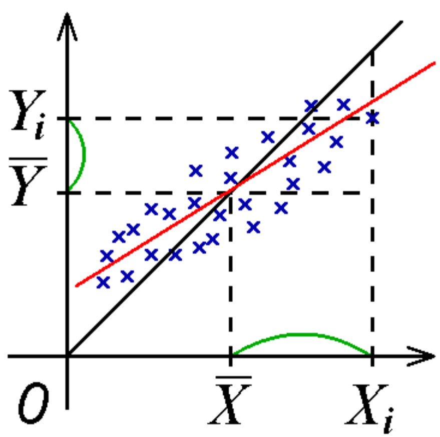
Регрессия

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

Н. Дрейпер, Г. Смит

Происхождение термина

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 $\mu \approx \overline{X} \approx \overline{Y}$, 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



$$Y - \overline{Y} = (2/3)(X - \overline{X}).$$

«Живучесть» термина

В примечании переводчиков книги Дрейпер Н., Смит Г. *"Прикладной регрессионный анализ"* (кн. 1, с. 26) 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".

Подгонка прямой

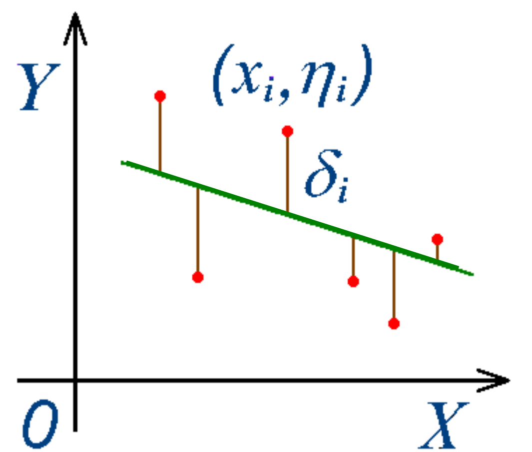
Пусть точки (x_i, η_i) получены в соответствии с моделью

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

Здесь коэффициенты прямой a и b — неизвестные параметры, x_i — (неслучайные) значения переменной X, ε_i — независимые и одинаково распределенные случайные ошибки, $\mathbf{M}\varepsilon_i = 0$. Для нахождения оценок коэффициентов a и b применим *метод наименьших квадратов* (МНК).

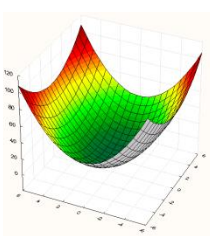
МНК был впервые опубликован Лежандром в 1805 г. Однако Гаусс утверждал, что он использовал МНК ещё в 1803 г.

Естественным условием точности подгонки *пробной прямой* $y = \alpha + \beta x$ служит близость к нулю всех *остатков*



$$\delta_i(\alpha, \beta) = \eta_i - \alpha - \beta x_i.$$

Наиболее простые формулы для оценок $\widehat{a}$ и $\widehat{b}$ получаются, если в качестве *меры качества подгонки* взять



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

МНК-оценка $(\widehat{a}, \widehat{b})$ есть точка минимума функции $F(\alpha, \beta)$.

Оценки коэффициентов a и b

Для вычисления МНК-оценок коэффициентов подгоняемой прямой используются следующие формулы:

$$\widehat{b} = \sum_{i=1}^{n} (\eta_i - \overline{\eta})(x_i - \overline{x}) \Big/ \sum_{i=1}^{n} (x_i - \overline{x})^2 , \quad \widehat{a} = \overline{\eta} - \widehat{b}\,\overline{x}.$$

Более устойчивым к выделяющимся наблюдениям является альтернативный метод оценивания коэффициентов a и b, предложенный Тейлом (H. Theil) в 1950 г. Согласно этому методу оценки вычисляются по формулам

$$\tilde{b} = MED\{(\eta_j - \eta_i)/(x_j - x_i), \quad 1 \leqslant i < j \leqslant n\},$$

$$\tilde{a} = MED\{\eta_i - \tilde{b}x_i, \quad i = 1, \ldots, n\}.$$

Причина устойчивости метода Тейла заключается в том, что одиночный «выброс» может исказить самое большее $(n - 1)$ оценку из $n(n - 1)/2$ оценок $(\eta_j - \eta_i)/(x_j - x_i)$ коэффициента наклона b.

Линейная регрессионная модель

Теперь рассмотрим зависимость признака η от $m \geq 2$ признаков $X_1, X_2, \ldots, X_m$. Предположим, что эта зависимость (приблизительно) линейная в некотором диапазоне значений признаков с точностью до случайных ошибок. Иными словами, пусть

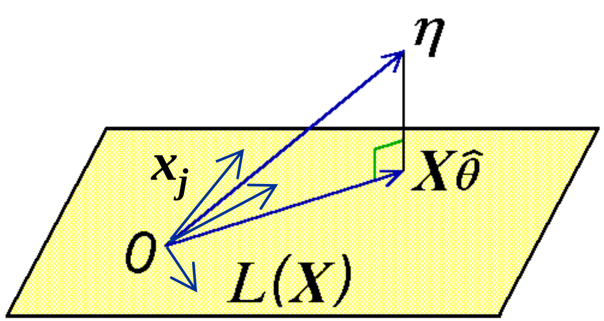
$$\eta = \theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_m X_m + \varepsilon,$$

где $\theta_1, \theta_2, \ldots, \theta_m$ — неизвестные коэффициенты (веса признаков), ε — случайная ошибка. Обозначим через x_{ij} значение признака X_j, где $j = 1, 2, \ldots, m$ для объекта с номером i, где $i = 1, 2, \ldots, n$. Тогда оценивание неизвестных коэффициентов методом наименьших квадратов заключается в минимизации по переменным $\theta_1, \theta_2, \ldots, \theta_m$

$$F(\theta_1, \theta_2, \ldots, \theta_m) = \sum_{i=1}^{n} (\eta_i - \theta_1 x_{i1} - \theta_2 x_{i2} - \ldots - \theta_m x_{im})^2.$$

На следующем слайде объясняется, как эта задача сводится к простой процедуре — решению системы линейных уравнений.

