Прикладной статистический анализ данных. 7. Регрессионный анализ.

Ольга Добролюбова
Юлиан Сердюк
cs.msu.psad@gmail.com

25.03.2022

Постановка задачи линейной регрессии

$1, \ldots, n$ — объекты;
$x_1, \ldots, x_k, y$ — признаки, значения которых измеряются на объектах;
$x_1, \ldots, x_k$ — объясняющие переменные (предикторы, регрессоры, факторы, признаки);
y — зависимая переменная, отклик.

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

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

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

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

Здесь и далее $n > k$ $(n \gg k)$.

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:

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

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

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

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

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

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:

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

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

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:

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

Предположения модели

1. Линейность отклика: $y = X\beta + \varepsilon$.
2. Случайность выборки: наблюдения $(x_i, y_i), i = 1, \ldots, n$ независимы.
3. Полнота ранга: ни один из признаков не является константой или линейной комбинацией других признаков ни в популяции, ни в выборке $(\operatorname{rank} X = k + 1)$.
4. Случайность ошибок: $\mathbb{E}(\varepsilon | X) = 0$.

В предположениях (1-4) МНК-оценки коэффициентов β являются несмещёнными:

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

и состоятельными:

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

Предположения модели

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

Дисперсия $\hat{\beta}_j$

В предположениях (1-5) дисперсии МНК-оценок коэффициентов β задаются следующим образом:

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

где $TSS_j = \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2$, 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:

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

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:

	Dummy		Deviation	
Тип должности	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.

Вопросы

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

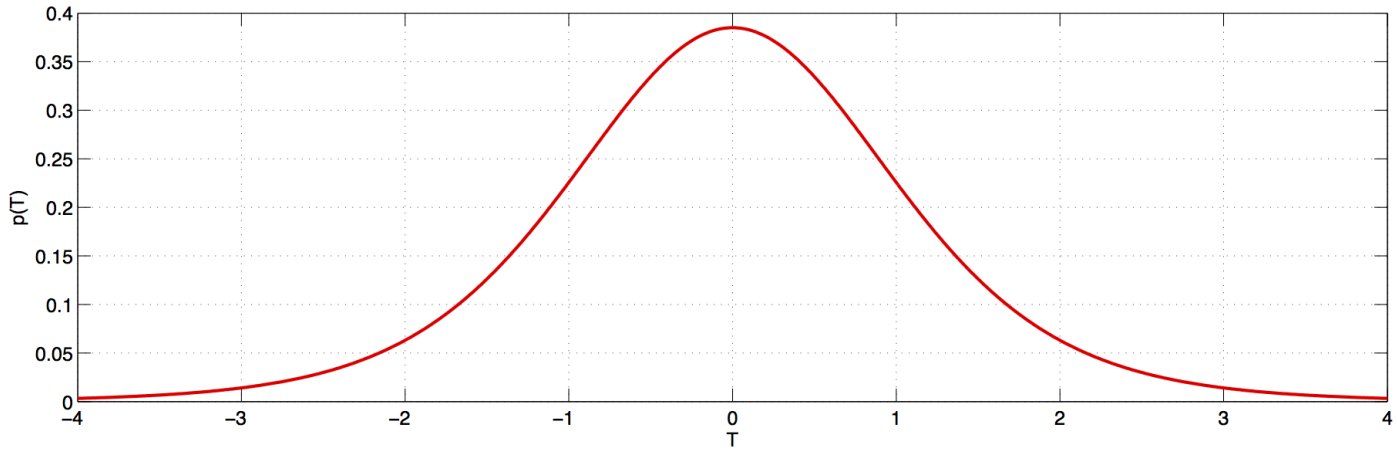
t-критерий Стьюдента

нулевая гипотеза: $H_0 : \beta_j = a$

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

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

нулевое распределение: $St(n-k-1)$



Критерий Фишера

$$\underset{n\times(k+1)}{X} = \left(\underset{n\times(k+1-k_1)}{X_1}, \underset{n\times k_1}{X_2}\right); \quad \underset{(k+1)\times 1}{\beta^T} = \left(\underset{(k+1-k_1)\times 1}{\beta_1^T}, \underset{k_1\times 1}{\beta_2^T}\right)^T;$$

