

Прикладной статистический анализ данных. 7. Регрессионный анализ.

Михаил Хальман
psad.homework@gmail.com

2017

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:

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

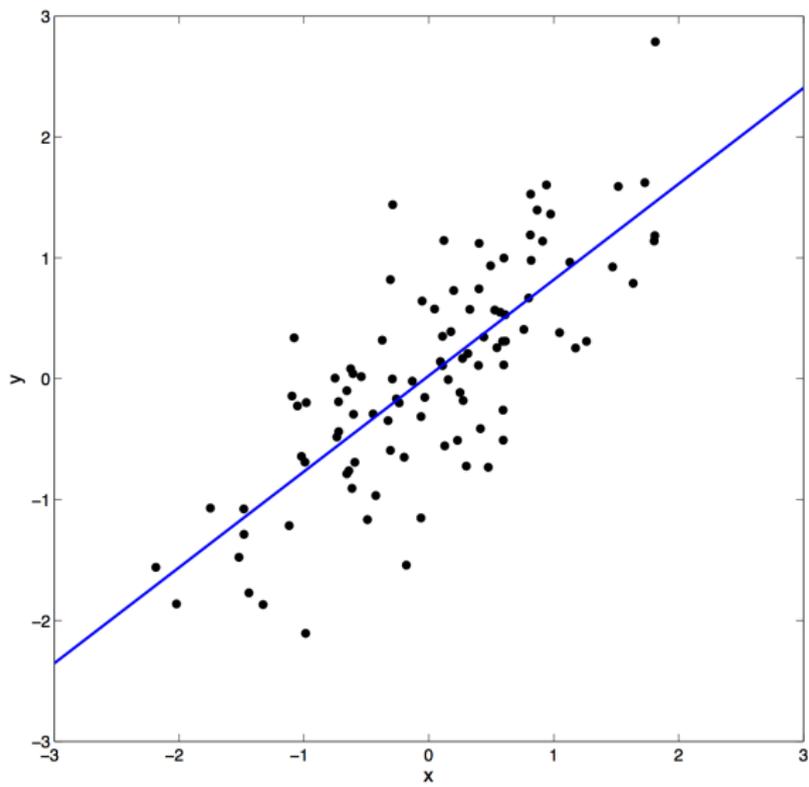
$$\|y - X\beta\|_2^2 \rightarrow \min_{\beta};$$

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

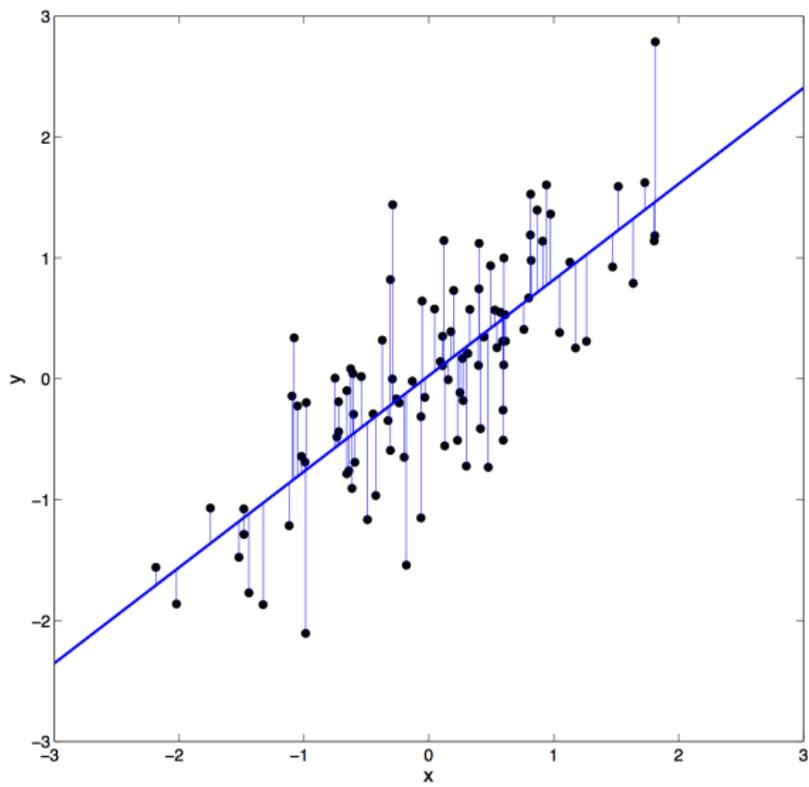
$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

