

РЕГРЕССИЯ

К регрессионному анализу относятся задачи выявления искаженной случайным «шумом» функциональной зависимости интересующего исследователя показателя Y от измеряемых переменных X_1, \dots, X_m . Данными служит таблица экспериментально полученных «зашумленных» значений Y на разных наборах x_1, \dots, x_m . Основной целью обычно является как можно более точный прогноз (предсказание) Y на основе измеряемых (*предикторных*) переменных.

Регрессионный анализ по праву может быть назван основным методом современной математической статистики.

Н. Дрейлер, Г. Смит

Predict (англ.) — предсказывать.

§ 1. ПОДГОНКА ПРЯМОЙ

Термин «регрессия» ввел Ф. Гальтон в своей статье «Регрессия к середине в наследовании роста» (1885 г.), в которой он сравнивал средний рост детей Y со средним ростом их родителей X (на основе данных о 928 взрослых детях и 205 их родителях). Гальтон заметил, что рост детей у высоких (низких) родителей обычно также выше (ниже) среднего роста популяции $\mu \approx \bar{X} \approx \bar{Y}$, но при этом отклонение от μ у детей меньше, чем у родителей. Другими словами, экстремумы в следующем поколении сглаживаются, происходит возвращение назад (*регрессия*) к середине. По существу, Гальтон показал, что зависимость Y от X хорошо выражается уравнением $Y - \bar{Y} = (2/3)(X - \bar{X})$ (рис. 1).

Если кто-то способен предсказать, чем закончатся его исследования, то эта проблема не очень глубока и, можно сказать, практически не существует.

А. Шильд

Позднее регрессией стали называть любую функциональную зависимость между случайными величинами, даже в тех ситуациях, когда предикторные переменные являются неслучайными. В примечании переводчиков на с. 26 книги [23] высказано интересное мнение по поводу «живучести» термина «регрессия».

«Можно предположить, что его удивительная устойчивость связана с переосмыслением значения. Постепенно исходная антропометрическая задача, занимавшая Гальтона, была забыта, а интерпретация вытеснилась благодаря ассоциативной связи с понятием «регресс», т. е. движение назад. Сначала берутся данные, а уж потом, задним числом, проводится их обработка. Такое понимание пришло на смену традиционной, еще средневековой, априорной модели, для которой данные были лишь инструментом подтверждения. Негативный оттенок, присущий понятию «регресс», думается и вызывает психологический дискомфорт, поскольку воспринимается одновременно с понятиями, описывающими такой прогрессивный метод, как регрессионный анализ.»

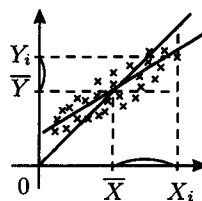


Рис. 1

Проиллюстрируем основные идеи регрессии на примере подгонки прямой под «облако» экспериментальных точек (x_i, η_i) , полученных в соответствии с моделью

$$\eta_i = a + bx_i + \varepsilon_i, \quad i = 1, \dots, n. \quad (1)$$

Здесь коэффициенты прямой a и b — неизвестные параметры, x_i — (неслучайные) значения предиктора X (для простоты допустим, что $x_i \neq x_j$), ε_i — независимые и одинаково распределенные случайные ошибки, $M\varepsilon_i = 0$. Для нахождения оценок коэффициентов a и b применим

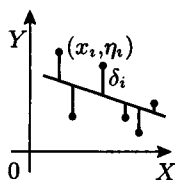


Рис. 2

Метод наименьших квадратов (МНК).

Естественным условием точности подгонки *пробной* прямой $y = \alpha + \beta x$ служит близость к нулю всех *остатков* $\delta_i(\alpha, \beta) = \eta_i - \alpha - \beta x_i$ (рис. 2). Общую меру близости к нулю можно выбирать по-разному (например, $\max |\delta_i|$ или $\sum |\delta_i|$), но наиболее простые формулы для оценок \hat{a} и \hat{b} получаются, если в качестве такой меры взять

$$F(\alpha, \beta) = \sum_{i=1}^n \delta_i^2 = \sum_{i=1}^n (\eta_i - \alpha - \beta x_i)^2. \quad (2)$$

Минимум $F(\alpha, \beta)$ достигается (см. задачу 1) в точке (\hat{a}, \hat{b}) , где

$$\hat{b} = \frac{\sum_{i=1}^n (\eta_i - \bar{\eta})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{a} = \bar{\eta} - \hat{b}\bar{x}. \quad (3)$$

Метод наименьших квадратов был впервые опубликован в 1805 г. Лежандром в работе, посвященной нахождению орбит комет. К. Гаусс утверждал, что использовал МНК еще до 1803 г.

Замечание 1. Прямая, подогнанная под «облако» точек МНК, вообще говоря, отличается от первой главной компоненты (см. § 1 гл. 20). Дело в том, что при построении главной компоненты переменные X и Y считаются равноправными, и минимизируется сумма квадратов длин отрезков, перпендикулярных компоненте, а не направленных вдоль оси Y , как в случае МНК (рис. 3).

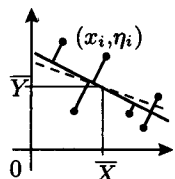


Рис. 3

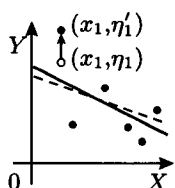


Рис. 4

Недостатком метода наименьших квадратов является излишняя *чувствительность* МНК-оценок к выделяющимся наблюдениям («выбросам»), возникающая вследствие зависимости меры F не от самих остатков δ_i , а от их квадратов. Стремление уменьшить остатки в точках «выбросов» может привести к значительному *смещению* оценок параметров. Так, если переместить вверх точку (x_1, η_1) на рис. 4, то, чтобы уменьшить δ_1^2 , МНК-прямая довольно сильно повернется.

Одной из устойчивых к «выбросам» (*робастных*) альтернатив МНК может служить

А. М. Лежандр
(1752–1833), французский математик.

Некоторые детали спора о приоритете между Лежандром и Гауссом приведены в примечании переводчиков к с. 32 книги [23].

Метод Тейла (см. [88, с. 215])

В этом методе оценки коэффициентов прямой задаются следующими формулами:

$$\begin{aligned}\tilde{b} &= MED\{(\eta_j - \eta_i)/(x_j - x_i), \quad 1 \leq i < j \leq n\}, \\ \tilde{a} &= MED\{\eta_i - \tilde{b}x_i, \quad i = 1, \dots, n\}.\end{aligned}\quad (4)$$

Причина робастности метода Тейла кроется в том, что оди-ночный «выброс» может исказить самое большее $(n - 1)$ оценок $(\eta_j - \eta_i)/(x_j - x_i)$ коэффициента наклона b , в то время как медиана вычисляется по $n(n - 1)/2$ оценкам.

Пример 1 ([86, с. 234]). Закон Хаббла в астрономии гласит: скорость удаления галактики прямо пропорциональна расстоянию до нее. В таблице ниже указаны расстояния Y (в миллионах световых лет) и скорости X (в сотнях миль в секунду) для 11 галактических созвездий (см. [66]).