Геометрическая интерпретация линейной регрессионной модели



$$X\theta = \theta_1 x_1 + \ldots + \theta_m x_m$$

Линейная модель

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix} = \begin{array}{c} \quad\quad x_j \quad\quad\quad\quad \\ \begin{pmatrix} x_{11} & \boxed{x_{12}} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \end{array} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \cdots \\ \theta_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

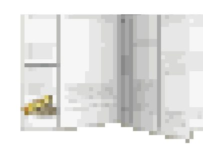
$$\boxed{\eta = X\theta + \varepsilon.}$$

Вычисление МНК-оценок

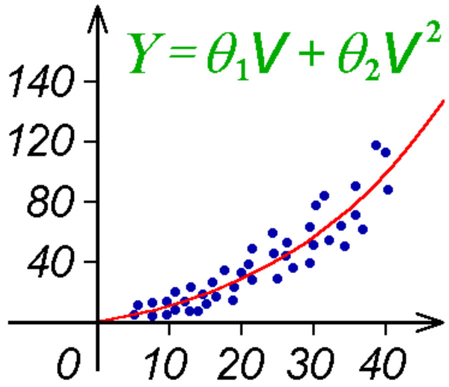
$$X^T(\eta - X\hat{\theta}) = 0 \quad \text{или} \quad (X^T X)\,\hat{\theta} = X^T \eta.$$

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

Система линейных уравнений относительно $\hat{\theta}$



Длина тормозного пути



На рисунке 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 физическим законам теоретическое расстояние, пройденное до остановки с момента торможения, *пропорционально квадрату скорости*.

Таким образом, в качестве модели годится $Y = \theta_1 V + \theta_2 V^2$. Для экспериментальных данных были подсчитаны МНК-оценки $\widehat{\theta}_1 = 0{,}76$ и $\widehat{\theta}_2 = 0{,}056$. График параболы $Y = \widehat{\theta}_1 V + \widehat{\theta}_2 V^2$ приведен на рисунке.

Изучение бедности

В 30 американских округах были собраны следующие демографические характеристики:

POP_CHNG — Population change (1960-1970) [прирост населения];

N_EMPLD — No. of persons employed in agriculture [число жителей, занятых в сельском хозяйстве];

PT_POOR — Percent of families below poverty level [процент жителей, находящихся за чертой бедности];

TAX_RATE — Residential and farm property tax ratex [местные налоги на землю и недвижимость];

PT_PHONE — Percent residence with telephones [доля телефонизации];

PT_RURAL — Percent rural population [доля сельского населения];

AGE — Median age [средний возраст жителей].

Изучим влияние на отклик PT_POOR остальных переменных (*предикторов*).

Визуальный анализ данных

1) Импортируйте файл Poverty.txt в RStudio под именем p

2) Постройте диаграммы размахов для всех признаков с помощью команды boxplot, а также для таблицы без 2-го столбца, и выявите <u>явные</u> «выбросы»

3) Замените явные «выбросы» на пропуски NA командой ifelse

4) Постройте матричную диаграмму рассеяния командой plot (нажмите Zoom, затем растяните окно во весь экран). Если обнаружите явные (т. е. не «хвостовые») двумерные «выбросы», то замените их на пропуски NA

5) Обратите внимание на 3-ю строку матричной диаграммы рассеяния и выясните, для каких предикторов наблюдается заметный наклон «облака» точек на диаграмме рассеяния предиктора с откликом PT_POOR (p[,3])

6) Выясните, какие из 6 предикторов значимо на уровне 0,05 коррелируют с откликом PT_POOR (запишите результат функции cor.test в переменную r и узнайте r$p.value).

7) Выясните, какие <u>предикторы</u> значимо на уровне 0,05 (0,01) коррелируют между собой (напишите двойной цикл и запишите индикаторы r$p.value<0.05 в матрицу m). Постройте на бумаге граф значимых связей, соединив рёбрами номера (или имена) значимо связанных предикторов

Множественная регрессия

Для построения *линейной регрессионной модели* и вывода на экран отчёта о ней используйте команды

m=*lm*(p[,3]~., data=p[,-3]); summary(m)

[Символ ~. означает, что в модели будут использованы все столбцы из таблицы, присвоенной аргументу data]

[Формула модели в общем случае имеет вид

Y ~ A + B + …]

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	30.483988	12.973554	2.350	0.028173
POP_CHNG	-0.348976	0.084190	-4.145	0.000423
N_EMPLD	0.002121	0.001043	2.034	0.054184
TAX_RATE	3.189461	3.399279	0.938	0.358290
PT_PHONE	-0.137053	0.132899	-1.031	0.313628
PT_RURAL	0.161840	0.060489	2.676	0.013815
AGE	-0.362263	0.252678	-1.434	0.165720

