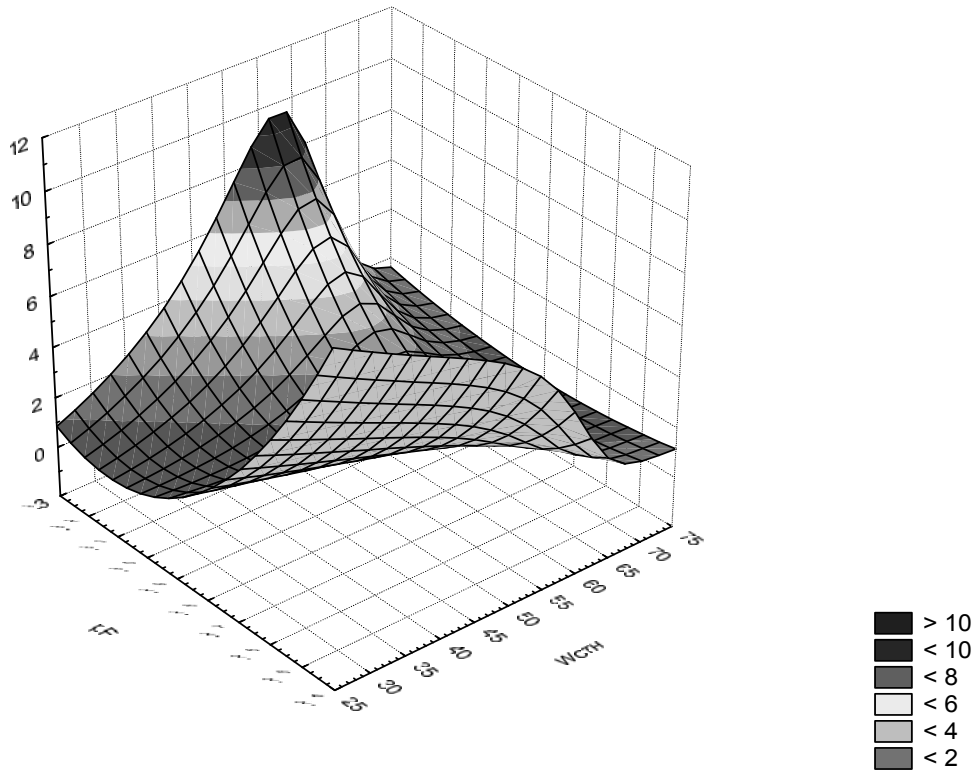


Рис. 8. График множественной регрессии T/T_0 от pF и от $W_{отн}$

Рассчитаем уравнение множественной регрессии для данных по относительной транспирации в зависимости от pF и относительной влажности воздуха.

Используем в программе STATISTICA раздел «Множественная регрессия», в результате получаем табл. 9.

Таблица 9

Итоговая таблица

	Regression Summary for Dependent Variable: T/T_0 (Пример 3 отн. транспир.) R = .89901628 R^2 = .80823026 Adjusted R^2 = .71234540 $F(2,4)$ = 8.4292 p < .03678 Std.Error of estimate: .04660					
	Beta	Std.Err.	B	Std.Err.	$t(4)$	p-level
Intercept			0.955250	1.046450	0.912848	0.412965
pF	-0.263579	0.802285	-0.096926	0.295027	-0.328535	0.758987
$W_{отн}$	0.642561	0.802285	0.003423	0.004274	0.800914	0.468055

Согласно табл. 9 уравнение множественной регрессии будет выглядеть следующим образом:

$$T/T_0 = 0.955 - 0.097 \text{ pF} + 0.003\ W_{отн}.$$

Статистический анализ этого уравнения проведем, выделив три основных итоговых пункта:

1. Получено уравнение множественной линейной регрессии, связывающее относительную транспирацию T/T_0 с давлением влаги pF и влажностью воздуха $W_{отн}$ в виде уравнения $T/T_0 = 0.955 - 0.0969 \times$ $\times$ pF + 0/0034 $W_{отн}$ для области давлений влаги 2.5 – 3.2 pF и влажности воздуха от 30 до 72 %.

2. Полученное уравнение согласно критерию Фишера высокодостоверно (уровень значимости менее 0.05, p = 0.03678).

3. Анализ значимости параметров уравнения регрессии с помощью t-критерия показал, что параметры b_1 (свободный член), b_2 (коэффициент перед значением pF) и b_3 (коэффициент перед значением влажности воздуха $W_{отн}$) не являются значимыми при принятом (менее 0.05) уровне значимости. Это означает, что мы имеем право использовать рассчитанное уравнение для определения относительной транспирации по данным о pF и относительной влажности воздуха $W_{отн}$, указывая, что оно достоверно лишь при уровне значимости 0.5. По-видимому, от использования такого уравнения лучше отказаться, а попробовать взять другой набор данных по относительной транспи-

рации в соответствии с величинами pF и влажности воздуха. Вполне вероятно, увеличив число дат, удастся повысить значимость параметров регрессионного уравнения.

2.2. Одновременный набор данных и построение уравнений множественной линейной регрессии

Регрессионный анализ наиболее часто применяется при анализе количественной взаимосвязи независимо от полученных почвенных характеристик и действующих факторов. Например, мы изучаем ряд свойств почвы в почвенной траншее длиной несколько десятков метров, определяя через каждые 20 см (или другое расстояние) ряд свойств: плотность почвы, содержание физической глины, водопроницаемость, сопротивление пенетрации, полевую влажность. Конечно, кроме пространственных закономерностей, связанных с изменением почвенного покрова, нас должны интересовать взаимосвязи основных свойств. К примеру, как зависит плотность почвы и сопротивление пенетрации от влажности. Влажность и сопротивление пенетрации можно довольно быстро и просто измерить и с помощью ранее полученной регрессионной формулы получить пространственное изображение величин плотности почвы: выявить на сельскохозяйственном поле зоны уплотнения и переуплотнения, рыхлые зоны, в которых влага не застаивается и быстро фильтруется в глубинные слои. Таким образом, восстановив с помощью регрессионного уравнения (по траншейным данным плотности, сопротивления пенетрации, влажности почвы) и рассчитав пространственное распределение плотности почвы по полю, мы можем высказать гипотезы о формировании уплотненных зон и их связи с влажностью почвы всего изучаемого поля. Итак, как же получаются эти регрессионные уравнения по полевым сопряженным данным?

Пример 5. Одновременный набор данных.

В результате исследования свойств почв траншеи мы получили (табл. 10) выборку пространственно распределенных данных о плотности почвы, влажности W и сопротивлении пенетрации (СП). Данные получены независимо, выборка организована так, что в почве одновременно определялись три свойства. Мы хотим выяснить зависимость плотности почвы от влажности и сопротивления пенетрации, чтобы в дальнейшем для данного поля пользоваться этой зависимостью и с её помощью (обладая пространственно распределенными

данными влажности и сопротивления пенетрации) восстановить (рассчитать) распределение плотности по полю.

Таблица 10

Экспериментальные данные влажности, плотности и сопротивления пенетрации почвы (траншея глубиной 10 – 15 см). Почва дерново-подзолистая, окультуренная (Московская обл., Пушкинский р-н Зеленоградский опорный пункт Почвенного института имени В. В. Докучаева)

№ п/п	*W*, %	Плотность, г/см^3	Сопротивление пенетрации, МПа
1	20.61	1.54	2.66
2	22.14	1.37	1.77
3	20.68	1.45	1.99
4	22.43	1.48	2.19
5	21.29	1.47	2.15
6	20.17	1.45	1.96
7	20.33	1.61	2.71
8	20.66	1.58	2.64
9	23.53	1.50	2.32
10	23.25	1.48	2.14
11	24.83	1.46	2.18
12	24.62	1.38	1.97
13	24.24	1.44	2.44
14	23.89	1.42	1.98
15	26.36	1.36	1.34
16	27.97	1.36	1.71
17	26.94	1.37	1.35
18	27.35	1.41	1.97
19	27.13	1.42	1.78
20	27.57	1.41	1.75
21	26.44	1.38	1.49
22	26.10	1.38	1.52
23	27.63	1.41	1.64
24	27.34	1.37	1.55
25	27.54	1.36	1.68
26	27.80	1.41	1.74
27	26.04	1.39	1.59
28	24.90	1.40	1.75

Мы собираемся выяснить, где в поле наблюдаются участки с повышенной плотностью, а где – с пониженной и т. д. Для этого используем множественный регрессионный анализ. В итоге получаем следующую результирующую таблицу с параметрами регрессии и их статистиками (табл. 11, аналог «выдачи» программы STATISTICA).

Таблица 11

Параметры уравнения множественной регрессии зависимости плотности почвы от её влажности и сопротивления пенетрации, статистики уравнения

Regression Summary for Dependent Variable: Плотность (Пример_транш) R = .91929904 R^2 = .84511072 Adjusted R = .83271958 $F(2,25)$ = 68.203 p < .00000 Std.Error of estimate: .02703						
	Beta	Std.Err.	B	Std.Err.	$t(25)$	p-level
Intercept			1.268084	0.101561	12.48589	0.000000
W, %	-0.161331	0.115504	-0.003958	0.002834	-1.39676	0.174763
Сопротивление пенетрации, МПа	0.794631	0.115504	0.137393	0.019971	6.87969	0.000000

Уравнение связи имеет вид

$$y = 1.268 - 0.004x_1 + 1.137x_2,$$

где y – плотность почвы, г/см^3; x_1 – влажность, % мас.; x_2 – сопротивление пенетрации, МПа.

Как видно из приведенного уравнения и статистической таблицы, полученное регрессионное уравнение согласно критерию Фишера достоверно и может быть использовано в указанной области влажности и сопротивления пенетрации. Совершенно очевидно, (при уровне значимости меньше 0.05) мы можем утверждать, что сопротивление пенетрации – значимый фактор, а влажность в диапазоне экспериментальных данных – нет. Это означает, что только при диапазоне влажности от 20.2 до 28 % уравнение может быть использовано. Очень узкий диапазон для суглинистых почв неприемлемый для оценок. Но если эти ограничения учитывать, то плотность почвы можно достоверно определить по показаниям пенетрометра, по величинам сопротивления пенетрации. Это очень практичный вывод.

Глава 3. НЕЛИНЕЙНАЯ РЕГРЕССИЯ

3.1. Основные функциональные зависимости, используемые в естествознании, их классификация

Конечно, следует снова повторить уже указанное правило: линейных зависимостей в природе весьма немного: «Природа не терпит линейных зависимостей». Вопрос «Почему?» сложный. По-видимому, вследствие того, что в процессе изменения фактора (предиктора) переменная отклика тоже меняется. Это связано с её структурой, внутренними взаимосвязями и прочим, что приводит к нелинейной зависимости. Примеров можно привести множество. Даже в случае разобранных нами примеров 1 и 2 мы всегда оговаривали область, где зависимости были близки к линейным. И в то же время мы прекрасно представляем, что зависимость плотности агрегата от влажности (см. пример 1) нелинейна, так как при некоторой влажности агрегат достигает предела усадки, когда объем уже не уменьшается и линейная зависимость меняется на экспоненциальную. Вот и получается, что в природе нелинейные зависимости встречаются значительно чаще, и при аппроксимации экспериментальных данных необходимо использовать нелинейную регрессию.

Основная задача данной главы – подобрать вид нелинейной функции и научиться некоторым методам определения параметров этой функции для конкретной экспериментальной выборки, т.е. проводить процедуру аппроксимации.

Таким образом, задача аппроксимации сводится к последовательному выполнению двух операций:

1. Установлению вида зависимости $y = \varphi(x)$ – процедура, которая осуществляется, как правило, опытным путем (т. е. на основе своего опыта или опыта коллег), но по определенным правилам.

2. Определение численных значений неизвестных параметров выбранной функции (b_1, b_2, b_3, …, b_n), при которых задача сглаживания экспериментальных данных решается наилучшим образом (анализ ошибок аппроксимации для точек y_1, y_2, y_3, …, y_n).

Согласно общепринятому определению функция – одно из основных понятий математики – выражает зависимость одних переменных от других. Если величины x и y связаны так, что каждому значению x соответствует определённое значение y, то y называют (однозначной) функцией аргумента x.

В экспериментах, как правило, мы задаем величину x и получаем соответствующую ей величину функции y. Или, как говорят математики, функцию отклика. Например, мы определяем основную гидрофизическую характеристику (ОГХ) методом мембранного пресса, когда первоначально насыщенный образец находится на керамической пластине. Затем мы задаем над образцом почвы повышенное газовое давление, и из образца вытекает вода до тех пор, пока внешнее заданное давление не сравняется с давлением почвенной влаги. В этот момент мы определяем влажность образца. Мы получили одну из точек на кривой ОГХ – зависимость влажности почвы y от давления влаги x. Повышая ступенчато давление влаги x над образцом, мы получаем соответствующие значения влажности y. Эти пары значений будут характерными именно для испытуемого образца. И теперь мы должны аппроксимировать полученные данные некоторой функцией. Начнем с первой части процедуры аппроксимации – установления вида зависимости $y = \varphi(x)$.

Все множество функций, применимых к описанию явлений в природе, можно разделить на несколько больших групп в зависимости от направленности процесса:

I. Монотонные:
   1. Возрастающие.
   2. Убывающие.

II. С одним экстремумом.

III. С несколькими экстремумами.

IV. С изломом.

Монотонные (убывающие и возрастающие) функции

Убывающие и возрастающие функции могут быть:

1) линейными;

2) степенными (за исключением параболы);

3) показательными и экспоненциальными;
4) логарифмическими;
5) логистическими.

Линейная функция

Линейная функция уже нами рассматривалась (см. рис. 1). Напомним, что она имеет вид $y = b_1 + b_2 x$, графически это прямая линия.

Линейная функция употребляется во всех областях науки для описания пропорциональной зависимости. Однако для описания зависимостей многих почвенных процессов, таких как ОГХ, функция влагопроводности, зависимости температуропроводности от влажности и множества других этот вид функции непригоден: эти зависимости нелинейны и для их описания требуются другие виды. В науках об окружающей нас природе линейные функции используются нечасто.

Степенная функция

Самым простой и распространенной функцией среди степенных является квадратичная с графиком, называемым параболой (рис. 9). Здесь надо отметить, что крутизна ветвей этой параболы будет тем выше, чем выше степень (приведены примеры для степени, равной 2 и 4). График этот будет симметричен относительно оси *y*, но только в том случае, если показатель степени четный, тогда и функция называется четной, как это представлено на рис 10.

В почвоведении, где предикторы представлены, как правило, положительными числами, степенная функция выглядит несколько иначе: $y = x^{b_1}$ или $y = \left(\frac{x}{b_2} \right)^{b_1}$. При этом в самом общем случае *x* может быть любым действительным числом при показателе степени больше 0 и не равно 0 при показателе степени меньше 0. Для любых b_1 график функции проходит через точку (1; 1). В уравнениях степенной функции первую называют однопараметрической, а вторую – двухпараметрической.

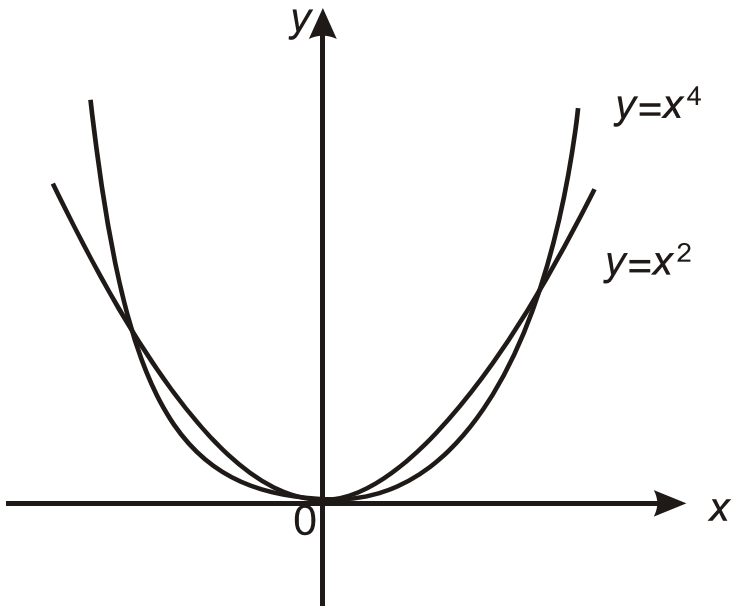


Рис. 9. График зависимости степенной функции (парабола)

Графики степенной функции при положительном показателе b_1 называются параболами порядка b_1, а при отрицательном – гиперболами порядка b_1. Таким образом, смена знака показателя при аргументе с положительного на отрицательный превращает функцию возрастающую в убывающую. Это правило нам уже встречалось выше (см. линейную регрессию): «Смена знака перед аргументом превращает функцию из убывающей (если был минус) в возрастающую».