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

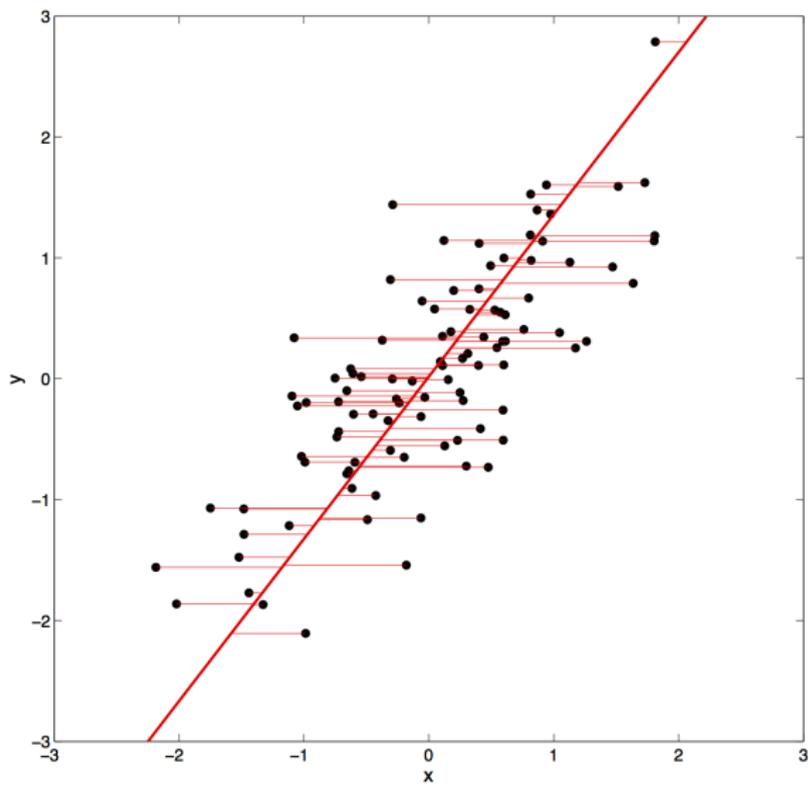
Инверсия задачи регрессии



Инверсия задачи регрессии



Инверсия задачи регрессии



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:

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

Предположения модели

- 1 Линейность отклика: $y = X\beta + \varepsilon$.
- 2 Случайность выборки: наблюдения $(x_i, y_i), i = 1, \dots, n$ независимы.
- 3 Полнота ранга: ни один из признаков не является константой или линейной комбинацией других признаков ни в популяции, ни в выборке ($\text{rank } X = k + 1$).
- 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 β , линейных по y .

Неправильное определение модели

Недоопределение: если зависимая переменная определяется моделью

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

а вместо этого используется модель

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

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 β , но дисперсия её возрастает.

Дисперсия $\hat{\beta}_j$

В предположениях (1-5) дисперсии МНК-оценок коэффициентов β 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 σ^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 $\hat{\beta}_j$.

Дисперсия $\hat{\beta}_j$

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

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**.

Проблема мультиколлинеарности решается с помощью отбора признаков или использования регуляризаторов.

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 $y = \beta_0 + \beta_1 x$. 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 β_1, β_2 оценивают среднюю разницу в уровнях зарплат инженера и управляющего с рабочим.
- При deviation-кодировании коэффициенты β_1, β_2 оценивают среднюю разницу в уровнях зарплат рабочего и инженера со средним по всем должностям.

