

# Прикладная статистика 7. Регрессионный анализ.

11 октября 2013 г.

## Постановка задачи линейной регрессии

$x_1, \dots, x_n$  — объекты;

$f_1, \dots, f_k, y$  — признаки, значения которых измеряются на объектах;

$f_1, \dots, f_k$  — объясняющие переменные, предикторы, регрессоры,  
факторы, признаки;

$y$  — зависимая переменная, отклик.

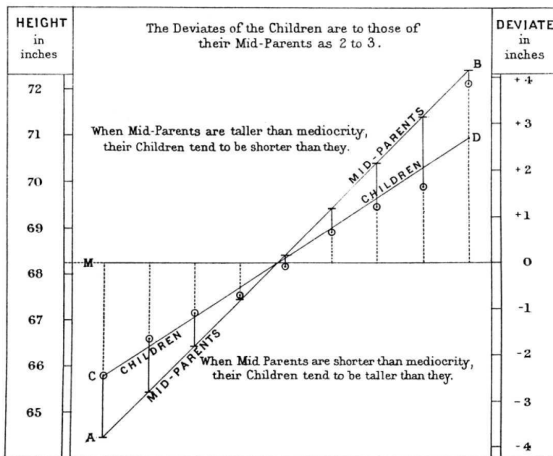
$y \approx F(f_1, \dots, f_k)$  — модель регрессии;

$y \approx \theta_0 + \sum_{j=1}^k \theta_j f_j$  — модель линейной регрессии.

Здесь и далее  $n > k$  (или даже  $n \gg k$ ).

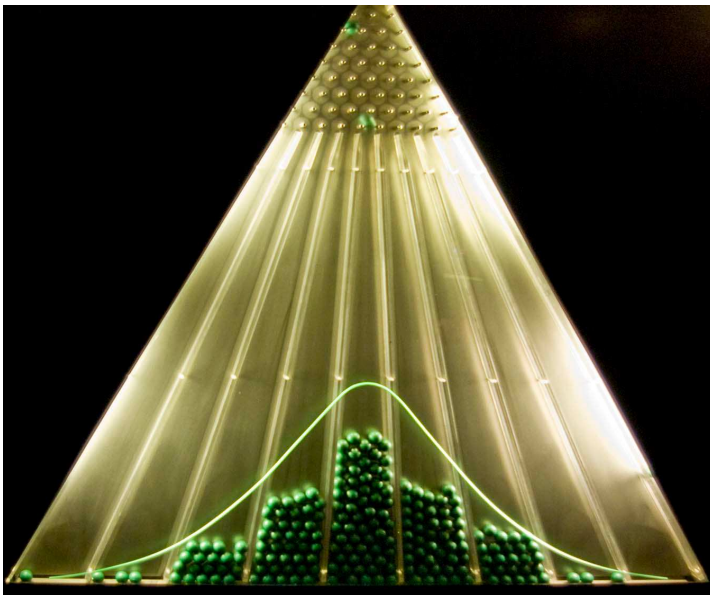
# Первое появление

Впервые такая постановка появляется в работе Гальтона 1885 г. «Регрессия к середине в наследственности роста».



$$y - \bar{y} \approx \frac{2}{3} (x - \bar{x}) .$$

# Машина Гальтона



Матричные обозначения:

$$X = \begin{pmatrix} x_{10} = 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} = 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}; \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}; \quad \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}.$$

Метод наименьших квадратов:

$$\sum_{i=1}^n \left( y_i - \sum_{j=0}^k \theta_j x_{ij} \right)^2 \rightarrow \min_{\theta};$$

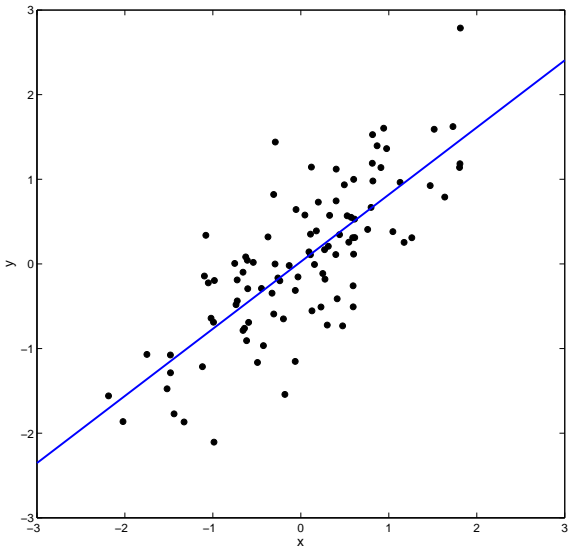
$$\|y - X\theta\|_2^2 \rightarrow \min_{\theta};$$