Estimate — коэффициенты линейной модели, *Intercept* — константа θ_0, добавленная в модель, т. е. модель имеет вид

$$\eta = \theta_0 + \theta_1 X_1 + \ldots + \theta_m X_m + \varepsilon$$

t value — характеристика степени влиятельности предиктора на отклик

Pr(>|*t*|) — фактические уровни значимости (отличия от нуля) соответствующих коэффициентов модели, которые вычисляются в предположении *нормальности распределения* ошибок наблюдений

Характеристики качества модели

```
Residual standard error: 3.382 on 21 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7805,    Adjusted R-squared:  0.7178
F-statistic: 12.45 on 6 and 21 DF,  p-value: 5.588e-06
```

Residual standard error — оценка для стандартного отклонения σ ошибки ε, которая вычисляется на основе регрессионных остатков δ_i по формуле

$$\hat{\sigma} = \sqrt{\sum \delta_i^2 / (n - m - 1)}$$

p-value — фактический уровень значимости всей модели, определяемый на основе F-статистики, который также вычисляется в предположении *нормальности распределения* ошибок наблюдений

Multiple R-squared — коэффициент детерминации (см. следующем слайд)

Для сохранения значений характеристик качества используйте команды
s=summary(m); s$sigma; s$r.squared

Коэффициент детерминации

Основным показателем силы связи между откликом и всеми предикторами служит *коэффициент детерминации*

$$R^2 = 1 - \sum_{i=1}^{n} (\eta_i - \widetilde{\eta}_i)^2 \Big/ \sum_{i=1}^{n} (\eta_i - \overline{\eta})^2 , \quad \text{где } \widetilde{\boldsymbol{\eta}} = \boldsymbol{X}\widehat{\boldsymbol{\theta}}.$$

Таким образом, чем ближе R^2 к 1, тем в большей степени предикторы определяют (детерминируют) отклик. Если же значение R^2 близко к 0, то сглаживание наблюдений константой (их выборочным средним) мало отличается от сглаживания с помощью наилучшей линейной функции от предикторов.

Показатель R^2 также называют *корреляционным отношением,* потому что он равен квадрату коэффициента корреляции R между откликом и *прогнозом,* который является проекцией отклика на пространство линейных комбинацией предикторов.

Избыточность (redundancy)

Чтобы решить, какие именно из тесно связанных между собой предикторов следует оставить в модели полезно учесть следующее:

а) если предиктор связан с откликом причинно-следственной связью (при которой, изменяя предиктор, можно управлять откликом), то, как правило, следует оставить в модели именно его;

б) если характер связи неизвестен, то лучше оставить предиктор, имеющий наибольший коэффициент корреляции с откликом;

в) *степень избыточности* предиктора можно охарактеризовать его коэффициентом детерминации R^2 на основе остальных предикторов.

Пример анализа избыточности

Рассмотрим сильно связанные (на уровне 0,01) предикторы:

№ 1. POP_CHNG — прирост населения;
№ 2. N_EMPLD — число жителей, занятых в сельском хозяйстве;
№ 5. PT_PHONE — доля телефонизации;
№ 6. PT_RURAL — доля сельского населения.

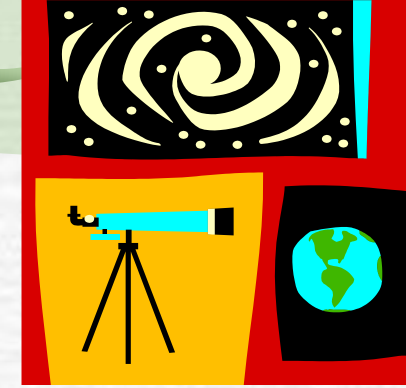
1) Какие из этих предикторов связаны с откликом PT_POOR (№ 3) причинно-следственной связью? Признак влияет на отклик или наоборот?

2) Какой из рассматриваемых предикторов сильнее всего коррелирует с откликом? (примените функцию cor.test)

3) Вычислите степени избыточности предикторов, прогнозируя каждый из предикторов на основе остальных (используйте функции типа

summary(lm(p[,j]~., data=p[,-c(j, 3, 4 ,7)])) (для j = 1, 2, 5, 6)

4) На основе пунктов 1-3 решите, какой (какие) из предикторов предпочтительнее оставить в модели