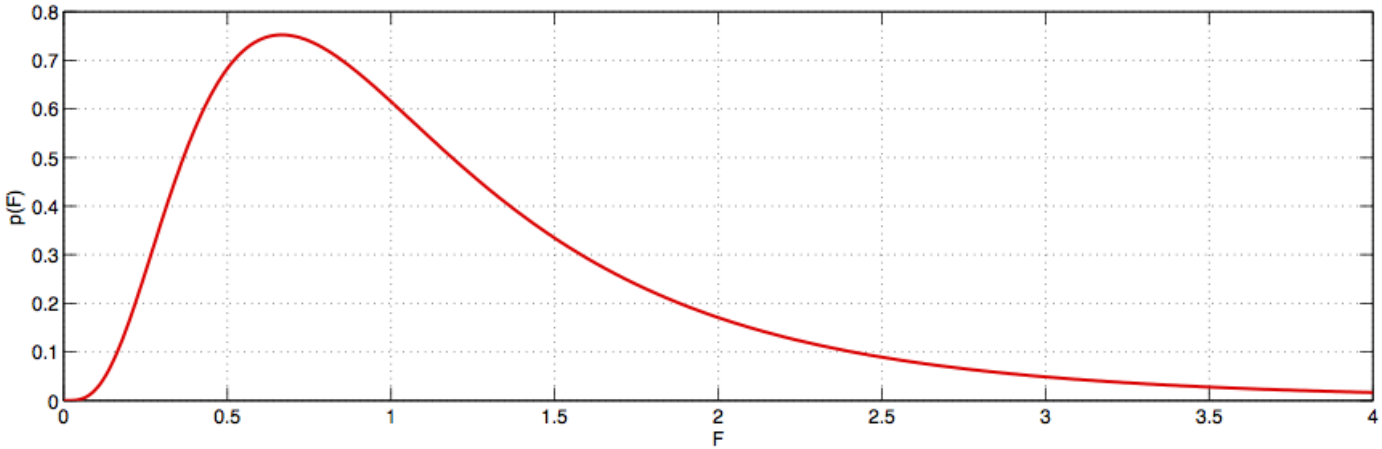
нулевая гипотеза: $H_0 : \beta_2 = 0$

альтернатива: $H_1 : H_0$ неверна

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

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

нулевое распределение: $F(k_1, n-k-1)$



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:

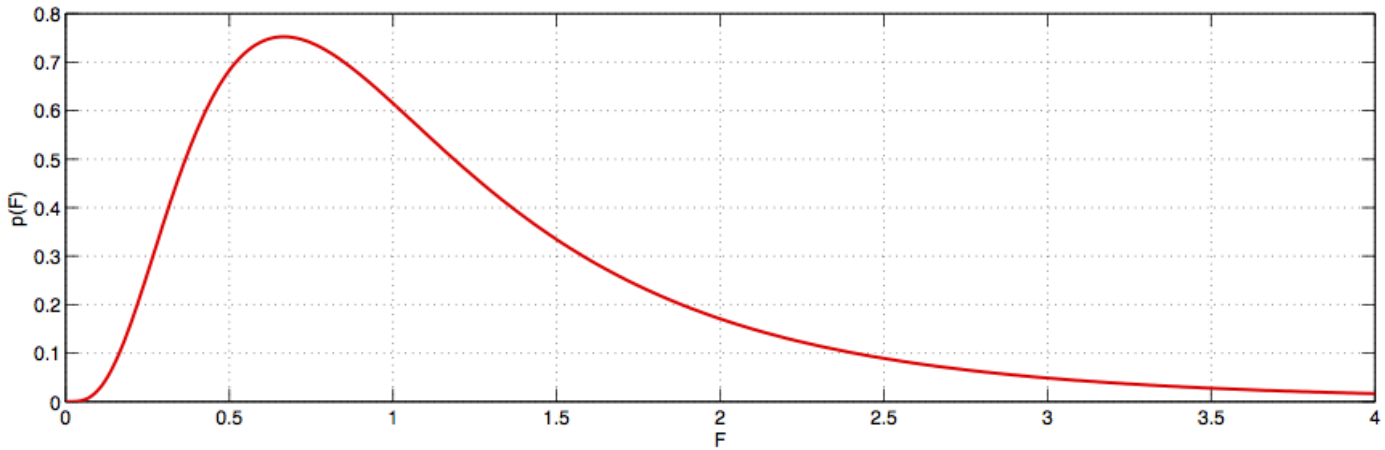
- отдельные признаки из X_2 недостаточно хорошо объясняют y, но совокупный эффект значим;
- признаки в X_2 мультиколлинеарны.

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:	$H_0 : \beta_1 = \cdots = \beta_k = 0$
альтернатива:	$H_1 :$ H_0 неверна
статистика:	$F = \frac{R^2/k}{\left(1-R^2\right)/(n-k-1)}$
нулевое распределение:	$F(k, n-k-1)$



Критерий Фишера

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

$H_0 : \beta_1 = \cdots = \beta_5 = 0.$
$H_1 : H_0$ неверна $\Rightarrow p = 6.0331 \times 10^{-9}.$

Сравнение невложенных моделей

Пример: имеются две модели:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \tag{1}$$

$$y = \gamma_0 + \gamma_1 \log x_1 + \gamma_2 \log x_2 + \varepsilon. \tag{2}$$

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:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \hat{\hat{y}} + \varepsilon,$$
$$y = \gamma_0 + \gamma_1 \log x_1 + \gamma_2 \log x_2 + \gamma_3 \hat{y} + \varepsilon.$$