$$2X^T (y - X\theta) = 0,$$

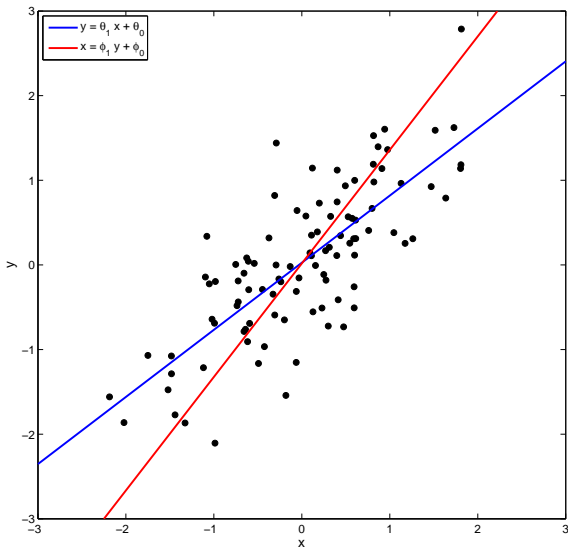
$$\hat{\theta} = (X^T X)^{-1} X^T y,$$

$$\hat{y} = X (X^T X)^{-1} X^T y.$$

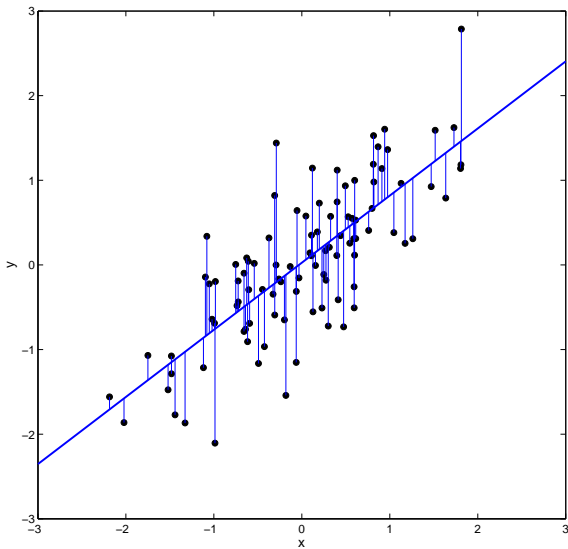
## Инверсия задачи регрессии



## Инверсия задачи регрессии

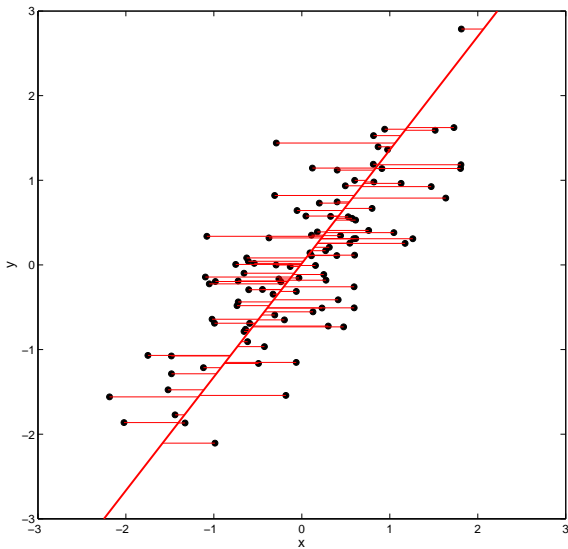


## Инверсия задачи регрессии

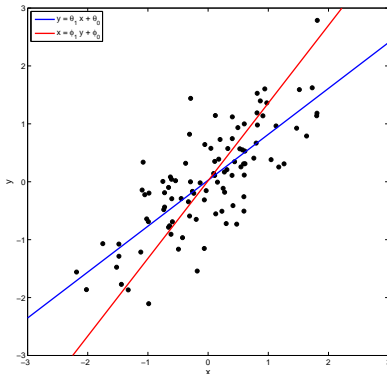




## Инверсия задачи регрессии



## Инверсия задачи регрессии



- Две прямые пересекаются в точке  $(\bar{x}, \bar{y})$ .
- Косинус угла между прямыми, осуществляющими линейную МНК-регрессию  $x$  на  $y$  и  $y$  на  $x$ , равен значению выборочного коэффициента корреляции между  $x$  и  $y$ .

## Goodness-of-fit

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (\text{Total Sum of Squares});$$

$$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (\text{Explained Sum of Squares});$$

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (\text{Residual Sum of Squares});$$

$$TSS = ESS + RSS.$$

Коэффициент детерминации:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

$R^2 = \text{corr}^2(y, \hat{y})$  — квадрат коэффициента множественной корреляции  $y$  с  $X$ .