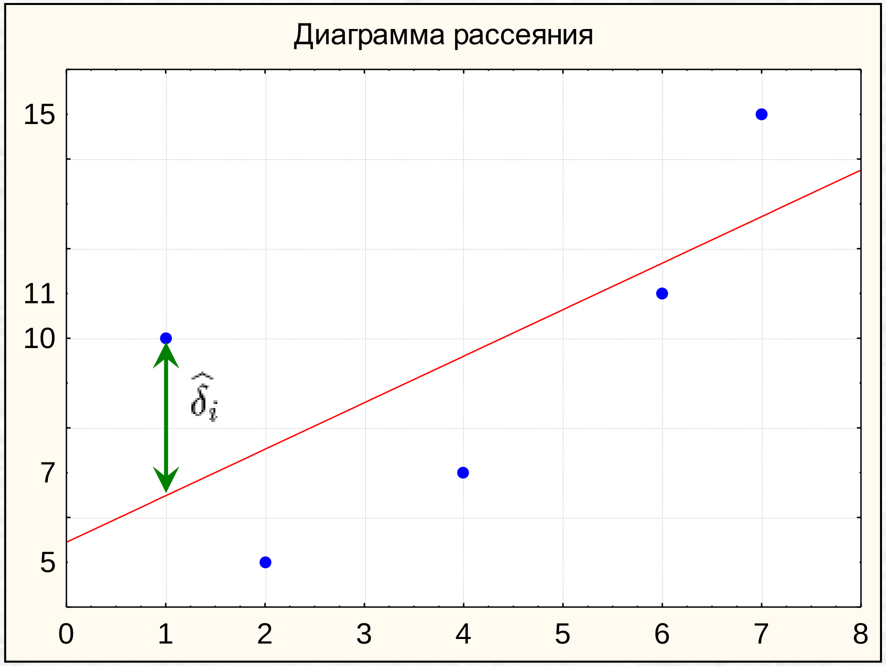
Анализ остатков

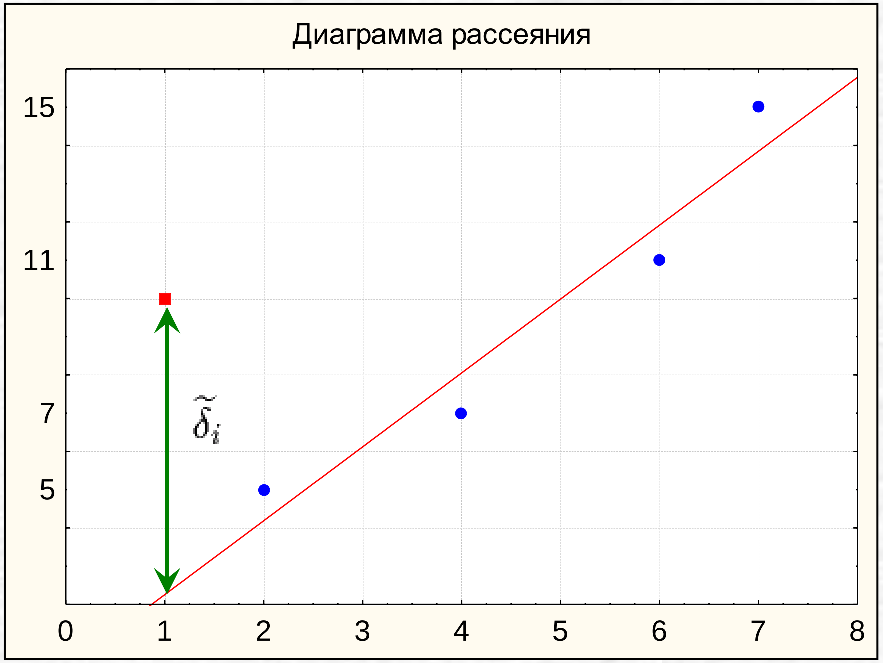
Почти все величайшие открытия в астрономии вытекают из рассмотрения того, что мы уже раньше назвали качественными или численными *остаточными феноменами*, иначе говоря, они вытекают из анализа той части числовых или качественных результатов наблюдения, которая «торчит» и остается необъясненной после выделения и учёта всего того, что согласуется со строгим применением известных методов.

Дж. Гершель, «Основы астрономии», 1849 г.

Далее рассматриваются методы анализа остатков и поиска скрытых «выбросов»

Deleted residuals (остатки после удаления отдельных наблюдений)



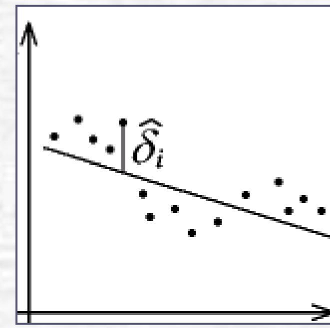


Критерий Дарбина — Уотсона

Данный критерий применяется для подтверждения значимости наблюдаемой сериальной корреляции остатков $\widehat{\delta}_i$ (см. рисунок справа сверху). К подобному поведению остатков приводит справедливость следующей альтернативы H_1 для гипотезы H_0 независимости ошибок наблюдений ε_i:

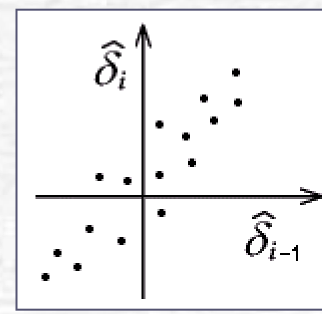
$$H_1:\ \varepsilon_i = \rho\,\varepsilon_{i-1} + \zeta_i,$$

где $\zeta_i \sim \mathcal{N}(0,\,\sigma^2)$ и независимы, а $\rho \neq 0$, $|\rho| < 1$.



Статистикой критерия Дарбина — Уотсона служит

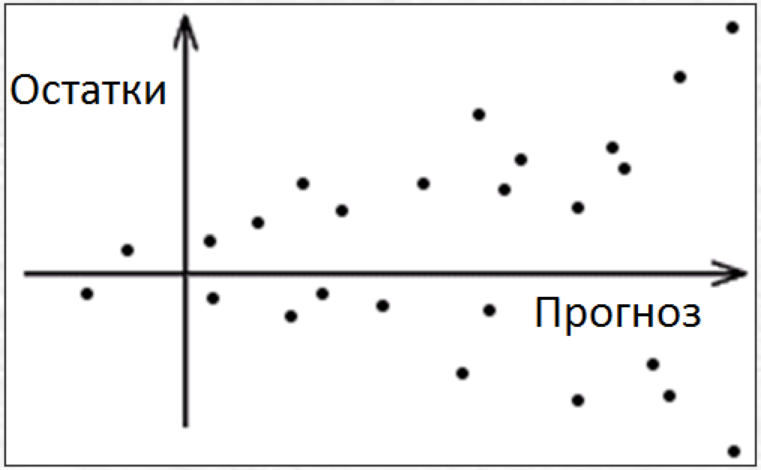
$$d = \sum_{i=2}^{n}(\widehat{\delta}_i - \widehat{\delta}_{i-1})^2 \Bigg/ \sum_{i=1}^{n}\widehat{\delta}_i^2.$$