При помощи критерия Стьюдента проверим

$$H_{01}: \beta_3 = 0, \;\; H_{11}: \beta_3 \neq 0,$$
$$H_{02}: \gamma_3 = 0, \;\; H_{12}: \gamma_3 \neq 0.$$

H_{01} \ H_{02}	Принята	Отвергнута
Принята	Обе модели хороши	Модель (1) значимо лучше
Отвергнута	Модель (2) значимо лучше	Обе модели плохи

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**:

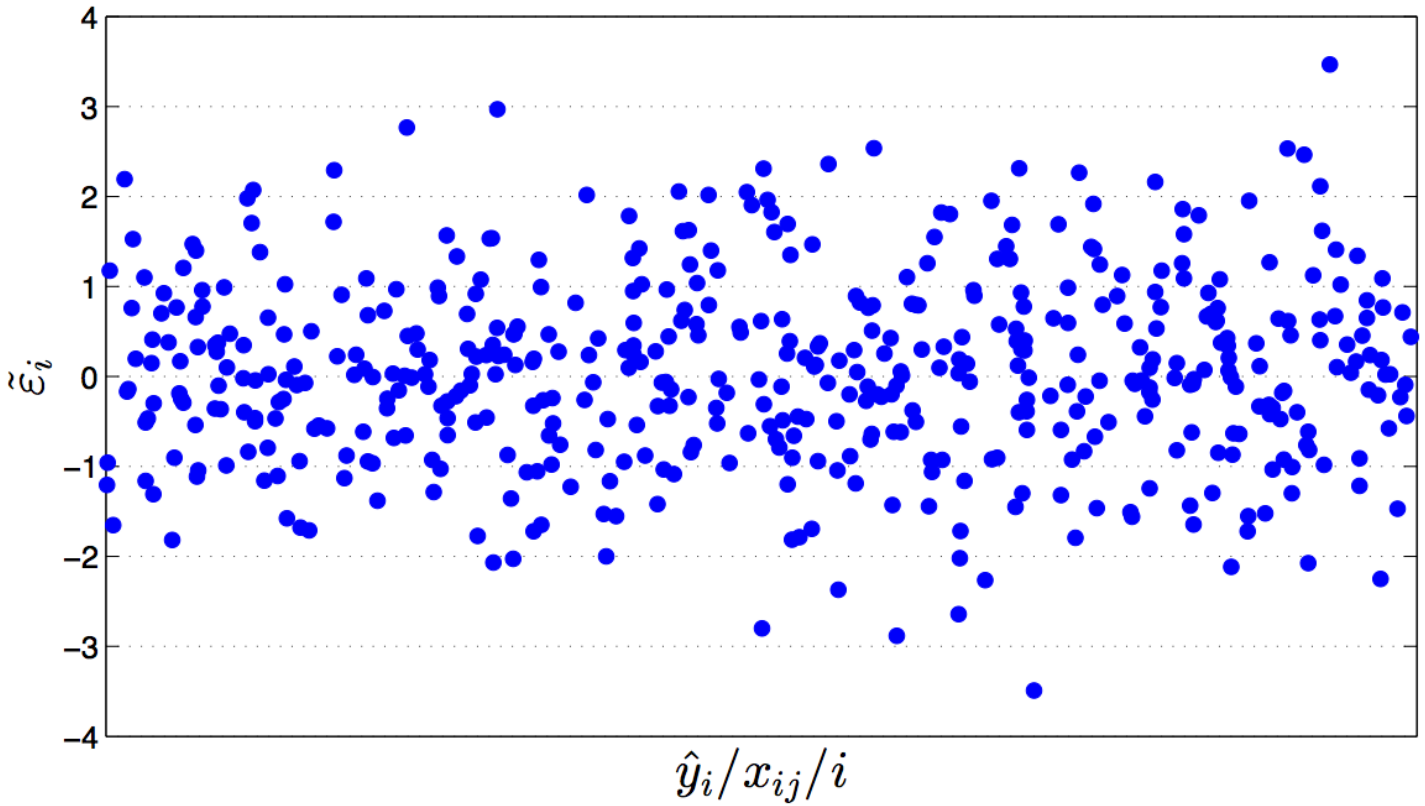
$$\hat{\varepsilon}_i = y_i - \hat{y}_i,\ i = 1, \ldots, n.$$

Стандартизированные остатки:

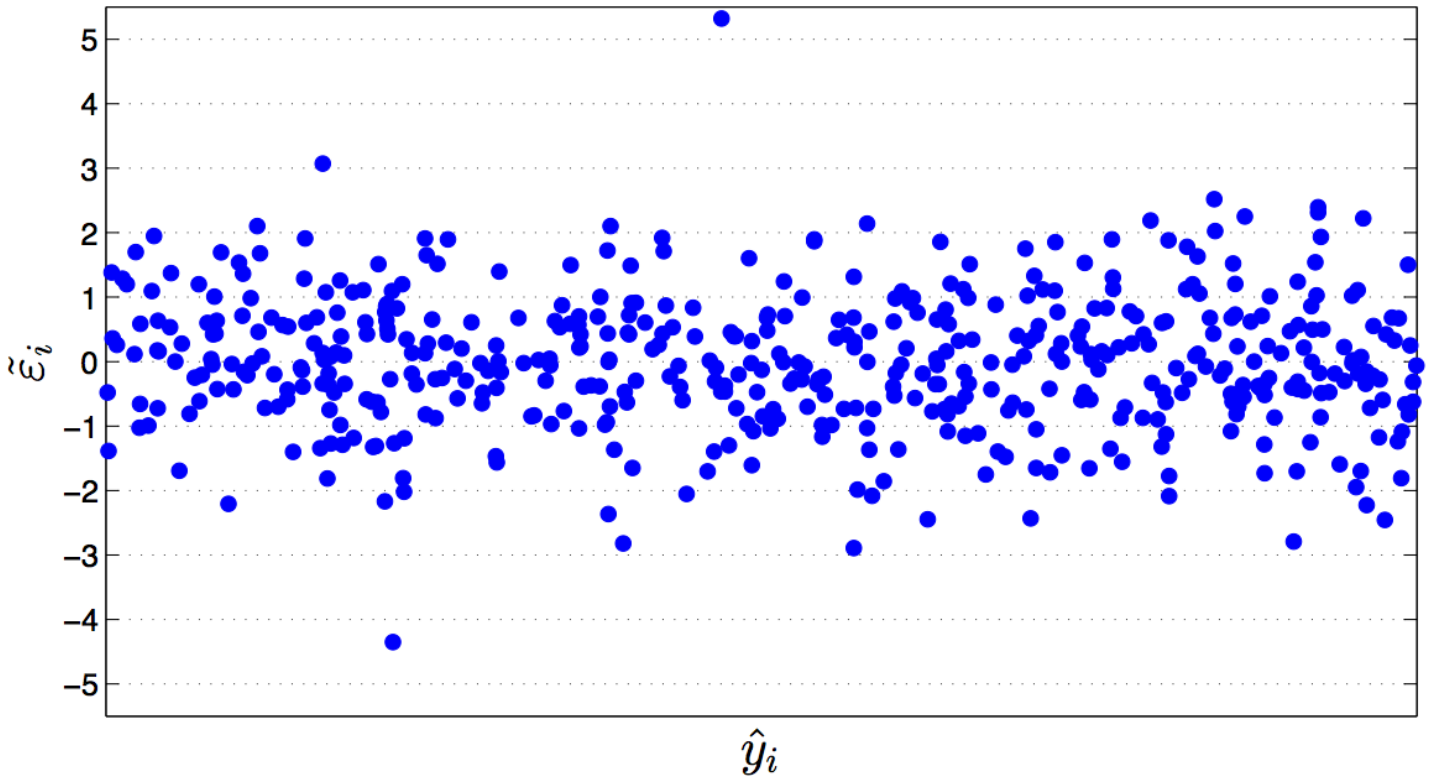
$$\tilde{\varepsilon}_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}},\ i = 1, \ldots, n.$$