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 ?

$$H_0: \beta_1 = 0.$$

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

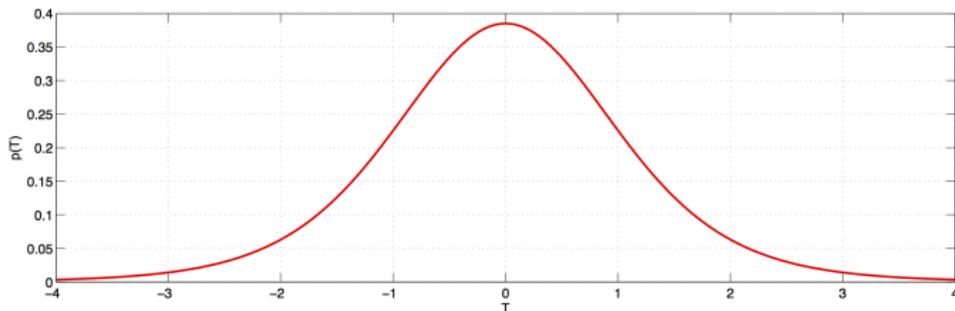
t-критерий Стьюдента

нулевая гипотеза: $H_0: \beta_j = a$

альтернатива: $H_1: \beta_j < \neq > a$

статистика:
$$T = \frac{\hat{\beta}_j - a}{\sqrt{\frac{RSS}{n-k-1} (X^T X)^{-1}_{jj}}}$$

нулевое распределение: $St(n - k - 1)$



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.

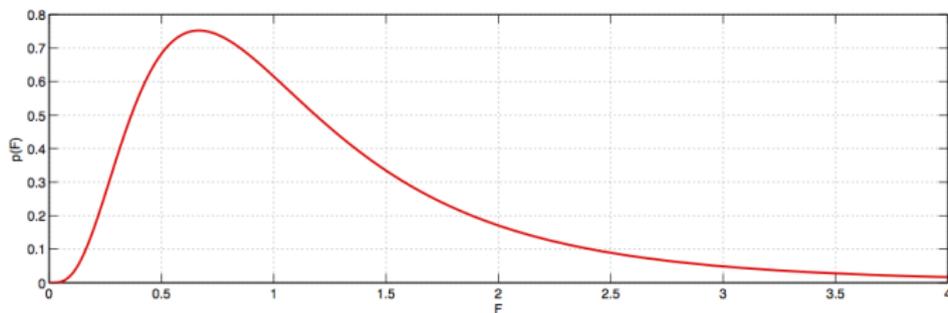
Критерий Фишера

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

альтернатива: $H_1: H_0$ неверна

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

нулевое распределение: $F(k, n - k - 1)$



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:

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

$$y = \gamma_0 + \gamma_1 \log x_1 + \gamma_2 \log x_2 + \gamma_3 \hat{\hat{y}} + \varepsilon.$$

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	Обе модели хороши	Модель (1) значительно лучше
Отвергнута	Модель (2) значительно лучше	Обе модели плохи

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

(Freedman, 1983): пошаговая регрессия несовместима с проверкой гипотез о значимости коэффициентов: критерии Фишера и Стьюдента антиконсервативны, если вычисляются на той же самой выборке, на которой настраивалась модель.