H_0 отвергается | Неизвестно | H_0 принимается

0 | d_L | d_U

Критерий Голдфелда — Квандта



Упорядочим n наблюдений в порядке возрастания значений прогноза $\widetilde{\eta}$ и выберем k первых и k последних наблюдений ($k < n/2$). Гипотеза гомоскедастичности отвергается на уровне значимости α, если статистика

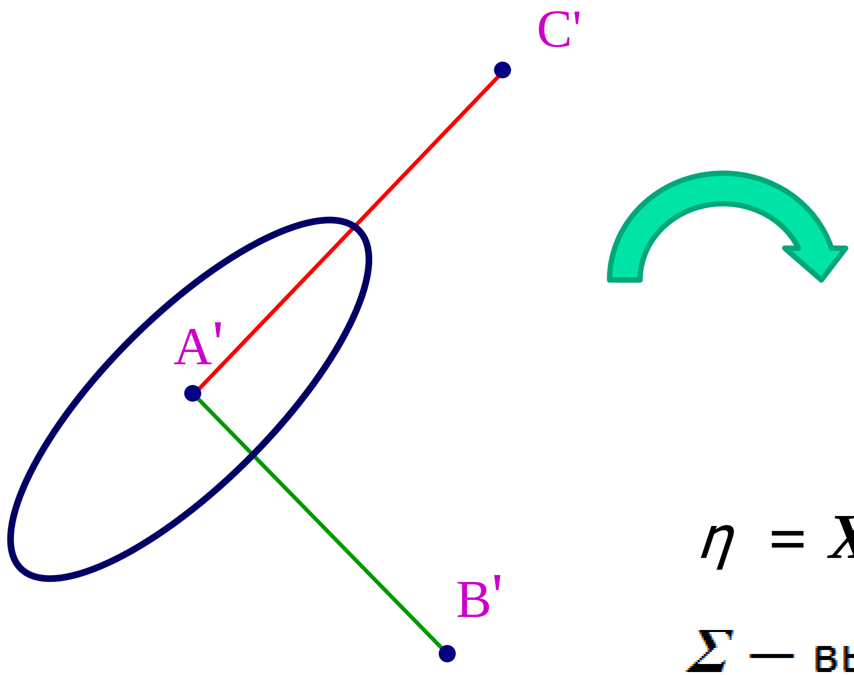
$$F = \sum_{i=n-k+1}^{n} \widehat{\delta}_i^2 \Big/ \sum_{i=1}^{k} \widehat{\delta}_i^2 > f_{\alpha,k-m+1,k-m+1},$$

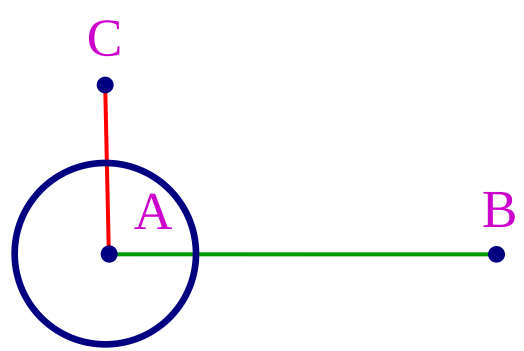
где $f_{\alpha,k-m+1,k-m+1}$ обозначает $(1-\alpha)$-квантиль распределения Фишера с $k-m+1$ и $k-m+1$ степенями свободы, m — число предикторов в модели (включая константу).

Отметим, что *мощность* (т. е. единица минус вероятность ошибки II рода, см. §3 в теме 5) критерия Голдфелда—Квандта оказывается максимальной, если взять $k \approx n/3$.

Расстояние Махаланобиса

Предложено индийским статистиком Махаланобисом (Prasanta Chandra Mahalanobis) в 1936 году.





$$\eta = X\theta + \varepsilon$$

Σ — выборочная ковариация матрица столбцов матрицы X.

$$d^2(x, y) = (y - x)^T \Sigma^{-1} (y - x)$$

Поиск «влиятельных» объектов: расстояния Махаланобиса и Кука

Некоторые строки из таблицы данных могут оказаться «нетипичными по предикторам» в том смысле, что они не попадают внутрь 95%-доверительного эллипсоида рассеяния. Такие строки $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})$ будут иметь большие *расстояния Махаланобиса*

$$M_i = d^2(\boldsymbol{x}_i, \bar{\boldsymbol{x}}) = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T,$$

где $\bar{\boldsymbol{x}}$ — вектор-строка из выборочных средних всех предикторов, $\boldsymbol{\Sigma}$ — выборочная ковариация матрица предикторов, T — операция транспонирования. Эти строки, как рычаг (англ. *leverage*), способны отклонять регрессионную гиперплоскость от истинного положения.

Расстояние Кука (Cook's distance) определяются формулой

$$C_i = \left| \boldsymbol{X}\hat{\boldsymbol{\theta}}_{-i} - \boldsymbol{X}\hat{\boldsymbol{\theta}}) \right|^2 \Big/ \left(m\hat{\sigma}^2 \right),$$

где $\hat{\boldsymbol{\theta}}$ — вектор МНК-оценок, $\hat{\boldsymbol{\theta}}_{-i}$ — новый вектор МНК-оценок, получаемый при исключении i-й строки аналогично *deleted residuals*, $\hat{\sigma}$ — средняя ошибка. Расстояние Кука выражает степень влияния исключения i-й строки на изменение вектора прогноза $\boldsymbol{X}\hat{\boldsymbol{\theta}}$.