Требуется подогнать прямую $Y = bX$ к этим данным.

Дистанции огромного
размера...

Сказлужб в «Горе от ума»
А. С. Грибоедова

Созвездие	X	Y	Y/X
Дева (Virgo)	22	7,5	0,341
Пегас (Pegasus)	68	24	0,353
Персей (Perseus)	108	32	0,296
Волосы Вероники (Coma Berenices)	137	47	0,343
Большая Медведица (Ursa Major No. 1)	255	93	0,365
Большая Медведица (Ursa Major No. 2)	700	260	0,371
Лев (Leo)	315	120	0,381
Северная Корона (Corona Borealis)	390	134	0,344
Близнецы (Gemini)	405	144	0,356
Волопас (Bootes)	685	245	0,358
Гидра (Hydra)	1100	380	0,345

Имеется $11 \times 10/2 = 55$ попарных наклонов, начиная с наименьшего 0,187 для Льва и Северной Короны. Медиана наклонов $\tilde{b} = 0,359$, оценка метода наименьших квадратов $\hat{b} = 0,353 \approx \tilde{b}$.

Замечание 2. Интересно, что оценка \tilde{b} в формуле (4) связана с ранговым коэффициентом Кендэла τ (см. § 7 гл. 20), причем эта связь аналогична связи МНК-оценки \hat{b} в формуле (3) с обычным выборочным коэффициентом корреляции $\hat{\rho}$.

Действительно, занумеруем наблюдения так, что $x_1 < \dots < x_n$. Если из η_i вычесть истинные значения bx_i , то $\eta_i - bx_i = a + \varepsilon_i$ ($i = 1, \dots, n$) образуют выборку (набор независимых и одинаково распределенных случайных величин). Не зная b , будем вычитать из η_i величины βx_i , где β меняется по нашему произволу. Чем ближе β к b , тем больше $\eta_i - \beta x_i$ будут похожи на выборку. В противном случае, они будут проявлять тенденцию к возрастанию (или убыванию) вместе с номером i (это зависит

от знака $b - \beta$). В этом легко убедиться, записав $\eta_i - \beta x_i$ в виде $\eta_i - \beta x_i = \eta_i - b x_i + x_i(b - \beta) = a + \varepsilon_i + x_i(b - \beta)$, из которого понятно, что в силу возрастания x_i с увеличением i у $\eta_i - \beta x_i$ появляется положительный (или отрицательный) «снос».

Тенденцию к изменению $\eta_i - \beta x_i$ с ростом i (или ее отсутствие) можно исследовать с помощью коэффициентов корреляции. Возьмем сначала *обычный выборочный коэффициент корреляции* $\hat{\rho}(\beta)$ между рядами (x_1, \dots, x_n) и $(\eta_1 - \beta x_1, \dots, \eta_n - \beta x_n)$:

$$\hat{\rho}(\beta) = \frac{\sum (x_i - \bar{x})[(\eta_i - \bar{\eta}) - \beta(x_i - \bar{x})]}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum [(\eta_i - \bar{\eta}) - \beta(x_i - \bar{x})]^2}}.$$

Наименьшей зависимости $\eta_i - \beta x_i$ от x_i ($i = 1, \dots, n$) соответствует значение $\hat{\rho} = 0$. По отношению к β это дает уравнение

$$\sum_{i=1}^n (x_i - \bar{x})(\eta_i - \bar{\eta}) = \beta \sum_{i=1}^n (x_i - \bar{x})^2,$$

решением которого и является МНК-оценка \hat{b} из формулы (3).

В случае *коэффициента Кендэла* τ заменим x_i и $\eta_i - \beta x_i$ их рангами i и T_i соответственно. Тогда (см. формулу (24) гл. 20)

$$\tau = \frac{2}{n(n-1)} \sum_{i < j} \text{sign}(T_j - T_i) = \frac{2}{n(n-1)} \sum_{i < j} \text{sign}(\eta_j - \beta x_j - \eta_i + \beta x_i).$$

Правую часть можно переписать также в виде

$$\tau(\beta) = \frac{2}{n(n-1)} \sum_{i < j} \text{sign} \left(\frac{\eta_j - \eta_i}{x_j - x_i} - \beta \right).$$

Рассуждения, аналогичные проведенным в комментарии 3 к критерию знаков из § 2 гл. 15, показывают, что решением уравнения $\tau(\beta) = 0$ является оценка \hat{b} , задаваемая формулой (4).

Какими свойствами обладают МНК-оценки? Как вытекает из задачи 2, в случае, когда вектор ошибок $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ *нормально распределен*, МНК совпадает с методом максимального правдоподобия (см. § 4 гл. 9) и, следовательно, является наиболее точным для больших выборок (асимптотически эффективным). Другие статистические свойства МНК-оценок обсуждаются в § 3.

§ 2. ЛИНЕЙНАЯ РЕГРЕССИОННАЯ МОДЕЛЬ

Предположим, что (с точностью до случайных ошибок) целевая переменная Y есть *линейная комбинация* $\theta_1 X_1 + \dots + \theta_m X_m$ предикторных переменных X_1, \dots, X_m с неизвестными коэффициентами $\theta_1, \dots, \theta_m$.

Измерения η_i переменной Y ($i = 1, \dots, n$, где $n \geq m$), отвечающие заданным (не обязательно различным) значениям x_{i1}, \dots, x_{im}

предикторных переменных, имеют вид

$$\eta_i = \theta_1 x_{i1} + \dots + \theta_m x_{im} + \varepsilon_i,$$

где ε_i — случайные ошибки.

Вводя для векторов и матриц обозначения $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^T$, $\mathbf{X} = \|x_{il}\|_{n \times m}$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$, можно записать модель в матричной форме:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}. \quad (5)$$

Матрицу \mathbf{X} называют *матрицей плана эксперимента*.

Будем предполагать, что для модели (5) выполняется допущение

Д1. Столбцы $\mathbf{x}_l = (x_{1l}, \dots, x_{nl})^T$, $l = 1, \dots, m$, матрицы \mathbf{X} линейно независимы. Иными словами, ввиду выполнения неравенства $n \geq m$ матрица \mathbf{X} имеет ранг m (см. П10).

Пример 2. Подгонка полинома. Рассмотрим $\mathbf{x}_1 = (1, \dots, 1)^T$ и $\mathbf{x}_{l+1} = (u_1^l, \dots, u_n^l)^T$, $l = 1, \dots, m-1$. Здесь $u_1 < \dots < u_n$ — так называемые «узлы», в которых вычисляются значения многочлена

$$p(u) = \theta_1 + \theta_2 u + \theta_3 u^2 + \dots + \theta_m u^{m-1}$$

и «запугмляются» случайными ошибками ε_i (рис. 5 для $m = 4$).