## Предположения модели

- 1 Линейность:  $y = X\theta + \varepsilon$ .
- 2 Случайность выборки: имеется независимая выборка наблюдений  $(x_i, y_i), i = 1, \dots, n$ .
- 3 Полнота ранга: ни в популяции, ни в выборке ни один из признаков не является линейной комбинацией других ( $\text{rank } X = k$ ).
- 4 Случайность ошибок:  $\mathbb{E}(\varepsilon | X) = 0$ .

В предположениях (1-4) МНК-оценки коэффициентов  $\theta$  являются несмещёнными:

$$\mathbb{E}\hat{\theta}_j = \theta_j, \quad j = 0, \dots, k,$$

и состоятельными:

$$\forall \gamma > 0 \quad \lim_{n \rightarrow \infty} P\left(|\theta_j - \hat{\theta}_j| < \gamma\right) = 1, \quad j = 0, \dots, k.$$

## Предположения модели

- 1 Линейность:  $y = X\theta + \varepsilon$ .
- 2 Случайность выборки: имеется независимая выборка наблюдений  $(x_i, y_i), i = 1, \dots, n$ .
- 3 Полнота ранга: ни в популяции, ни в выборке ни один из признаков не является константой; кроме того, ни один из признаков не является линейной комбинацией других ( $\text{rank } X = k$ ).
- 4 Случайность ошибок:  $\mathbb{E}(\varepsilon | X) = 0$ .
- 5 Гомоскедастичность ошибок: дисперсия ошибки не зависит от значений признаков:  $\mathbb{D}(\varepsilon | X) = \sigma^2$ .

(предположения Гаусса-Маркова).

Теорема Гаусса-Маркова: в предположениях (1-5) МНК-оценки имеют наименьшую дисперсию в классе оценок  $\theta$ , линейных по  $y$ .

Дисперсия  $\hat{\theta}_j$ 

В предположениях (1-5) дисперсии МНК-оценок коэффициентов  $\theta$  задаются следующим образом:

$$\mathbb{D}\hat{\theta}_j = \frac{\sigma^2}{TSS_j (1 - R_j^2)},$$

где  $TSS_j = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$ ,  $R_j^2$  — коэффициент детерминации при регрессии  $x_j$  на все остальные признаки из  $X$ .

- Чем больше дисперсия ошибки  $\sigma^2$ , тем больше дисперсия оценки  $\hat{\theta}_j$ .
- Чем больше вариация значений признака  $x_j$  в выборке, тем меньше дисперсия оценки  $\hat{\theta}_j$ .
- Чем лучше признак  $x_j$  объясняется линейной комбинацией оставшихся признаков, тем больше дисперсия оценки  $\hat{\theta}_j$ .

$R_j^2 < 1$  по предположению (3); тем не менее, может быть  $R_j^2 \approx 1$ .

Дисперсия  $\hat{\theta}_j$ 

В матричном виде:

$$\text{cov } \hat{\theta} = \sigma^2 (X^T X)^{-1}.$$

Если столбцы  $X$  почти линейно зависимы, то матрица  $X^T X$  плохо обусловлена, и дисперсия оценок  $\hat{\theta}_j$  велика.

Близкая к линейной зависимость между двумя или более признаками  $x_j$  называется **мультиколлинеарностью**.

## Неправильное определение модели

**Недоопределение:** если зависимая переменная определяется моделью

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_{j-1} x_{j-1} + \theta_j x_j + \theta_{j+1} x_{j+1} + \cdots + \theta_k x_k,$$

а вместо этого используется модель

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_{j-1} x_{j-1} + \theta_{j+1} x_{j+1} + \cdots + \theta_k x_k,$$

то МНК-оценки  $\hat{\theta}_0, \dots, \hat{\theta}_{j-1}, \hat{\theta}_{j+1}, \dots, \hat{\theta}_k$  являются смещёнными и несостоятельными оценками  $\theta_0, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_k$ .

**Переопределение:** если признак  $x_j$  не влияет на  $y$ , т.е.  $\theta_j = 0$ , то МНК-оценка  $\hat{\theta}$  остаётся несмещённой состоятельной оценкой  $\theta$ , но дисперсия её возрастает.



## Бинарные признаки

Если  $x_j$  принимает только два значения, то они кодируются нулём и единицей. Например, если  $x_j$  — пол испытуемого, то можно задать  $x_j = [\text{пол} = \text{мужской}]$ .

Механизм построения регрессии не меняется.

## Категориальные признаки

Как кодировать дискретные признаки  $x_j$ , принимающие более двух значений?

