



## Постановка задачи линейной регрессии

$1, \dots, n$  — объекты;

$x_1, \dots, x_k, y$  — признаки, значения которых измеряются на объектах;

$x_1, \dots, x_k$  — объясняющие переменные (предикторы, регрессоры, факторы, признаки);

$y$  — зависимая переменная, отклик.

Хотим найти такую функцию  $f$ , что  $y \approx f(x_1, \dots, x_k)$ ;

$$\operatorname{argmin}_f \mathbb{E} (y - f(x_1, \dots, x_k))^2 = \mathbb{E} (y | x_1, \dots, x_k).$$

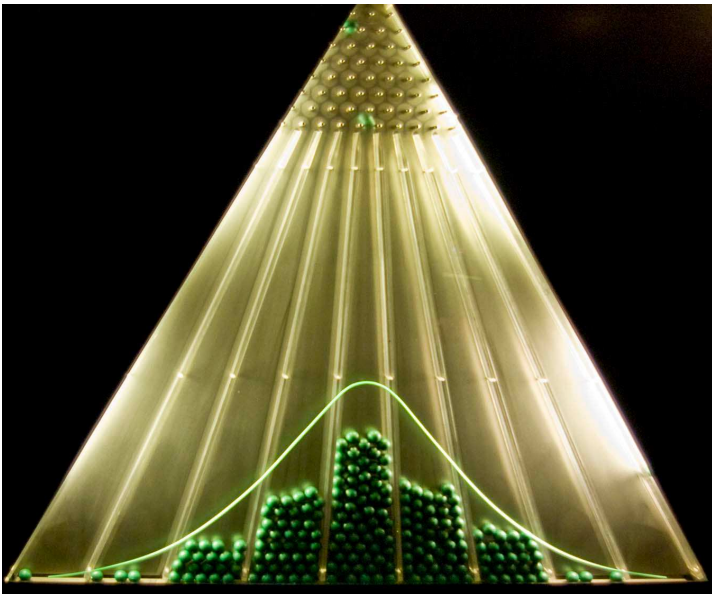
$\mathbb{E} (y | x_1, \dots, x_k) = f(x_1, \dots, x_k)$  — модель регрессии;

$\mathbb{E} (y | x_1, \dots, x_k) = \beta_0 + \sum_{j=1}^k \beta_j x_j$  — модель линейной регрессии.

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



# Машина Гальтона



## Метод наименьших квадратов (МНК)

Матричные обозначения:

$$X = \begin{pmatrix} x_{10} = 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} = 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}; \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}; \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Метод наименьших квадратов:

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

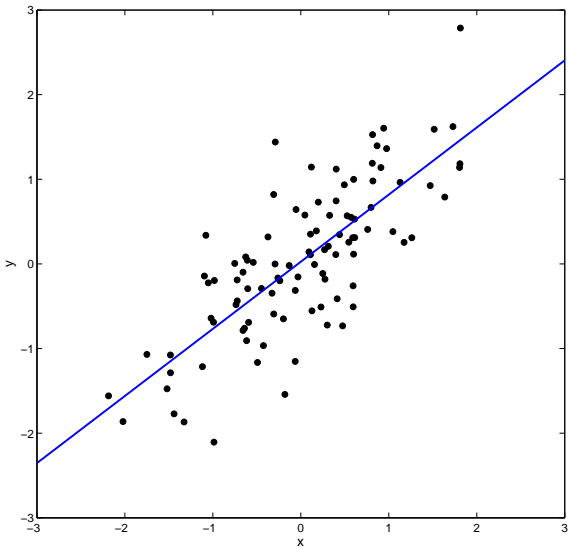
$$\|y - X\beta\|_2^2 \rightarrow \min_{\beta};$$

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

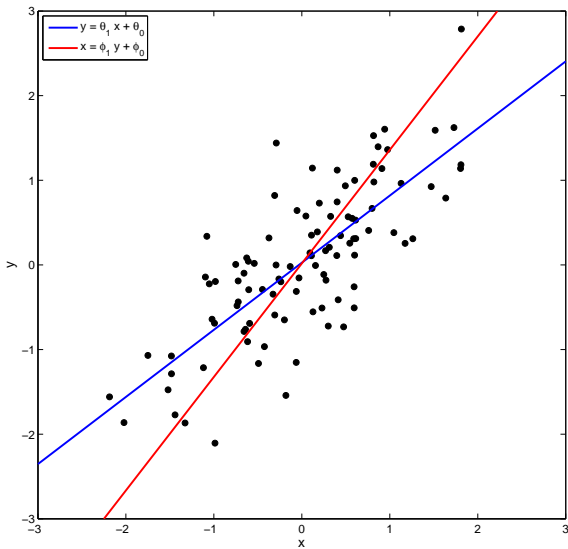
$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