Матрица \mathbf{X} имеет вид

$$\begin{pmatrix} 1 & u_1 & u_1^2 & \dots & u_1^{m-1} \\ 1 & u_2 & u_2^2 & \dots & u_2^{m-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & u_n & u_n^2 & \dots & u_n^{m-1} \end{pmatrix}.$$

Определитель подматрицы $\tilde{\mathbf{X}}$, образованной первыми m строками \mathbf{X} ($m \leq n$), является известным из линейной алгебры *определителем Вандермонда*: $\det \tilde{\mathbf{X}} = \prod_{1 \leq i < j \leq m} (u_j - u_i)$. Он отличен от нуля, поскольку «узлы» различны. Поэтому ранг подматрицы $\tilde{\mathbf{X}}$ (и матрицы \mathbf{X}) равен m , а столбцы $\mathbf{x}_1, \dots, \mathbf{x}_m$ линейно независимы.

В дальнейшем при изучении статистических свойств оценок параметров $\theta_1, \dots, \theta_m$ (см. § 3) будем считать выполняющимся также допущение

Д2. Случайные величины $\varepsilon_1, \dots, \varepsilon_n$ одинаково распределены с $M\varepsilon_i = 0$, $D\varepsilon_i = \sigma^2$ (параметр $0 < \sigma < \infty$ также неизвестен) и некоррелированы: $M\varepsilon_i \varepsilon_j = 0$ при $i \neq j$.

Оценим параметры $\theta_1, \dots, \theta_m$ методом наименьших квадратов, минимизируя по $\boldsymbol{\theta}$ функцию $F(\boldsymbol{\theta}) = \sum_{i=1}^n (\eta_i - \theta_1 x_{i1} - \dots - \theta_m x_{im})^2$.

Точка ее минимума $\hat{\boldsymbol{\theta}}$ называется *МНК-оценкой*, вектор $\hat{\boldsymbol{\delta}}$ с

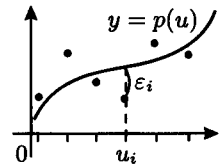


Рис. 5

компонентами $\hat{\delta}_i = \eta_i - \hat{\theta}_1 x_{i1} - \dots - \hat{\theta}_m x_{im}$, $i = 1, \dots, n$, — вектором остатков, а значение в точке минимума $RSS = F(\hat{\theta})$ — остаточной суммой квадратов.*)

Для нахождения оценки $\hat{\theta}$ интерпретируем задачу минимизации функции $F(\theta)$ в терминах пространства \mathbb{R}^n . Рассмотрим в \mathbb{R}^n подпространство $L(X)$, порождаемое столбцами x_1, \dots, x_m матрицы X . Очевидно, что вектор $X\theta = \theta_1 x_1 + \dots + \theta_m x_m$ пробегает $L(X)$, когда θ пробегает \mathbb{R}^m . Поскольку

$$F(\theta) = |\eta - X\theta|^2,$$

видим, что минимизация по θ равносильна нахождению в подпространстве $L(X)$ вектора $X\hat{\theta}$, наименее удаленного от η . Как известно, таким вектором служит ортогональная проекция η на $L(X)$ (рис. 6). Следовательно, вектор остатков $\hat{\delta} = \eta - X\hat{\theta}$ должен быть ортогонален векторам x_1, \dots, x_m , порождающим подпространство $L(X)$, т. е.

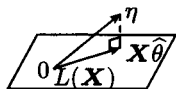


Рис. 6

$$X^T(\eta - X\hat{\theta}) = 0 \quad \text{или} \quad (X^T X) \hat{\theta} = X^T \eta. \quad (6)$$

Соотношение (6) представляет собой систему линейных уравнений относительно $\hat{\theta}$ с положительно определенной (задача 3) матрицей $B = X^T X$, которую называют информационной. Решить систему (6) можно, например, методом Холецкого (П10). Так как ввиду положительной определенности матрица B является невырожденной, для МНК-оценки $\hat{\theta}$ получаем представление

$$\hat{\theta} = B^{-1} X^T \eta. \quad (7)$$

Пример 3 ([2, с. 177], [27, с. 57]). На рис. 7 точками изображены результаты эксперимента по изучению зависимости между скоростью автомобиля V (в милях/час) и расстоянием Y (в футах), пройденным им после сигнала об остановке. Для каждого отдельного случая результат определяется в основном тремя факторами: скоростью V в момент подачи сигнала, временем реакции θ_1 водителя на этот сигнал и тормозами автомобиля. Автомобиль успеет проехать путь $\theta_1 V$ до момента включения водителем тормозов и еще $\theta_2 V^2$ после этого момента, поскольку согласно элементарным физическим законам теоретическое расстояние, пройденное до остановки с момента торможения, пропорционально квадрату скорости (убедитесь!).

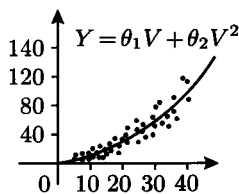


Рис. 7

Таким образом, в качестве модели годится $Y = \theta_1 V + \theta_2 V^2$. Для экспериментальных данных по формуле (7) были подсчитаны значения $\hat{\theta}_1 = 0,76$ и $\hat{\theta}_2 = 0,056$ (график параболы $Y = \hat{\theta}_1 V + \hat{\theta}_2 V^2$ приведен на рис. 7).

*) В обозначении использованы первые буквы соответствующего английского термина «Residual Sum of Squares».

дисперсию в классе линейных несмещенных оценок, построенных по порядковым статистикам $X_{(1)}, \dots, X_{(n)}$. Приведем некоторые результаты вычислений для равномерного и показательного законов (подробности см. в [38, с. 89]).

Для U_i , равномерно распределенных на отрезке $[0, 1]$, в задаче 1 гл. 5 были подсчитаны $\alpha_i = \mathbf{M}U_{(i)} = i/(n+1)$. В [38, с. 93] найдены

$$v_{ij} = \mathbf{M}U_{(i)}U_{(j)} - \alpha_i\alpha_j = \frac{i(n-j+1)}{(n+1)^2(n+2)}, \quad i \leq j.$$

Там же проверяется, что \mathbf{V}^{-1} имеет элементы $(n+1)(n+2)p_{ij}$, где $p_{ii} = 2$, $p_{i,i-1} = p_{i-1,i} = -1$, $p_{ij} = 0$ при $|j-i| > 1$,

а МНК-оценки параметров масштаба и сдвига выглядят так:

$$\hat{\sigma} = \frac{n+1}{n-1} (X_{(n)} - X_{(1)}), \quad \hat{\mu} = \frac{X_{(1)} + X_{(n)}}{2} - \frac{\hat{\sigma}}{2} = \frac{nX_{(1)} - X_{(n)}}{n-1}.$$

Они (с точностью до репараметризации модели) совпадают с оценками метода спейсингов из задачи 4 гл. 9.