«Плечо» (leverage) объекта

С расстоянием Махаланобиса тесно связано понятие «плеча» («рычага»). В линейной регрессионной модели «плечо» i-го объекта определяется как диагональный элемент h_{ii} *проекционной матрицы*

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

(матрица так называется потому, что

$$\boldsymbol{H}\eta = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\eta = \boldsymbol{X}\hat{\theta} = \tilde{\eta},$$

т. е. в результате умножения $\boldsymbol{H}$ на отклик η получается прогноз $\tilde{\eta}$ — проекция вектора η на подпространство $\boldsymbol{X}\theta$).

Известно, что $0 \le h_{ii} \le 1$, а также, что дисперсия i-го остатка

$\mathsf{D}\hat{\delta}_i = (1 - h_{ii})\sigma^2$, где σ^2 — дисперсия ошибок ε_i.

Стьюдентизованным остатком называется $t_i = \dfrac{\delta_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$,

где оценка $\hat{\sigma}$ была определена ранее. «Плечо» линейно зависит от расстояния Махаланобиса: $M_i = (n-1)(h_{ii} - 1/n)$.

Основные шаги анализа остатков

- 1) Плохо прогнозируемые значения:
постройте командой plot(m) диаграмму рассеяния «Прогноз – Остаток» и проверьте, есть ли регрессионные остатки, оказавшиеся снаружи полосы с границами $\pm 2\hat{\sigma}$.

- 2) Визуальный анализ остатков (проверка адекватности модели):
внимательно рассмотрите «облако» точек на построенной диаграмме с целью обнаружения возможных «тенденций» в поведении остатков: наличия длинных серий остатков одного знака, увеличения (уменьшения) абсолютных величин остатков с ростом предсказанных значений, колебаний дисперсии остатков и др.

- 3) Нормальность распределения остатков:
нажав клавишу Enter, проверьте предположение о нормальности распределения ошибок с помощью диаграммы нормальных квантилей (Q-Q plot)

- 4) Поиск сильно влияющих объектов:
нажав клавишу Enter, выясните, присутствуют ли в таблице данных строки с очень большими расстояниями Махаланобиса и Кука

Продолжение анализа остатков

- 5) Визуальный контроль отсутствия сериальных корреляций остатков:

выполните следующие команды:

```
f=m$fitted.values            [записать прогноз в вектор f для краткости]
ord=order(f)                 [такая перестановка, что вектор f[ord] упорядочен]
r=m$residuals[ord]           [упорядочить остатки по росту значений прогноза]
x=c(NA, r); y=c(r, NA)       [сдвиг значений остатков на 1 вправо с добавкой NA]
plot(x, y)                   [построить диаграмму рассеяния]
abline(v=0, h=0)             [вывести на диаграмму рассеяния оси координат]
```

проверьте, не концентрируются ли точки в I и III координатных углах [положительная автокорреляция] или в II и IV координатных углах [отрицательная автокорреляция]

- 6) Проверка адекватности модели критериями Дарбина — Уотсона и Голдфелда — Квандта: установите пакет *lmtest* с помощью кнопки Install Packages и подключите его, поставив «галку» на вкладке Packages; с его помощью проверьте модель m (точнее говоря, её регрессионные остатки) на положительную и отрицательную автокорреляцию (функция dwtest) и на увеличение дисперсии остатков (функция gqtest). Внимание: в обеих функциях необходимо задать аргумент order.by=ord

Профессионализм

Я хотел бы спросить: «Что такое профессионал?» Многие, возможно, ответят, что профессионал — это человек, который очень много знает о своем предмете. Однако с этим определением я не мог бы согласиться, потому что никогда нельзя знать о каком-либо предмете действительно много. Я предпочёл бы такую формулировку: профессионал — это человек, которому известны грубейшие ошибки, обычно совершаемые в его профессии, и который, поэтому умеет их избегать.

В. Гейзенберг, «Физика и философия. Часть и целое».

«Ловушки» регрессии

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

Марк Твен

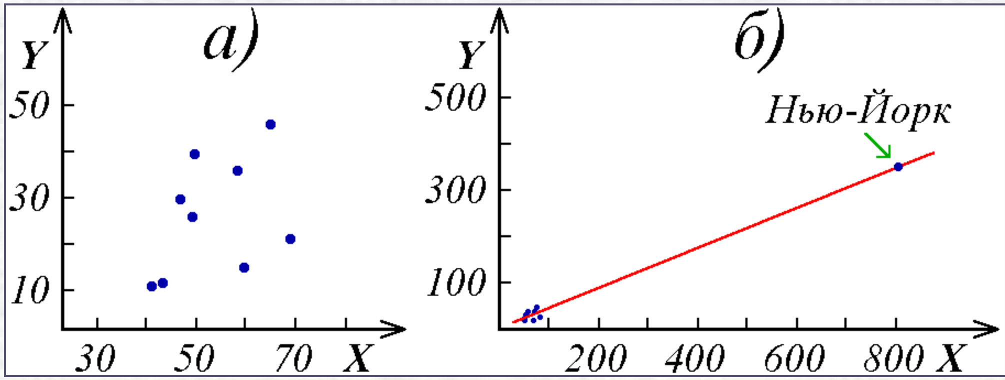
Есть несколько **типичных ошибок** (“тонких мест”), которые следует иметь в виду, применяя регрессионный анализ. Сами по себе, они достаточно очевидны. Тем не менее, о них часто забывают при работе с реальными данными и в результате приходят к неверным выводам.