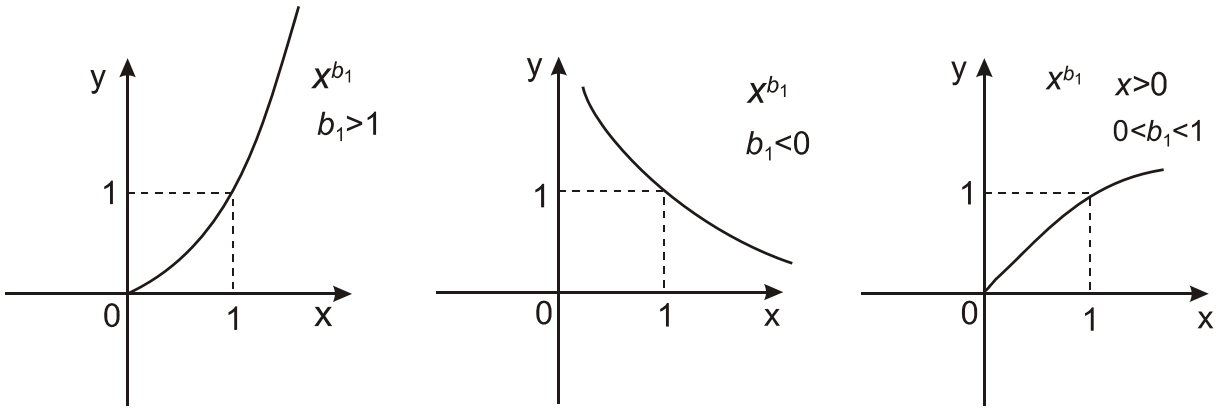


Рис. 10. Примеры графиков степенной функции

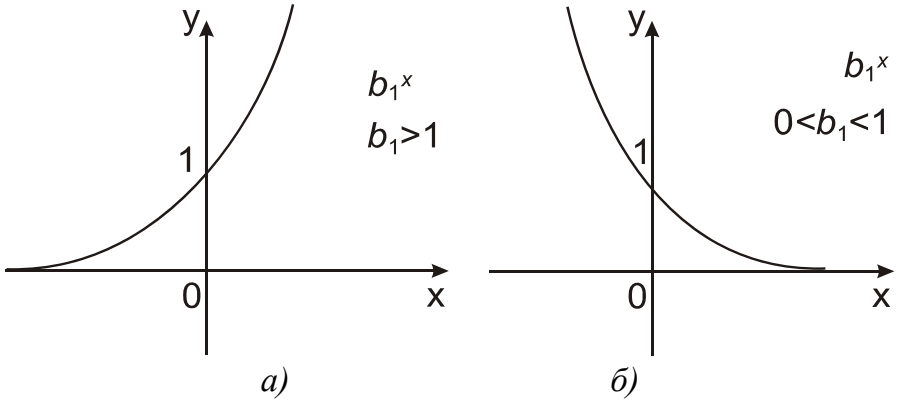
Степенная функция очень широко используется в почвоведении. Например, уравнение Фрейндлиха для описания процесса сорбции (Орлов, 1992 г.): $Q = mC^n$, где Q – концентрация сорбированного иона [моль/100 г почвы]; C – концентрация иона в контактирующем растворе, моль/л; m и n – эмпирические параметры, характерные для

каждого почвенного образца. Важен и еще один момент. Во многих случаях эмпирические параметры носят физический смысл. В частности, в указанном степенном уравнении Фрейндлиха, несмотря на его эмпирический характер, степенной параметр *n* (показатель степени), как было показано Г. Спозито, можно рассматривать как показатель неоднородности сорбционных центров, он приближается к нулю по мере возрастания неоднородности и стремится к 1 при увеличении их однородности. Это означает, что определение параметров аппроксимации имеет не только практическое значение (использование в математических моделях), но и важно для понимания и сравнения физических явлений. Однако, чтобы понять физический смысл и значение регрессионных коэффициентов, надо проводить специальные эксперименты, анализировать поведение функции в различных условиях. В самом общем случае эти коэффициенты, как и любые регрессионные коэффициенты, безразмерны. При использовании регрессионного уравнения нужно указывать не только его вид, но и размерность функции отклика и факторов (предикторов), а также границы области определения факторов (обязательно!).

Показательная и экспоненциальная функции

При рассмотрении многих природных процессов и явлений может оказаться полезным анализ показательной функции b_1^x, где b_1 – основание показательной функции, имеет только положительные значения, а в качестве переменной выступает показатель степени.

Частный (но наиболее употребительный!) случай показательной функции – функция экспоненциальная: $y = b_1 \exp(b_2 x)$, где exp – основание натуральных логарифмов (e = 2.7182...). Особым свойством экспоненты является то, что она возрастает (рис. 11, *а*) или убывает (рис. 11, *б*) быстрее, чем степенная функция.



а) *б)*

Рис. 11. Пример графиков показательной функции

Используя функцию натурального логарифма ln$x$, можно выразить показательную функцию с произвольным положительным основанием b_1 через экспоненту: $b_1^{\,x} = e^{x \ln b_1}$. Эта связь позволяет ограничиться изучением свойств экспоненты в большинстве аппроксимаций.

Опять-таки работает правило «знака перед аргументом»: «плюс» – возрастающая функция, «минус» – убывающая.

Экспоненциальная функция сначала довольно быстро возрастает (или убывает при минусе перед аргументом), а затем постепенно и плавно «выходит на плато», что указывает на наступление некоторого равновесия. Эти свойства показательной и экспоненциальной функций позволяют использовать их во многих областях науки, в том числе при описании химических реакций, роста численности микроорганизмов и др. Экспоненциальные функции очень важны, поскольку они описывают такие физические явления, как радиационный распад. Эту функцию используют для определения количества радиоактивного вещества, оставшегося к моменту t по формуле $N = N_0 2^{-\frac{t}{T_{1/2}}}$, где N_0 – начальное количество вещества; $T_{1/2}$ – период полураспада радиоактивного вещества, т. е. промежуток времени, в течение которого распадается половина начального количества ядер радиоактивного изотопа. В агрохимии для токсикантов, агрохимикатов и других используют термин «период полуразложения», хотя английский вариант этого термина один: *Half-life.* Мы еще неоднократно столкнемся с этой функцией при описании кинетических явлений разложения, распада, отмирания микроорганизмов и др. Запомним эту функцию, она очень полезна.

Описывая экспоненциальные функции, нельзя не упомянуть часто встречающееся распределение (или функцию) Вейбулла. В несколько упрощенном виде эта функция записывается так:

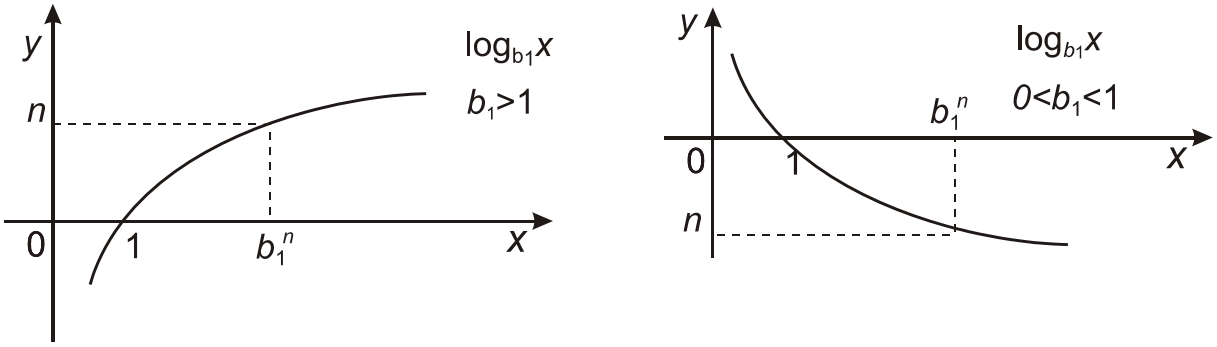
$$y = 1 - \exp(-(x / b_1)^{b2}).$$

Она встречается в почвоведении нередко, например, для описания скорости распада агрегатов в воде и распределения гранулометрических частиц, хотя строгого, подходящего для большинства почв уравнения для гранулометрического состава не предложено. Все же упомянем распределение Вейбулла и для распределения частиц по гранулометрическому составу

$$y = b_1 + (1 - b_1)\left[1 - \exp\ (-b_2 x^{b_3})\right].$$

Логарифмическая функция

Логарифмическая функция с основанием b_1 – это функция вида $\log_{b_1} x$, где $b_1 > 0$ и $\neq 1$. Если $b_1 > 1$, то функция на всей области определения возрастает (рис. 12, *а*), а если $0 < b_1 < 1$ – убывает (рис. 12, *б*). Особенностью логарифмической функции является то, что нулевое значение она принимает в точке $x = 1$ при любом $b_1 > 0$.



а) *б)*

Рис. 12. Примеры графиков логарифмической функции

При работе с логарифмической функцией может оказаться полезным тот факт, что графики показательной вида b_1^x и логарифмической функций, имеющих одинаковое основание, симметричны относительно биссектрисы $y = x$, т. е. функция $y = \log_b x$ обратна показательной функции $y = b^x$. Поэтому в большинстве случаев используют показательную или экспоненциальную функции (рис. 13).

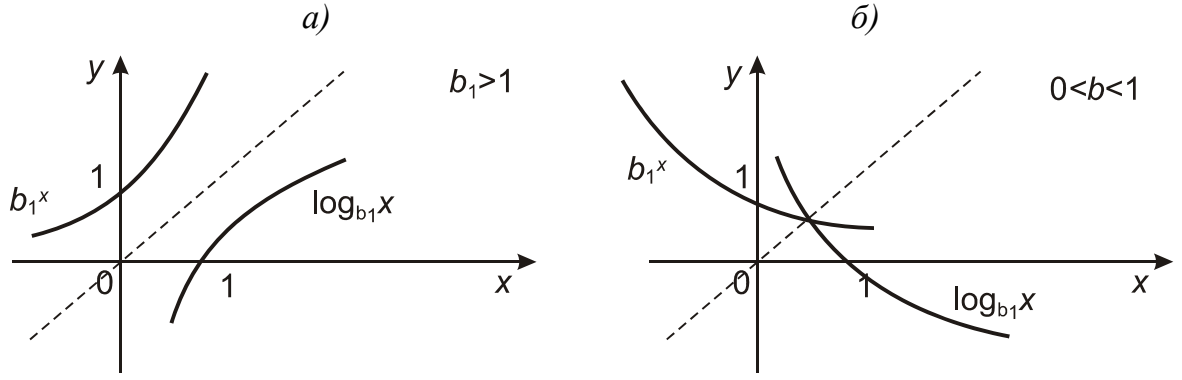


Рис. 13. Примеры графиков функций: *а* – показательной; *б* – логарифмической

В научных исследованиях логарифмическую функцию используют широко: в астрономии – для измерения величины блеска звезд, описания спиральной формы галактик; в биологии – при описании формы ракушек улиток, рогов животных, паутин и т. д.; в физике – для оценки громкости шума; в экономике – при различных расчетах роста денежных капиталов и др. Но наиболее часто логарифмическая функция используется в почвоведении при характеристике реакций первого порядка $\ln\frac{C_0}{C_i} = k_1 t$ или в записи экспоненциальной функции $C_i = C_0 \exp\ (k_1 t)$, где C_0 – начальная концентрация некоторого вещества; C_i – концентрация этого вещества в момент времени t, а k_1 – постоянная первого порядка. Все эти названия («кинетика первого порядка», «постоянная первого порядка» и др.) широко используются при математическом описании процессов разложения или трансформации некоторых веществ в почве (например, сложных органических веществ, радионуклидов и др.). Сейчас же для нас важно то, что и показательная, и логарифмическая функции могут быть в той или иной мере сведены к экспоненциальной, пользоваться которой при аппроксимации многих природных явлений бывает проще.

Логистическая функция

Логистическая функция, или логистическая кривая, – самая общая сигмоидальная (S-образная) кривая. Простейшая логистическая функция может быть описана формулой $y = \frac{1}{1+e^{-x}}$.

Для логистической функции область допустимых значений x совпадает с множеством всех действительных чисел. Более того, она уникальна своей формой (рис. 14): сначала медленно возрастает, затем возрастает уже ускоренно, напоминая показательную функцию, а после второй фазы возрастания уже медленно и постепенно приближается к некоторой максимальной величине. Из-за своей формы, напоминающей греческую прописную букву «сигма», эту кривую нередко называют «сигмоидной».

Благодаря своей форме, указанным трем фазам (медленного, ускоренного возрастания и постепенного выравнивания), эта сигмоидная функция применима для очень многих природных процессов, которые сначала развиваются медленно (лаг-фаза в биологических

процессах), потом ускоряются, а в завершающей стадии постепенно замедляются.

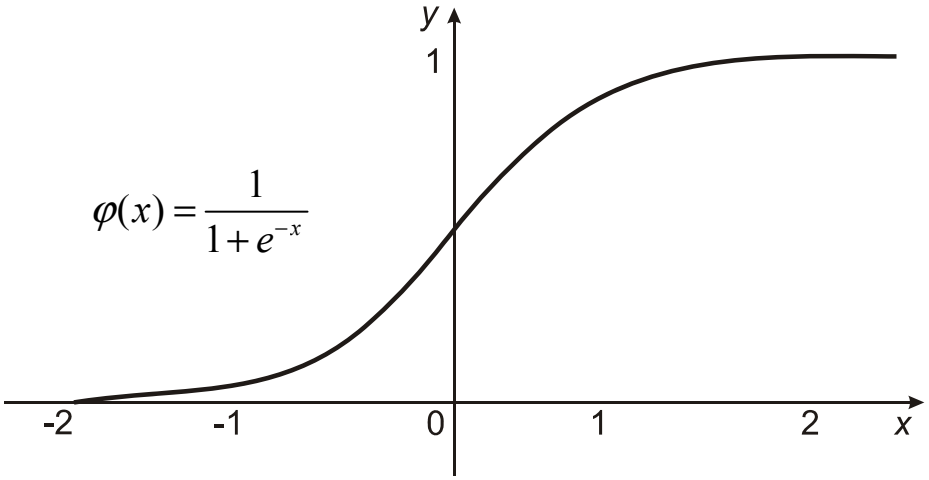


Рис. 14. График логистической функции

Вторая стадия – стадия быстрого роста логистической кривой приблизительно соответствует экспоненте (показательной функции). Затем по мере насыщения рост замедляется, проходит линейную фазу и, наконец, в «зрелом периоде» практически останавливается. Так можно описать многие процессы роста при обратном знаке перед аргументом функции – это процессы разложения, распада и многие другие.

Эта функция очень широко используется в экологии, где носит название по имени впервые сформулировавшего его бельгийского математика – «уравнение Ферхюльста». При ограничении процессов размножения организмов в популяции, каким-либо ресурсом, например, количеством доступной пищи, удельная скорость роста популяции зависит от ее численности (плотности). Математические модели, учитывающие данный эффект, называются моделями плотностно-зависимого роста. Логистическое уравнение Ферхюльста является простейшей моделью из этого ряда. В данной модели предполагается, что удельная скорость роста популяции линейно уменьшается с ростом численности, и имеется также некая предельная численность популяции K, при достижении которой добавление к популяции новых особей возможно лишь при условии определенной гибели уже имеющихся. Эта предельная численность K имеет название «емкость среды». Данный параметр известен также как ресурсный. Уравнение Ферхюльста в дифференциальном виде имеет следующий вид:

$$\frac{dx}{dt} = xr\left(1 - \frac{x}{K}\right),$$

где r – мальтузианский параметр; K – ресурсный параметр.

Уравнение Ферхюльста имеет аналитическое решение

$$x(t) = \frac{x_0 K e^n}{K - x_0 + x_x e^n},$$

где x_0 – начальная численность популяции.

Для многих биологических процессов применяется также уравнение логистического типа, которое тоже носит название по имени великого ученого, впервые применившего этот тип уравнения. Это был крупнейший французский биохимик Жаком Моно (1912 – 1976). Уравнение Моно в самой общей записи имеет следующий вид:

$$y = \frac{b_1 x}{K_\text{М} + x},$$

где $K_\text{М}$ – константа Михаэлиса, равная концентрации субстрата, при которой скорость роста равна половине максимальной; b_1 – максимальная скорость роста, равная величине r в формуле Ферхюльста.

В почвоведении этот тип функций также широко распространен. Вспомним уравнение Ленгмюра, которое почвоведы используют для описания процессов сорбции (Д. С. Орлов, 1992 г.): $A = A_\infty \dfrac{K_\text{л} C}{1 + K_\text{л} C}$, где A – равновесная концентрация вещества в поглощенном состоянии, моль/г, в соответствии с концентрацией вещества в растворе, С, моль/л; A_∞ – максимальная концентрация поглощенного вещества, моль/г; $K_\text{л}$ – константа Ленгмюра, размерность которой совпадает с размерностью концентрации, моль/л, а физический смысл её аналогичен константе Михаэлиса: это концентрация вещества в растворе, при которой концентрация вещества в сорбированном состоянии равна половине максимальной $A_\infty / 2$.

Нередко сигмоидную логит-функцию записывают, используя экспоненциальный вид, например так: $y = \dfrac{\exp(b_0 + b_1 x)}{1 + \exp(b_0 + b_1 x)}$.

В этом случае в большей степени выполняется условие скорости подъема функции на второй стадии и более плавного «торможения» подъема на третьей. Это, как правило, бывает ближе к описанию некоторых природных процессов.

С этой сигмоидной функцией, чрезвычайно распространенной в математическом моделировании природных процессов и в почвоведении, в частности, мы еще многократно столкнемся в нашем курсе.

Из истории вопроса…

Закон Вебера – Фехнера

Основной закон психофизиологии, связывающий интенсивность воздействия и интенсивность ощущения, носит название закона Вебера – Фехнера. Э. Г. Вебер (E. H. Weber, 1795 – 1878) – немецкий физиолог и анатом, исследовал органы чувств человека, т. е. слух, зрение и осязание. Он впервые обнаружил, что ощущение растет нелинейно с увеличением воздействующего раздражителя. Например, если к 60 горящим свечам добавить еще одну, то человек заметит увеличение яркости. А вот если горят 120 свечей, то человек не заметит добавления еще одной. Требуется добавить две свечи, чтобы испытуемый заметил увеличение яркости. А если горят 300 свечей, то надо добавить уже пять свечей, чтобы почувствовать различие. Так Вебером было предложено понятие «порог различения». Вполне понятно, что порог ощущения нелинейно зависит от интенсивности (прироста) раздражающего фактора. Этот факт стал основой для всей экспериментальной физиологии органов чувств и получил название «правило Вебера». Густав Теодор Фехнер (G. Th. Fechner, 1801 – 1887) – немецкий физик, проанализировал данные Вебера и предложил логарифмическую зависимость ощущения от величины воздействующего сигнала. Вот и получился закон Вебера – Фехнера, который математически выразился в соответствующей логарифмической зависимости. Все было бы хорошо, если бы в середине 20-го века американский исследователь С. С. Стивенс не усомнился именно в логарифмической форме и не предложил степенную функцию, которая больше, по его мнению, подходит для описания поведения сенсорной системы, т. е. в

этом случае в логарифмических координатах зависимость ощущения от воздействующего фактора (стимула) становится простой прямой. Так какой же формулой закона пользоваться: экспоненциальной или степенной? Какая лучше описывает процесс? Какое уравнение использовать, скажем, в светотехнике? Мы уже знаем, что для различных процессов можно использовать разные формулы, которые могут лучше или хуже описывать этот процесс. И к выбору можно привлечь некоторые формальные критерии. Например, приведенный ниже непараметрический критерий Вильямса – Клюта. Но, к сожалению, он тоже не всегда дает однозначный ответ, и, кроме того, мы ведь всегда хотим разобраться в механизме явления, т. е. понять, что происходит, в том числе и с помощью математического описания явления. Так, видимо, оказалось и в случае закона Вебера – Фехнера. В определенном диапазоне воздействующего фактора экспериментальные данные описывает лучше экспонента, а в другом – степенная зависимость. Вывод из этой интересной истории, конечно, должен быть такой: надо понимать различие между физикой процесса, которая может быть очень сложной и никогда не опишется в широком диапазоне одним лишь уравнением, и простой аппроксимацией экспериментальных данных, которая, хотя и полезна в работе, но чаще формальна. Мы снова возвращаемся к основному девизу математического моделирования: «Цель математического моделирования – не цифры, а понимание».

Функции с одним экстремумом

Среди функций, имеющих один максимум или минимум, можно выделить две, наиболее употребимые в естественных науках:

1. Параболическая.
2. Гауссовская функция и гауссовская логит-функция.