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

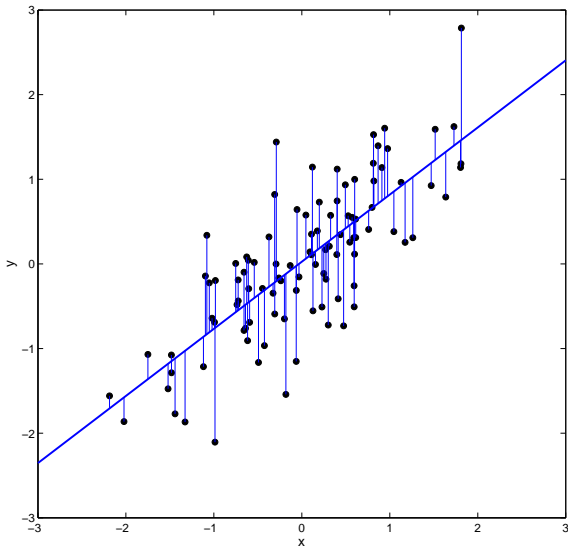
# Инверсия задачи регрессии



## Инверсия задачи регрессии

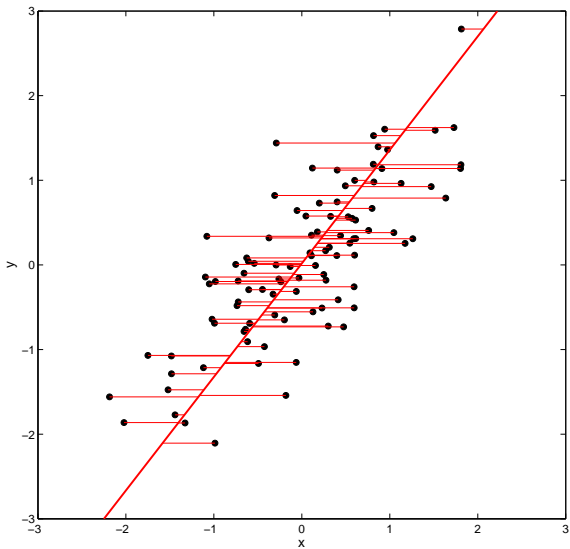


## Инверсия задачи регрессии

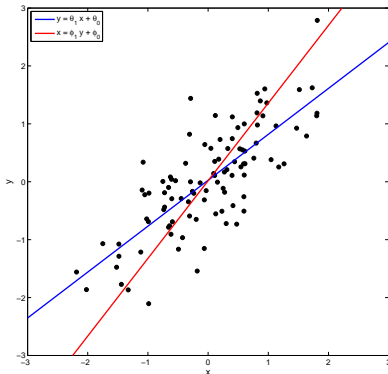




# Инверсия задачи регрессии



# 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:

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

$R^2 = r_{y\hat{y}}^2$  — квадрат коэффициента множественной корреляции  $y$  с  $X$ .

## Предположения модели

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

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

$$\mathbb{E}\hat{\beta}_j = \beta_j, j = 0, \dots, k,$$

и состоятельными:

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

## Предположения модели

- 1 Линейность:  $y = X\beta + \varepsilon$ .
- 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  $\beta$ , линейных по  $y$ .

Дисперсия  $\hat{\beta}_j$ 

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

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

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

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

Дисперсия  $\hat{\beta}_j$ 

В матричном виде:

$$\mathbb{D}(\hat{\beta} | X) = \sigma^2 (X^T X)^{-1}.$$

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

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

## Неправильное определение модели

**Недоопределение:** если зависимая переменная определяется моделью

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_{j-1} x_{j-1} + \beta_j x_j + \beta_{j+1} x_{j+1} + \dots + \beta_k x_k,$$

а вместо этого используется модель

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_{j-1} x_{j-1} + \beta_{j+1} x_{j+1} + \dots + \beta_k x_k,$$

то МНК-оценки  $\hat{\beta}_0, \dots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \dots, \hat{\beta}_k$  являются смещёнными и несостоятельными оценками  $\beta_0, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k$ .

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





## 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  $y = \beta_0 + \beta_1 x$ . Тогда для рабочего, инженера и управляющего ожидаемые средние уровни заработной платы определяются следующим образом:

$$y_{bc} = \beta_0 + \beta_1,$$

$$y_{pr} = \beta_0 + 2\beta_1,$$

$$y_{wc} = \beta_0 + 3\beta_1.$$

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- При dummy-кодировании коэффициенты  $\beta_1, \beta_2$  оценивают среднюю разницу в уровнях зарплат инженера и управляющего с рабочим.
- При deviation-кодировании коэффициенты  $\beta_1, \beta_2$  оценивают среднюю разницу в уровнях зарплат инженера и управляющего со средним по всем должностям.



## Предположение о нормальности ошибок

ⓘ Нормальность ошибок:  $\varepsilon \sim N(0, \sigma^2)$ .

- В предположениях (1-6) МНК-оценки совпадают с оценками максимального правдоподобия.

ММП:

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

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

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

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

## Предположение о нормальности ошибок

- Эквивалентная запись предположения (6):

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

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

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

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

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

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

Доверительные интервалы для  $\sigma$  и  $\beta$ 

100(1 -  $\alpha$ )% доверительный интервал для  $\sigma$ :

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

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

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

## Доверительный и предсказательный интервалы для отклика

Для нового объекта  $x_0$  возьмём  $c = x_0$ ;  $100(1 - \alpha)\%$  доверительный интервал для среднего отклика  $\bar{y}(x_0) = x_0^T \beta$ :

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

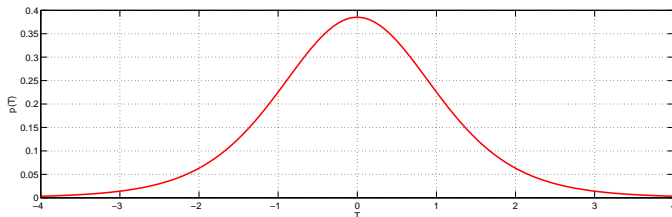
Чтобы построить предсказательный интервал для  $y(x_0) = x_0^T \beta + \varepsilon(x_0)$ , учтём ещё дисперсию ошибки:

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



## t-критерий Стьюдента

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



достигаемый уровень значимости:

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

## 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

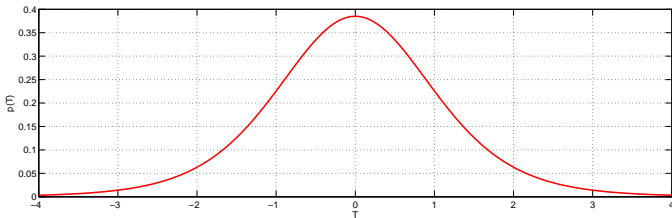
$$y = \beta_0 + \beta_1 x + \varepsilon.$$

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

$$H_0: \beta_1 = 0.$$

$$H_1: \beta_1 \neq 0 \Rightarrow p = 2.2021 \times 10^{-5}.$$

## t-критерий Стьюдента

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

достигаемый уровень значимости:

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

## t-критерий Стьюдента

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

$$\ln price = \beta_0 + \beta_1 \ln nox + \beta_2 \ln dist + \beta_3 rooms + \beta_4 stratio + \varepsilon,$$

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  $\beta_1$  имеет смысл эластичности цены по признаку  $nox$ . По экономическим соображениям интерес представляет гипотеза о том, что эластичность равна  $-1$ .

$$H_0: \beta_1 = -1.$$

$$H_1: \beta_1 \neq -1 \Rightarrow p = 0.6945.$$

## Критерий Фишера

$$X_{n \times (k+1)} = \begin{pmatrix} X_1 & X_2 \\ n \times (k+1-k_1) & n \times k_1 \end{pmatrix}; \quad \beta^T_{(k+1) \times 1} = \begin{pmatrix} \beta_1^T & \beta_2^T \\ (k+1-k_1) \times 1 & k_1 \times 1 \end{pmatrix}^T;$$

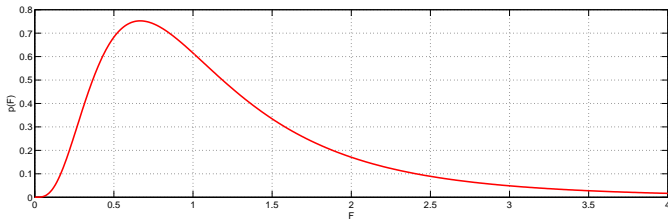
нулевая гипотеза:  $H_0: \beta_2 = 0$ ;

альтернатива:  $H_1: H_0$  неверна;

статистика:  $RSS_r = \|y - X_1\beta_1\|_2^2$ ,  $RSS_{ur} = \|y - X\beta\|_2^2$ ,

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

$F \sim F(k_1, n - k - 1)$  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:

$$weight = \beta_0 + \beta_1cigs + \beta_2parity + \beta_3inc + \beta_4med + \beta_5fed + \varepsilon,$$

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?

$$H_0: \beta_4 = \beta_5 = 0.$$

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

## 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.

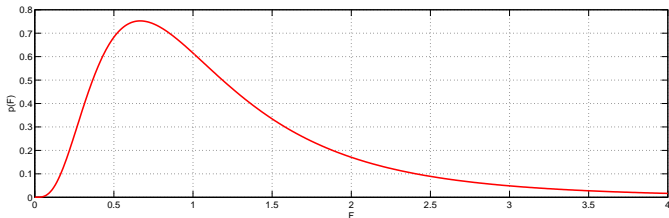
# Критерий Фишера

нулевая гипотеза:  $H_0: \beta_1 = \dots = \beta_k = 0;$

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:

$$p(f) = f \text{cdf}(1/f, n - k - 1, k).$$



# Критерий Фишера

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

$$H_0: \beta_1 = \dots = \beta_5 = 0.$$

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

# Сравнение невложенных моделей

**Пример:** имеются две модели:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \quad (1)$$

$$y = \gamma_0 + \gamma_1 \log x_1 + \gamma_2 \log x_2 + \varepsilon. \quad (2)$$

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:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \hat{y} + \varepsilon,$$

$$y = \beta_0 + \gamma_1 \log x_1 + \gamma_2 \log x_2 + \gamma_3 \hat{\hat{y}} + \varepsilon.$$

При помощи критерия Стьюдента проверим

$$H_{01}: \beta_3 = 0, \quad H_{11}: \beta_3 \neq 0,$$

$$H_{02}: \gamma_3 = 0, \quad H_{12}: \gamma_3 \neq 0.$$

$H_{01} \backslash H_{02}$	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  $p_R = 0.1$ .
- ...

# Эксперимент Фридмана

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

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

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

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:

```
[beta,~,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(beta)), ...  
                    repmat(abs(stats.t), 1, length(beta)), ...  
                    Cor, stats.dfe);
```

Работает при  $k + 1 \leq 25$ .



## Отбор признаков с учётом эффекта множественной проверки гипотез

R, длинный способ:

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

R, короткий способ:

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

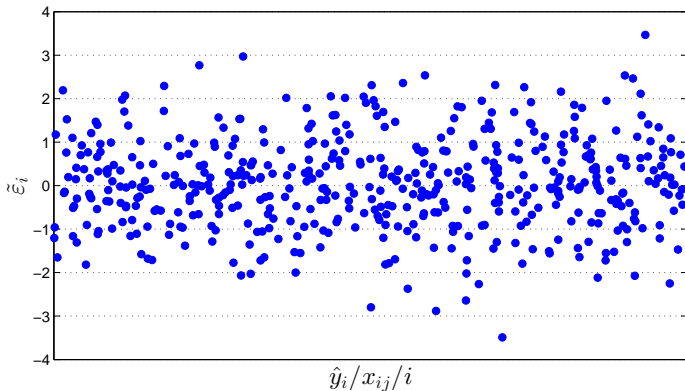
$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = 1, \dots, n.$$