Для *показательно распределенных* U_i ($F(u) = (1 - e^{-u})I_{\{u \geq 0\}}$) при решении задачи 5 гл. 4 было установлено, что

$$Z_i = (n-i+1)(U_{(i)} - U_{(i-1)}), \quad U_{(0)} = 0, \quad i = 1, \dots, n,$$

независимы и распределены так же, как и U_i . Поскольку $\mathbf{M}Z_i = 1$,

$$\alpha_i = \mathbf{M} \sum_{k=1}^i \frac{Z_i}{n-k+1} = \sum_{k=1}^i \frac{\mathbf{M}Z_i}{n-k+1} = \sum_{k=1}^i \frac{1}{n-k+1},$$

$$v_{ij} = \sum_{k=1}^i \sum_{l=1}^j \frac{\text{cov}(Z_i, Z_j)}{(n-k+1)(n-l+1)} = \sum_{k=1}^{\min(i,j)} \frac{1}{(n-k+1)^2}.$$

В [38, с. 96] проверяется, что матрица \mathbf{V}^{-1} имеет элементы q_{ij} , где

$$q_{ii} = (n-i+1)^2 + (n-i)^2, \quad q_{i,i-1} = q_{i-1,i} = -(n-i+1)^2$$

($q_{ij} = 0$ при $|j-i| > 1$), а МНК-оценки параметров σ и μ таковы:

$$\hat{\sigma} = n(\bar{X} - X_{(1)})/(n-1), \quad \hat{\mu} = \bar{X} - \hat{\sigma} = (nX_{(1)} - \bar{X})/(n-1).$$

В случае *нормального закона* для моментов порядковых статистик $U_{(i)}$ нет простых формул. Известно, что $\hat{\mu} = \bar{X}$, однако для вычисления $\hat{\sigma}$ приходится пользоваться таблицами (см. [70]).

§ 6. ПАРАДОКСЫ РЕГРЕССИИ

Существуют три вида
лжи: ложь, наглая ложь и
статистика.

Марк Твен

Есть несколько **типичных ошибок** («тонких мест»), которые следует иметь в виду, применяя регрессионный анализ. Сами по себе они достаточно очевидны. Тем не менее, о них часто забывают при работе с реальными данными и в результате приходят к неверным выводам.

1. Неоднородность данных. Существенно исказить вид регрессионной зависимости могут не только выделяющиеся наблюдения («выбросы») по оси отклика Y , но и аномальные значения предиктора X .

Пример 9 ([2, с. 64], [54]). На рис. 11, а представлены данные о числе телевизионных точек Y (в десят. тыс.), установленных в 1953 г. в девяти городах США (Денвере, Сан-Антонио, Канзас-Сити, Сиэтле, Цинцинати, Буффало, Нью-Орлеане, Милуоки, Хьюстоне) и о численности населения X (в десят. тыс.) этих городов.

Выборочный коэффициент корреляции между наборами x_1, \dots, x_9 и y_1, \dots, y_9 $\hat{r} = 0,403$, что при $n = 9$ свидетельствует о весьма малой степени линейной связи между X и Y .

Если же к этим данным добавить соответствующие сведения о Нью-Йорке ($x_{10} = 802, y_{10} = 345$), то пересчитанный для $n = 10$ коэффициент $\hat{r} = 0,995$. На рис. 11, б изображены в более мелком масштабе точки с координатами (x_i, y_i) , $i = 1, \dots, 10$, и проведена регрессионная прямая, по сути, соединяющая центр масс первых девяти точек (\bar{x}, \bar{y}) с (x_{10}, y_{10}) .

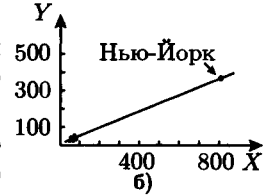
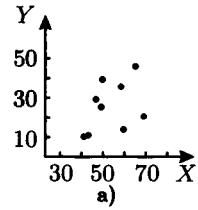


Рис. 11

2. Коррелированность предикторов. В случае, когда регрессионная модель включает много предикторов, некоторые из них могут оказаться приблизительно линейно связанными между собой. Обсудим связанные с этим проблемы.

При подгонке полинома на отрезке $[0, 1]$ (см. пример 2) уже для предикторов $X_3 = U^2$ и $X_4 = U^3$, измеряемых в «узлах» $u_i = i/20, i = 0, 1, \dots, 20$, выборочный коэффициент корреляции \hat{r} между соответствующими столбцами матрицы X составляет 0,986 (рис. 12).

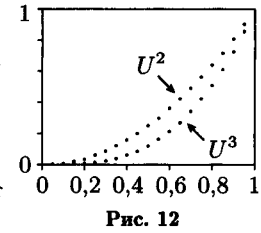


Рис. 12

Сильная коррелированность предикторов опасна тем, что приводит к неустойчивости МНК-оценок, вычисляемых по формуле (7), к малым возмущениям наблюдений η_i . Дело в том, что в этом случае столбцы матрицы X оказываются практически линейно зависимыми, вследствие чего матрица $B = X^T X$ становится почти вырожденной, а задача поиска решения линейной системы (6) — плохо обусловленной (см. [6, с. 131]).

Интерполяция — частный случай регрессии при $n = m$.

Рассмотрим подробнее влияние возмущений на интерполяционный полином степени $n - 1$, проходящий через точки плоскости с координатами (u_i, η_i) , $i = 1, \dots, n$. В форме Лагранжа он выглядит так:

$$p(u) = \sum_{i=1}^n \eta_i L_i(u), \quad \text{где } L_i(u) = \prod_{\substack{j=1 \\ j \neq i}}^n \frac{u - u_j}{u_i - u_j}.$$

То, что $p(u)$ интерполирует заданные точки, вытекает из равенств $L_i(u_i) = 1$ и $L_i(u_j) = 0$ при $j \neq i$.

На рис. 13, а, заимствованном из [53, с. 32], построен полином степени 8. На рис. 13, б продемонстрированы последствия добавления новой (десятой) точки: график даже не поместился в окне рисунка.

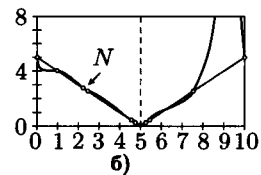
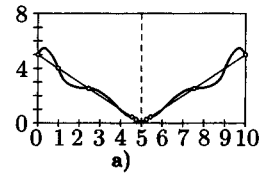


Рис. 13

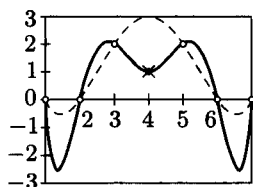


Рис. 14

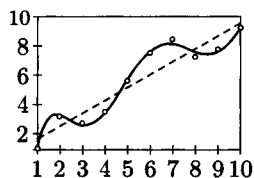


Рис. 15

Другой пример: даже для полинома степени 6 удаление точки с координатами (4,1) на рис. 14 приводит к значительным изменениям — возникновению осцилляций.

В регрессионных задачах полиномы высоких степеней имеют тенденцию *сглаживать ошибки наблюдений*, вместо того, чтобы выражать истинный вид зависимости отклика от предиктора. На рис. 15 десять точек, разбросанных вблизи прямой, сглажены для иллюстрации МНК-полиномом степени 6. Если зависимость на самом деле достаточно сложная, то для ее обнаружения могут оказаться полезными методы непараметрической регрессии, обсуждаемые в гл. 22.