Параболические функции

График с одним экстремумом можно получить в частном случае степенной функции или при использовании любой квадратичной функции. Например, очень часто можно встретить запись классической параболической функции в виде полинома 2-й степени, а именно $y = b_0 + b_1 x + b_2 x^2$.

Действительно, случай квадратичной функции мы будем относить к более общему случаю полиномиальной зависимости (многочлена степени *n,* где $n = 2$), а под параболой будем понимать функцию вида $y = x^{b_1}$ или $y = b_2 - \left(\frac{x}{b_3}\right)^{b_1}$, где b_1 – четное число (рис. 15).

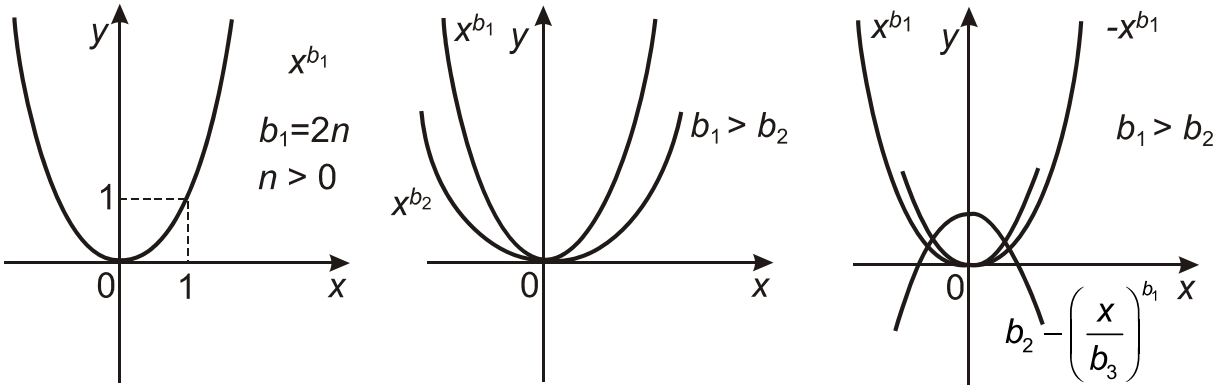


Рис. 15. Примеры графиков параболических функций

Гауссовская функция

Кривая **нормального распределения**, или гауссовская кривая, напоминает параболу и чаще всего описывает распределение частот в выборке (например, гистограмма распределения частот ошибок). Если выборкой является популяция, многие ее свойства можно описать с помощью кривой нормального распределения (например, изменение численности при изменении условий среды)

$$\phi(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}^e} - \frac{(x-\mu)^2}{2\sigma^2} = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

где μ и σ – параметры гауссовского распределения, интерпретируемые обычно как математическое ожидание (среднее) и дисперсия; x – значение случайной величины.

Вид гауссовской функции с одним максимумом для случая описания роста живых организмов $y = b_1 \exp\left[-0.5\frac{(x-b_2)^2}{b_3^2}\right]$, где b_1, b_2, b_3 – параметры аппроксимации.

Вид гауссовской функции весьма специфичен: очень напоминает симметричный колокольчик (рис. 16), симметричный в отношении

значения b_2, среднего. В природе случаи строгой симметрии не так уж часты, нередко для лучшего понимания происходящих процессов и встречающихся явлений указанным параметрам придают физический или эколого-биологический смысл: b_1 – обилие вида; b_2 – биологический оптимум и b_3 – толерантность (мера экологической амплитуды).

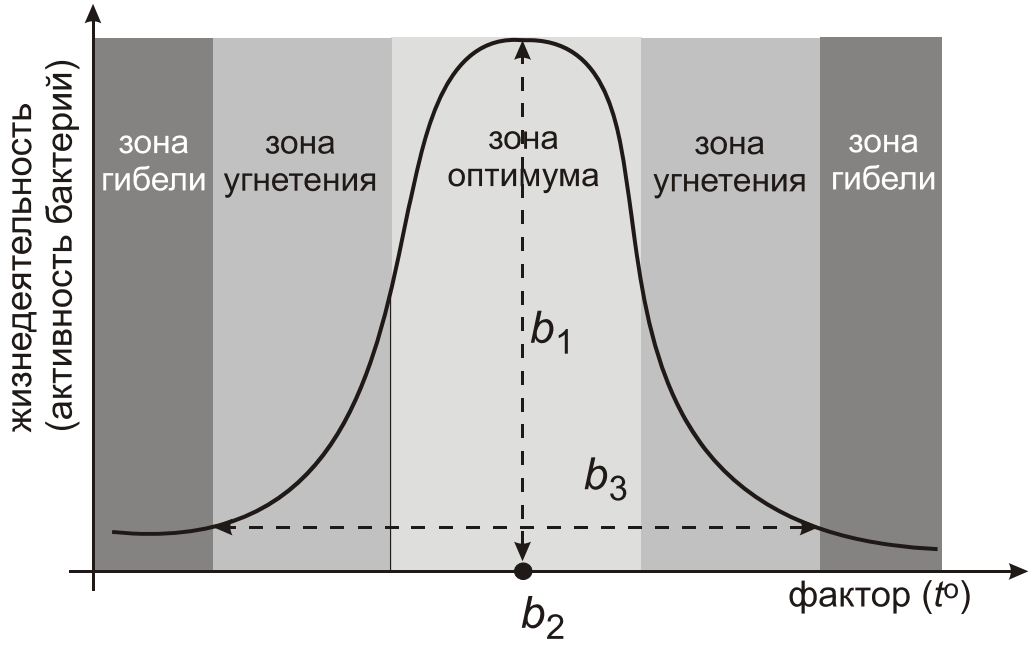


Рис. 16. Вид гауссовского распределения (Джонгман и др., 1999): b_1 – обилие вида (максимальная численность); b_2 – точка оптимума фактора, b_3 – предел толерантности (биологическая валентность)

Иногда используют несколько другой вид гауссовской кривой, так называемую «гауссовскую логит-кривую» (рис. 17). В этом случае в качестве предиктора берут параболу (или полином 2-й степени). Эта кривая имеет более плоскую вершину, чем гауссовская, что нередко используется для описания зависимости продуктивности от того или иного природного фактора.

Итак, можно подвести некоторые итоги рассмотрения нелинейных функций, применяемых в естествознании. Ниже, на рис. 17, приведены их аналитические уравнения и графики, которые помогут выбрать ту или иную функцию для описания экспериментальных данных.

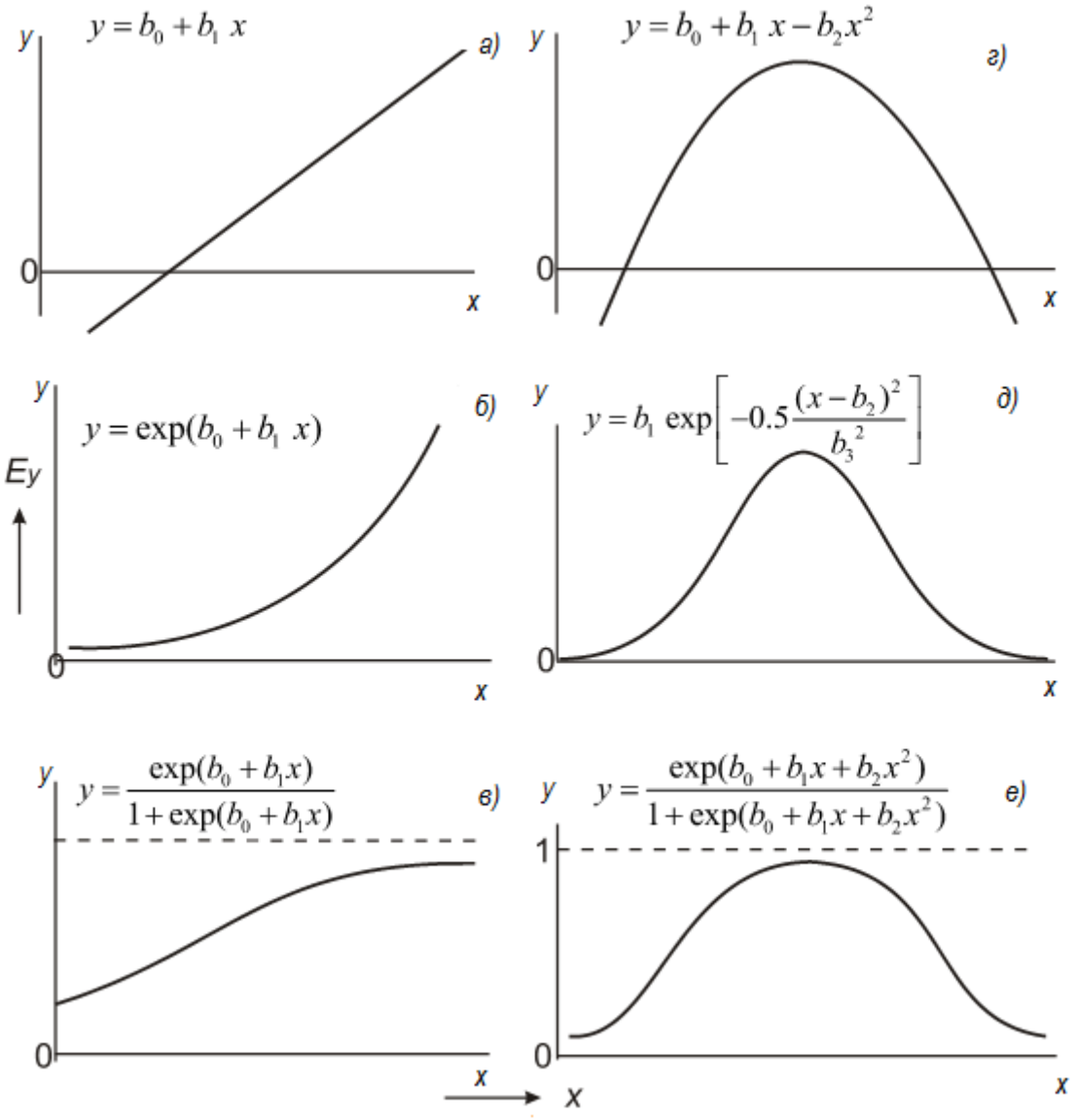


Рис. 17. Примеры графиков для описания экспериментальных данных: *а* – прямая; *б* – экспоненциальная кривая; *в* – сигмоидная кривая; *г* – парабола; *д* – гауссовская кривая; е – гауссовская логит-кривая

Функции с несколькими экстремумами

Итак, перед нами стоит задача попытаться найти функциональное выражение числовой последовательности, имеющей сложный вид. Это может быть динамика какого-либо свойства, т. е. его изменение во времени, тогда аргументом будет время, а переменной отклика – какое-либо свойство почвы, например, влажность, температура, содержание иона и прочее. Или одномоментное распределение какого-либо свойства по профилю почвы. Тогда аргументом будет глубина *z*, а переменной отклика – какое-либо свойство почвы, например содер-

жание солей. В любом случае мы имеем распределение некой функции отклика, имеющее несколько экстремумов. Попытаемся найти подходы для математического описания этого явления.

Для описания кривых с несколькими экстремумами применяются следующие функции:

1. Полиномы 3-й степени и более высокой.
2. Сплайн-функция.

Полиномиальная функция

Полиномиальная функция имеет вид многочлена степени n $y = b_0 + b_1x + b_2 x^2 + b_3 x^3 + \ldots + b_n x^n$ и применяется для описания экспериментальных данных в случае, если ни одна из вышеописанных функций не применима. Полиномиальная функция высокой степени способна описать практически любые данные, однако интерпретация коэффициентов аппроксимации при этом затруднительна. В почвоведении, впрочем, нередко бывает необходимо описать довольно сложный процесс или явление. Например, математически описать динамику температуры на поверхности почвы для того, чтобы наиболее удобным образом далее представить это описание в математических моделях теплопереноса в почвах. Здесь приходит на помощь именно полиномиальная функция.

Для математического описания сложных и неясных по физической сути явлений нередко ничего не остается, как попробовать начать это описание именно с полинома, причем высокой степени. Потом, возможно, удастся это уравнение (полином) упростить и, более того, даже выяснить физический смысл некоторых параметров (см. п. 3.2 «Элиминирование параметров аппроксимации»). Однако подчеркнем, в большинстве случаев параметры полиномов не имеют физического смысла и в физически обоснованных моделях их использование возможно, но без раскрытия физической основы процесса и соответственно управления им.

Например, в моделях распространения тепла в почве нередко используют зависимость теплопроводности λ от объемной влажности почвы Θ, которая имеет нелинейный характер, полиномиальную модель, которой предложили Чанг и Хортон (Chung and Horton, 1987 г.): $\lambda(\Theta) = b_1 + b_2 \cdot \Theta + b_3 \cdot \Theta^{0.5}$, где b_1, b_2 и b_3 – некоторые эмпирические па-

раметры, не имеющие строгого физического смысла. Это означает, что при использовании модели сложно будет сказать, что происходит с почвой, когда изменяется тот или иной параметр, поэтому затруднительно интерпретировать результаты изменения температурного режима почв при изменении параметров указанного уравнения. Значительно более «физично» (т. е. с приданием физического смысла и последующей физической интерпретации) уравнение Т. А. Архангельской, которое тоже связывает температуропроводность с влажностью w следующим образом:

$$\kappa = \kappa_0 + a\exp\left[-0.5\left(\frac{\ln\left(\frac{w}{w_0}\right)}{b}\right)^2\right],$$

где w – влажность почвы; κ – соответствующая ей температуропроводность; κ_0, a, w_0 и b – параметры кривой. Эти параметры имеют ясный физический смысл: κ_0 – температуропроводность сухой почвы; w_0 – влажность, при которой достигается максимум температуропроводности; $\kappa_0 + a$ – максимальная температуропроводность при $w = w_0$. Параметр b характеризует ширину пика кривой.

Как видно, параметры κ_0, a, w_0 и b имеют ясный физический смысл, и это дает возможность объяснить, к чему может привести изменение того или иного параметра, или вследствие чего, каких процессов этот параметр изменился. В этом случае, когда уравнение получено на основании рассмотрения физических законов, параметры нелинейной регрессии могут иметь не только физическое обоснование, но и размерности. Конечно, это уже понимание процессов, что и требует математическое моделирование.

Или, например, в работе А. В. Смагина описание запасов гумуса в профиле черноземов производилось с помощью комбинаций двух экспоненциальных функций типа $C(z) = A\exp(-mz) + B\exp(-bz) + C_0$, где A, B, C и m, b – параметры аппроксимации распределения органического вещества по глубине профиля чернозема (z). Приведенное уравнение, по сути, представляет собой полином с использованием экспоненциальных функций. Применяя указанную аппроксимацию и анализируя парамет-

ры модели, А. В. Смагин приходит к очень интересным выводам. В частности, расчет модельной динамики запасов гумуса при агродеградации показывает, что за 100 – 200-летний период эксплуатации черноземные почвы теряют от 30 до 85 т/га органического вещества. В дальнейшем процесс хоть и замедляется, но со временем деградация затрагивает все более глубокие слои почвы. Характерный для объемного содержания органического вещества черноземов экстремум смещается с глубины 5 – 10 см до 40 см и постепенно сглаживается.

Заметим, что это результат моделирования, точнее, использования полиномов для описания пространственного внутрипрофильного распределения свойств почвы. Можно считать, что это определенный метод анализа профильных распределений, поиска закономерностей и, возможно, предсказания процессов. Использование этого метода находит широкое применение, в частности, при анализе профильного распределения различных параметров поровой структуры почв и в ряде других исследований. Подчеркнем, в данном случае рассмотрены лишь метод анализа распределения того или иного свойства по профилю и поиск с помощью этого метода пространственных внутрипочвенных закономерностей. В любом случае этот метод перспективен, но лишь при строгом учете всех условий аппроксимации и элиминирования параметров.

Все указанные процедуры легко воспроизводятся в стандартных пакетах STATISTICA, SIGMAPLOT и некоторых других. Надо быть внимательным при их использовании и тщательно разбираться со статистической выдачей, в которой есть все указанные выше результирующие величины, поскольку на их основании и делают выбор и выводы.

Сплайн-функция

Сплайн-функции не имеют конкретного математического выражения, это кусочно-заданная функция, в которой для каждого отрезка между экспериментальными точками подбирается свой вид полинома 3-й степени или более высокой, хорошо описывающий прохождение кривой через многочисленные максимумы и минимумы. Этот вид функции используется при моделировании различных поверхностей, в том числе при картографировании (горизонтали на картографиче-

ских картах). Однако для других целей применение этой функции ограничено, так как физического смысла такой вид зависимости не несет и последующая (после аппроксимации экспериментальных данных сплайн-функцией) физическая интерпретация такого рода аппроксимации невозможна.

Не всегда можно подобрать к экспериментальным данным некую простую по форме зависимость, например, при действии на экосистему нескольких внешних факторов. В этом случае приходится иметь дело со сложными полиномиальными функциями с множеством параметров. Но так ли уж важны для целей адекватного моделирования все коэффициенты в подобных непростых случаях? С большой долей вероятности можно утверждать, что не все параметры имеют физический смысл, а поэтому значимо влияют на результат моделирования. Для упрощения сложных функций в таких случаях применяется процедура *элиминирования* параметров.

3.2. Элиминирование параметров аппроксимации

В практике экспериментальных исследований очень часто определяют кривую сорбции (десорбции) паров воды почвами. Это довольно стандартное для физиков почв определение. Получились весьма разнообразные, но в целом характерные S-образные кривые (рис. 18).

Естественно, S-образный вид кривых сорбции предполагал единую функцию для описания этого явления. Однако из литературы известно, что единого уравнения для описания этого явления не существует. Большинство предлагаемых уравнений применимо к отдельным частям изотерм сорбции паров воды почвами (это уравнение БЭТ, Ленгмюра, Фаррера и др.) и не в состоянии описать изотермы во всем интервале относительных давлений $0 < p/p_0 < 1$. Однако была необходимость предложить модель для диапазона относительных давлений паров $0.1 < p/p_0 < 1$ и использовать её для конкретных почв. Как же поступить в этом случае? Было предложено это Г. В. Харитоновой. Разберем случай, характерный для использования процедуры элиминирования параметров в целях единого и вполне аккуратного описания такого вида кривых.