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

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))
```

Короткий способ:

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

Работает при $k \lesssim 100$.

Значимость категориальных предикторов

Категориальный предиктор, кодируемый несколькими фиктивными переменными, необходимо включать или исключать целиком. Значимость соответствующих фиктивных переменных лучше проверять в совокупности.

В случае, когда по отдельности какие-то фиктивные переменные не значимы, допустимо объединять уровни категориального предиктора, основываясь на интерпретации.

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 ε необходимо проверять.

Оценивать ошибку ε будем при помощи **остатков**:

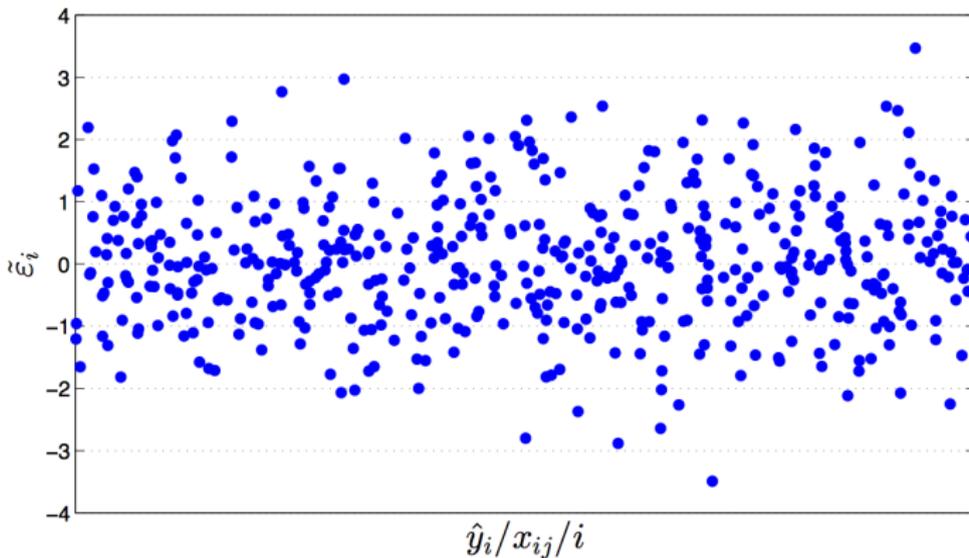
$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = 1, \dots, n.$$

Стандартизированные остатки:

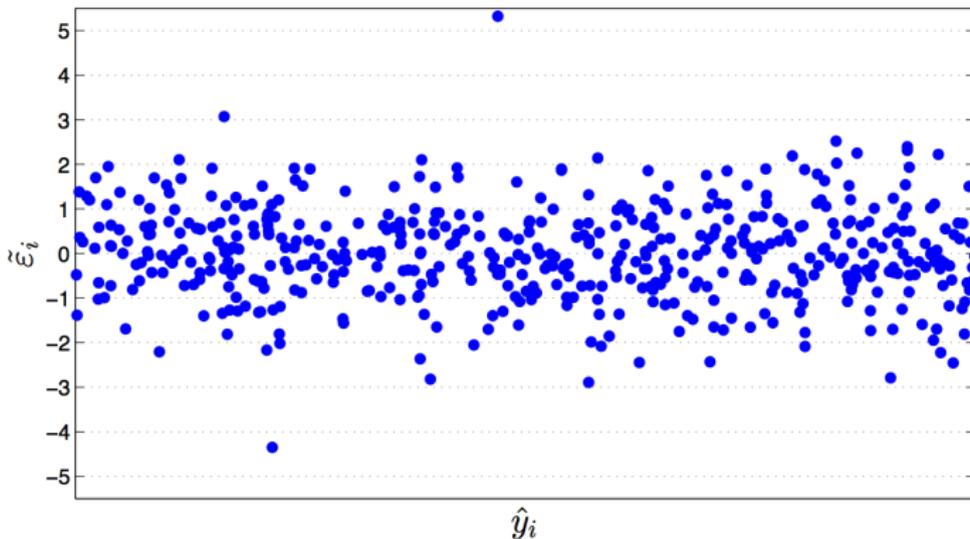
$$\tilde{\varepsilon}_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}}, \quad i = 1, \dots, n.$$

Визуальный анализ

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

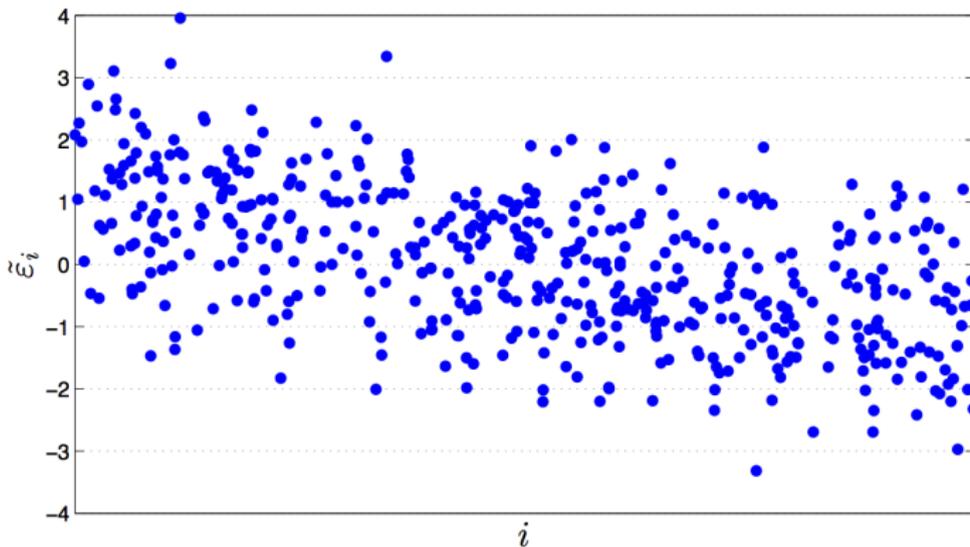


Визуальный анализ