# Визуальный анализ

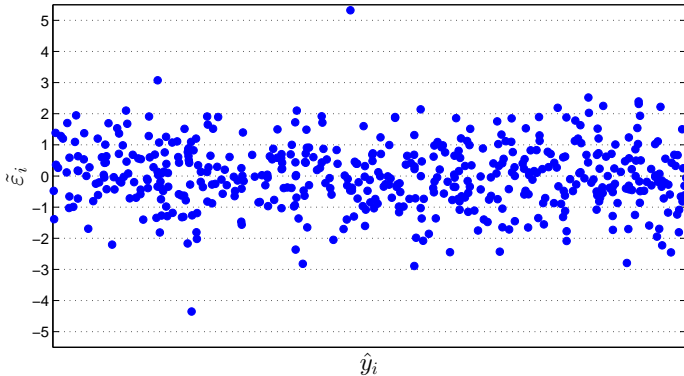
Стандартизированные остатки:

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

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

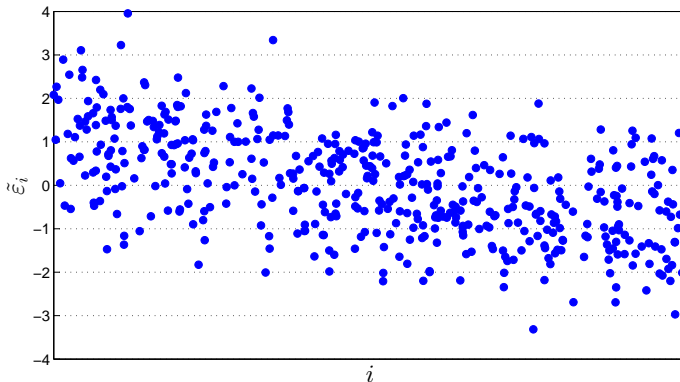


## Визуальный анализ



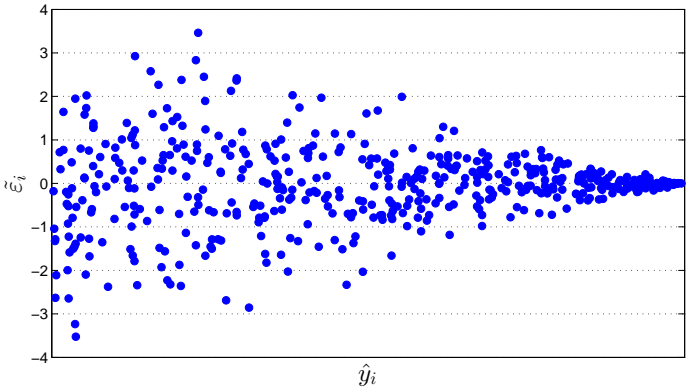
Возможно, присутствуют выбросы

## Визуальный анализ



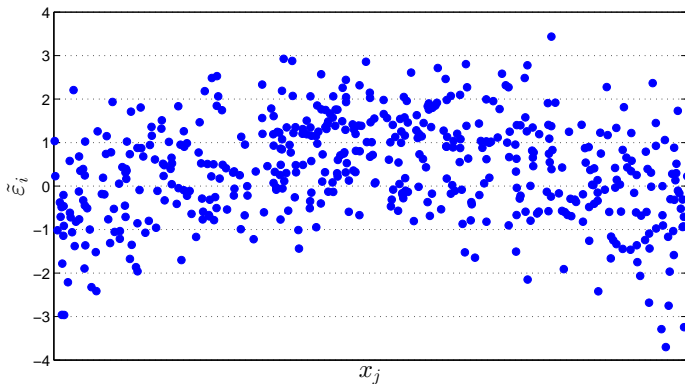
В данных имеется тренд

# 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).
- Выбросы — расстояние Кука.
- Проверка гомоскедастичности: критерий Бройша-Пагана.



## Расстояние Кука

Остатки недостаточно полно характеризуют наличие выбросов, так как регрессия сильно подстраивается под большие отклонения.

Расстояние Кука — мера воздействия  $i$ -го наблюдения на регрессионное уравнение:

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{k \cdot RSS} = \frac{\hat{\varepsilon}_i^2}{k \cdot RSS} \frac{h_i}{(1 - h_i)^2},$$

$\hat{y}_{j(i)}$  — предсказания модели, настроенной по наблюдениям  $1, \dots, i-1, i+1, \dots, n$ , для наблюдения  $j$ ;

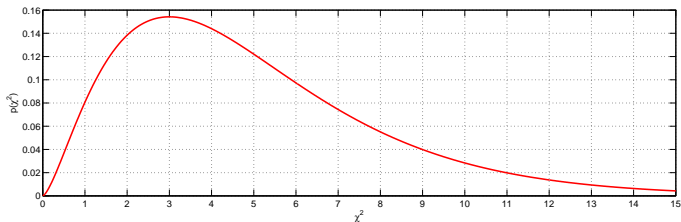
$h_i$  — диагональный элемент матрицы  $H = X(X^T X)^{-1} X^T$  (hat matrix).

Варианты порога на  $D_i$ :

- $D_i = 1$ ;
- $D_i = 4/n$ ;
- $D_i = 3\bar{D}$ ;
- визуально по графику зависимости  $D_i$  от  $\hat{y}_i$ .

## 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:

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

# Гетероскедастичность

Гетероскедастичность может быть следствием недоопределения модели.

Последствия гетероскедастичности:

- нарушаются предположения критериев Стьюдента и Фишера и методов построения доверительных интервалов для  $\sigma$  и  $\beta$  (независимо от объёма выборки);
- МНК-оценки  $\beta$  и  $R^2$  остаются несмещёнными и состоятельными.

Варианты:

- переопределить модель, добавить признаки, преобразовать отклик;
- использовать модифицированные оценки дисперсии коэффициентов для оценки значимости;
- настроить параметры методом взвешенных наименьших квадратов.