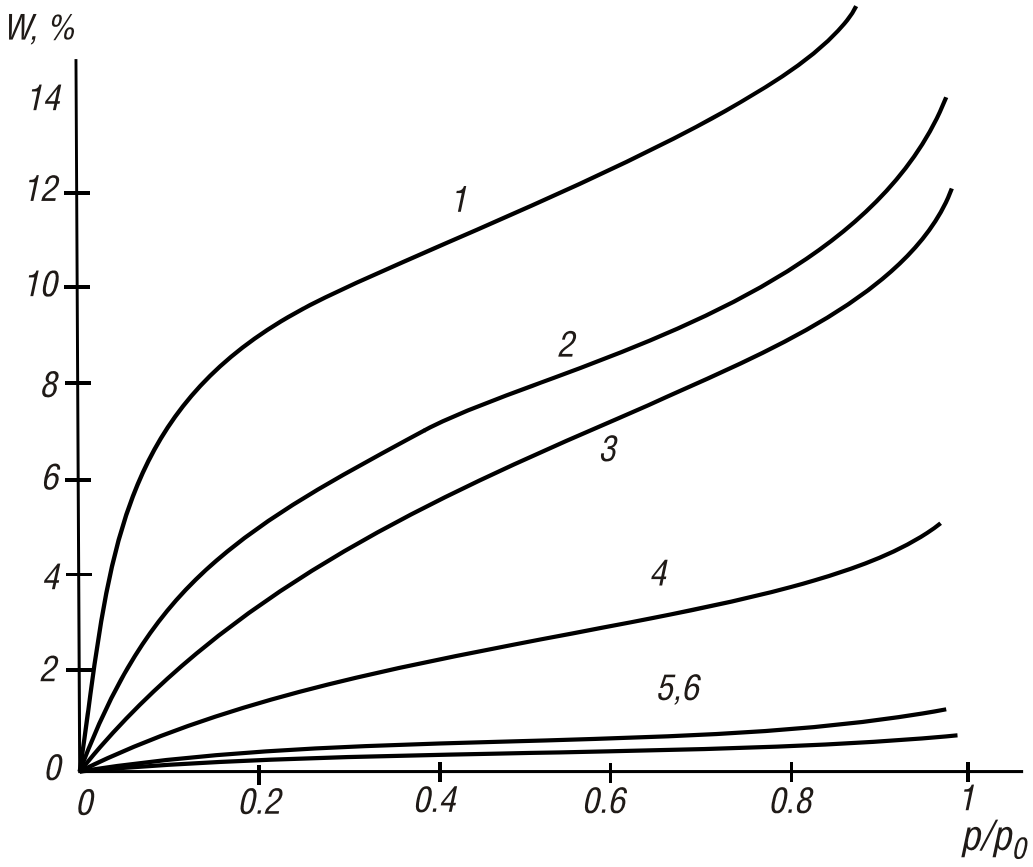


Рис. 18. Примеры кривых сорбции паров воды для различных почвенных объектов (по «Полевым и лабораторным методам исследования свойств и режимов почв», 2001)

Элиминирование (от лат. *elimino* – выношу за порог, изгоняю) – исключение из уравнения аппроксимации параметров, не имеющих физического смысла, и за счет этого упрощение вида зависимости.

Г. В. Харитонова применила весьма общий прием: сначала использовала полином высокой степени (например, 6-й), а затем избавилась (элиминировала) от некоторых составляющих полинома без потери качества описания им экспериментальных данных. Для анализа изотерм сорбции паров воды почвами зависимостью вида $W = f(p/p_0)$, где W – влажность, % от массы сухой почвы; p/p_0 – относительное давление паров воды, она за основу приняла уравнение

$$W = C_0 + C_1(p / p_0) + C_2(p / p_0)^2 + C_3(p / p_0)^3 + ... + C_n(p / p_0)^n$$

как наиболее полно соответствующее изотермам сорбции на всем их протяжении (Карпачевский, 1985 г.).

Естественно, для полученного массива данных сорбционных измерений коэффициент детерминации R^2, или оценка корреляции фак-

тических значений с моделью с возрастанием n от 1 до 8, монотонно возрастает от 0.9 при n = 1 и приближается к величинам, близким 1 при n = 6 – 8. Среднее значение коэффициента детерминации R^2, вычисленное для экспериментальных и расчетных данных при n = 6, составило R^2 = 0.996 – 0.999. Стандартная ошибка оценки влажности S_w по модели (S_{ey} – стандартная ошибка для оценки Y) с возрастанием n от 1 до 6 существенно уменьшается. Статистическая значимость указанного полинома проверялась по критерию Фишера: $F_{\text{регрессии}} = S^2_{\text{рег}} / S^2_{\text{ост}}$, где $S^2_{\text{рег}}$ – дисперсия, обусловленная использованием полинома; $S^2_{\text{ост}}$ – остаточная дисперсия, т. е. дисперсия, связанная с ошибками измерения факторов-предикторов (Химмельблау, 1973 г.): F-критерий для регрессии варьировал в интервале 200 – 3000, что много больше, чем $F_{\text{таб}}$ = 5 – 10 при принятом уровне значимости α = 0.05 ($F_{\text{таб}}$ для каждого метода определяется числом экспериментальных точек). Следовательно, на основании статистической значимости регрессии можно в уравнении принять n = 6.

Окончательный вид уравнения связи влажности и относительного давления паров воды для изотерм сорбции был выбран согласно следующим критериям (критериям элиминации): численное совпадение экспериментальных и расчетных данных; минимизация числа факторов модели на основании анализа критерия Фишера и согласие параметров модели с измеряемыми величинами. Выбор «наилучшего» уравнения регрессии проводился методом исключения, согласно которому элиминирование предикторной переменной проводится на основании величины ее частного F-критерия (Дрейпер, Смит, 1987). Если наименьшая величина частного F-критерия меньше заранее выбранного критического значения F, то переменная исключается из рассмотрения. Далее производятся перерасчет уравнения регрессии с учетом остающихся переменных и переход к следующему шагу, если новое уравнение регрессии статистически значимо (сравнение $F_{\text{рег}}$ и $F_{\text{таб}}$ при заданном уровне значимости p) и значение R^2 сохраняются на прежнем уровне.

Этим критериям удовлетворяет уравнение

$$W = A(p/p_0) + B(p/p_0)^3 + C(p/p_0)^6 + D,$$

где W – влажность, % от массы сухой почвы; p/p_0 – относительное давление паров воды; A, B, C и D – расчетные коэффициенты. Анализ

величин *A*, *B*, *C* и *D* показал, что коэффициенты уравнения, или параметры модели, сравнимы с влажностями изотерм по величине, при этом коэффициент *A* близок к величине максимальной гигроскопической влажности $W_{мг}$, которая относится к числу почвенно-гидрологических констант (Мичурин, 1975 г.). *F*-критерий для регрессии варьировал в интервале 140 – 2200, что говорит о возможности использования полученного уравнения для описания изотерм сорбции паров воды почвами в изученном интервале относительных давлений водяного пара $0.1 < p/p_0 < 0.98$ (табл. 12).

Таблица 12

Коэффициенты *A*, *B*, *C* и *D* уравнения при сорбционно-статическом определении изотерм адсорбции паров воды почвами (Харитонова, Шеин, Воронов, 2012)

Горизонт	Глубина, см	*A*	*B*	*C*	*D*	r^2	S_w
Дерново-сильноподзолистая легкосуглинистая почва							
Ad	0 – 2	4.91	–4.73	4.96	0.42	0.992	0.14

Итак, процедура элиминирования основана на сравнении аппроксимаций с различным количеством параметров с помощью критерия Фишера *F*. Еще раз напомним, что в статистике критерий Фишера используют для сравнения дисперсий двух выборок, соответствующих критерию нормальности: $F = \frac{\sigma_1^2}{\sigma_2^2}$ (H_0: $\sigma_1^2 = \sigma_2^2$), где σ_1^2 – бóльшая дисперсия; σ_2^2 – меньшая дисперсия.

Если для определенного уровня значимости и соответствующих чисел степеней свободы для первой и второй выборок вычисленное значение критерия *F* больше табличного, то нулевая гипотеза отвергается, а дисперсии считаются различными. При оценке сложного влияния фактора на функцию отклика (в данном случае на влажность) можно статистически грамотно записать: $y = b_1 + b_2 x + ... + b_n x^n + s$, где *s* – ошибки. Указанные ошибки относятся к результатам измерения наших аргументов, они также являются составной частью модели. Эти ошибки тоже имеют свою дисперсию σ_s, и модель при своем использовании дает некоторые ошибки $\sigma_м$. Поэтому можно использовать вышеприведенное дисперсионное отношение *F,* или критерий Фише-

ра, для характеристики качества нашей модели. Действительно, мы используем средние значения аргумента и функции отклика и поочередно проверяем, насколько сильно эти аргументы влияют на функцию отклика с учетом вариабельности. Вот для этого и используется критерий Фишера. Напомним, что критерий Фишера F – односторонний и как каждый статистический критерий имеет в своей основе нулевую гипотезу. В данном случае нулевая гипотеза формулируется как равенство выборочных (т. е. экспериментальной и модельной) дисперсий $\sigma_{м=} \sigma_s$. Если отношение выборочных дисперсий не превышает табличного значения F, то с заданной вероятностью и соответствующей степенью свободы нулевая гипотеза не отвергается. Это значит, что различия между сравниваемыми дисперсиями недостоверны. Значит, и наше уравнение может при указанной вероятности считаться недостоверным. Но если рассчитанное отношение оказывается больше критического, то с соответствующей вероятностью принимается альтернативная гипотеза, т. е. дисперсии неодинаковы: дисперсия наших измерений больше, чем дисперсия ошибок модели, и в этом случае мы имеем право использовать модель при данном уровне значимости.

Тот же принцип используется и для исключения параметров из уравнения аппроксимации. Так, если в результате исследования была подобрана некоторая зависимость (в нашем примере это был полином 6-й степени), нулевая гипотеза в этом случае исходит из того, что дисперсии экспериментальная и обусловленная моделью не различаются. Следовательно, модель является недостоверной. Если же вычисленное значение критерия F больше, чем табличное (для соответствующего уровня значимости и степеней свободы), это означает, что нулевая гипотеза неверна, а данное уравнение действительно объясняет исследуемое явление. Так и получилось, когда использовался полином 6-й степени. Далее необходимо упростить найденную зависимость с помощью процедуры элиминирования путем сравнения дисперсий изначального уравнения и уравнения с исключением одного из параметров. Если критерий покажет значимость новой зависимости и более того, критерий F возрастет, исключенный параметр не имел существенного влияния, и для целей аппроксимации его можно опустить.

В результате поочередного исключения параметров, анализа критериев Фишера и проверки качества полученного *нового* уравнение по R^2 и было получено уравнение для характеристики процесса сорбции паров почвами, имеющее всего четыре параметра (вместо начальных семи) – *A, B, C* и *D*: $W = A(p/p_0) + B(p/p_0)^3 + C(p/p_0)^6 + D$.

Строго говоря, метод элиминирования параметров применяется наиболее часто и наиболее точно при иной организации выборки или ином построении эксперимента. Так, мы использовали принцип построения эксперимента, когда следили за переменной отклика (функцией) при активном изменении фактора-предиктора (аргумента) и находили между ними соответствие. Это классический физический эксперимент. Но выборка может быть построена и по-другому: мы можем следить за переменной отклика, которая формируется при воздействии множества природных факторов и на этом основании построить множественную полиномиальную модель. В этом случае также правомерно будет использовать метод элиминирования для упрощения и подробного анализа модели. Более того, появится возможность вычленить наиболее «весомые» значимые факторы-предикторы. Таким способом формируются выборки при мониторинговых наблюдениях, когда надо следить за важной переменной отклика при вариабельности (временной, пространственной) различных факторов, используемых как факторы-предикторы. Именно так поступают, когда состояние здоровья населения меняется при изменении факторов окружающей среды, и выясняют значимость того или иного фактора. Это иной способ построения выборки данных, для которого применение метода элиминирования параметров считается более обоснованным.

Подведем итоги. Построение математической модели в общем случае сводится к определению вида зависимости и с помощью информации, полученной в результате опыта, – определению параметров модели, т. е. числовых значений коэффициентов аппроксимации.

Если вид зависимости $y = \varphi(x)$ принципиально неизвестен, то при выборе необходимо учитывать, что построенная математическая модель, кроме того, что она точно описывает объект исследования, должна быть при этом наиболее простой для того, чтобы иметь физический смысл коэффициентов, а также быть по возможности единой для схожих явлений, а значит иметь предсказательную способность в задачах прогнозирования.

Данная задача решается во многом благодаря знаниям и личному опыту исследователя. Если установлен вид зависимости (например, сорбционные явления, явления увеличения или уменьшения числа микроорганизмов, накопления веществ при определенных условиях, которые лучше всего описываются логистическими кривыми), если вид зависимости интересующих нас функций и предикторов знаком по литературе, то можно воспользоваться известным видом функции и переходить непосредственно к определению значений параметров аппроксимации, т. е. к операции аппроксимации. Если же вид функции неизвестен, то необходимо построить график зависимости переменной от действующего фактора (эмпирический способ). Это покажет общий вид зависимости (убывающая, с одним максимумом и др.) и поможет выбрать функцию из указанных четырех типов. Также необходимо заранее продумать, какие могут быть ограничения у функции (некоторые свойства не могут принимать отрицательные значения, некоторые – превышать какое-то значение, например, 100 %, зависимости могут обязательно начинаться из точки 0; 0 или проходить через начало координат и пр.). Ещё раз напомним функции, которые помогут вам ускорить их выбор для описания экспериментальных данных в вашем конкретном исследовании (табл. 13).

Таблица 12

Функциональные зависимости, используемые в естествознании

Раздел почвоведения, использование	Функция	Вид	Автор уравнения
Почвоведение Накопление почвенного органического вещества	Показательная	$y(t) = y_0\left(1 - e^{-kt}\right)$	Уравнение Костычева – Иенни по накоплению орг. вещества почв
Почвенная гидрология Основная гидрофизическая характеристика.	Степенная	$S_e = \frac{\theta_i - \theta_r}{\theta_s - \theta_r} = \begin{cases} \left\|\alpha P_{к-с}\right\|^{-n} \text{ для } P_{к-с} \le -\frac{1}{\alpha}, \\ 1 \text{ для } P_{к-с} \ge -\frac{1}{\alpha} \end{cases}$	Brooks and Corey, 1964 г.
	Логистическая	$S_e = \frac{\theta_i - \theta_r}{\theta_s - \theta_r} = \begin{cases} \left(\frac{1}{1+(\alpha P_{к-с})^n}\right)^m \text{ для } P_{к-с} < 0, \\ \theta_s \text{ для } P_{к-с} \ge 0 \end{cases}$	van Genuchten, 1980 г.
Химия почв Адсорбция	Степенная	$Q = mC^n$	Фрейндлих

Окончание табл. 12

Раздел почвоведения, использование	Функция	Вид	Автор уравнения
Химия почв	Логит-функция	$A = A_{\infty}\dfrac{K_{\text{л}}C}{1+K_{\text{л}}C}$	Ленгмюр
Агрохимия Поглощение веществ растениями	То же	$J_r = \dfrac{J_{\max}K_m\text{c}}{1+K_m\text{c}} - \text{c}_{\min}$	Уравнение Михаэлиса-Ментен
Химия, физика почв Кинетика 1-го порядка	Экспоненциальная	$C_i = C_0\exp(-k_1 t)$	–
Кинетика 2-го порядка	Логит-функция	$C_i = \dfrac{C_0}{1+k_2 C_0 t}$	–
Кинетика нулевого порядка	–	$C_i = k_0\text{c}$,	–
Физика почв Зависимость электросопротивления от влажности	Экспоненциальная	$y = b_1\exp(-b_2 x)$, где y – электросопротивление; x – влажность	Поздняков, 2011 г.
Земледелие и агрохимия	Логарифмическая	$\lg(A-y) = \lg A - b(x-\text{c})$	М. А. А. Мичерлих

Из истории вопроса…

Правило Мичерлиха

Макс Айдхард Альфред Мичерлих родился в Берлине 29 августа 1874 года. Образование получил в Кильском университете и Берлинском сельскохозяйственном институте. Физикам почв хорошо известно о связанной воде по Мичерлиху, а также о теплоте сжигания. Этому физическому почвенному параметру была посвящена диссертация Мичерлиха, в которой он предложил использовать его для оценки плодородия. Известен Мичерлих прежде всего по работам математического описания зависимости урожая от дозы удобрений. На основании собственных наблюдений он усомнился в точности закона Либиха, которого придерживались большинство ученых того времени, и доказал, что между дозами удобрений и урожаем существует не прямая пропорциональная, зависимость а логарифмическая. В результате он предложил такую функцию:

$$\lg(A-y)=\lg A-b(x-c),$$

где A – максимальный урожай; y – урожай при внесении того или иного химического вещества в почву в количестве x; b – фактор действия этого вещества; c – его содержание в почве. (Отметьте (!), как Мичерлих точно заметил и предвидел не просто зависимость урожая, а именно прибавку (т. е. разницу максимального и реального) от добавления в почву удобрения (внесенного вещества минус содержание его в почве). Более того, он установил и эмпирические значения коэффициентов действия питательных веществ (того самого параметра b): N – 0.2, P_2O_5 – 0.6, K_2O – 0.4, Mg – 2.0. *В дальнейшем дискуссии кипели именно вокруг этих параметров уравнения Мичерлиха, в частности, константности величины b (параметра действия фактора). Эти параметры в опытах самого М. А. А. Мичерлиха и многих других исследователей колебались в довольно широких пределах. Мичерлих объяснял изменчивость параметров тем, что почва – очень динамичная система, в ней происходят процессы химической и микробиологической трансформаций как питательных веществ, так и самой почвы. Но вот что примечательно: многие ученые сосредоточились на определении параметров уравнения и математическом анализе расчетов Мичерлиха. Но при этом появлялись новые факты и даже теории. Так, великий российский агрохимик Д. Н. Прянишников тоже принял участие в верификации коэффициента действия азотных удобрений. И анализируя результаты, выдвинул и подтвердил гипотезу о равноценности нитратов и солей аммония в питании растений при благоприятном соотношении других факторов. Прекрасное доказательство того, что теоретическая модель при её экспериментальном изучении может быть мощным стимулом для появления и развития новых гипотез и теорий. В этом тоже огромная сила и привлекательность математических моделей в изучении природы (по статье В. И. Папасяна «К 140-летию со дня рождения М. А. А. Мичерлиха». «Агрохимический вестник», № 1, 2015`. С. 38 – 40).*

Нередко при специальных исследованиях используют и некоторые другие (кроме среднеквадратической ошибки) выражения для ошибок (погрешностей) моделирования. Они (для информации) све-

дены в таблице (табл. 14). Мы же в основном будем использовать среднеквадратичную ошибку (RMSE).