Визуальный анализ

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

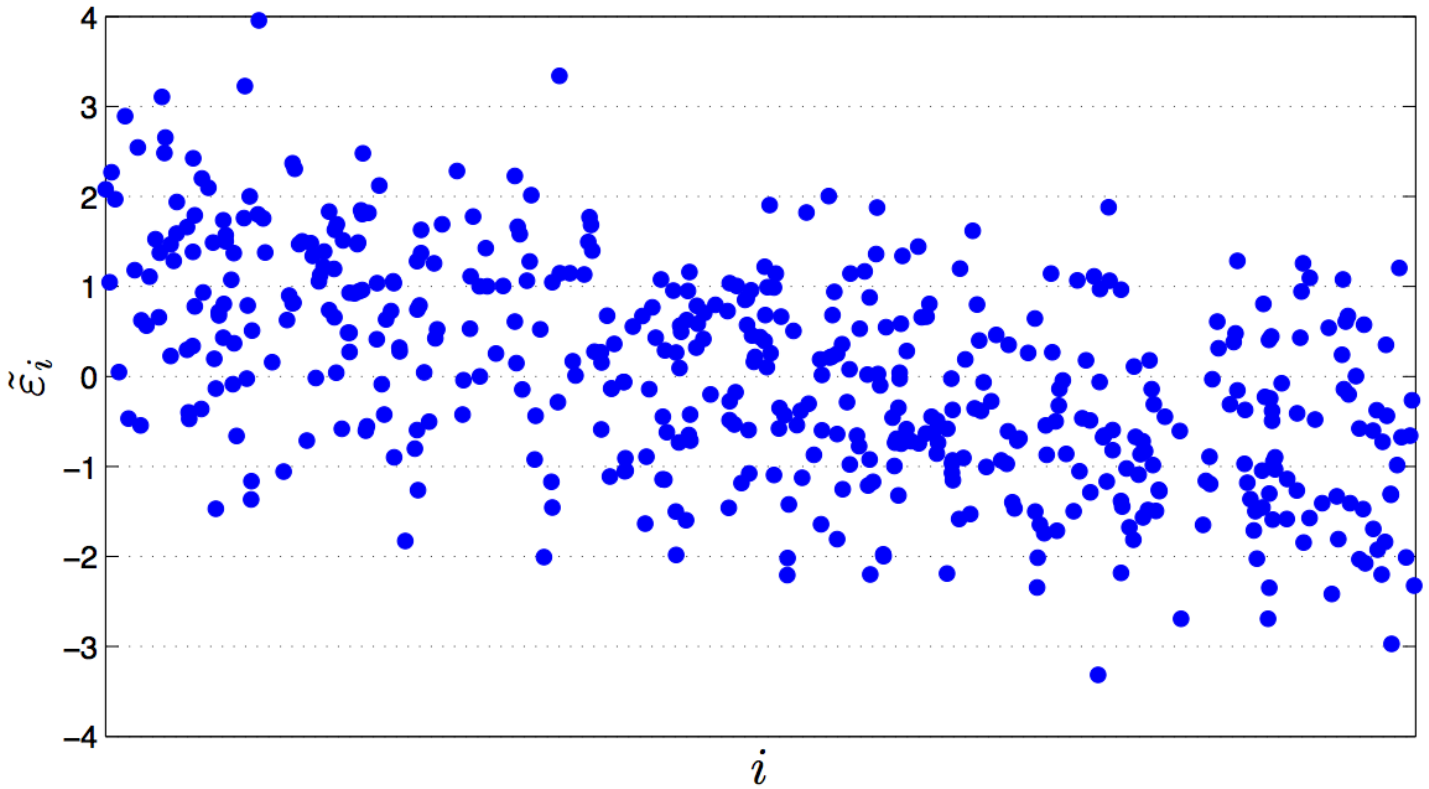


Визуальный анализ



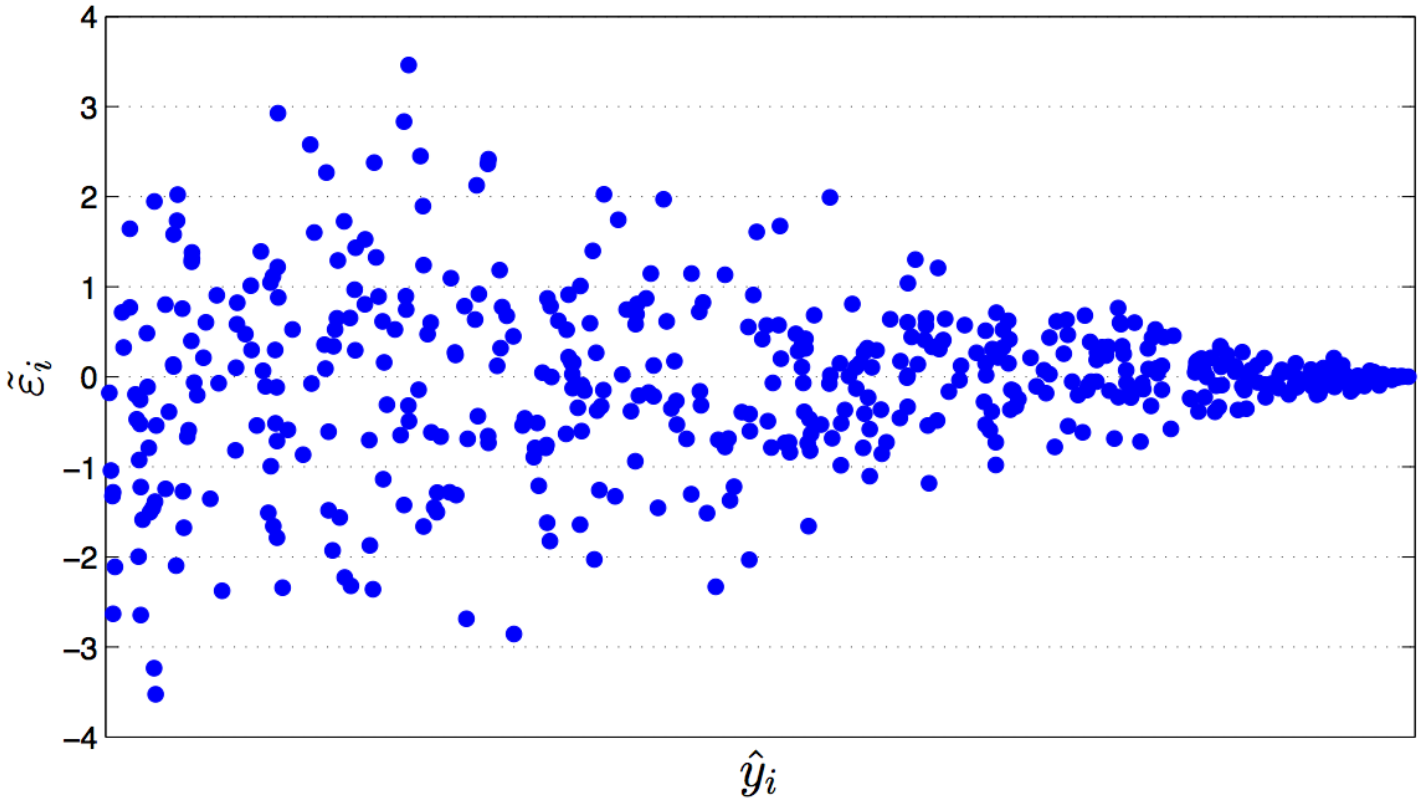
Возможно, присутствуют выбросы

Визуальный анализ



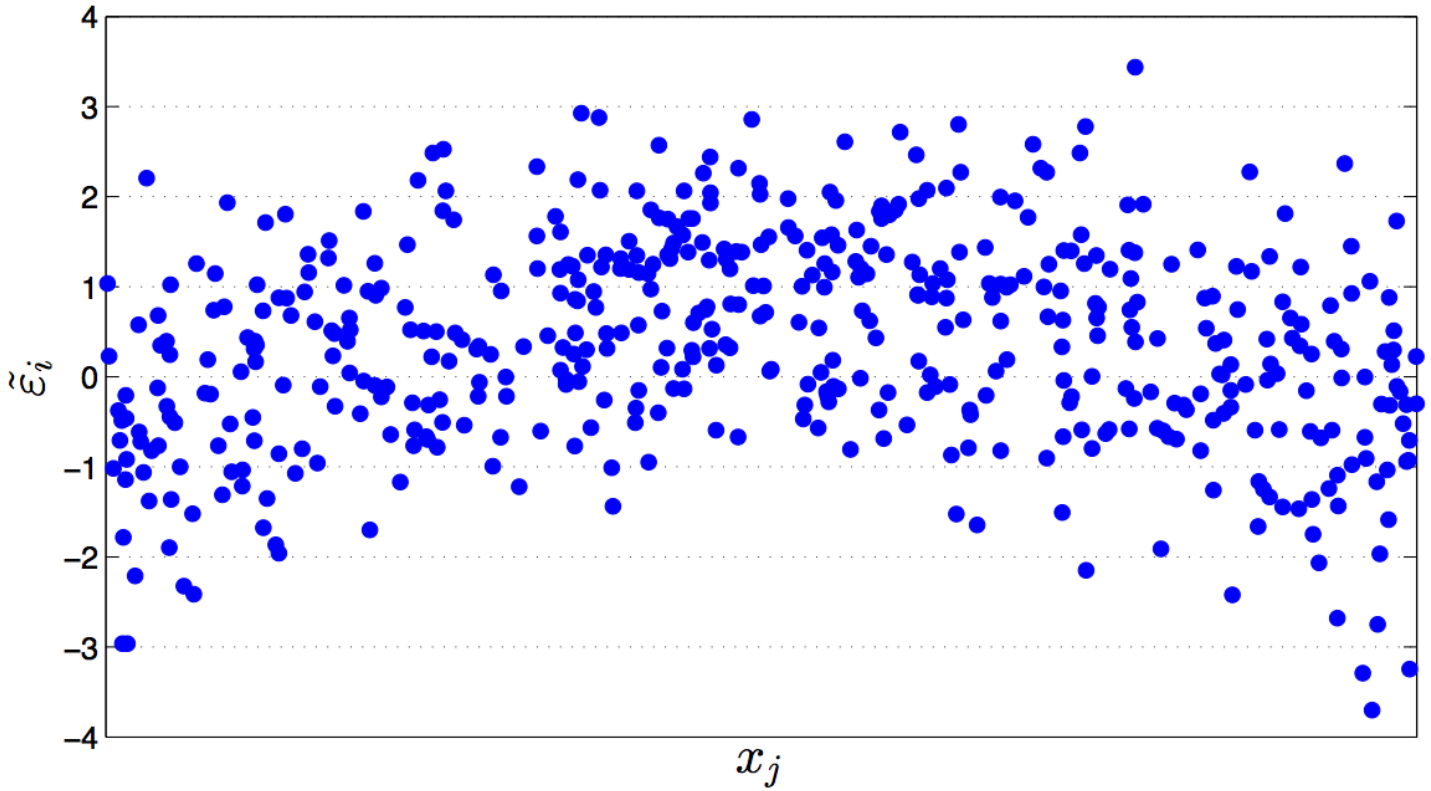
В данных имеется тренд

Визуальный анализ



Гетероскедастичность

Визуальный анализ



Стоит добавить квадрат признака x_j

Формальные критерии

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

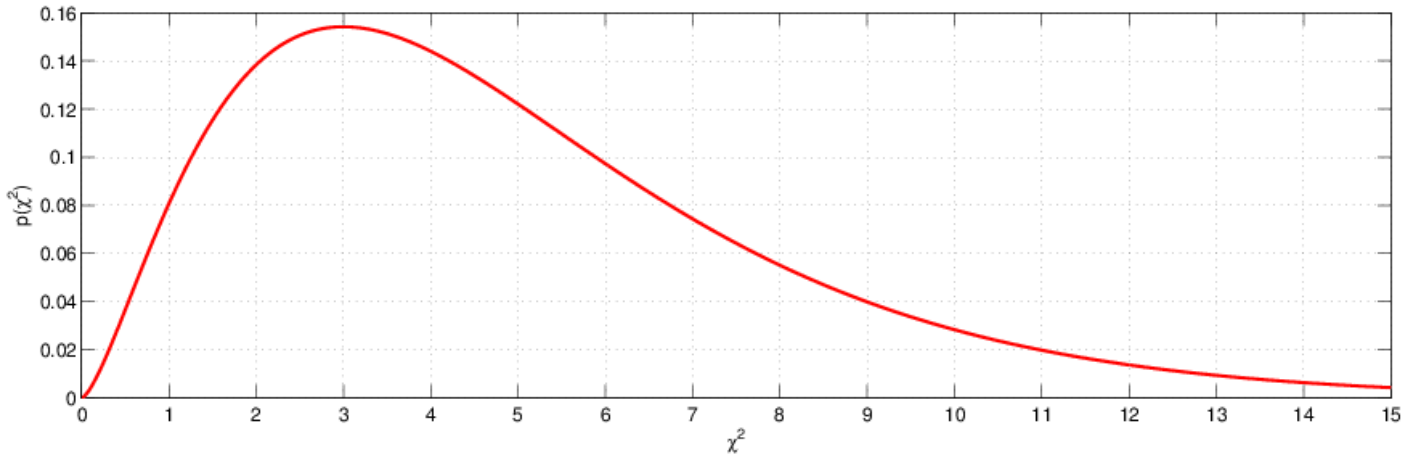
Критерий Бройша-Пагана

нулевая гипотеза: $H_0 : \mathbb{D}\varepsilon_i = \sigma^2$

альтернатива: $H_1 : H_0$ неверна