Пусть  $y$  — средний уровень заработной платы,  $x$  — тип должности (рабочий / инженер / управляющий). Допустим, мы закодировали эти должности следующим образом:

Тип должности	$x$
рабочий	1
инженер	2
управляющий	3

и построили регрессию  $y = \theta_0 + \theta_1 x$ . Тогда для рабочего, инженера и управляющего ожидаемые средние уровни заработной платы определяются следующим образом:

$$y_{bc} = \theta_0 + \theta_1,$$

$$y_{pr} = \theta_0 + 2\theta_1,$$

$$y_{wc} = \theta_0 + 3\theta_1.$$

Согласно построенной модели, разница в средних уровнях заработной платы рабочего и инженера в точности равна разнице между зарплатами инженера и управляющего.

## Фиктивные переменные

Верный способ использования категориальных признаков в регрессии — введение бинарных фиктивных переменных (dummy variables).

Пусть признак  $x_j$  принимает  $m$  различных значений, тогда для его кодирования необходима  $m - 1$  фиктивная переменная.

Способы кодирования:

Тип должности	Dummy		Deviation	
	$x_1$	$x_2$	$x_1$	$x_2$
рабочий	0	0	1	0
инженер	1	0	0	1
управляющий	0	1	-1	-1

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

- При dummy-кодировании коэффициенты  $\theta_1, \theta_2$  оценивают среднюю разницу в уровнях зарплат инженера и управляющего с рабочим.
- При deviation-кодировании коэффициенты  $\theta_1, \theta_2$  оценивают среднюю разницу в уровнях зарплат инженера и управляющего со средним по всем должностям.

## Вопросы

- 1 Как найти доверительные интервалы для  $\theta_j$  и проверить гипотезу  $H_0: \theta_j = 0$ ?
- 2 Как найти доверительный интервал для значений отклика на новом объекте  $y(x_0)$ ?
- 3 Как проверить адекватность построенной модели?

# Предположение о нормальности ошибок

- Нормальность ошибок:  $\varepsilon \sim N(0, \sigma^2)$ .
- В предположениях (1-6) МНК-оценки совпадают с оценками максимального правдоподобия.

ММП:

$$p(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}\varepsilon_i^2},$$

$$\ln \prod_{i=1}^n p(\varepsilon_i) \rightarrow \max_{\theta},$$

$$\sum_{i=1}^n \left( -\frac{1}{2} \ln(2\pi\sigma) - \frac{1}{2\sigma^2} \varepsilon_i^2 \right) \rightarrow \max_{\theta},$$

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \left( y_i - \sum_{j=0}^k \theta_j x_{ij} \right)^2 \rightarrow \min_{\theta}.$$

# Предположение о нормальности ошибок

- Эквивалентная запись предположения (6):

$$y|X \sim N(X\theta, \sigma^2).$$

- МНК-оценки  $\hat{\theta}$  имеют наименьшую дисперсию среди всех несмещённых оценок  $\theta$ .
- МНК-оценки  $\hat{\theta}$  имеют нормальное распределение  $N(\theta, \sigma^2 (X^T X)^{-1})$ .
- Несмещённой оценкой  $\sigma^2$  является

$$\hat{\sigma}^2 = \frac{1}{n - k - 1} RSS;$$

кроме того,  $\frac{1}{\sigma^2} RSS \sim \chi_{n-k-1}^2$ .

- $\forall c \in \mathbb{R}^{k+1}$

$$\frac{c^T (\theta - \hat{\theta})}{\hat{\sigma} \sqrt{c^T (X^T X)^{-1} c}} \sim St(n - k - 1).$$

# Доверительные интервалы

100(1 - α)% доверительный интервал для σ:

$$\sqrt{\frac{RSS}{\chi_{n-k-1, 1-\alpha/2}^2}} \leq \sigma \leq \sqrt{\frac{RSS}{\chi_{n-k-1, \alpha/2}^2}}.$$

Возьмём  $c = \begin{pmatrix} 0 \dots 0 & 1 & 0 \dots 0 \\ & j & \end{pmatrix}$ ; 100(1 - α)% доверительный интервал для  $\theta_j$ :

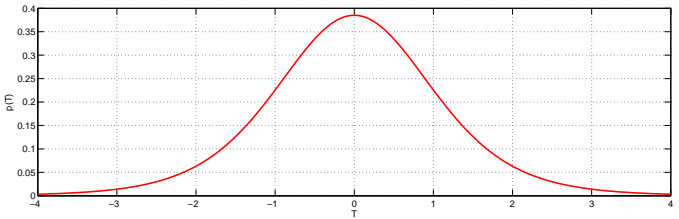
$$\hat{\theta}_j \pm t_{n-k-1, 1-\alpha/2} \hat{\sigma} \sqrt{(X^T X)^{-1}_{jj}}.$$

Для нового объекта  $x_0$  возьмём  $c = x_0$ ; 100(1 - α)% доверительный интервал для  $y(x_0)$ :

$$x_0^T \hat{\theta} \pm t_{n-k-1, 1-\alpha/2} \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0}.$$