## Преобразование Бокса-Кокса

Пусть значения отклика  $y_1, \dots, y_n$  положительны. Если  $\frac{\max y_i}{\min y_i} > 10$ , стоит рассмотреть возможность преобразования  $y$ . В каком виде его искать?

Часто полезно рассмотреть преобразования вида  $y^\lambda$ , но оно не имеет смысла при  $\lambda = 0$ .

Вместо него можно рассмотреть семейство преобразований

$$W = \begin{cases} (y^\lambda - 1) / \lambda, & \lambda \neq 0, \\ \ln y, & \lambda = 0, \end{cases}$$

но оно сильно варьируется по  $\lambda$ .

Вместо него можно рассмотреть семейство преобразований

$$V = \begin{cases} (y^\lambda - 1) / (\lambda \dot{y}^{\lambda-1}), & \lambda \neq 0, \\ \dot{y} \ln y, & \lambda = 0, \end{cases}$$

где  $\dot{y} = (y_1 y_2 \dots y_n)^{1/n}$  — среднее геометрическое наблюдений отклика.

## Метод Бокса-Кокса

Процесс подбора  $\lambda$ :

- 1 выбирается набор значений  $\lambda$  в некотором интервале, например,  $(-2, 2)$ ;
- 2 для каждого значения  $\lambda$  выполняется преобразование отклика  $V$ , строится регрессия  $V$  на  $X$ , вычисляется остаточная сумма квадратов  $RSS(\lambda)$ ;
- 3 строится график зависимости  $RSS(\lambda)$  от  $\lambda$ , по нему выбирается оптимальное значение  $\lambda$ ;
- 4 выбирается ближайшее к оптимальному удобное значение  $\lambda$  (например, целое или полуцелое);
- 5 строится окончательная регрессионная модель с откликом  $y^\lambda$  или  $\ln y$ .

Доверительный интервал для  $\lambda$  определяется как пересечение кривой  $RSS(\lambda)$  с линией уровня  $\min_{\lambda} RSS(\lambda) \cdot e^{\chi_{1,1-\alpha}^2/n}$ . Если он содержит единицу, возможно, не стоит выполнять преобразование.

## Устойчивая оценка дисперсии Уайта

Если не удаётся избавиться от гетероскедастичности, для оценки значимости признаков можно использовать критерии, основанные на устойчивой оценке дисперсии.

White's heteroscedasticity-consistent estimator (HCE):

$$\mathbb{D}(\hat{\beta} | X) = (X^T X)^{-1} (X^T \text{diag}(\hat{\varepsilon}_1^2, \dots, \hat{\varepsilon}_n^2) X) (X^T X)^{-1}.$$

Асимптотика устойчивой оценки:

$$\sqrt{n}(\beta - \hat{\beta}) \xrightarrow{d} N(0, \Omega),$$

$$\hat{\Omega} = n (X^T X)^{-1} (X^T \text{diag}(\hat{\varepsilon}_1^2, \dots, \hat{\varepsilon}_n^2) X) (X^T X)^{-1}.$$



## Использование устойчивых оценок дисперсии

R, пакет sandwich:

```
m <- lm(y ~ X)
library("sandwich")
library("lmtest")

#significance of every predictor
coeftest(m, df = Inf, vcov = vcovHC(m, type = "HCO"))

#significance of the group of predictors
waldtest(m1, m2, vcov = vcovHC(m1, type = "HCO")) #m1 - bigger model

#significance of the whole equation
waldtest(m, vcov = vcovHC(m, type = "HCO"))
```

Matlab:

```
beta = glmfit(X,y,'normal');
Cov = hac(X5,log(y),'type','HC','weights','HCO');

%significance of every predictor
p = 2 * pnorm(abs(stats5.beta ./ sqrt(diag(Cov))))

%significance of the group of predictors
test_ind = [9]; %indices of predictors to be tested
p = 1-chi2cdf(beta(test_ind)' * inv(Cov(test_ind, test_ind)) * beta(test_ind), ...
             length(test_ind))

%significance of the whole equation
p = 1-chi2cdf(beta(2:end)' * inv(Cov(2:end, 2:end)) * beta(2:end), length(beta)-1)
```



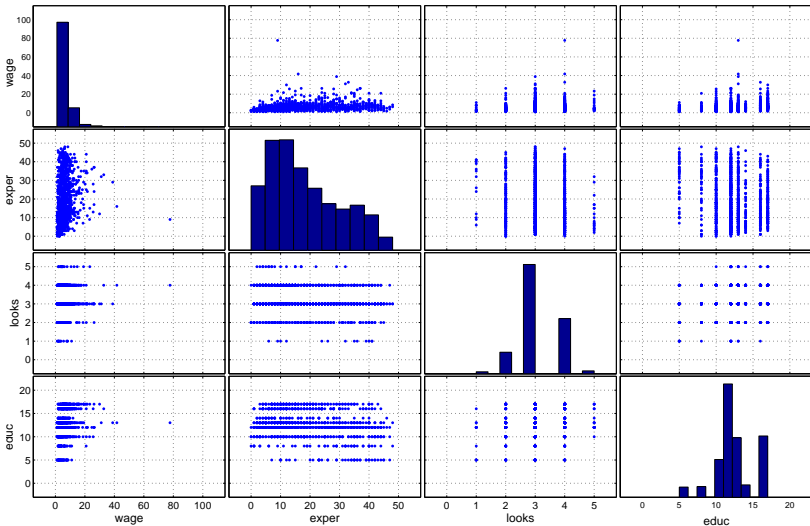
## Влияние внешней привлекательности на уровень заработка

Hamermesh, D. S., and J. E. Biddle (1994), *Beauty and the Labor Market*, *American Economic Review* 84, 1174–1194: по 1260 опрошенным имеются следующие данные:

- заработная плата за час работы, \$,
- опыт работы, лет,
- образование, лет,
- внешняя привлекательность, в баллах от 1 до 5,
- бинарные признаки: пол, семейное положение, состояние здоровья (хорошее/плохое), членство в профсоюзе, цвет кожи (белый/чёрный), занятость в сфере обслуживания (да/нет).