статистика: $LM = nR^2_{\hat{\varepsilon}^2}$, $R^2_{\hat{\varepsilon}^2}$ — коэффициент детерминации при регрессии квадратов остатков на признаки

нулевое распределение: χ^2_k



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 $y_1, \ldots, y_n$ 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

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

но оно сильно варьируется по λ.
Вместо него можно рассмотреть семейство преобразований

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

где $\dot{y} = (y_1 y_2 \ldots y_n)^{1/n}$ — 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 $\min\limits_{\lambda} RSS\,(\lambda) \cdot e^{\chi^2_{1,1-\alpha}/n}$. Если он содержит единицу, возможно, не стоит выполнять преобразование.

Устойчивая оценка дисперсии Уайта

Если не удаётся избавиться от гетероскедастичности, при анализе моделей (дальше) можно использовать устойчивые оценки дисперсии.
White's heteroscedasticity-consistent estimator (HCE):

$$\mathbb{D}\left(\hat{\beta}\middle| X\right) = \left(X^T X\right)^{-1} \left(X^T \operatorname{diag}\left(\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2\right) X\right) \left(X^T X\right)^{-1}.$$

Асимптотика устойчивой оценки:

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

$$\hat{\Omega} = n\left(X^T X\right)^{-1} \left(X^T \operatorname{diag}\left(\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2\right) X\right) \left(X^T X\right)^{-1}.$$