Таблица 14

Используемые критерии совпадения рассчитанных по модели и экспериментальных данных (ошибки или погрешности модели)

Ошибка	Обозначение	Расчетная формула
Средняя (mean error)	*ME*	$\frac{1}{N}\sum_{N}(y_{\text{р}} - y_{\text{э}})$
Среднеквадратичная (root mean square error)	*RMSE*	$\sqrt{\frac{\sum_{N}(\varsigma' - \varsigma)^2}{N}}$
Средняя абсолютная (absolute mean error)	*AME*	$\frac{1}{N}\sum_{N}\lvert\varsigma' - \varsigma\rvert$
Относительная (relative mean error)	*RME*	$\frac{1}{N}\sum_{N}\frac{\varsigma' - \varsigma}{\varsigma}$
Несмещенная среднеквадратичная (unbiased root mean square error)	*URMSE*	$\sqrt{\frac{\sum_{N}(\varsigma' - \varsigma - ME)^2}{N}}$

Примечание. N – размер массива проверочных данных (объем выборки); $y_{\text{р}}$, $y_{\text{э}}$ – рассчитанные и экспериментально полученные значения искомой функции отклика; ς' – расчетное, ς – экспериментальное значение отклика соответственно.

Средняя ошибка характеризует среднее расхождение между вычисленными и измеренными данными и служит критерием наличия систематической погрешности в работе модели. Вместо средней ошибки можно рассчитывать среднюю абсолютную ошибку *(AME)*, это позволяет избежать взаимной компенсации систематических ошибок противоположного знака, например, когда модель дает завышенную оценку в одной области значений аргумента и заниженную – в другой. Иногда бывает полезно использовать относительную ошибку *(RME)*. Среднеквадратичная ошибка *(RMSE)* включает как систематическую, так и случайную составляющие. Для того чтобы разделить эти составляющие, рассчитывают несмещенную среднеквадратичную ошибку *(URMSE)*.

3.3. Адекватность нелинейной аппроксимации

Обычно оценка моделей основывается как на визуальном графическом анализе, так и на использовании статистических показателей. Для визуального анализа применяют сравнение измеренных и прогнозируемых данных. Этот вид сравнения дает возможность заметить аномалии в наблюдаемых и прогнозных величинах, различия между ними: насколько удовлетворительно модель описывает искомую величину и есть ли отклонения расчетной величины от реальной, есть ли наличие систематической погрешности. Существует правило Сайерта (Cyert, 1966), который один из первых сформулировал условия приемлемости результатов расчета по качественному анализу динамических расчетных и реальных данных. Эти условия следующие: правило совпадения экстремумов и правило совпадения средних.

Конечно, графические интерпретации субъективны, и поэтому необходимо дополнить такой анализ использованием статистических критериев, которые дают количественную меру соответствия между прогнозными и измеренными величинами.

Адекватность модели (*adequacy of a model*) – соответствие модели моделируемому объекту или процессу. Адекватность – условное понятие, так как полного соответствия модели реальному объекту быть не может, иначе это была бы не модель, а сам объект. (Лопатников, 2003).

Для сравнения моделей и оценки их адекватности в настоящее время существует относительно небольшое число общепринятых критериев. Необходимо отметить несколько важных моментов. Экспериментальный материал должен по возможности захватывать крайние значения изучаемого явления (Пачепский, 1992), т. е. массив данных должен представлять всю область величин искомого свойства, которые могут наблюдаться в естественных условиях. Фактический материал для проверки модели должен быть достаточно разнообразным (влажные с обильными осадками периоды должны быть представлены наряду с засушливыми; воздействие токсиканта на организм должно изучаться при малых и высоких концентрациях).

Немаловажно и то, что применение классических статистических подходов требует прежде всего доказательства нормальности распределения (Дмитриев, 2009). Для этого необходимы большой массив данных и соответствующий выбор критерия «нормальности»

распределения. Чаще всего рекомендуется использовать непараметрические статистические критерии. Однако начинают исследование модели на адекватность с простых качественных и полуколичественных критериев, которые хотя и не могут дать достоверный и весомый ответ об адекватности, но позволяют очень многое сказать о работе модели, возникновении ошибок при моделировании, и как правило дать аргументированный ответ об адекватности модели (Сметник, Спиридонов, Шеин, 2005).

Также надо помнить, что не все критерии способны показать наличие систематических ошибок. Так, обычно исследователь заканчивает свой эксперимент, предлагая ту или иную формулу аппроксимации. Он подсчитывает **коэффициент детерминации R^2**, с помощью которого сравнивает экспериментальные и расчетные данные. Мы уже указывали, что высокий коэффициент R^2 совсем не означает, что аппроксимация прошла удачно, так как он указывает на случайные погрешности и не способен описать погрешности систематические.

Лучшее начало проверки работы модели – это **построение зависимости экспериментальных значений от расчетных**. При качественной аппроксимации получаем биссектрису (рис. 19).

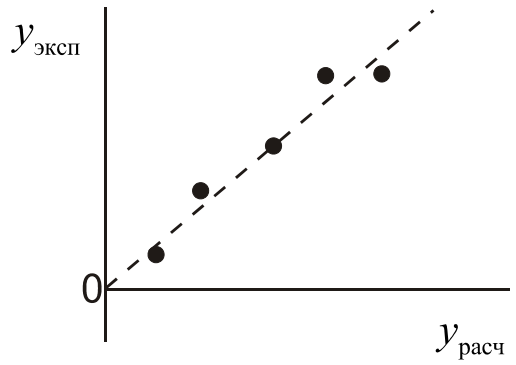
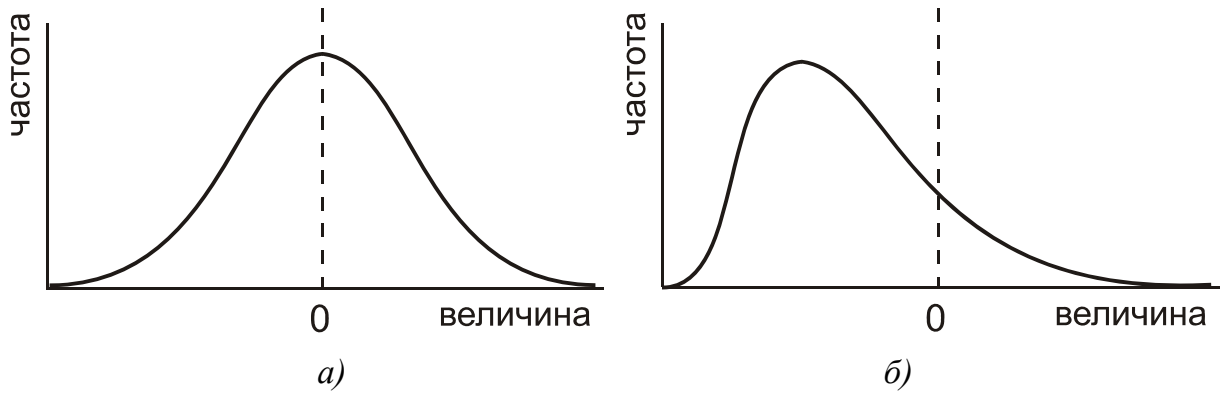


Рис. 19. График зависимости экспериментальных значений от расчетных

Кроме того, построение такого графика может, пусть пока и качественно, указать на наличие систематических ошибок. На их наличие укажет положение точек по одну сторону от биссектрисы или же положение этого графика не вдоль биссектрисы угла, а под некоторым другим (не 45º) углом.

Кроме того, к описательным статистическим критериям относятся **оценка распределения погрешностей** (гистограмма распределения погрешностей) и **характеристика разброса погрешностей** (в виде "*box & whisker plots*" в пакете Statistika 6.0). По гистограмме судят о нормальности распределения погрешностей и о том, что средняя величина погрешности близка к нулю. Если распределение заметно

отличается от нормального, значит есть систематическая погрешность (рис. 20).



а) *б)*

Рис. 20. Графики распределения погрешностей: *а* – распределение близко к нормальному; *б* – систематическая погрешность в отрицательной области

Характеристика разброса погрешностей («*box & whisker plots*» в пакете Statistika 6.0), которая включает медиану, квартильный размах и размах варьирования, дает представление в том, насколько медиана близка к нулю, распределение симметрично относительно медианы (близко к нормальному) и велик разброс ошибок.

На основании **анализа линейной регрессии ошибок от реальной величины функции отклика** удобно проводить оценку систематических ошибок (Пачепский, 1992), если на это указывает качественный анализ.

Возможен и аналитический анализ наличия систематических погрешностей. Для этого надо рассчитать уравнение регрессии ошибок от реальной (экспериментальной) величины $\Delta_{ij} = a_i + b_i y_{ij}$.

Далее можно проанализировать значимость параметров регрессии с использованием *t*-критерия. Напомним, что при получении регрессионного уравнения необходима его проверка с помощью критерия Фишера. Для этого сравнивают рассчитанный критерий с табличным для соответствующего уровня значимости и степени свободы. Если рассчитанное значение для данного уравнения превышает критическое (табличное), то с соответствующей вероятностью можно утверждать, что проверяемая зависимость значима, а это означает, что при аппроксимации присутствуют систематические ошибки. Если есть необходимость дать оценку этим систематическим ошибкам, следует проанализировать на значимость соответствующие регресси-

онные коэффициенты с помощью *t*-критерия. Например, если рассчитанный *t*-критерий для коэффициента a_i оказывается меньше табличного, то с определенной вероятностью можно утверждать что предполагаемая нулевая гипотеза неверна и коэффициент a_i значимо отличается от нуля. Это означает, что ошибки возрастают с ростом реальной (экспериментальной) величины, что тоже характеризует наличие систематических ошибок. Если ситуация с коэффициентом b_i аналогичная, то систематические ошибки присутствуют во всем диапазоне реальных величин переменной отклика. Отметим, что подобный анализ в программах STATISTICA весьма прост и нагляден (см. п. 3.6 «Оценка параметров аппроксимации и процедура элиминирования»).

3.4. Подбор параметров аппроксимации для выбранной функции и процедура сканирования для поиска параметров

Итак, вид функции, которым предполагается аппроксимировать экспериментальные данные, определен, следующим шагом в поиске модели станет подбор численных значений параметров аппроксимации.

Параметр – это числовой коэффициент, или свободный член уравнения, полученный при аппроксимации экспериментальных данных выбранной функцией.

Практически всегда для целей нахождения оптимальных параметров выбранной зависимости используется метод наименьших квадратов, когда сумма квадратов ошибок моделирования (или среднеквадратического отклонения) минимальна:

$$S_r = \sqrt{\frac{1}{N}\sum_{j=1}^{n} n_j \Delta_j^{\,2}} = \min.$$

Математических способов нахождения минимума среднеквадратической ошибки весьма много. Безусловно, мы не можем здесь ознакомиться со всеми, но ведь у нас конкретная задача: определить численные значения параметров аппроксимации или получить конкретный математический вид функции для нашей экспериментальной выборки. Более того, надо подобрать числовые значения параметров аппроксимации выбранной нами функции, дать статистическую оценку

полученному уравнению, параметрам аппроксимации и провести статистический анализ ошибок моделирования (погрешностей моделирования) для выявления их типа (случайные или систематические, нормально ли распределенные и т. д.). Наша проблема на данном этапе аппроксимации распадается на четыре задачи:

1. Определение числовых значений параметров аппроксимации.
2. Статистическая оценка полученного уравнения.
3. Статистическая оценка полученных параметров аппроксимации, их достоверность.
4. Анализ полученных ошибок моделирования (ошибок аппроксимации).

Начнем разбираться с самого простого и наглядного примера аппроксимации данных – линейного уравнения, хотя прекрасно помним, что линейные функции в математическом моделировании природных процессов используются нечасто, более того, крайне редко.

Процедура аппроксимации нелинейных уравнений. Метод сканирования

Выше неоднократно отмечалось, что в почвоведении и других естественных науках линейные уравнения применяются редко, так как в природе процессы как правило нелинейны. Соответственно и модели этих процессов должны быть нелинейными. Разберем, как в случае нелинейных аппроксимаций при использовании всего набора нелинейных функций (табл. 15) будет протекать процесс аппроксимации и как будут выполняться выделенные четыре этапа проверки.

Рассмотрим пример, приведенный в книге Я. А. Пачепского (Пачепский, 1992).

Яков Аронович рассматривает результаты эксперимента по изучению зависимости критической глубины грунтовых вод (y, м) и их минерализацией (x, г/л). Эти результаты приведены в табл. 15.

Таблица 15

Данные критической глубины грунтовых вод (y, м) и их минерализация (x, г/л)

x	0.8	2.0	5.0	10.0	20.0
y	1.1	1.5	2.1	3.0	4.0

Необходимо определить вид зависимости и параметры аппроксимации, подобрать вид аппроксимирующей функции и провести все четыре этапа статистической проверки полученного уравнения взаимосвязи глубины грунтовых вод и их минерализации.

Начинаем как всегда с построения графика зависимости (рис. 21). Безусловно, мы имеем дело с нелинейной зависимостью. Надо выбрать из предложенных выше монотонно возрастающую. Начинаем с функции с наименьшим количеством параметров, например, степенной $\left(\frac{x}{b_2}\right)^{b_1}$.

Однако теперь мы имеем дело с нелинейной функций, и требуется специальный метод аппроксимации, получения параметров аппроксимации. Критерий такого получения параметров нам известен: это нахождение минимума среднеквадратической ошибки $S_r = \sqrt{\frac{1}{N}\sum_{j=1}^{n} n_j \Delta_j^2} = \min$. В данном случае можно представить среднеквадратическую ошибку в следующем виде:

$$S_r = \frac{1}{\sqrt{4}}\left\{\left[1.5-\left(\frac{2}{b_1}\right)^{b_2}\right]^2+\left[2.4-\left(\frac{5}{b_1}\right)^{b_2}\right]^2+\left[3.0-\left(\frac{10}{b_1}\right)^{b_2}\right]^2+\left[4.0-\left(\frac{20}{b_1}\right)^{b_2}\right]^2\right\}^{1/2}.$$

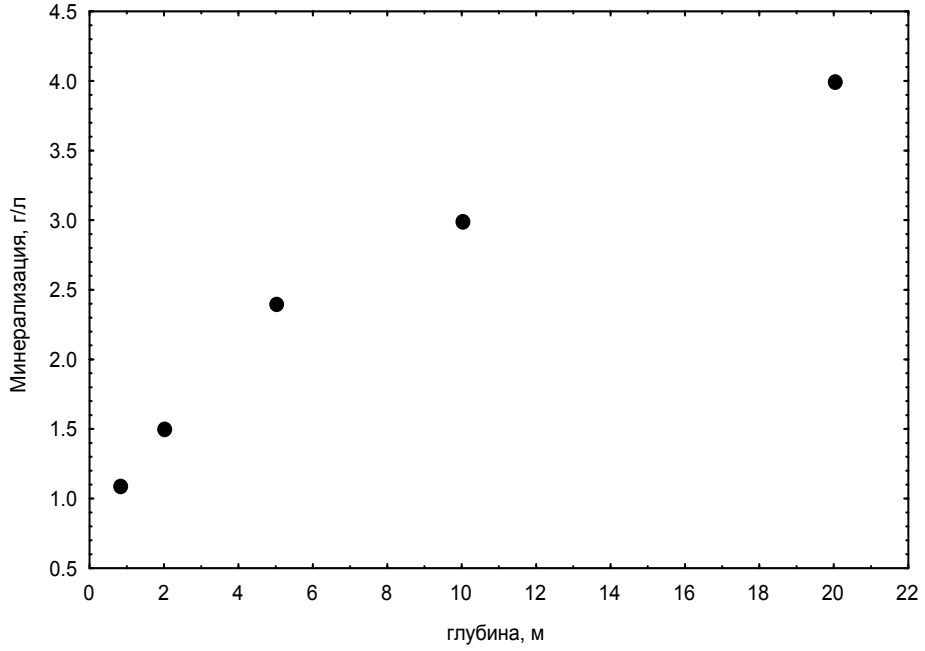


Рис. 21. Экспериментальная зависимость глубины грунтовых вод от их минерализации (Пачепский, 1992)

Решением данного уравнения будут конкретные численные значения для параметров b_1 и b_2. Это уравнение не удастся решить обычным способом: два неизвестных в одном уравнении. Необходимо использовать другие методы. Для нахождения значений параметров в многопараметрических нелинейных уравнениях используют алгоритмы решения вычислительных задач методами высшей математики, которые можно разделить на детерминистические и стохастические. Давайте разберем наиболее показательный детерминистический метод сканирования, или симплекс-метод. Прежде всего представим поле параметров b_1 и b_2 в виде крупной сетки размерами 5 × 5 см (рис. 22). В узлах сетки рассчитывается значение S_r. Начнем, например, со значения $b_2 = 0.2$ и $b_1 = 0.2$ (точка A). Тогда можно рассчитать S_r в этой точке

$$S_r = \frac{1}{\sqrt{4}}\left\{\left[1.5-\left(\frac{2}{0.2}\right)^{0.2}\right]^2+\left[2.4-\left(\frac{5}{0.2}\right)^{0.2}\right]^2+\left[3.0-\left(\frac{10}{0.2}\right)^{0.2}\right]^2+\left[4.0-\left(\frac{20}{0.2}\right)^{0.2}\right]^2\right\}^{1/2}=0.28.$$

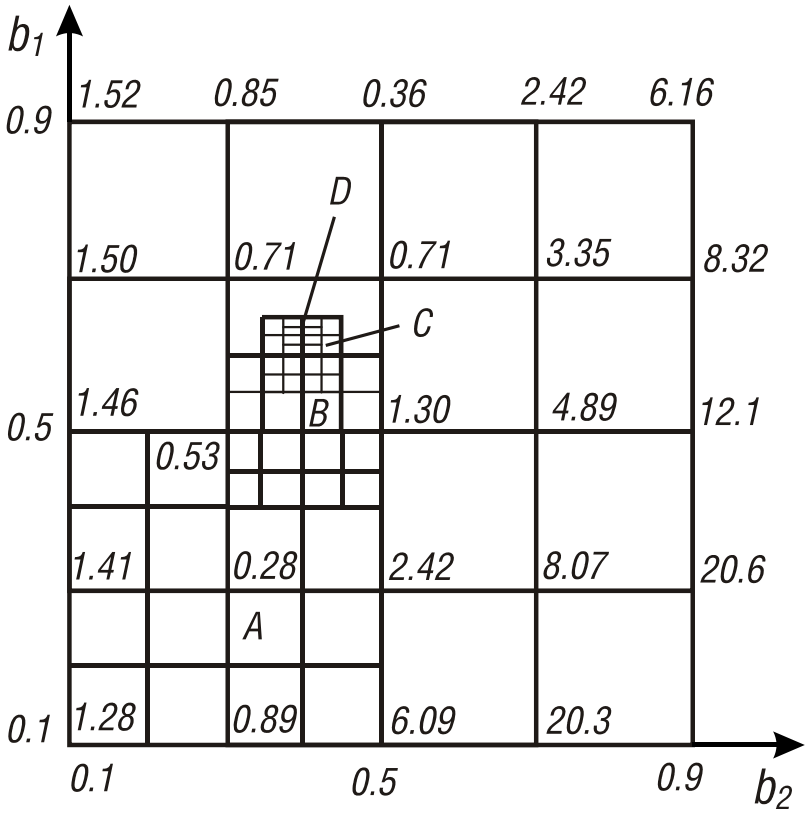


Рис. 22. Пример использования симплекс-метода для подбора параметров b_1 и b_2

Казалось бы, на всем пространстве (поле значений b_1 и b_2) значение 0.28 и является минимальным, так как в других узлах значения