Оценить влияние внешней привлекательности на уровень заработка с учётом всех остальных факторов.

# Данные

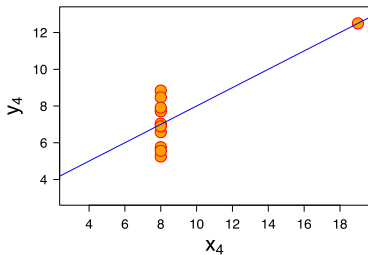
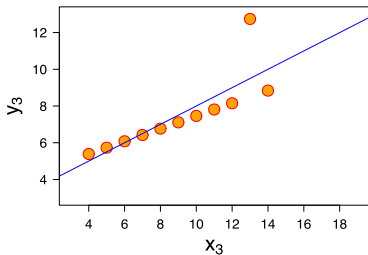
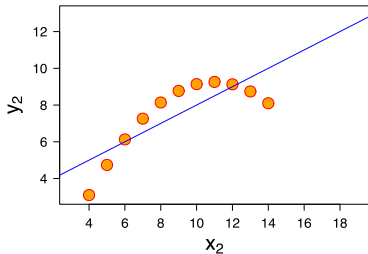
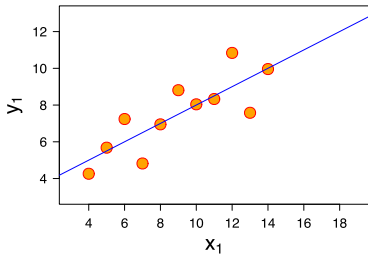


# О необходимости визуализации данных

Пример: Anscombe's quartet

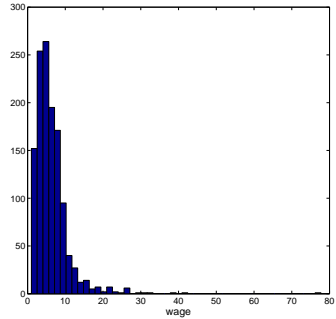
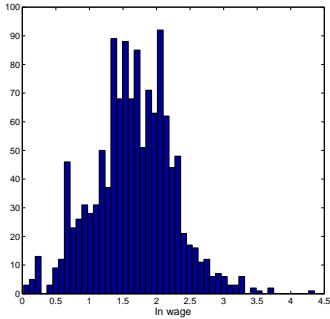
№	1	2	3	4
$\bar{x}$	9	9	9	9
$S(x)$	11	11	11	11
$\bar{y}$	7.5	7.5	7.5	7.5
$S(y)$	4.127	4.127	4.128	4.128
$r_{xy}$	0.816	0.816	0.816	0.816

## О необходимости визуализации данных





# Выбросы

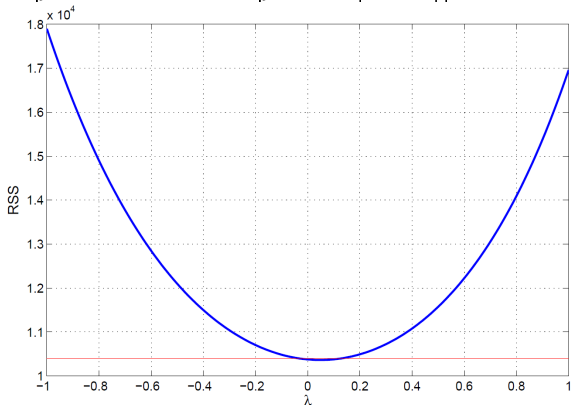


Больше 30 долларов в час в выборке получают только 5 человек.  
Исключим их.

## Преобразование отклика

$$\frac{\max y}{\min y} = 29.4.$$

Найдём преобразование отклика при помощи метода Бокса-Кокса:



95% доверительный интервал:  $(-0.028, 0.124)$ .

Возьмём  $\lambda = 0$ , т. е. будем делать регрессию логарифма отклика.

## Модель 1

Построим линейную модель:

$$\begin{aligned} \ln wage = & 0.43 + 0.01exper + 0.19union + 0.12goodhlth - 0.10black - \\ & - 0.39female + 0.04married - 0.15service + 0.08educ - \\ & - 0.01aboveavg - 0.13belowavg. \end{aligned}$$

$$F = 78.63, p = 6.8 \times 10^{-125}, R^2 = 0.387, R_a^2 = 0.382.$$

Критерий	p-value
Шапиро-Уилка (нормальность)	$1.1 \times 10^{-4}$
знаковых рангов (несмещённость)	0.8944
Бройша-Пагана (гомоскедастичность)	$5.7 \times 10^{-4}$

Признаки, коэффициенты при которых значимо отличаются от нуля (множественная проверка с дисперсиями Уайта): *exper*, *union*, *female*, *service*, *educ*, *belowavg*.