# t-критерий Стьюдента

нулевая гипотеза:  $H_0: \theta_j = 0;$   
 альтернатива:  $H_1: \theta_j < \neq > 0;$   
 статистика:  $T = \frac{\hat{\theta}_j}{\sqrt{\frac{RSS}{n-k-1} (X^T X)^{-1}_{jj}}};$   
 $T \sim St(n - k - 1)$  при  $H_0;$



достигаемый уровень значимости:

$$p(t) = \begin{cases} 1 - tcdf(t, n - k - 1), & H_1: \theta_j > 0, \\ tcdf(t, n - k - 1), & H_1: \theta_j < 0, \\ 2(1 - tcdf(|t|, n - k - 1)), & H_1: \theta_j \neq 0. \end{cases}$$



## t-критерий Стьюдента

**Пример:** имеется 12 испытуемых,  $x$  — результат прохождения испытуемым составного теста скорости реакции,  $y$  — результат его теста на симулятора транспортного средства. Проведение составного теста значительно проще и требует меньших затрат, поэтому ставится задача предсказания  $y$  по  $x$ , для чего строится линейная регрессия согласно модели

$$y = \theta_0 + \theta_1 x + \varepsilon.$$

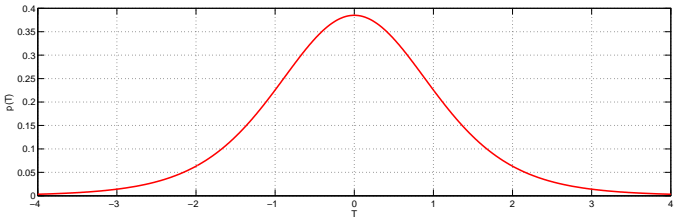
Значима ли переменная  $x$  для предсказания  $y$ ?

$$H_0: \theta_1 = 0.$$

$$H_1: \theta_1 \neq 0 \Rightarrow p = 2.2021 \times 10^{-5}.$$

# t-критерий Стьюдента

нулевая гипотеза:  $H_0: \theta_j = a;$   
 альтернатива:  $H_1: \theta_j < \neq > a;$   
 статистика:  $T = \frac{\hat{\theta}_j - a}{\sqrt{\frac{RSS}{n-k-1} (X^T X)^{-1}_{jj}}};$   
 $T \sim St(n - k - 1)$  при  $H_0;$



достигаемый уровень значимости:

$$p(t) = \begin{cases} 1 - tcdf(t, n - k - 1), & H_1: \theta_j > a, \\ tcdf(t, n - k - 1), & H_1: \theta_j < a, \\ 2(1 - tcdf(|t|, n - k - 1)), & H_1: \theta_j \neq a. \end{cases}$$

## t-критерий Стьюдента

**Пример:** по выборке из 506 жилых районов, расположенных в пригородах Бостона, строится модель средней цены на жильё следующего вида:

$$\ln price = \theta_0 + \theta_1 \ln nox + \theta_2 \ln dist + \theta_3 rooms + \theta_4 stratio + \varepsilon,$$

где  $nox$  — содержание в воздухе двуокиси азота,  $dis$  — взвешенное среднее расстояние от жилого района до пяти основных мест трудоустройства,  $rooms$  — среднее число комнат в доме жилого района,  $stratio$  — среднее отношения числа студентов к числу учителей в школах района.

Коэффициент  $\theta_1$  имеет смысл эластичности цены по признаку  $nox$ . По экономическим соображениям интерес представляет гипотеза о том, что эластичность равна  $-1$ .

$$H_0: \theta_1 = -1.$$

$$H_1: \theta_1 \neq -1 \Rightarrow p = 0.6945.$$

# Критерий Фишера

$$X_{n \times (k+1)} = \begin{pmatrix} X_1 & X_2 \\ n \times (k+1-k_1) & n \times k_1 \end{pmatrix}; \quad \theta^T_{(k+1) \times 1} = \begin{pmatrix} \theta_1^T & \theta_2^T \\ (k+1-k_1) \times 1 & k_1 \times 1 \end{pmatrix}^T;$$

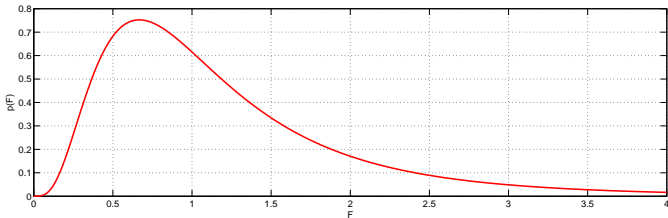
нулевая гипотеза:  $H_0: \theta_2 = 0;$

альтернатива:  $H_1: H_0$  неверна;