выше, а нам надо найти минимум среднеквадратической ошибки. Однако можно «измельчить» сетку и рассчитать значения параметров в ближайших точках и, наконец, найти такую точку, вокруг которой все значения на рассматриваемом поле параметров будут увеличиваться. Достигнут минимум S_r в рассматриваемом поле параметров. Для точки достигнутого минимума мы и определяем b_1 и b_2. В данном случае $b_1 = 0.628$ и $b_2 = 0.401$. Таким образом, искомое уравнение после операции сканирования приобрело конкретный вид $y = \left(\frac{x}{0.401} \right)^{0.628}$. Заметим, что во время операции сканирования сделали очень важный шаг: мы начали наши расчеты с конкретных значений b_1 и b_2. Мы их задали в виде значений 0.2 и 0.2. Эти значения называются «начальные приближения». Задавать их – большое искусство. Если их задать совсем другими, мы окажемся совсем в ином месте рассматриваемого поля и по мере движения («дробления сетки») можем быть совсем не в генеральном минимуме, а в так называемом «локальном минимуме», что будет серьезной ошибкой. Даже современные расчетные программы в этом случае «зависают» и не находят нужного решения. Значит, нужно научиться подбирать правильные (близкие к реальным) начальные приближения. В расчетных программах обычно «зашиты» некие начальные приближения. Но они могут привести к «локальному минимуму» и сбою в решении. Как же достичь решения в этом случае? Лучше всего самому задать предполагаемые начальные приближения, например, анализируя график. В степенном уравнении параметр b_2 отвечает за угол наклона, а b_1 – за положение кривой («выше-ниже»). Из графика видно, что по оси ординат он пересекает область значений 0.5 – 1.0. Соответственно начальное значение b_1 нужно взять из этой области и также из графика определить примерный угол наклона. Если взять эти значения как начальное приближение, аппроксимация пройдет более успешно и мы получим более реальные ее параметры. В данном случае надо использовать свое умение «читать» графики и (примерно, весьма примерно) определять параметры уравнения. Тем более, что выше мы уже вспомнили для некоторых функций физическое значение входящих параметров. Это необходимо помнить при использовании аппроксимации в современных математических пакетах: далеко не всегда предложенные («зашитые») в пакетах начальные приближения приводят к успешному

решению; возможны сбои, и чтобы их не было, примените другие начальные приближения. Опора для их выбора – использование *уже известных литературных данных*, приближенное нахождение параметров из графиков и, наконец, Ваш личный опыт.

Теперь для приведенного примера решим четыре статистические задачи (см. п. 3.3) по оценке достоверности самого уравнения, его параметров и анализа полученных ошибок моделирования.

Прежде всего, значение *F*-критерия весьма велико (больше 2500) и достоверно при уровне значимости 0.000014. Конечно же, нулевая гипотеза неверна, и мы с высокой степенью достоверности можем использовать полученное уравнение.

Весьма достоверны и параметры аппроксимации (табл. 16). Результаты, приведенные в таблице, определенно указывают на то, что нулевая гипотеза отвергается в пользу альтернативной с уровнем значимости 0.0034 и 0.00025 соответственно, что явно ниже используемого в науках о Земле уровня значимости ($\alpha = 0.05$).

Таблица 16

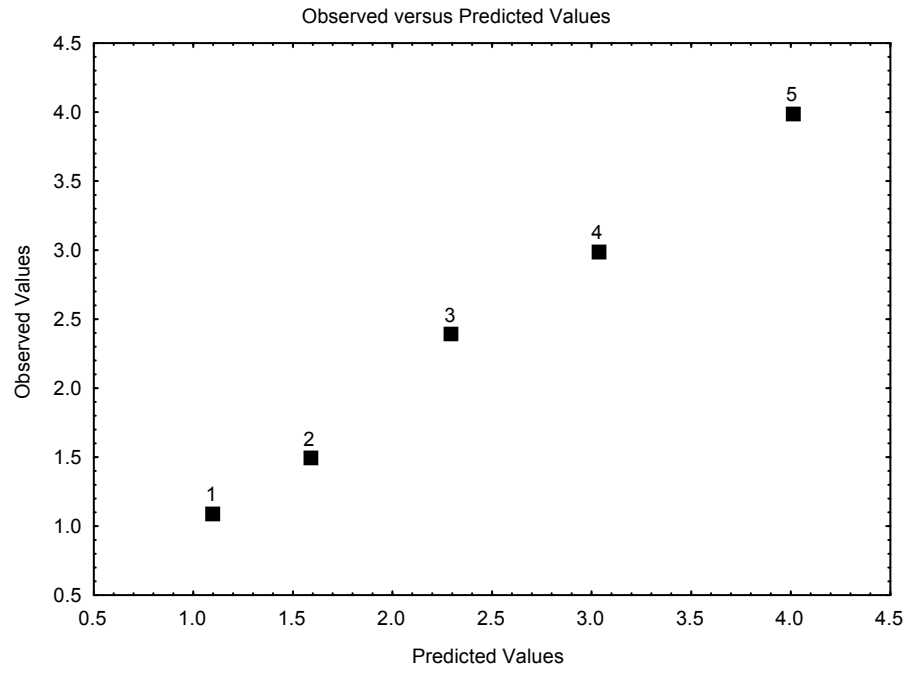
Параметры аппроксимации

Параметр	Оценка	Стандартное отклонение	Значение *t*-критерия	Уровень значимости
b_2	0.640359	0.075407	8.49207	0.003429
b_1	0.403406	0.016453	24.51896	0.000149

Последний шаг, который необходимо сделать для проверки модели, – проанализировать ошибки аппроксимации и ответить на вопрос: имеются ли систематические ошибки? Для этого нужно построить графики и проанализировать две зависимости: *зависимости реальной переменной отклика от расчетной и ошибок аппроксимации от расчетной величины y*. Ниже приведены графики этих зависимостей (рис. 23). Они указывают на достаточно хорошее совпадение расчетных и реальных величин и на то, что ошибки не имеют систематического характера, не зависят от функции-отклика, хотя и весьма велики. Действительно, график зависимости расчетной величины от реальной близок к линейному и представляет собой биссектрису угла начала координат, что говорит об отсутствии систематических ошибок. Второй график показывает, что ошибки не имеют систематиче-

ского характера и не зависят от величины отклика *y*. Но графики определенно демонстрируют, что для уверенных утверждений необходимо получить еще немалое число экспериментальных данных.

а)



б)

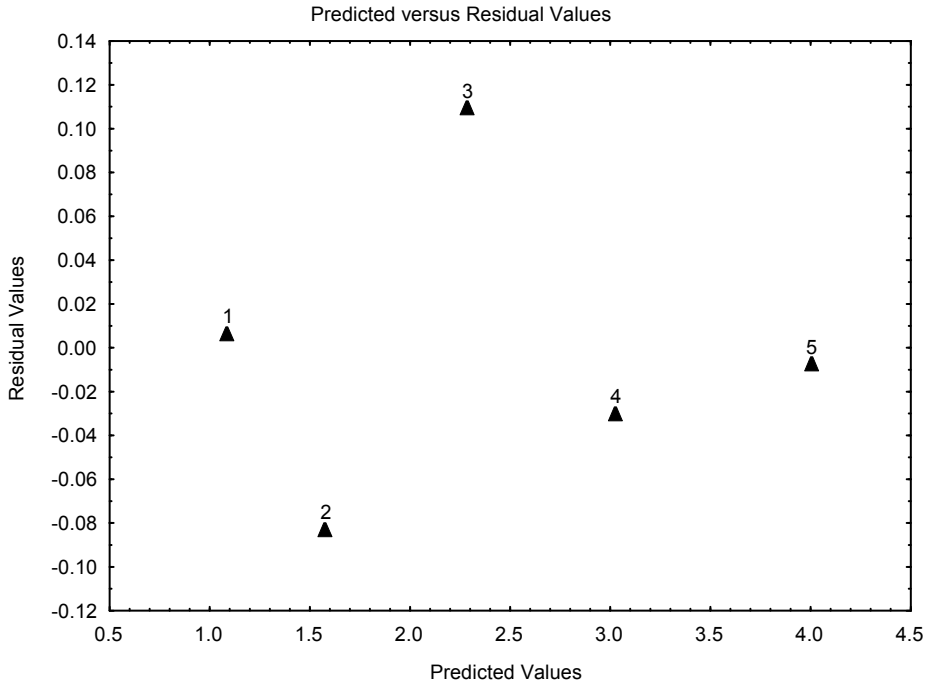


Рис. 23. Графики зависимости: *а* – реальных величин от расчетных; *б* – погрешностей от расчетной величины глубины грунтовых вод

При подборе параметров функции более сложного вида, чем степенная, расчеты будут еще больше усложняться. Для нахождения значений параметров в таких уравнениях были адаптированы алгоритмы вычислительных операций, которые по сути остаются теми же. Так что кратко описанная процедура аппроксимации и последующего анализа ее параметров остается той же и будет включать указанные этапы, независимо от вида функции.

Для того чтобы верифицировать процедуры аппроксимации реальных данных нелинейной функцией, нахождения и статистическую оценку параметров аппроксимации, можно оценить возможности трансформации нелинейных уравнений для лучшего подбора аппроксимирующей функции. Рассмотрим получение уравнения методом нелинейной регрессии поэтапно на конкретном примере.

Пример 6. Исследуем распад агрегатов в воде, анализируя во времени количество оставшихся агрегатов в воде (по методу Андрианова). Первоначально мы взяли 30 агрегатов и каждые 30 с вели счет оставшимся в течение 5 мин. Получили следующие данные (табл. 17 и рис. 24).

Таблица17

Данные распада агрегатов в стоячей воде (метод Андрианова)

№ п/п	Агрегаты водоустойчивые, шт.	Время, с
1	30	0
2	22	30
3	17	60
4	14	90
5	12	120
6	10	150
7	9	180
8	9	210
9	8	240
10	8	270
11	8	300

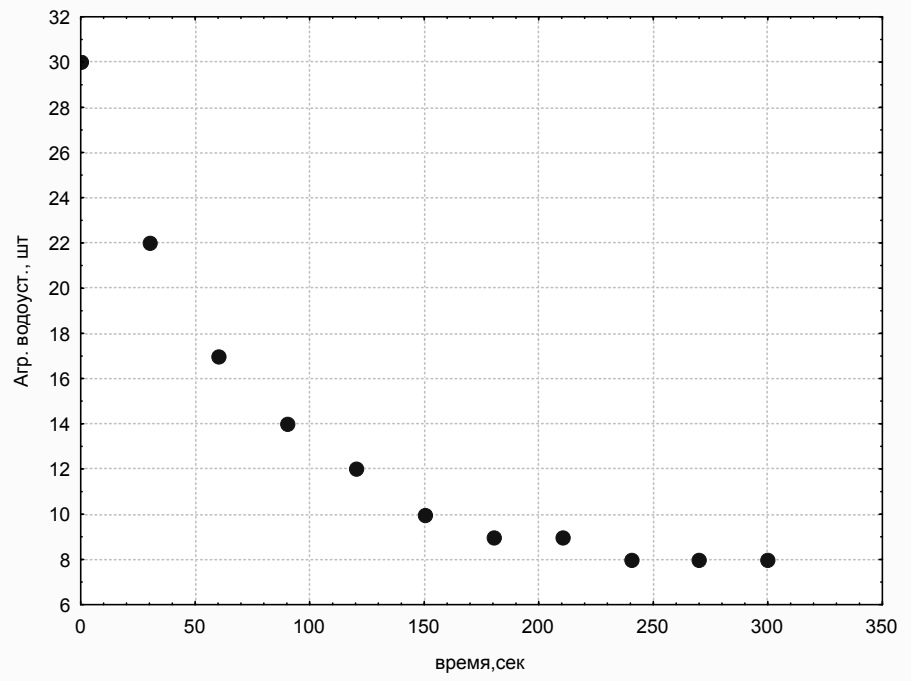


Рис. 24. Распад водоустойчивых агрегатов во времени (количество оставшихся агрегатов)

Требуется подобрать вид зависимости распада агрегатов во времени и получить количественные параметры аппроксимации.

Решение

Из графика зависимости водоустойчивых (оставшихся к определенному времени) агрегатов от времени видно, что эта зависимость нелинейная, убывающая, причем убывающая быстро, скорее всего, экспоненциально. Поэтому для начала имеет смысл применить экспоненциальную функцию вида $y = b_1 \exp(-b_2 x)$, где y – количество водоустойчивых агрегатов; x – время, с.

Использование программы Statistica

Первый шаг – открываем программу Statistica, далее выбираем меню «Статистика», затем команду «Дополнительные Линейные/Нелинейные модели», далее «Нелинейная оценка» и вид уравнения (экспоненциальный убывающий), попадаем в диалоговое окно, где выбираем предпочтительный метод (можно рекомендовать метод Марквардта). Пытаемся провести расчет, но программа «зависает», не считает. Смотрим данное пособие и понимаем, что дело в начальных приближениях. Идем в раздел «Start Values». Машина автоматически предлагает для параметров b_1 и b_2 величины 0.1. Теперь понятно: для величины b_1, которая по графику и смыслу должна быть близка к 50 (коли-

чество нераспавшихся агрегатов в начале опыта), в расчетах принимается 0.1. Исправляем начальные значения. Ставим в «Start Values» для параметра b_1 значение 50, для b_2 оставляем предыдущее значение (0.1).

Получаем аппроксимацию, которая дает относительно хорошее (визуальное) совпадение расчетной кривой с реальными данными.

Результаты представлены в табл. 18 и на рис 25.

Таблица 18

Результаты аппроксимации данных

Model is: $v_1 = b_1$*exp(-b_2*v_2) Dep. Var. Агр. водоуст., шт. Level of confidence: 95.0% (alpha = 0.050)						
	Estimate	Standard	*t*-value	*p*-level	Lo. Conf	Up. Conf
b_1	26.85538	1.585394	16.93925	0.000000	23.26897	30.44179
b_2	0.00572	0.000614	9.31483	0.000006	0.00433	0.00711

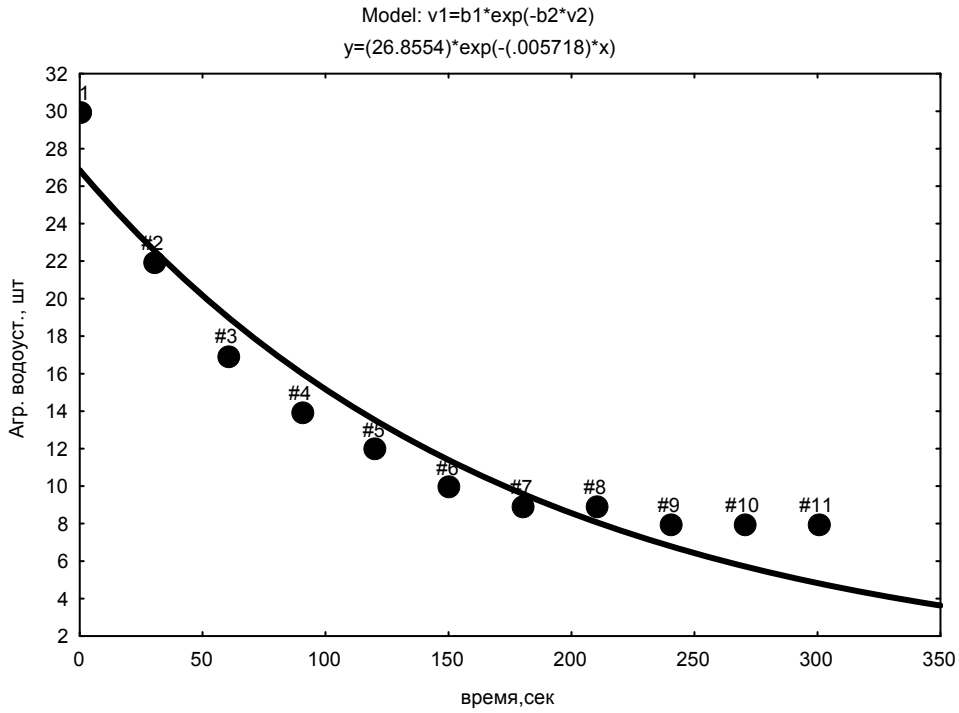
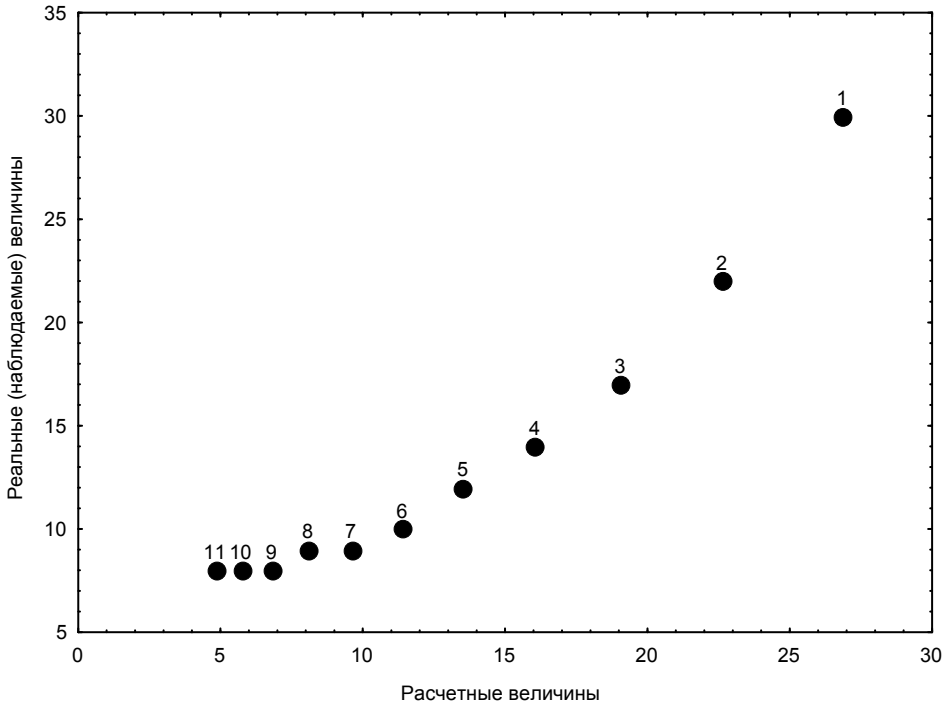


Рис. 25. График экспоненциальной зависимости числа распавшихся агрегатов от времени их нахождения в стоячей воде

Далее мы обязаны проанализировать расчеты на наличие систематических ошибок, хотя бы визуально, т. е. графически. Строим зависимость расчетных величин от реальных (должна быть линия,

близкая к биссектрисе) и зависимость ошибок регрессии от расчетной величины (рис 26).