# Модель 2

Редуцированная модель:

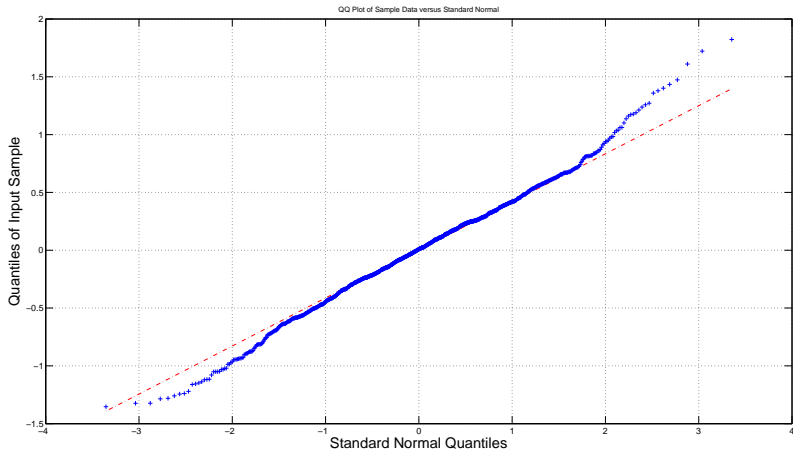
$$\ln wage = 0.54 + 0.01exper + 0.18union - 0.41female - 0.15service + 0.08educ - 0.008aboveavg - 0.12belogavg.$$

$$F = 110.04, p = 1.5 \times 10^{-125}, R^2 = 0.382, R_a^2 = 0.378.$$

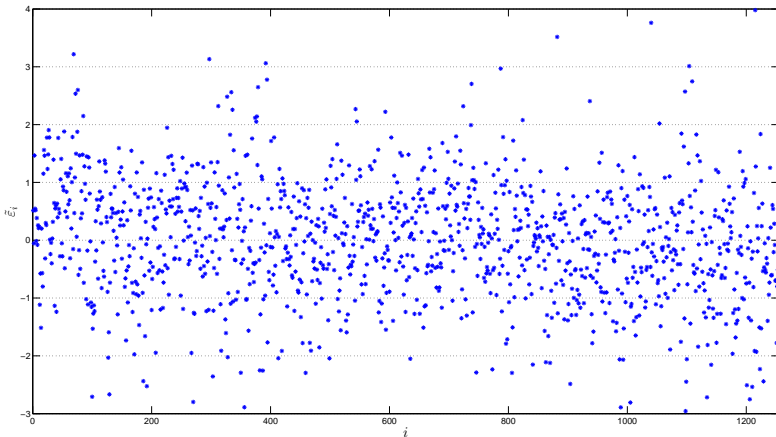
Критерий	p-value
Шапиро-Уилка (нормальность) знаковых рангов (несмещённость)	$1.6 \times 10^{-4}$ 0.8480
Бройша-Пагана (гомоскедастичность)	$4.0 \times 10^{-5}$

Значимы все признаки, кроме *aboveavg* (множественная проверка с дисперсиями Уайта).