1) Неоднородность данных

2) Коррелированность предикторов

3) Неадекватность модели

4) Скрытый фактор



«Ловушки» 2 – 4 подробно обсуждаются на следующих слайдах.

Плохая обусловленность матрицы

Ради краткости введём новые обозначения в уравнениях для вычисления МНК-оценок, полученных ранее:

$$\boldsymbol{B} = \boldsymbol{X}^T \boldsymbol{X}, \quad \boldsymbol{d} = \boldsymbol{X}^T \boldsymbol{\eta}.$$

Таким образом, поиск МНК-оценок сводится к решению системы линейных уравнений с матрицей коэффициентов $\boldsymbol{B}$ и правой частью $\boldsymbol{d}$.

В случае сильной коррелированности предикторов матрица ***B*** оказывается *плохо обусловленной* (см. файл Векторы и матрицы.pdf). Для решений таких систем характерна катастрофическая неустойчивость к возмущению правой части. Классическим примером служит линейная система с *матрицей Гильберта* $\boldsymbol{H}$, имеющей элементы $h_{ij} = 1/(i + j - 1)$. Для численного эксперимента возьмём $\boldsymbol{H}$ размерности $m = 5$. Возмутим нулевую правую часть, положив последнюю компоненту вектора $\boldsymbol{d}$ равной 0,001. В результате вместо нулевого решения получим

1	0,5	0,333333	0,25	0,2		0,63		0
0,5	0,333333	0,25	0,2	0,166667		-12,6		0
0,333333	0,25	0,2	0,166667	0,142857	x	56,7	=	0
0,25	0,2	0,166667	0,142857	0,125		-88,2		0
0,2	0,166667	0,142857	0,125	0,111111		44,1		0,001

$\boldsymbol{H}$ $\qquad$ $\boldsymbol{d}$

Пример: тесты на профпригодность

Импортируйте Job_prof.txt в RStudio под именем d, поставив переключатель Heading в положение Yes и убрав «галку» перед надписью Strings as factors.

Задача заключается в изучении характера зависимости уровня профессионализма JOB_PROF от результата выполнения теста TEST4.

- 1) Постройте диаграмму рассеяния признаков TEST4 и JOB_PROF:

  x=d[,4]; y=d[,5]; plot(x, y)

- 2) Добавьте на диаграмму линию «тренда» с помощью команды

  lines(loess.smooth(x, y), col="magenta", lwd=3)   [lwd — line width]

  Видим, что зависимость JOB_PROF от TEST4 слегка нелинейная.

- 3) Вычислите новый признак x2=x^2 и постройте линейную модель:

  summary(lm(y~x+x2))

  Убедитесь, что сама регрессионная модель значима [*p-value* $< 10^{-7}$], но все коэффициенты модели оказались незначимыми из-за очень сильной коррелированности предикторов x и x^2 [cor(x, x2) = 0,99876]

  Какой из предикторов предпочтительнее оставить в модели?

Пошаговая (stepwise) регрессия

Этот метод используется в случае большого числа предикторов и в случае коррелированности предикторов.

Суть процедуры заключается в постепенном увеличении числа предикторов в модели. Опишем один из используемых подходов.

Сначала определяется предиктор, имеющий наибольший коэффициент корреляции с откликом. Если при его добавлении в модель коэффициент при нём значимо (скажем, на уровне 5%) отличается от 0, то предиктор оставляется.

На очередном шаге среди предикторов, ещё не включенных в модель, определяется тот, который имеет наибольшую *частную корреляцию* с откликом при устранении влияния предикторов, уже присутствующих в модели. Если при его добавлении в модель коэффициент при нём значимо отличается от 0, то он оставляется.

Затем производится «чистка»: из модели удаляются все ранее включенные в нее предикторы, которые после добавления нового предиктора стали незначимыми (скажем, на уровне 10%).

Добавление предикторов прекращается, когда на очередном шаге коэффициент при новом предикторе окажется незначимым.

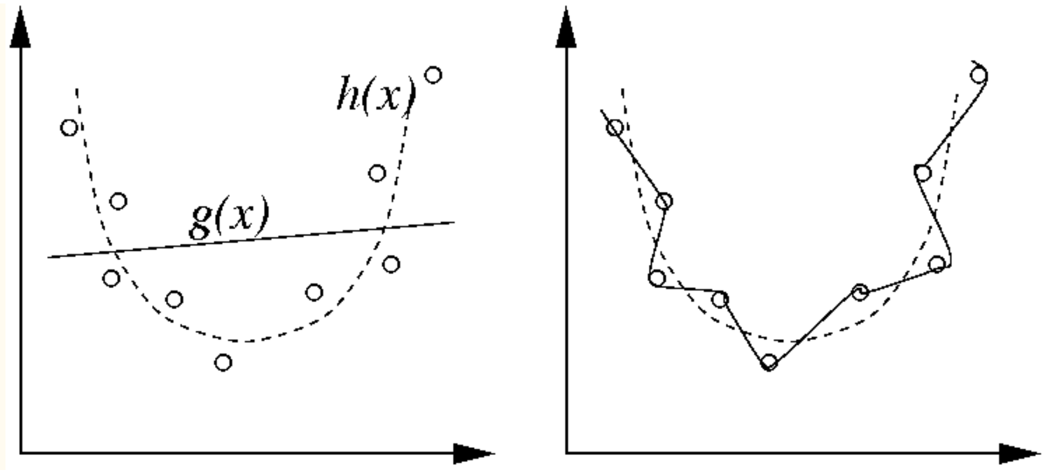
Альтернативный средством против коррелированности предикторов является переход к новым признакам — ортогональным главным компонентам.