а)



б)

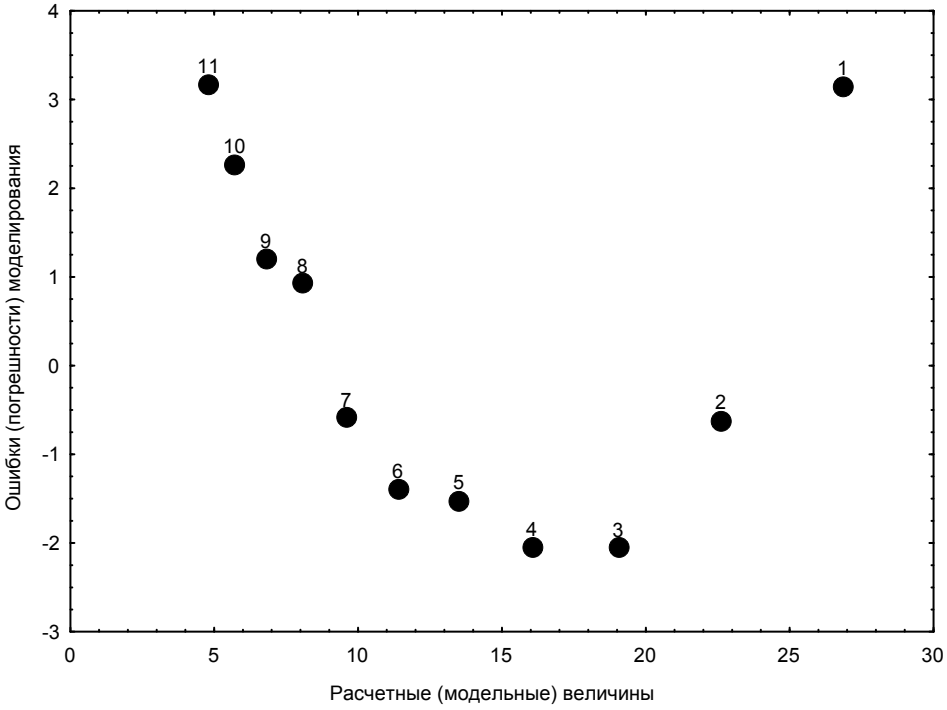


Рис. 26. Зависимости реальных величин от расчетных *(а)* и распределение погрешностей от расчетной величины модели распада агрегатов в воде *(б)*

Графики (см. рис. 26) показывают, что рассчитанные и реальные величины лежат не строго вдоль биссектрисы угла. Зависимость ошибок от расчетной величины носит весьма сложный характер: ошибка велика при малых и больших значениях расчетной величины, т. е. ошибки максимальны в начальный период времени и в конечный. Значит наша зависимость дает существенные погрешности для описания процесса распада агрегатов во времени для слабоустойчивых агрегатов и весьма устойчивых, которые остаются к концу эксперимента.

Как поступить? С одной стороны, полученное уравнение аппроксимации значимое, параметры аппроксимации тоже значимы. Но систематические ошибки для описания процесса распада агрегатов в воде весьма значительные в начале эксперимента и в конце, что, безусловно, вызывает сомнения в возможности использования выбранного экспоненциального уравнения.

Если мы хотим добиться лучшего совпадения, необходимо проанализировать подробнее расхождения, возникшие при аппроксимации. На следующем этапе мы можем: 1) попытаться изменить вид функции (с экспоненциальной на показательную или логарифмическую); 2) увеличить количество параметров (сделать функцию трехпараметрической) и попытаться произвести трансформацию аргумента или функции, т. е. для начала прибавить третий параметр.

Из первого предлагаемого шага ничего не выйдет: показательная функция не даст лучших результатов (можете проверить). А вот второй предлагаемый вариант вполне возможен. Действительно, на начальном и конечном этапах динамики распада агрегатов в воде возникают ошибки превышения расчетных данных. Возможно, совпадение расчетных и реальных данных улучшится, если мы введем дополнительный параметр – свободный член уравнения b_3. Предлагается испробовать такое уравнение: $y = b_1\exp(-b_2(x)) - b_3$. Как видите, добавлен третий параметр, b_3, который должен улучшить совпадение и аппроксимацию. Из рис. 27 видно, что аппроксимация прошла значительно лучше. Практически во всем диапазоне данных достигнуто хорошее совпадение, и ясно, что второе уравнение лучше подходит для описания процесса

распада агрегатов. Но все-таки необходимо построить графики зависимости реальных данных от расчетных и зависимости погрешностей аппроксимации от расчетной величины для того, чтобы убедиться, что явные систематические ошибки отсутствуют. Эти графики приведены на рис. 28. Графики показывают, что систематические ошибки уменьшились и зависимость погрешностей аппроксимации от расчетной величины отсутствует. Следовательно, нет необходимости применять критерий Вильямса – Клюта, чтобы выяснить, какая же функция лучше (см. п. 3.5). Ясно, что трехпараметрическая.

Добавлением одного параметра было достигнуто вполне приемлемое решение для описания процесса динамики распада агрегатов в стоячей воде, процесса водоустойчивости агрегатов.

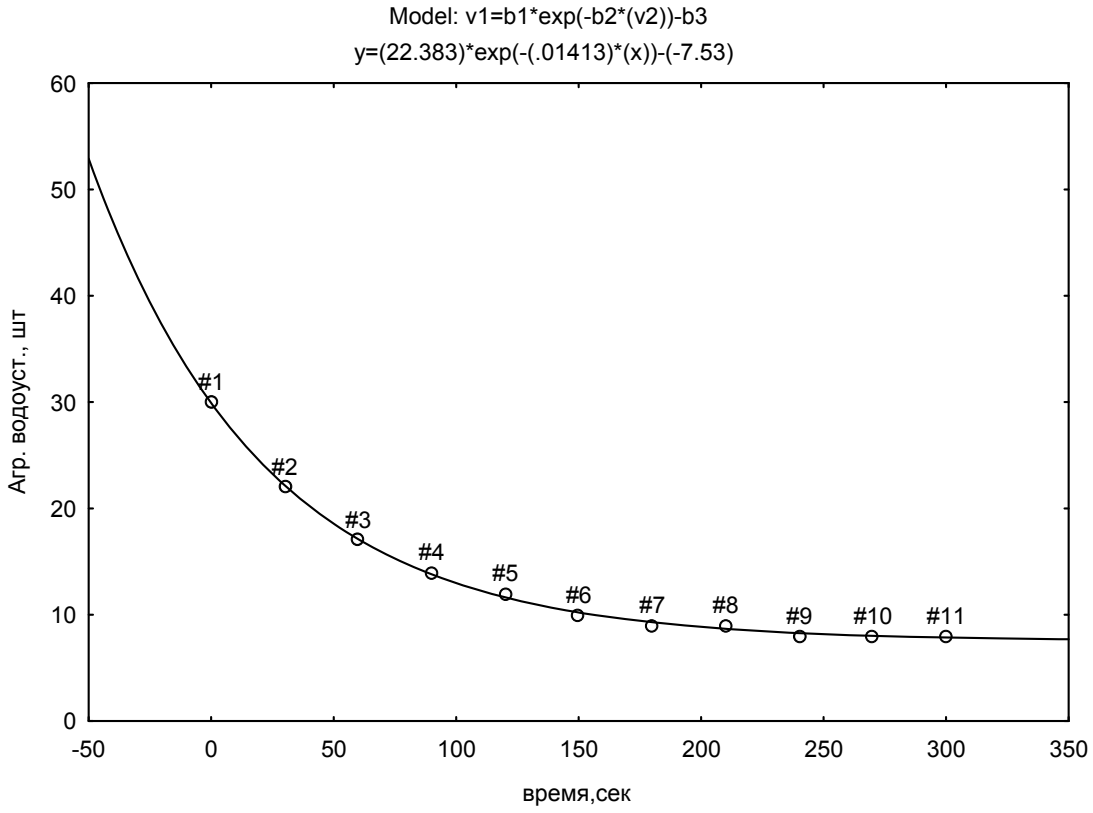
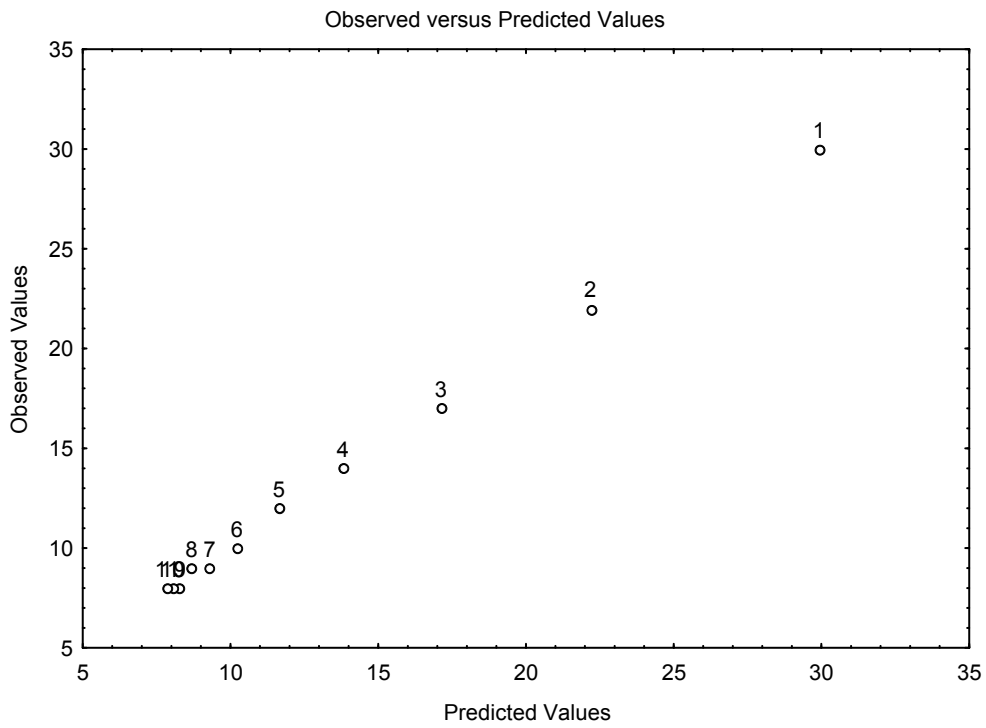


Рис. 27. График экспоненциальной трансформированной зависимости числа распавшихся агрегатов от времени их нахождения в стоячей воде

а)



б)

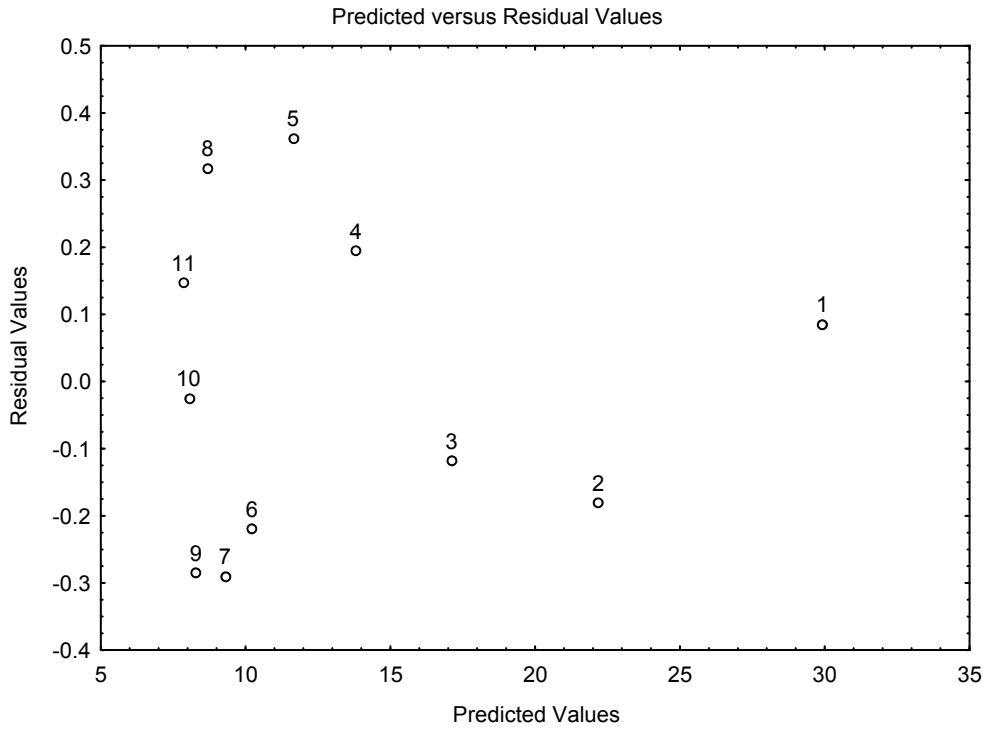


Рис. 28. Зависимости реальных величин от расчетных *(а)* и распределение погрешности от расчетной величины для трансформированной модели распада агрегатов в воде *(б)*

3.5. Анализ различия моделей и выбор лучшей. Непараметрический критерий Вильямса – Клюта

Для того чтобы ответить на вопрос о большей адекватности той или иной модели, часто используется критерий, предложенный Вильямсом и Клютом. Отметим, что этот критерий является непараметрическим, т. е. не требует строгого соблюдения закона нормального распределения для исходных данных и пр. То есть этот критерий достаточно демократичен, и его применение возможно при небольшом (не столь большом, как в классической статистике) количестве данных, но и надежность выводов, полученная с помощью непараметрических критериев, несколько слабее по сравнению с классическими параметрическими. Итак, остановимся на следующей проблеме: у нас есть массив экспериментальных данных «отклик – предиктор», для этого массива мы можем подобрать несколько (скажем, две) нелинейные функции. Далее стоит вопрос: различаются ли описания экспериментальных данных с помощью этих функций или они равнозначны (не имеет значение, какую использовать)? И если они различаются, то какая функция лучше описывает экспериментальные данные?

Вполне понятно, что в зависимости от цели исследования, требований точности и надежности моделирования можно использовать тот или иной набор критериев оценки качества моделирования. Проблема должна решаться исходя из целей моделирования и требований, предъявляемых к расчетным данным. Для познания реальный объект бесконечен.

Итак, алгоритм расчета критерия Вильямса – Клюта для определения лучшей модели:

1. Найти абсолютные погрешности моделирования по модели № 1 (уравнение аппроксимации 1) $\Delta_{\text{абс}_1} = y_{\text{эксп}} - y_{\text{расч}}$ и по модели № 2 (уравнение аппроксимации 2) $\Delta_{\text{абс}_2} = y_{\text{эксп}} - y_{\text{расч}}$.
2. Для каждой пары ошибок первой и второй модели для соответствующей экспериментальной точки в программе Excel найти их полусумму и полуразность.
3. В пакете STATISTIKA 6.0 построить регрессионную зависимость вида $V = kU$, где V – полусумма; U – полуразность; k – коэффициент регрессии.

4. Оценить достоверность коэффициента регрессии k по t-критерию.
5. Если k достоверен, то модели достоверно различаются.
6. Далее обратить внимание на знак t-критерия. Если $t > 0$, то первая модель лучше, если же $t < 0$, то вторая.
7. Сделать вывод о том, какая модель более качественно аппроксимирует экспериментальные данные.

Пример 7

Давайте вспомним экспериментальные данные из п. 3.4, представляющие собой зависимость глубины грунтовых вод от их минерализации. Они представлены в таблице ниже. Тогда для того чтобы разобраться в процедуре аппроксимации данных нелинейной функцией, мы использовали степенную функцию и получили соответствующую аппроксимацию и достоверное уравнение. Зададимся вопросом: а возможно экспоненциальное уравнение лучше? Надо выбрать лучшее уравнение для описания экспериментальных данных.

Решение

Степенная модель $y_1 = \left(\frac{x}{b_1}\right)^{b_2}$, а экспоненциальная – $y_2 = b_1 \exp(b_2 x)$.

Используя непараметрический критерий Вильямса – Клюта, попытаемся выбрать лучшую функцию для описания приведенных экспериментальных данных. Алгоритм расчета представлен выше, приведем его в табличной форме, обозначив символом (1) использование степенного уравнения, а символом (2) – экспоненциального.

Это исходные экспериментальные данные.

x	0.8	2.0	5.0	10.0	20.0
y	1.1	1.5	2.1	3.0	4.0

Расчетные величины с соответствующими погрешностями, получающимися при использовании указанных зависимостей, и линейное регрессионное уравнение $V = k \cdot U$ представлены в табл. 19.

Таблица 19

Расчет линейной регрессии между степенной и экспоненциальной моделью (*V* – полусумма, *U* – разность ошибок моделей)

X	*Y*	Δ_1	Δ_2	*V*	*U*	*k*	*t*-value для *k*	*p*-level для *k*
2	1.5	–0.08	–0.431	–0.25560	–0.3512	0.6210	4.250	0.02385
5	2.1	0.1112	0.2100	0.160635	0.09885			
10	3	–0.0294	0.2995	0.13506	0.3289	ВЫВОД: 1. Модели достоверно различаются, так как *k* достоверно. 2. Степенная функция лучше, так как *t*-value положителен		
20	4	–0.0097	–0.106	–0.05797	–0.0965			

Таким образом, с помощью критерия Вильямса – Клюта мы доказали, что степенная и экспоненциальная модели для описания наших данных различаются и степенная модель лучше, так как *t*-значение коэффициента регрессии положительно. Поэтому степенную модель и надо использовать в дальнейших расчетах.

Еще один пример из области физики почв.

Исследовали сопротивление расклиниванию (т.е. внедрение конуса в почвенные агрегаты) как характеристики механической устойчивости агрегатов от влажности. Почвенные агрегаты увлажняли, потом медленно подсушивали. В процессе иссушения агрегатов определяли сопротивление расклиниванию. Получили следующую зависимость сопротивления расклиниванию (*Pm*, кг/см^2) от влажности почвы (*W*, %) (рис. 29).