статистика:  $RSS_r = \|y - X_1\theta_1\|_2^2, \quad RSS_{ur} = \|y - X\theta\|_2^2,$

$$F = \frac{(RSS_r - RSS_{ur})/k_1}{RSS_{ur}/(n-k-1)};$$

$F \sim F(k_1, n - k - 1)$  при  $H_0;$



достигаемый уровень значимости:

$$p(f) = fcdf(1/f, n - k - 1, k_1).$$

## Критерий Фишера

**Пример:** для веса ребёнка при рождении имеется следующая модель:

$$weight = \theta_0 + \theta_1 cigs + \theta_2 parity + \theta_3 inc + \theta_4 med + \theta_5 fed + \varepsilon,$$

где *cigs* — среднее число сигарет, выкуривавшихся матерью за один день беременности, *parity* — номер ребёнка у матери, *inc* — среднемесячный доход семьи, *med* — длительность в годах получения образования матерью, *fed* — отцом. Данные имеются для 1191 детей.

Зависит ли вес ребёнка при рождении от уровня образования родителей?

$$H_0: \theta_4 = \theta_5 = 0.$$

$$H_1: H_0 \text{ неверна} \Rightarrow p = 0.2421.$$

## Связь между критериями Фишера и Стьюдента

Если  $k_1 = 1$ , критерий Фишера эквивалентен критерию Стьюдента для двусторонней альтернативы.

Иногда критерий Фишера отвергает гипотезу о незначимости признаков  $X_2$ , а критерий Стьюдента не признаёт значимым ни один из них.

Возможные объяснения:

- отдельные признаки из  $X_2$  недостаточно хорошо объясняют  $y$ , но совокупный эффект значим;
- признаки в  $X_2$  мультиколлинеарны.

Иногда критерия Фишера не отвергает гипотезу о незначимости признаков  $X_2$ , а критерий Стьюдента признаёт значимыми некоторые из них.

Возможные объяснения:

- незначимые признаки в  $X_2$  маскируют влияние значимых;
- значимость отдельных признаков в  $X_2$  — результат множественной проверки гипотез.

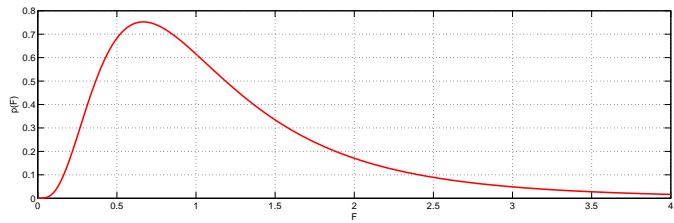
# Критерий Фишера

нулевая гипотеза:  $H_0: \theta_1 = \dots = \theta_k = 0;$

альтернатива:  $H_1: H_0$  неверна;

статистика:  $F = \frac{R^2/k}{(1-R^2)/(n-k-1)};$

$F \sim F(k, n - k - 1)$  при  $H_0;$



достигаемый уровень значимости:

$$p(f) = fcd f(1/f, n - k - 1, k).$$

## Критерий Фишера

**Пример:** имеет ли вообще смысл модель веса ребёнка при рождении, рассмотренная выше?

$$H_0: \theta_1 = \dots = \theta_5 = 0.$$

$$H_1: H_0 \text{ неверна} \Rightarrow p = 6.0331 \times 10^{-9}.$$



# Сравнение невложенных моделей

**Пример:** имеются две модели:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \varepsilon, \quad (1)$$

$$y = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + \varepsilon. \quad (2)$$

Как понять, какая из них лучше?

# Критерий Давидсона-МакКиннона

Пусть  $\hat{y}$  — оценка отклика по первой модели,  $\hat{\hat{y}}$  — по второй.  
 Подставим эти оценки как признаки в чужие модели:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 \hat{y} + \varepsilon,$$

$$y = \theta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + \beta_3 \hat{\hat{y}} + \varepsilon.$$

При помощи критерия Стьюдента проверим

$$H_{01} : \theta_3 = 0, \quad H_{11} : \theta_3 \neq 0,$$

$$H_{02} : \beta_3 = 0, \quad H_{12} : \beta_3 \neq 0.$$

$H_{01} \backslash H_{02}$	Принята	Отвергнута
Принята	Модели одинаково хороши	Модель (1) значительно лучше
Отвергнута	Модель (2) значительно лучше	Модели одинаково плохи

## Приведённый коэффициент детерминации

Стандартный коэффициент детерминации всегда увеличивается при добавлении регрессоров в модель, поэтому для отбора признаков его использовать нельзя.