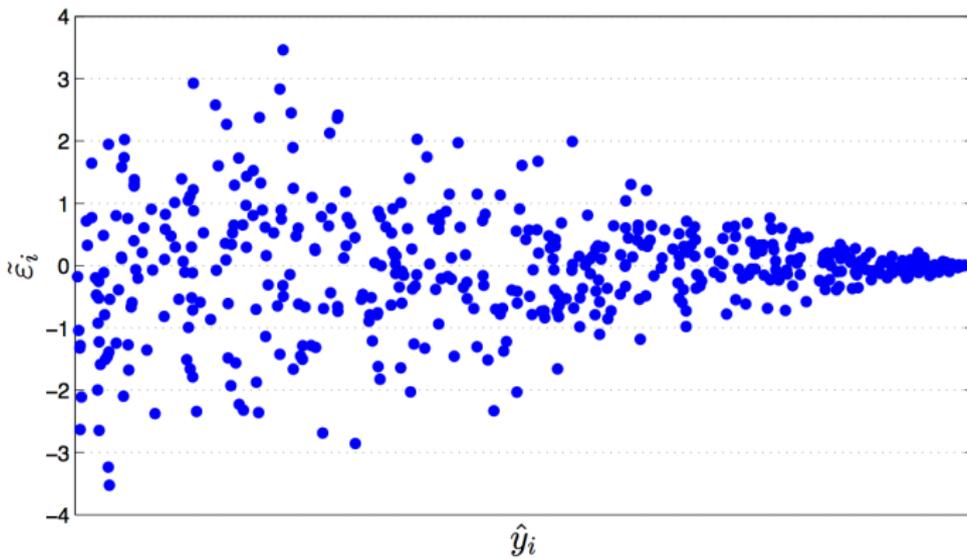
Возможно, присутствуют выбросы

Визуальный анализ



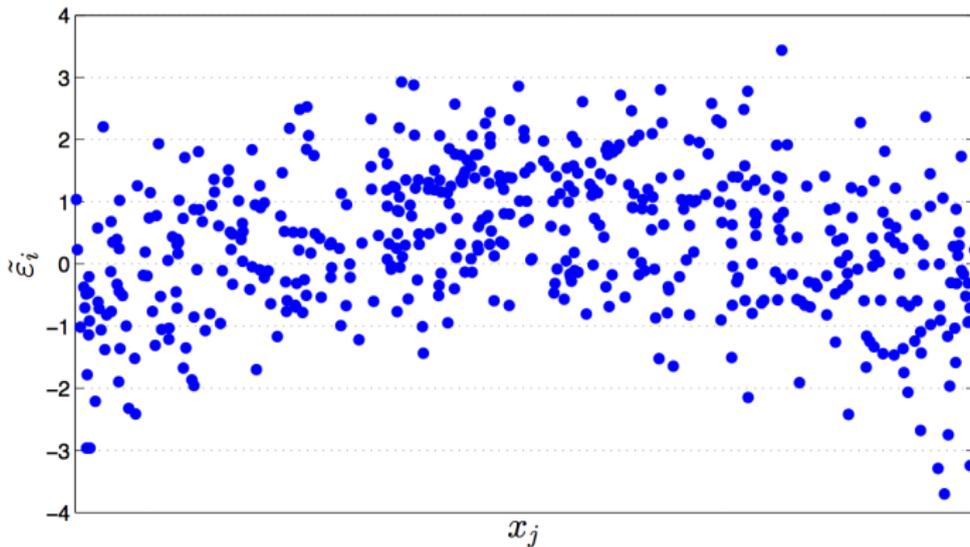
В данных имеется тренд

Визуальный анализ



Гетероскедастичность

Визуальный анализ



Стоит добавить квадрат признака x_j

Формальные критерии

- Проверка нормальности — занятие 4.
- Проверка несмещённости: если остатки нормальны — критерий Стьюдента (занятие 4), нет — непараметрический критерий (занятие 5).
- Проверка гомоскедастичности: критерий Бройша-Пагана.

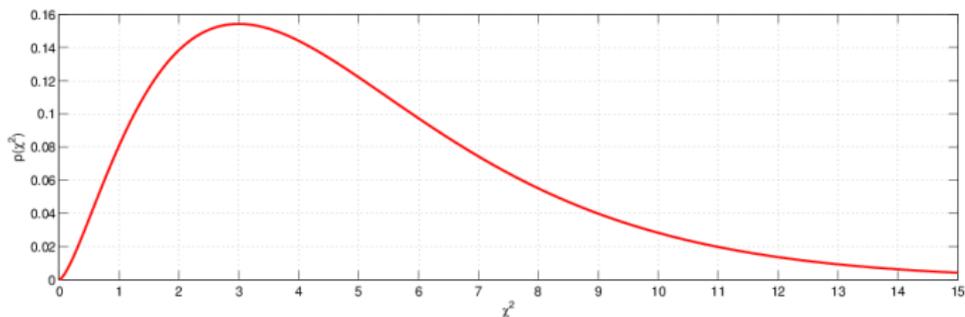
Критерий Бройша-Пагана

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

нулевое распределение: χ_k^2



Гетероскедастичность

Гетероскедастичность может быть следствием недоопределения модели.

Последствия гетероскедастичности:

- МНК-оценки β и R^2 остаются несмещёнными и состоятельными
- нарушаются предположения критериев Стьюдента и Фишера и методов построения доверительных интервалов для σ и β (независимо от объёма выборки)

Варианты:

- переопределить модель, добавить признаки, преобразовать отклик
- использовать модифицированные оценки дисперсии коэффициентов

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^λ , 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 λ .

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 λ :

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

Доверительный интервал для λ 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, при анализе моделей (далее) можно использовать устойчивые оценки дисперсии.

White's heteroscedasticity-consistent estimator (HCE):

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

Асимптотика устойчивой оценки:

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

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

Использование устойчивых оценок дисперсии

Пакет sandwich:

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

#significance of every predictor
coeftest(m, df = Inf, vcov = vcovHC(m, type = "HC0"))

#significance of the group of predictors
waldtest(m1, m2, vcov = vcovHC(m1, type = "HC0")) #m1 - bigger model

#significance of the whole equation
waldtest(m, vcov = vcovHC(m, type = "HC0"))
```


Расстояние Кука

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

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

$\hat{y}_{j(i)}$ — 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 .