Компромисс между смещением и дисперсией (bias-variance trade-off)

$$\mathbf{M}(\hat{\theta}-\theta)^2 = (\mathbf{M}\hat{\theta}-\theta)^2 + \mathbf{D}\hat{\theta}$$

Квадратичный риск = Квадрат смещения + Дисперсия

Регрессия



Штраф за сложность модели

Akaike's an Information Criterion (Akaike, 1973)

Подробности см. в книге K. P. Burnham, R. Anderson «Model Selection and Multimodel Inference»

$$\text{AIC} = -2\ln L + 2K$$

Здесь L — максим. правдоподобие, K — число параметров в модели. Для линейной регрессионной модели с нормальными ошибками имеем

$$-2\log L = n\log\hat{\sigma}^2, \quad \text{где} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\delta_i^2,$$

n — число наблюдений, δ_i^2 — регрессионные остатки, K включает в себя *intercept* (константу) и неизвестную дисперсию ошибок σ^2.

Bayesian Information Criterion (Schwarz, 1978)

$$\text{BIC} = -2\ln L + K\ln n$$

Поскольку $\ln n > 2$ при $n > 7$, штраф BIC больше, чем штраф AIC, поэтому BIC-модели проще

Пример: пошаговая регрессия на основе AIC и BIC

1) Удалите из таблицы p (данные из файла Poverty.txt) строку 25, содержащую «выброс»: q=p[-25,]

2) Постройте линейную регрессионную модель для таблицы q:

   m=lm(q[,3]~., data=q[,-3])

3) Подключите пакет MASS (поддерживающий монографию Venables W. N., Ripley B. D. "Modern Applied Statistics with S"), поставив перед ним «галку» на вкладке Packages

4) Примените к модели m алгоритм регрессии *обратной пошаговой регрессии*: полная модель упрощается до тех пор, пока уменьшается показатель AIC:

   stepAIC(m)

5) Сравните результаты пошаговой регрессии с показателем AIC с результатами для показателя BIC:

   stepAIC(m, k=log(nrow(q)))

6) Установите уровни значимости предикторов в модели, полученной с помощью алгоритма BIC

Нелинейные преобразования

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

[Последняя функция на…
$b > 0$ она возрастает, им…
$y = 1/a$ и перегиб в точке…

Переход к новым перем…
задачу к подгонке прямой y'…

В книге Дж. Литлвуда «Математическая смесь» содержится любопытная классификация углов из книги по альпинизму: «Перпендикулярно — 60º, мой дорогой сэр, абсолютно перпендикулярно — 65º, перпендикулярно — 65º, нависающе — 70º».

Содержательные модели

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

Влияние скрытого фактора

Скрытый фактор. Желание истолковывать регрессионную связь как причинно-следственную может приводить к парадоксам.

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 и ожидалось, Z увеличивалась при уменьшении H и V. Однако (что поначалу представлялось необъяснимым), точность бомбометания Z возрастала также и при увеличении X.

Дальнейший анализ позволил понять причину этого парадокса. Дело оказалось в том, что первоначально в модель не был включен такой важный фактор, как Y — *облачность*. Он сильно влияет и на Z (уменьшая точность), и на X (бессмысленно высылать истребители, если ничего не видно). Сильные отрицательные причинно-следственная связи в парах (Y,Z) и (X,Y) привели к появлению положительного коэффициента при X в линейной регрессионной модели для Z.

Основные этапы регрессионного анализа

1) Выявление одномерных безусловных «выбросов» с помощью диаграмм размахов и их удаление
2) Построение диаграмм рассеяния для поиска двумерных «выбросов», визуального изучения однородности данных и характера связи отклика с каждым предиктором в отдельности
3) Применение монотонных преобразований признаков в случае обнаружения нелинейной зависимости
4) Изучение корреляционных связей предикторов между собой и удаление избыточных предикторов из регрессионной модели
5) Проверка значимости всей линейной модели и каждого из коэффициентов модели в отдельности
6) Использование пошаговой регрессии, если имеет место коррелированность предикторов или их количество велико (во избежание переподгонки модели на обучающей выборке на один предиктор должно приходиться не менее 20 объектов)
7) Анализ остатков: скрытые «выбросы», адекватность модели
8) Выполнение перепроверки на случайной контрольной выборке

Домашнее задание

1) Импортируйте файл Poverty.txt в RStudio под именем p

2) Ради краткости введите переменные n=nrow(p); x=p[,-3]; y=p[,3]

3) Напишите программу для вычисления deleted residuals

(Чтобы набрать код программы, щёлкните по кнопке с белым крестиком внутри зелёного кружка, находящейся в левом верхнем углу экрана и выберите в меню пункт R Script Ctrl+Shift+N. В появившемся окне введите код программы. Для выполнения программы выделите весь код и нажмите кнопку Run с зелёной стрелкой, находящуюся над окном с кодом.)

4) Постройте линейную регрессионную модель m для самих y и x, затем постройте диаграмму рассеяния её регрессионных остатков и deleted residuals с номерами объектов (используйте команду plot с аргументами asp=1 и type="n", затем команду text с аргументом labels=1:n)

5) Выведите на диаграмму прямую с уравнением $y = x$ и нажмите кнопку Zoom. Объект из какой строки является явным «скрытым выбросом»?

6) Удалите из таблицы p строку этого объекта и повторите пункты 2-5. Есть ли другие объекты (строки таблицы), предположительно являющиеся «скрытыми выбросами»? Какие номера имеют соответствующие строки?