Для сравнения моделей, содержащих разное число признаков, можно использовать приведённый коэффициент детерминации:

$$R_a^2 = \frac{ESS/(n - k - 1)}{TSS/(n - 1)} = 1 - (1 - R^2) \frac{n - 1}{n - k - 1}.$$

## Пошаговая регрессия

- **Шаг 0.** Настраивается модель с одной только константой, а также все модели с одной переменной. Рассчитывается  $F$ -статистика каждой модели и достигаемый уровень значимости. Выбирается модель с наименьшим достигаемым уровнем значимости. Соответствующая переменная  $X_{e1}$  включается в модель, если этот достигаемый уровень значимости меньше порогового значения  $p_E = 0.05$ .
- **Шаг 1.** Рассчитывается  $F$ -статистика и достигаемый уровень значимости для всех моделей, содержащих две переменные, одна из которых  $X_{e1}$ . Аналогично принимается решение о включении  $X_{e2}$ .
- **Шаг 2.** Если была добавлена переменная  $X_{e2}$ , возможно,  $X_{e1}$  уже не нужна. В общем случае просчитываются все возможные варианты исключения одной переменной, рассматривается вариант с наибольшим достигаемым уровнем значимости, соответствующая переменная исключается, если он превосходит пороговое значение  $p_R = 0.1$ .
- ...

David A. Freedman, A Note on Screening Regression Equations. The American Statistician, Vol. 37, No. 2 (May, 1983), pp. 152-155.

## Отбор признаков с учётом эффекта множественной проверки гипотез

$$\forall c_1, \dots, c_{k_1} \in \mathbb{R}^{k+1}$$

$$t_j = \frac{c_j^T (\theta - \hat{\theta})}{\hat{\sigma} \sqrt{c_j^T (X^T X)^{-1} c_j}}, \quad j = 1, \dots, k_1$$

имеют совместное распределение Стьюдента с числом степеней свободы  $n - k - 1$  и корреляционной матрицей

$$R = DC^T (X^T X)^{-1} CD,$$
$$C = (c_1, \dots, c_{k_1}),$$
$$D = \text{diag} \left( c_j^T (X^T X)^{-1} c_j \right)^{-\frac{1}{2}}.$$

Для одновременной проверки значимости всех коэффициентов регрессии достаточно взять в качестве  $C$  единичную матрицу.

## Отбор признаков с учётом эффекта множественной проверки гипотез

Matlab:

```
[theta,~,stats] = glmfit(X,y,'normal');  
D      = diag(1 ./ sqrt(diag(stats.covb)));  
Cor    = D * stats.covb * D';  
p_adj  = 1 - mvtcdf(repmat(-abs(stats.t), 1, length(theta)), ...  
                    repmat(abs(stats.t), 1, length(theta)), ...  
                    Cor, stats.dfe);
```

Работает при  $k + 1 \leq 25$ .

## Отбор признаков с учётом эффекта множественной проверки гипотез

R, длинный способ:

```
m      <- lm(y ~ X)
theta  <- coef(m)
Vtheta <- vcov(m)
D      <- diag(1 / sqrt(diag(Vtheta)))
t      <- D %*% theta
Cor    <- D %*% Vtheta %*% t(D)
library("mvtnorm")
m.df   <- nrow(X) - length(theta)
p_adj  <- sapply(abs(t), function(x) 1-pmvt(-rep(x, length(theta)),
                                             rep(x, length(theta)),
                                             corr = Cor, df = m.df))
```

R, короткий способ:

```
m      <- lm(y ~ X)
theta  <- coef(m)
library("multcomp")
m.mc   <- glht(m, linfct = diag(length(theta)))
summary(m.mc)
```



## Проверка предположений Гаусса-Маркова

- Предположения (1-2) проверить нельзя.
- Предположение (3) легко проверяется, без его выполнения построить модель вообще невозможно.
- Предположения (4-6) об ошибке  $\varepsilon$  необходимо проверять.

Оценивать ошибку  $\varepsilon$  будем при помощи **остатков**:

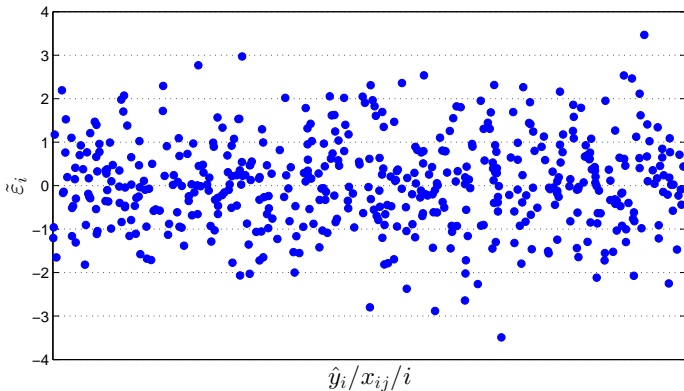
$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = 1, \dots, n.$$