Кроме вычислительной неустойчивости, коррелированность предикторов приводит к *затруднениям в интерпретации* результатов расчетов. Например, при исследовании зависимости *веса* Z студентов двух групп от их *роста* X и *размера обуви* Y методом наименьших квадратов (после предварительного выравнивания масштабов данных) в первой группе было получено регрессионное уравнение

$$Z - \bar{Z} = 0,9(X - \bar{X}) + 0,1(Y - \bar{Y}),$$

а для второй группы зависимость оказалась сильно отличающейся:

$$Z - \bar{Z} = 0,2(X - \bar{X}) + 0,8(Y - \bar{Y}).$$

Как объяснить существенное различие коэффициентов этих двух моделей?

На практике, подсчитав МНК-оценки $\hat{\theta}_1, \dots, \hat{\theta}_m$ в линейной регрессионной модели, исследователь в первую очередь обращает внимание на предикторы с *самыми большими* (по абсолютной величине) $\hat{\theta}_j$, так как именно их изменение сильнее всего сказывается на отклике. Исследователь нередко хочет не только точно предсказывать отклик для произвольных значений предикторов, но и желает, задавая эти значения, управлять откликом, надеясь, что регрессия отражает *причинно-следственную связь* между откликом и предикторами. Понятно, что «нажимать» надо на самые эффективные «рычаги».

В приведенном выше примере для первой группы важнейшим предиктором оказался X , а для второй — Y . Дело здесь в том, что X и Y сильно коррелируют друг с другом, вследствие чего общий «весовой» коэффициент при $(X - \bar{X}) + (Y - \bar{Y})$ случайным образом распределился между слагаемыми.

К счастью, рассмотренное затруднение нетрудно преодолеть. Достаточно проверить предикторы на наличие тесных линейных связей и каждую обнаруженную группу заменить в модели на ее единственного представителя.

3. Неадекватность модели. Когда простейшая линейная зависимость $Y = \theta_1 X_1 + \dots + \theta_m X_m$ неадекватно описывает данные,

<i>T</i>	1865	1866	1867	1868	1869	1870	1871	1872	1873	1874	1875	1876
<i>Y</i>	9,10	9,66	10,06	10,71	11,95	12,26	12,85	14,84	15,12	13,92	14,12	13,96
<i>T</i>	1877	1878	1879	1880	1881	1882	1883	1884	1885	1886	1887	1888
<i>Y</i>	14,19	14,54	14,41	18,58	19,82	21,56	21,76	20,46	19,84	20,81	22,82	24,03
<i>T</i>	1889	1890	1891	1892	1893	1894	1895	1896	1897	1898	1899	1900
<i>Y</i>	25,88	27,87	26,17	26,92	25,26	26,03	29,37	31,29	33,46	36,46	40,87	41,35
<i>T</i>	1901	1902	1903	1904	1905	1906	1907	1908	1909	1910		
<i>Y</i>	41,14	44,73	46,82	46,22	54,79	59,66	61,30	48,80	60,60	66,20		

Рис. 16

для построения более сложной модели обычно пытаются отдельно изучить влияние каждого предиктора X_j на отклик Y . Для этого сглаживают двумерное «облако» точек при помощи некоторой нелинейной функции. Список наиболее часто используемых *монотонных* функций содержится в следующей таблице.

Поведение отклика	Уравнение	Усл. на b	x'	y'
Очень быстрый рост ^{*)}	$y = e^{a+bx}$	$b > 0$	x	$\ln y$
Быстрый (степенной) рост	$y = e^{a+b \ln x}$	$b > 1$	$\ln x$	$\ln y$
Медленный рост	$y = e^{a+b \ln x}$	$0 < b < 1$	$\ln x$	$\ln y$
Очень медленный рост	$y = a + b \ln x$	$b > 0$	$\ln x$	y
Медленная стабилизация	$y = a + b/x$	$b \neq 0$	$1/x$	y
Быстрая стабилизация	$y = a + be^{-x}$	$b \neq 0$	e^{-x}	y
Кривая S-образной формы	$y = 1/(a + be^{-x})$	$b > 0$	e^{-x}	$1/y$

^{*)} В [51, с. 46] содержится любопытная классификация углов из книги по альпинизму (изданной около 1900 г.): «Перпендикулярно — 60° , мой дорогой сэр, абсолютно перпендикулярно — 65° , нависающе — 70° ».

[Последняя функция называется *логистической кривой*. При $a > 0$, $b > 0$ она возрастает, имеет две горизонтальные асимптоты $y = 0$, $y = 1/a$ и перегиб в точке с координатами $(\ln(b/a), 1/(2a))$.]

Переход к новым переменным x' и y' (см. таблицу) сводит задачу к подгонке прямой $y' = a + bx'$ из § 1.

Пример 10. В таблице на рис. 16 для каждого из годов с 1865 по 1910 (T — номер года) указано количество чугуна Y (в млн. тонн), которое выплавлялось за год во всем мире. Постараемся подогнать регрессионную кривую к этим данным.

Положим $X = T - 1864$. На рис. 17, а изображена кривая (ломаная), соединяющая точки плоскости с координатами (X_i, Y_i) , $i = 1, \dots, n$, где $n = 46$. Ван дер Варден (см. [13, с. 179]) пишет: «Эта кривая поднимается вверх значительно быстрее, чем прямая линия или квадратная парабола». Не соглашаясь с этим мнением, попытаемся сгладить кривую с помощью параболы. На рис. 17, а приведен график подогнанный параболы с помощью МНК. При этом остаточная сумма квадратов $RSS = 364,7$ и (согласно теореме 1) оценка стандартного отклонения ошибок $\hat{\sigma} = \sqrt{RSS/(n - m)} = 2,91$, где $m = 3$.

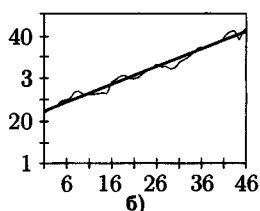
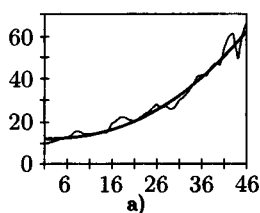


Рис. 17

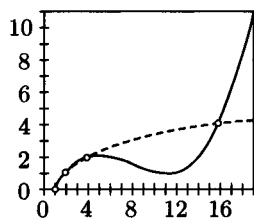


Рис. 18

Поскольку заметна некоторая тенденция усиления колебаний кривой относительно параболы с ростом X , применим, следуя Ван дер Вардену, преобразование $Y' = \ln Y$. На рис. 17, б построена кривая, соединяющая точки с координатами (X_i, Y'_i) , $i = 1, \dots, n$. Она хорошо согласуется с подогаданной МНК-прямой, имеющей уравнение $y = \hat{a} + \hat{b}x$, где $\hat{a} = 2,203$, $\hat{b} = 0,0413$ (оценки коэффициентов прямой вычисляются по формуле (3)).

Однако, если попытаться использовать последнюю подгонку для предсказания (прогноза) Y_i на основе формулы $\hat{Y}_i = \exp(\hat{a} + \hat{b}X_i)$, то получим значение $RSS = 381,8$, которое несколько больше, чем вычисленное ранее при подгонке параболы. Справедливости ради отметим, что *сумма модулей остатков* для параболы, наоборот, немного больше: $95,6 > 93,6$. В целом, можно считать подгонки примерно одинаковыми по точности.