## Остатки модели 2

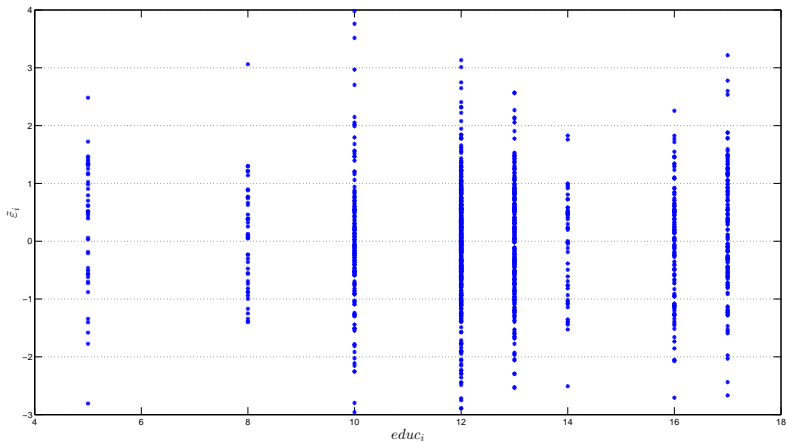


# Остатки модели 2

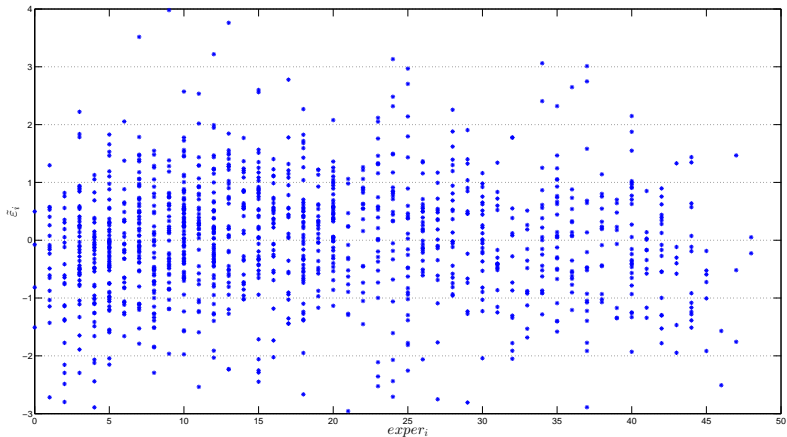




## Остатки модели 2



## Остатки модели 2

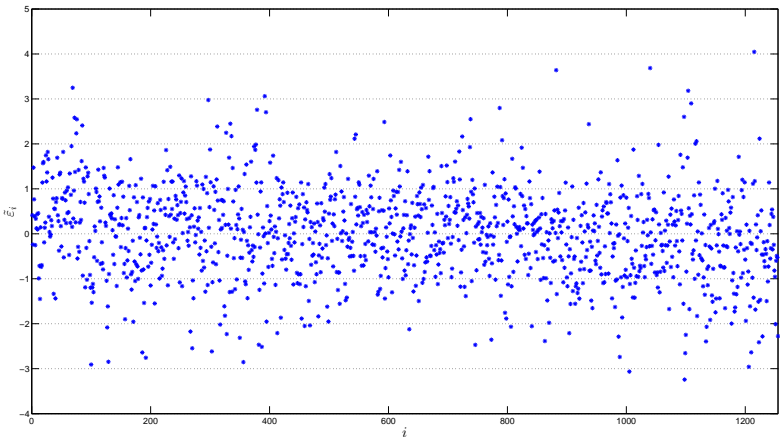








# Остатки модели 3

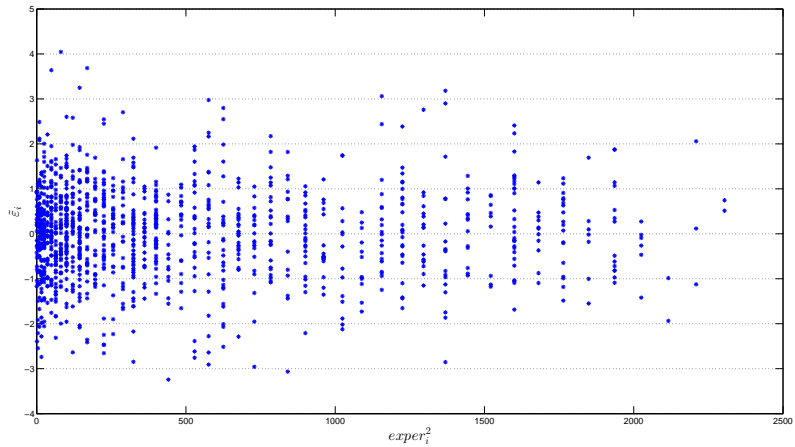








# Остатки модели 3





## Модель 5

Редуцированная модель:

$$\ln wage = 0.50 + 0.04exper - 0.0006exper^2 + 0.19union - 0.44female + \\ + 0.07educ + 0.004educ * belogavg - \\ - 0.005aboveavg - 0.19belogavg.$$

$$F = 99.72, p = 3.1 \times 10^{-128}, R^2 = 0.390, R_a^2 = 0.386.$$

Критерий	p-value
Шапиро-Уилка (нормальность)	$2.5 \times 10^{-5}$
знаковых рангов (несмещённость)	0.6959
Бройша-Пагана (гомоскедастичность)	$5.7 \times 10^{-4}$

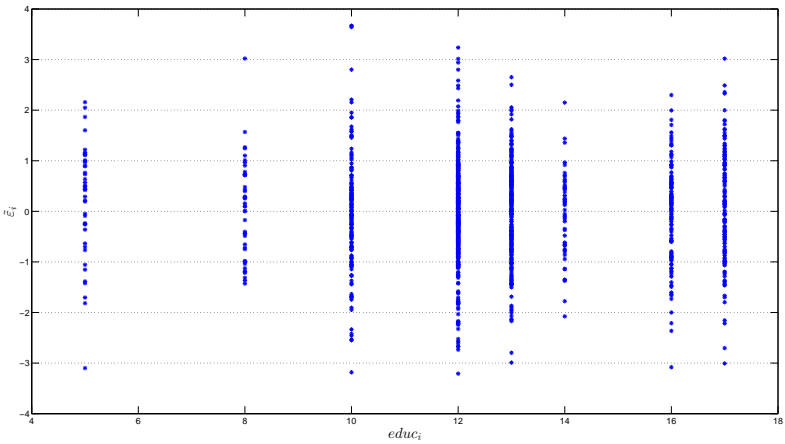








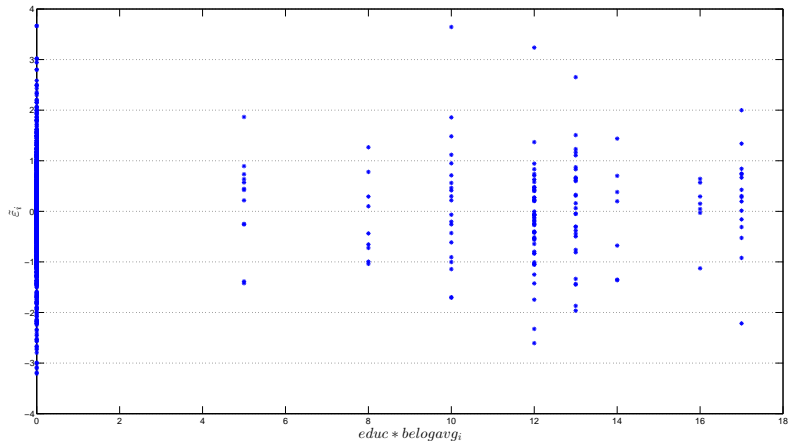
# Остатки модели 5







# Остатки модели 5



## Модель 3 против поделки 5

Критерий Давидсона-Маккиннона показывает превосходство модели 3 над моделью 5 ( $p_1 = 0.7404$ ,  $p_2 = 5.2 \times 10^{-7}$ ).







# Требования к решению задачи методом линейной регрессии

- визуализация данных, анализ распределения признаков (оценка необходимости трансформации), оценка наличия выбросов;
- оценка необходимости преобразования отклика и его поиск методом Бокса-Кокса;
- отбор признаков;
- визуальный анализ остатков;
- проверка гипотез об остатках: нормальность, несмещённость, гомоскедастичность;
- анализ необходимости добавления взаимодействий и квадратов признаков;
- расчёт расстояний Кука, возможное удаление выбросов, обновление модели;
- выводы.