Как видно из рис. 29, сопротивление расклиниванию агрегатов нелинейно возрастает с уменьшением влажности. Как правило, такой тип нелинейности выражают либо экспоненциальной зависимостью, либо степенной, т. е. можно использовать уравнения типа

$$Pm = b_1^{'} \exp(-b_2^{'} W), \text{ или } Pm = \left(\frac{W}{b_2^{''}} \right)^{-b_1^{''}}.$$

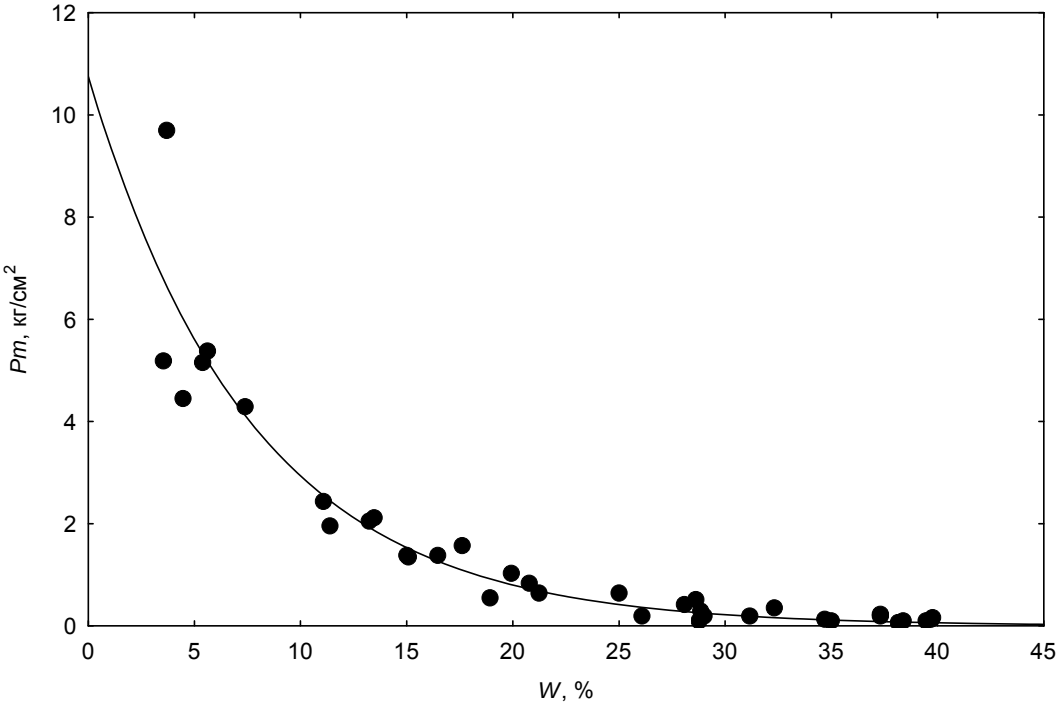


Рис. 29. Зависимость сопротивления расклиниванию (*Pm*, кг/см^2) от влажности (*W*, %) для агрегатов 3 – 5 мм чернозема обыкновенного (пашня). Кривая – аппроксимация уравнением вида $Pm = b_1^{'} \exp(-b_2^{'} W)$

Оба эти уравнения достаточно хорошо описывают сопротивление агрегатов расклиниванию при повышении влажности.

Сравнение моделей по методу Вильямса – Клюта показало, что уравнение экспоненциального типа $Pm = b_1^{'} \exp(-b_2^{'} W)$ лучше аппроксимирует экспериментальные данные (см. алгоритм расчета выше).

3.6. Оценка параметров аппроксимации и процедура элиминирования

С повсеместным распространением различных статистических пакетов рассчитать количественные значения коэффициентов в уравнении аппроксимации не составляет труда. И после нахождения численного значения необходимо этот параметр оценить. Напомним, что самой распространенной статистикой для оценки значений параметров аппроксимации является прежде всего *t*-статистика (см. п. 1.4), или критерий Стьюдента, который рассчитывается по формуле

(Дмитриев, 2009) $t = \dfrac{|b|}{S_b}$, где b – найденное значение параметра; S_b – стандартное отклонение параметра.

Рассчитанное значение *t*-критерия сравнивается с его табличным значением (при определенной вероятности и числе степеней свободы). Если рассчитанный критерий Стьюдента выше, чем табличный, то параметр значимо отличен от нуля с заданной вероятностью.

В зависимости от ситуации нередко требуется либо оценить достоверность различия двух выборок, описываемых одной и той же функцией, либо выбрать наиболее подходящую полученным результатам функцию. Рассмотрим первый случай на конкретном примере. Например, мы исследовали зависимость биомассы двух видов бактерий от температуры, т. е. изучали их термофильность. Получили соответствующие точки, которые могут быть представлены на графике (рис. 30).

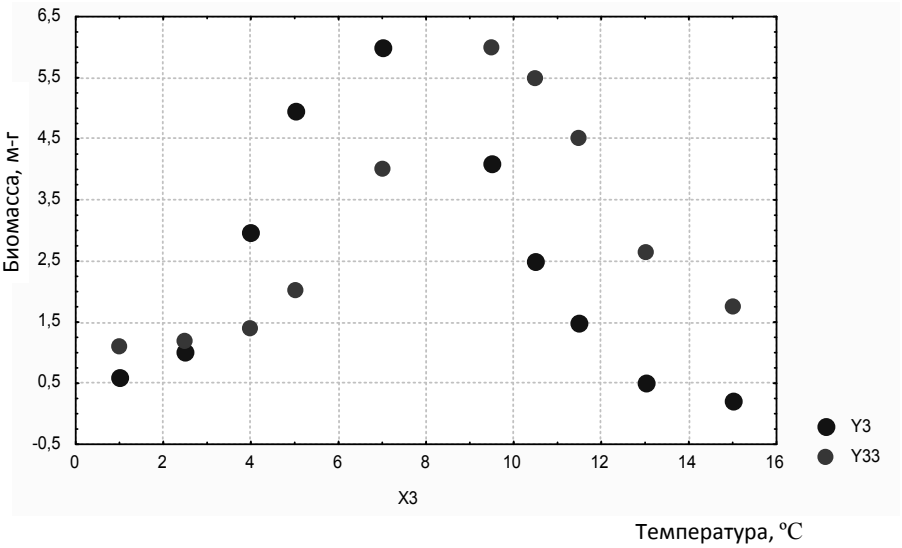


Рис. 30. Зависимость биомассы двух видов бактерий от температуры

Здесь явно экспериментальные точки зависимости биомассы от температуры для двух видов бактерий описываются функцией с одним максимумом, очевидно, гауссиадой:

$$y = b_1 \exp\left[-\left(\frac{(x - b_2)^2}{b_3} \right) \right].$$

Стоит вопрос: достоверно ли отличаются исследованные виды бактерий по термофильности?

На языке математического моделирования исследуемого процесса этот вопрос переформулировался бы так: достоверны ли параметры аппроксимации двух выборок, имеющих один вид и аппроксимируемых одной функцией? Вот его-то и решим. Мы проведем аппроксимацию указанных двух экспериментальных выборок гауссовой функцией, получим параметры аппроксимации, докажем их значимость. Теперь надо рассмотреть, достоверно ли различаются параметры, полученные при аппроксимации экспериментальных данных функцией одного вида. Зная параметры аппроксимации и их статистику (в частности, среднеквадратические ошибки S_b), по полученным параметрам можно сравнить исследованные объекты статистически. Для соответствующих параметров аппроксимации $b_n^{'}$ и $b_n^{''}$ разных выборок можно рассчитать *t*-критерий по следующей формуле:

$$t = \frac{\left| b_n^{'} - b_n^{''} \right|}{\sqrt{(S_{b_n^{'}})^2 + (S_{b_n^{''}})^2}},$$

где $S_{b_n^{'}}$ и $S_{b_n^{''}}$ – стандартные отклонения параметров $b_n^{'}$ и $b_n^{''}$.

Соответственно если *t*-критерий оказывается больше табличного для данной степени свободы и уровня значимости (традиционно 0.05), то параметры двух выборок значимо отличаются друг от друга. В этом случае можно говорить о достоверности различий соответствующих характеристик процесса. Например, если различаются параметры b_2 уравнения, описывающего зависимость биомассы бактерий от температуры, то можно утверждать, что исследованные два вида бактерий достоверно различаются по биологическому оптимуму, т. е. по оптимальным температурам своего развития, а если достоверно различаются, например, параметры b_3, то два вида различаются по толерантности (или по экологической амплитуде).

Такого рода задачи по доказательству различия двух функций постоянно возникают в почвоведении и экологии. Широко известно использование уравнений Фрейндлиха и Ленгмюра для описания процессов равновесной сорбции, параметры которых также активно применяются при моделировании процессов переноса агрохимикатов и различных токсикантов (см. табл. 13). Весьма интересными пред-

ставляются попытки описания кривых сорбции-десорбции паров воды почвами с помощью нелинейных функций (см. п. 3.2. «Элиминирование параметров аппроксимации»), процессов водоустойчивости агрегатов (см. п. 3.5 «Анализ различия моделей и выбор лучшей»), некоторых физико-механических свойств почв и почвенных агрегатов в зависимости от влажности. Кроме того, большинство параметров указанных функций имеют физическое обоснование. Случай использования функции Гаусса уже разобран.

Возьмем теперь пример из химии. В степенном уравнении Фрейндлиха, используемом для характеристики процессов сорбции, несмотря на его эмпирический характер, степенной параметр (показатель степени) можно рассматривать как показатель неоднородности сорбционных центров – он приближается к нулю по мере возрастания неоднородности и стремится к 1 при увеличении их однородности. Поэтому анализ параметров, их возможное сравнение для различных объектов могут дать большое количество информации об объекте и его функционировании.

Вторая проблема, возникающая при практическом использовании процедуры аппроксимации, – это выбор наиболее подходящей функции для описания одних и тех же экспериментальных данных. Эта проблема также весьма распространена в науке при количественном описании данных. Так, например, для описания основной гидрофизической характеристики (ОГХ) в почвенной гидрофизике используют более 20 типов математических моделей, которые имеют свои преимущества и недостатки. Это связано с разнообразием как ОГХ для разных почвенных объектов, так и физически обоснованных моделей, в которых используется аппроксимация ОГХ. Химики используют либо уравнение Фрейндлиха, либо Ленгмюра для описания процессов сорбции. Но какое из них лучше? Разберем на примере.

К вопросу о …

Что такое Монте-Карло у ядерных физиков?

В эпоху массовых открытий элементарных частиц, в 50 – 70-е годы прошлого столетия, физики в Дубне изучали появление новых частиц по их следам в различных трековых камерах, разглядывая пленки, смотря в микроскопы. Это время прошло. Современные ядер-

ные физики, исследуя фундаментальные основы материального мира, работают уже на совсем другом уровне. Прежде всего на другом уровне используемых энергий и на таких колоссальных ускорителях, когда происходит рождение новых частиц, их появляется очень большое количество. Наблюдать их прямыми физическими методами не удается, просто невозможно при столь огромном их количестве. И тогда физики-ядерщики используют так называемый «метод Монте-Карло». Они используют так называемую стандартную модель, которая должна описывать рождение частиц при определенных уровнях энергии взаимодействия. Проигрывают на модели то или иное событие. Получают в результате моделирования ответ, где и сколько частиц должно появиться, как они разлетаются и каковы особенности их поведения. Затем сравнивают то, что получили в эксперименте на ускорителе, с тем, что они насчитали. Если результаты сходятся, значит, многое правильно в теории образования частиц, они получили еще одно подтверждение на существование. Например, в результате эксперимента получилось 20 частиц, а по расчетам – другое количество, например, 17. По-видимому, модель не учитывает процесса образования некоторого количества частиц, в нее должен быть введен дополнительный процесс, при котором образуются еще три частицы. Физически обоснованная модель усовершенствуется, а мы начинаем больше понимать в физических процессах, которые исследуем. Но вот почему у физиков-ядерщиков этот метод получил жаргонное название «Монте-Карло», трудно представить. В моделировании почвенных процессов он называется методом адаптации и оптимизации. Но в конечном счете этот процесс сравнения экспериментальных и расчетных данных позволяет расширить наше понимание природы, дать начало новым знаниям.

ЗАКЛЮЧЕНИЕ

В предлагаемом издании изложены основные теоретические положения математической статистики. Регрессионный анализ является одним из наиболее распространенных в естественных науках. Его грамотное применение и оценка полученных с его помощью результатов дают широкие возможности для познания процессов, определяющих развитие почв и почвенного покрова, различных почвенных свойств и их взаимосвязей.

Учебное пособие написано авторами на основе ежегодно читаемых ими на протяжении последних десяти лет курсов лекций по применению математических методов для почвоведов. Пособие не имеет аналога в литературе. В нем систематизированы современные представления о почвах, их свойствах и возможностях математического анализа взаимосвязи процессов и свойств почв как в пространстве, так и во времени.

Пособие состоит из трех глав. В первой главе рассматриваются понятия линейного регрессионного анализа (переменная отклика и фактор) и использование их в различных моделях, которые распространены в почвоведении.

Во второй главе представлены положения о множественной регрессии и особенностях математических процедур при ее использовании.

В третьей главе предложены методы нелинейной регрессии, которые весьма актуальны для современного почвоведения, прежде всего с развитием прогнозных математических моделей.

Издание адаптировано к современным образовательным технологиям.

Авторы надеятся, что изучение представленного в учебном пособии материала поможет студентам лучше понять теоретический курс дисциплины «Статистические методы исследования в почвоведении».

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Cyert, R. M.* A description and evaluation of some firm simulations / R. M. Cyert // Proc. of the IBM Computing symposium on simulation models and gaming. – N.Y., 1966.

2. *Gottesbüren B.* Comparison of pesticide leaching models: results using the Weiherbach dataset. Agricultural Water Management/ B. Gottesbüren [et al.], 44:153-181.

3. StatSoft, Inc. (2001). Электронный учебник по статистике [Электронный ресурс]. – URL: http://www.statsoft.ru/home/textbook (дата обращения: 12.05.2015).

4. Большая российская энциклопедия : в 30 т. – М. : Большая Рос. энцикл., 2014. – Т. 25. – 764 с.

5. Большая советская энциклопедия : в 30 т. – М. : Сов. энцикл. : 1982. – Т. 25. – 624 с.

6. *Данилов, Ю. А.* Нелинейность / Ю. А. Данилов // Знание – сила. – 1982. – № 11. – С. 34 – 36 [Электронный ресурс]. – URL: http://spkurdyumov.narod.ru/Znakomstvo.htm (дата обращения: 03.07. 2015).

7. *Евдокимов, Е. В.* Динамика популяций в задачах и решениях : учеб. пособие / Е. В. Евдокимов. – Томск : Изд-во Том. гос. ун-та, 2001. – 72 с. [Электронный ресурс]. – URL: http://www.masters. donntu.edu.ua/ 2009/fizmet/kucherenko/library/4.htm (дата обращения: 21.08.2015).

8. *Дмитриев, Е. А.* Математическая статистика в почвоведении. 3-е изд., испр. и доп. – М. : Либроком, 2009. – 328 с. – ISBN 978-5-397-00039-0.

9. *Князев, Б. А.* Начала обработки экспериментальных данных : электрон. учеб. и программа обработки данных для начинающих : учеб. пособие / Б. А. Князев, В. С. Черкасский. // Новосибир. ун-т. – Новосибирск, 1996. – 93 с. [Электронный ресурс]. – URL: http://www.phys.nsu.ru/cherk/Metodizm_old.PDF (дата обращения: 16.04. 2015).

10. *Кобзарь, А. И.* Прикладная математическая статистика / А. И. Кобзарь. – М. : Физматлит, 2006. – 816 с. – ISBN 5-9221-0707-0.

11. Универсальный симулятор, базирующийся на технологии искусственных нейронных сетей, способный работать на параллельных

машинах / О. В. Крючин [и др.] // Вестник Тамб. ун-та. (Серия «Естественные и технические науки»). – Тамбов, 2008. – Т. 13. Вып. 5. – С. 372 – 375.

12. *Лопатников, Л. И.* Экономико-математический словарь. Словарь современной экономической науки / Л. И. Лопатников. – 5-е изд., перераб. и доп. – М. : Дело, 2003. – 520 с. – ISBN 5-7749-0275-7.

13. Моделирование и прогнозирование состояния окружающей среды : метод. указания к лаб. практикуму. В 2 ч. Ч. 1 / сост. А. Н. Гороховский – Донецк : ДонНТУ, 2009. – 130 с. [Электронный ресурс]. – URL: http://window.edu.ru/window_catalog/files/r69269/mainLab-MiPOC1.pdf (дата обращения: 30.07.2015).

14. *Орлов, Д. С.* Химия почв / Д. С. Орлов. – М. : Изд-во: МГУ, 1992. – 631 с.

15. *Пачепский, Я. А.* Математические модели физико-химических процессов в почвах / Я. А. Пачепский. – М. : Наука, 1992. – 120 с.

16. *Ризниченко, Г. Ю.* Лекции по математическим моделям в биологии : учеб. пособие для студентов биол. специальностей высш. учеб. заведений / Г. Ю. Ризниченко. – М. ; Ижевск : R & C Dynamics; РХД, 2002 [Электронный ресурс]. – URL: http://www.library. biophys.msu.ru/MathMod/BM.HTML (дата обращения: 18.08.2015).

17. *Она же.* Математические модели в биофизике и экологии / Г. Ю. Ризниченко. – М. ; Ижевск : Ин-т компьютер. исслед., 2003. – 183 с. [Электронный ресурс]. – URL: http://www.library.biophys. msu.ru/MathMod/EM.HTML (дата обращения: 15.03.2015).

18. *Сметник, А. А.* Миграция пестицидов в почвах / А. А. Сметник, Ю. Я. Спиридонов, Е. В. Шеин. – М. : РАСХН-ВНИИФ, 2005. – 326 с.

19. *Смиряев, А. В*. Моделирование: от биологии до экономики : учеб. пособие / А. В. Смиряев, А. В. Исачкин, Л. К. Харрасова. – М. : Изд-во МСХА, 2002. – 122 с. – ISBN 5-94327-123-6.

20. *Харитонова, Г. В.* Молекулярные межфазные взаимодействия в почвах / Г. В. Харитонова, Е. В. Шеин, Б. А. Воронов. – Владивосток : Дальнаука, 2012. – 172 с.

ОГЛАВЛЕНИЕ

Учебное издание

ШЕИН Евгений Викторович
МАЗИРОВ Михаил Арнольдович
КОРЧАГИН Алексей Анатольевич
и др.

РЕГРЕССИОННЫЙ АНАЛИЗ В ПОЧВОВЕДЕНИИ

Учебное пособие

Редактор Р. С. Кузина
Технический редактор С. Ш. Абдуллаева
Корректор Е. П. Викулова
Компьютерная верстка Е. А. Кузьминой

Подписано в печать 16.09.16.
Формат 60x84/16. Усл. печ. л. 5,11. Тираж 60 экз.
Заказ
Издательство
Владимирского государственного университета
имени Александра Григорьевича и Николая Григорьевича Столетовых.
600000, Владимир, ул. Горького, 87.