Следует иметь в виду, что в случае ошибки при выборе типа сглаживающей кривой результаты *экстраполяции**) могут оказаться совершенно неудовлетворительными. Это наглядно демонстрирует рис. 18, на котором зависимость $y = \log_2 x$ аппроксимируется полиномом степени 3, интерполирующим точки с координатами $(1,0)$, $(2,1)$, $(4,2)$ и $(16,4)$.

При построении модели нужно максимально учитывать всю имеющуюся информацию о *качественном поведении* регрессионной кривой: монотонность, выход на асимптоту и т. п. В идеале, желательно опираться на законы (физики, химии, экономики), лежащие в основе зависимости (как в примере 3). Следующий пример показывает, что в случае формальной подгонки кривой, взятой из некоторого класса функций, *необходима перепроверка*.

Пример 11 ([2, с. 177] по данным А. Я. Боярского). Рассмотрим в качестве отклика Z *вес коровы*, а в качестве предикторов — *окружность ее туловища* X и *расстояние от хвоста до холки* Y . Сравнительному анализу были подвергнуты три регрессионные модели:

- (а) *линейная*: $Z = \theta_1 + \theta_2 X + \theta_3 Y$,
- (б) *степенная*: $Z = \theta'_1 X^{\theta'_2} Y^{\theta'_3}$,
- (с) *учитывающая содержательный смысл задачи*: $Z = \theta_0 X^2 Y$.

Происхождение последней модели легко объяснить. Для этого следует представить себе приблизительно тушу коровы в форме цилиндра с длиной образующей, равной Y , и радиусом основания, равным $X/(2\pi)$ (рис. 19). Если ρ — средняя плотность, то вес $Z = \rho \pi [X/(2\pi)]^2 Y = \theta_0 X^2 Y$ с точностью до головы и ног («рогов и копыт»).

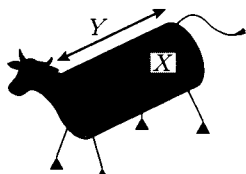


Рис. 19

*) То есть восстановления значений отклика по значениям предиктора, расположенным *вне* обследованного диапазона.

Вначале по всем ($n = 20$) имеющимся наблюдениям для каждой из моделей была вычислена МНК-оценка $\hat{\theta}$ векторного параметра θ и оценка $\hat{\sigma} = \sqrt{RSS/(n-m)}$ стандартного отклонения ошибок, где RSS — остаточная сумма квадратов (см. § 2), m — размерность вектора θ . Результаты расчетов приведены в левой стороне следующей таблицы из [2, с. 179].

Модель	По всем наблюдениям		По части наблюдений	
	$\hat{\theta}$	$\hat{\sigma}$	$\hat{\theta}_{\text{тяж}}$	$\hat{\sigma}_{\text{лег}}$
1	$\hat{\theta}_1 = -984,7$ $\hat{\theta}_2 = 4,73$ $\hat{\theta}_3 = 4,70$	25,9	$\hat{\theta}_1 = 453,2$ $\hat{\theta}_2 = 0,62$ $\hat{\theta}_3 = -0,22$	81
2	$\hat{\theta}'_1 = 0,0011$ $\hat{\theta}'_2 = 1,556$ $\hat{\theta}'_3 = 1,018$	24,5	$\hat{\theta}'_1 = 266,4$ $\hat{\theta}'_2 = 0,203$ $\hat{\theta}'_3 = -0,072$	79
3	$\hat{\theta}_0 = 1,13 \cdot 10^{-4}$	26,6	$\hat{\theta}_0 = 1,11 \cdot 10^{-4}$	28

Из них как-будто следует, что формальные модели 1 и 2 оказались несколько точнее содержательной модели 3. Однако, это лишь кажущееся благополучие, что сразу выявляется при проверке путем разбиения данных на обучающую и контрольную подвыборки.

В качестве *обучающей* была взята подвыборка из 10 самых тяжелых коров, а в качестве *контрольной* — из 10 оставшихся легких коров. На основе обучающей подвыборки для каждой из моделей была заново подсчитана МНК-оценка $\hat{\theta}_{\text{тяж}}$ (см. правую сторону таблицы). Видим, что модели 1 и 2 не выдержали испытание на *устойчивость коэффициентов* ($\hat{\theta}_3$ даже поменял знак).

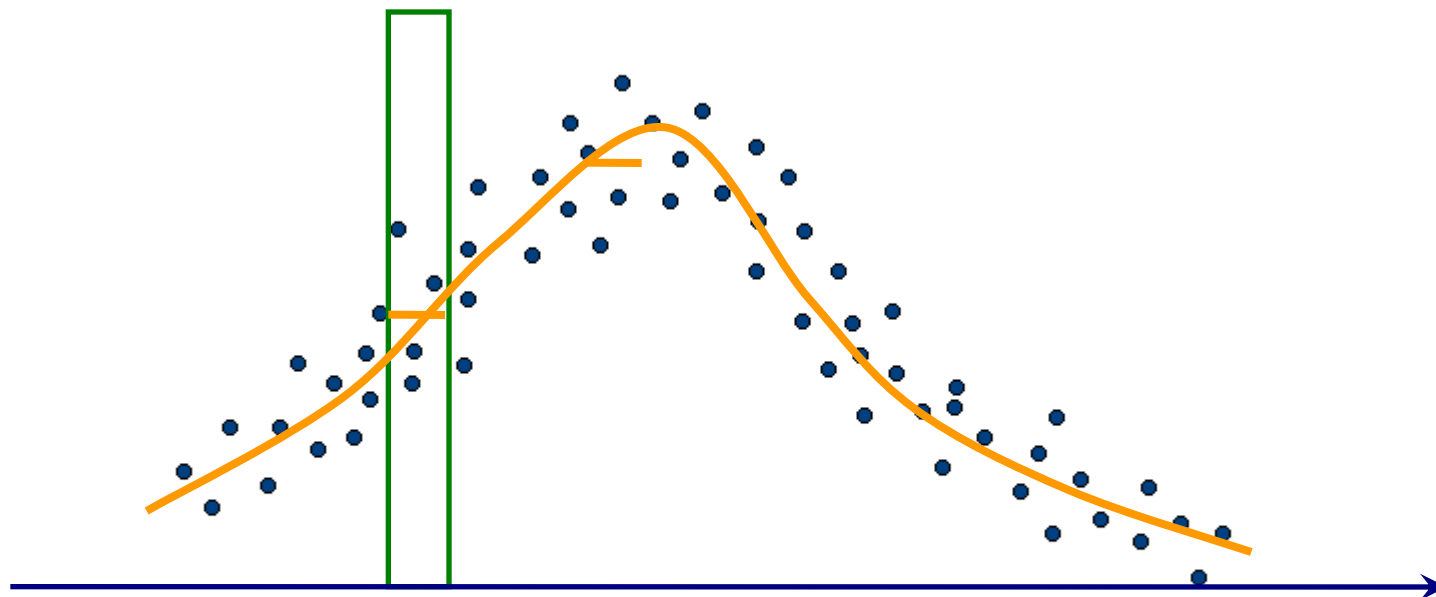
Кроме того, при попытке предсказания веса легких коров с помощью модели с такими коэффициентами, оценка стандартного отклонения ошибок $\hat{\sigma}_{\text{лег}}$ увеличилась для формальных моделей более чем в 3 раза!