# Визуальный анализ

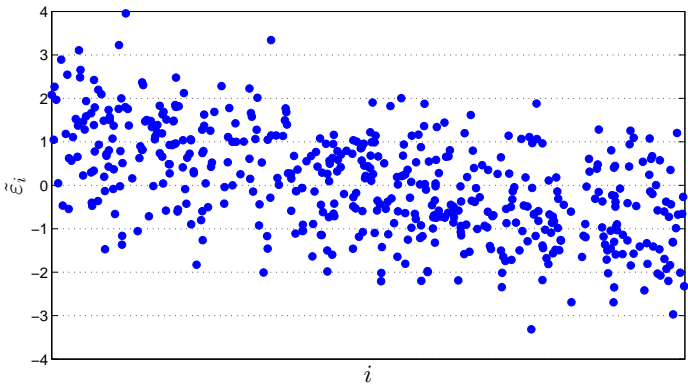
Стандартизированные остатки:

$$\tilde{\varepsilon}_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}}, \quad i = 1, \dots, n.$$

Строятся графики зависимости  $\tilde{\varepsilon}_i$  от  $\hat{y}_i$ ,  $x_{ij}, j = 1, \dots, k, i$ .

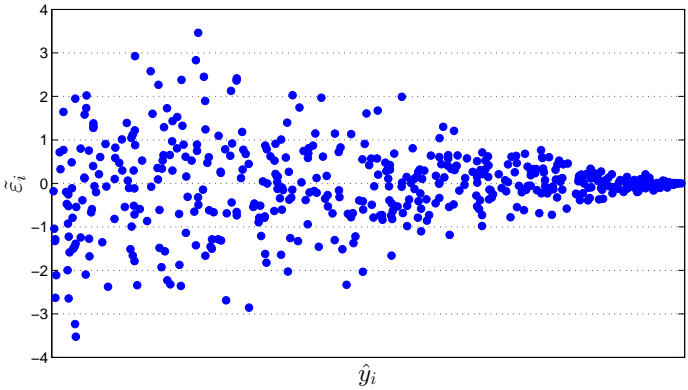


# Визуальный анализ



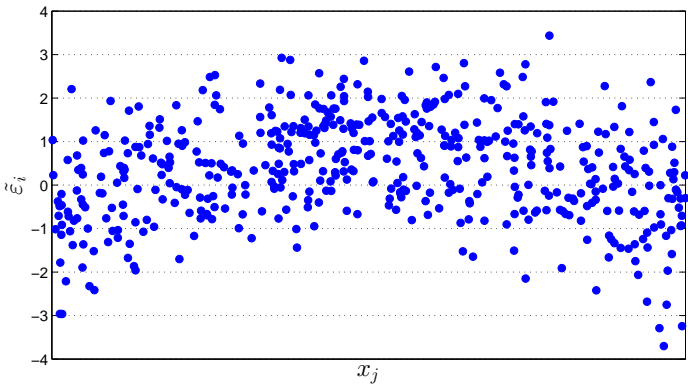
В данных имеется тренд

# Визуальный анализ



Гетероскедастичность

# Визуальный анализ



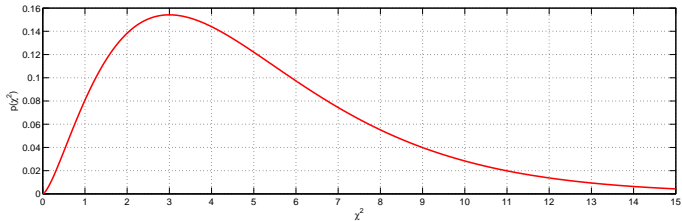
Стоит добавить квадрат признака  $x_j$

## Формальные критерии

- Проверка нормальности — занятие 2.
- Проверка несмещённости: если остатки нормальны — критерий Стьюдента (занятие 2), нет — непараметрический критерий (занятие 3).
- Проверка гомоскедастичности: критерий Бройша-Пагана.

# Критерий Бройша-Пагана

нулевая гипотеза:  $H_0: \mathbb{D}\varepsilon_i = \sigma^2$ ;  
 альтернатива:  $H_1: H_0$  неверна;  
 статистика:  $LM = nR_{\varepsilon^2}^2$ ,  $R_{\varepsilon^2}^2$  — коэффициент детерминации при регрессии квадратов остатков на признаки;  
 $LM \sim \chi_k^2$  при  $H_0$ ;



достигаемый уровень значимости:

$$p(f) = 1 - \text{chi2cdf}(LM, k).$$

Прикладная статистика  
7. Регрессионный анализ.

Рябенко Евгений  
riabenko.e@gmail.com