Этот пример убедительно демонстрирует, что не следует переусложнять модель, ориентируясь на минимизацию $\hat{\sigma}$: за счет трех управляемых параметров («рычагов») удалось «подогнать» формальные модели к данным лучше, чем однопараметрическую содержательную.

4. Скрытый фактор. Желание истолковывать регрессионную связь как причинно-следственную может приводить к парадоксам, подобным приведенным в двух следующих примерах (см. также близкое к этой теме понятие частной или «очищенной» корреляции из § 8 гл. 20).

Пример 12 (см. [57]). Во время второй мировой войны англичане исследовали зависимость *точности бомбометания* Z от ряда факторов, в число которых входили *высота бомбардировщика* H , *скорость ветра* V , *количество истребителей противника* X . Как

Непараметрическая регрессия



$$m(x) = \mathbf{M}(Y|X = x).$$

Здесь $\mathbf{M}(Y|X = x)$ — условное математическое ожидание случайной величины Y при условии события $\{X = x\}$. Если у вектора (X, Y) существует плотность $p(x, y)$, то

$$m(x) = \int y p(x, y) dy / p(x),$$

где $p(x) = \int p(x, y) dy$ — маргинальная плотность сл. в. X .

«Пики» скорости роста детей



Пример 10. Зависимость скорости роста от возраста. На рис. 20 представлены ядерные оценки кривых *средней скорости* V (в см/год) *роста* H (т. е. $V = H'$) в зависимости от *возраста* T (в годах) для мальчиков (штрихованная линия) и для девочек (сплошная линия). У девочек наблюдается локальный максимум скорости роста около 12 лет. У мальчиков аналогичный пик еще более выражен, но происходит примерно на 2 года позже.

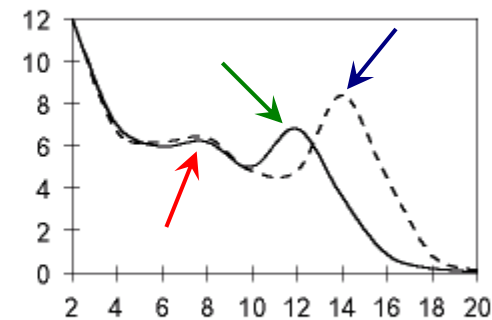


Рис. 20

Применение метода непараметрической регрессии позволило обнаружить и у мальчиков, и у девочек дополнительный локальный максимум V , так называемый *средний скачок роста*, в возрасте около 8 лет. Другие подходы, основанные на априорной фиксации параметрических моделей, приводят к значительным трудностям в обнаружении этого второго пика.



Алгоритм Lowess

В оригинале LOcally WEighted Scatter plot Smoothing (англ.) — локально взвешенное сглаживание диаграммы рассеяния.

- 1) Задаётся k от 1 до n . Для каждого $X_i = (X_{i1}, \dots, X_{im})$ вычисляется расстояние $h = h(X_i)$ до k -го ближайшего соседа среди точек $X_j, j = 1, \dots, n$.
- 2) Для заданного X_i и всех j вычисляются веса

$$w_j(X_i) = \frac{1}{h} q\left(\frac{|X_j - X_i|}{h}\right),$$

Этот алгоритм предложил W. S. Cleveland в 1979 году.

где $q(x)$ — некоторое ядро (например, *квартическое ядро* $q(x) = (15/16)(1 - x^2)^2 I_{\{|x| \leq 1\}} \cdot$)

Веса для всех i и j запоминаются в матрице.

- 3) В окрестности каждой точки X_i строится локальная линейная аппроксимация поверхности регрессии с помощью взвешенного метода наименьших квадратов путём минимизации по $\theta = (\theta_1, \dots, \theta_m)$ функции

$$R_i(\theta) = \sum_{j=1}^n w_j(X_i) (Y_j - \theta_1 X_{j1} - \dots - \theta_m X_{jm})^2.$$

- 4) Для каждого i определяется регрессионный остаток $\hat{\delta}_i$. Вычисляется выборочная медиана модулей остатков: $\hat{\sigma} = MED\{|\hat{\delta}_i|, i = 1, \dots, n\}$.

Плохо предсказываемые Y_i получают малые c_i

- 5) Для всех i определяются *коэффициенты предсказуемости* $c_i = q(\hat{\delta}_i / (6\hat{\sigma}))$.
- 6) Строится локальная линейная аппроксимация, как в пункте 3, но с весами $c_i w_j(X_i)$.
- 7) Пункты 4-6 повторяются до стабилизации величин c_i .

Пример применения метода Lowess

На рис. 6 из [85, с. 210] показано применение алгоритма «LOWESS» к моделированным данным. Точки генерировались в соответствии с моделью $Y_i = (1/50)x_i + \varepsilon_i$ ($n = 50$), $x_i = i$, $\varepsilon_i \sim \mathcal{N}(0, 1)$. Кривой регрессии является прямая с малым коэффициентом наклона, причем дисперсия ошибок $D\varepsilon_i$ настолько велика, что систематический прирост Y при увеличении X (так называемый *тренд*) почти не заметен из-за интенсивного «шума» (если оставить на рис. 6 только «крестики»).

Для $k = 25$, $m = 2$ и $X_i = (1, x_i)$ на рисунке представлен также результат сглаживания после двух итераций алгоритма. Построенная оценка уверенно игнорирует «выбросы», расположенные вблизи границ рисунка, и довольно точно воспроизводит теоретический линейный тренд.

Литература

85. Хардле В. Прикладная непараметрическая регрессия.
— М.: Мир, 1993

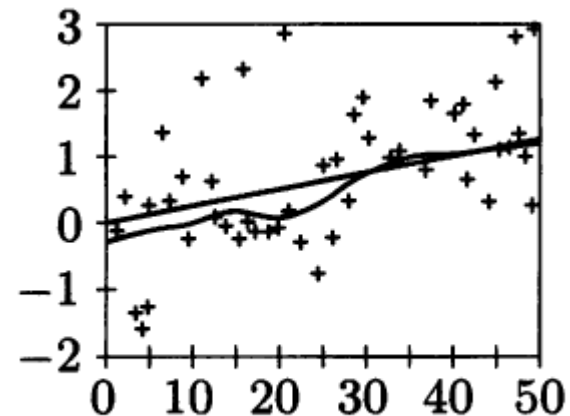
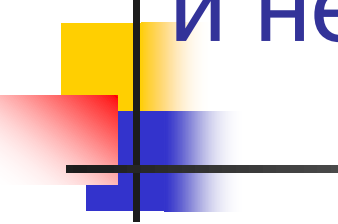


Рис. 6

Сравнение метода Lowess, линейной и непараметрической регрессий

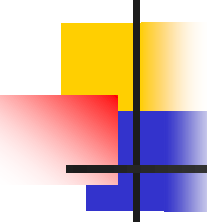


Цель исследования — изучить вид зависимости времени реакции до принятия лексического решения (т. е. решения, является ли набор букв словом английского языка) `RTlexdec` от частоты встречаемости слова в напечатанных текстах `WrittenFrequency`

- Установите пакет `languageR` (сопровождающий книгу Р. Н. Баауен «Analyzing Linguistic Data») с помощью кнопки `Install Packages`, находящейся в правом нижнем окне `RStudio` и подключите пакет, поставив перед ним «галку» на вкладке `Packages`
- Обозначьте ради краткости `e=english`
- Постройте диаграмму рассеяния изучаемых признаков из таблицы:
`plot(e[, c("RTlexdec", "WrittenFrequency")], pch = ".")`
- Отберите из таблицы `e` и запишите в таблицу `d` подмножество строк, относящихся к пожилым людям: `d=subset(e, AgeSubject=="old")`
- Введите следующие обозначения: `x=d$WrittenFrequency; y=d$RTlexdec`
- Подгоните кубический полином методом наименьших квадратов:
`m=lm(y~x+I(x^2)+I(x^3))`

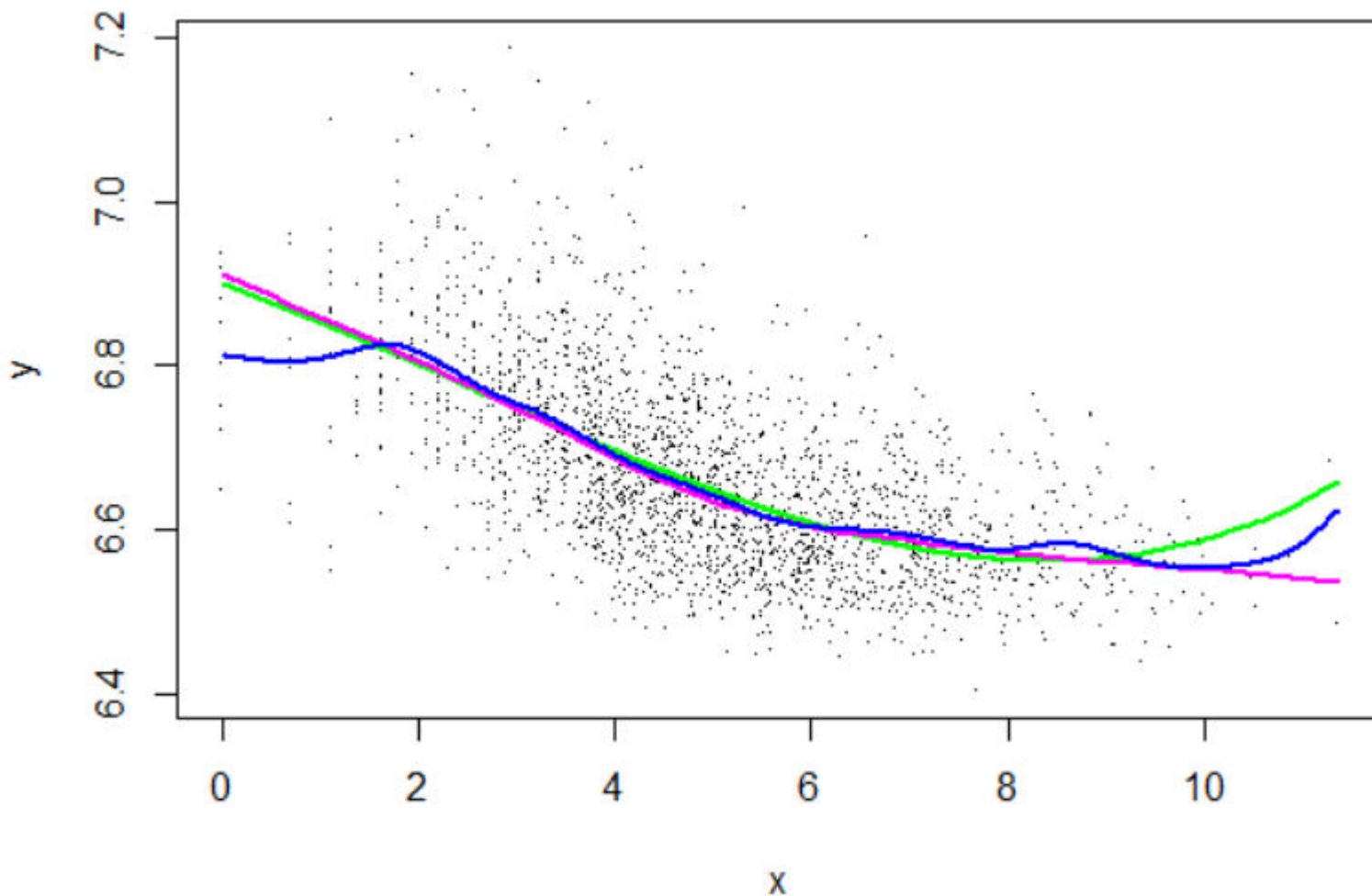
Так в моделях R обозначаются степени признаков

Продолжение исследования

- 
- Постройте график сглаживающей кривой, т. е. предсказанных значений:
`p=predict(m)` [прогноз]
`ord=order(x)` [такая перестановка, что `x[ord]` упорядочен]
`plot(y~x, pch = "."); lines(p[ord]~x[ord], col="green", lwd=2)`
 - Сгладьте «облако» точек методом `loess.smooth` (модификация алгоритма `Lowess`):
`lines(loess.smooth(x, y), col="magenta", lwd=2)`
 - Для применения непараметрической регрессии подключите пакет `KernSmooth` (ядерное сглаживание), поставив перед ним «галку» на вкладке `Packages`
 - Вычислите оптимальную ширину окна `h` и сгладьте «облако»:
`h=dpill(x, y); g=locpoly(x, y, degree=0, bandwidth=h)`
`lines(g, col="blue", lwd=2)`

Чем объяснить отличия графиков в начале и в конце диапазона значений `x`? Какой из методов сглаживания представляется наилучшим в данном случае?

Подогнанные кривые



МНК-подгонка
кубического
многочлена

Устойчивое loess-
сглаживание

Ядерное
сглаживание
(непараметри-
ческая регрессия)

Задачи

1. Вычислив частные производные, выведите формулы для МНК-оценок коэффициентов подгоняемой прямой $y = a + bx$.
2. Объясните, почему МНК-оценки коэффициентов линейной регрессионной модели совпадают с оценками максимального правдоподобия (при известной дисперсии ошибок σ^2).
3. Докажите, что для матрицы плана эксперимента X с линейно независимыми столбцами матрица $B = X^T X$ является положительно определённой.
4. Докажите, что МНК-оценка $\hat{\theta}$ в линейной регрессионной модели является несмещённой, и найдите её ковариационную матрицу.
5. Вычислите дисперсию прогноза на основе линейной регрессионной модели в произвольной точке $\mathbf{x} = (x_1, \dots, x_m)$.