Другие устойчивые оценки дисперсии

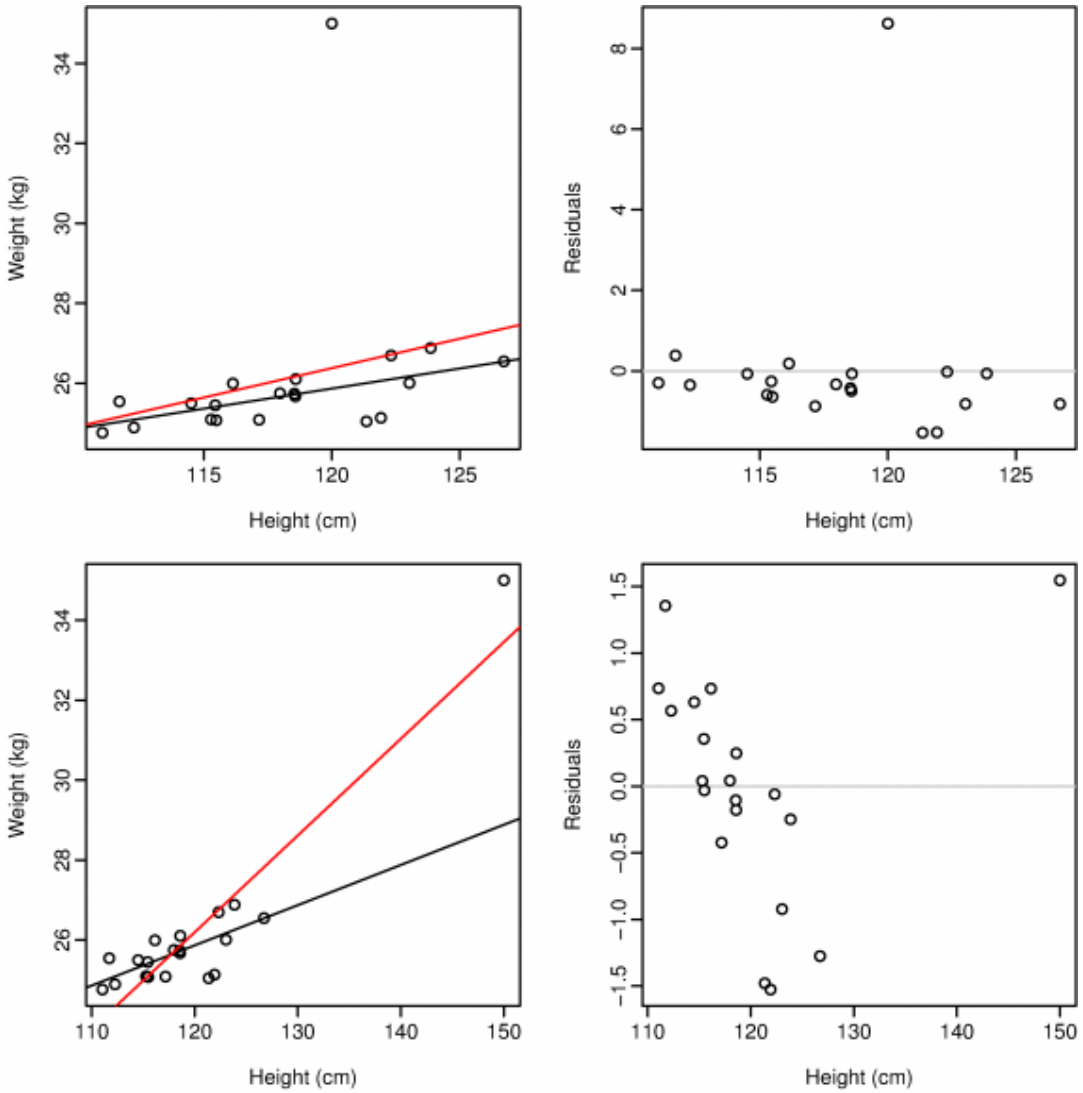
Элементы диагональной матрицы могут задаваться разными способами:

const	$\hat{\sigma}^2$
HC0	$\hat{\varepsilon}_i^2$
HC1	$\frac{n}{n-k}\hat{\varepsilon}_i^2$
HC2	$\frac{\hat{\varepsilon}_i^2}{1-h_i}$
HC3	$\frac{\hat{\varepsilon}_i^2}{(1-h_i)^2}$
HC4	$\frac{\hat{\varepsilon}_i^2}{(1-h_i)^{\min\left(4, \frac{nh_i}{k}\right)}}$

const — случай гомоскедастичной ошибки,
HC0 — оценка Уайта,
HC1–HC3 — модификации МакКиннона-Уайта,
HC4 — модификация Крибари-Нето.

Расстояние Кука

Регрессия сильно подстраивается под далеко стоящие наблюдения.



Расстояние Кука

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

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

$\hat{y}_{j(i)}$ — предсказания модели, настроенной по наблюдениям $1, \ldots, i-1, i+1, \ldots, n$, для наблюдения j;
h_i — диагональный элемент матрицы $H = X\left(X^T X\right)^{-1} X^T$ (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 $\hat{y}_i$.

Литература

- линейная регрессия в целом — Wooldridge (много примеров, без матричной алгебры);
- критерий Давидсона-Маккиннона (Davidson-MacKinnon test) — Davidson;
- множественная оценка значимости коэффициентов — Bretz, 4.4;
- преобразование Бокса-Кокса (Box-Cox transformation) — Дрейпер, гл. 14;
- устойчивые оценки дисперсии — White, MacKinnon, Cribari-Neto;
- расстояние Кука (Cook's distance) — Cook.

Дрейпер Н.Р., Смит Г. *Прикладной регрессионный анализ*, 2007.

Кобзарь А.И. *Прикладная математическая статистика*, 2006.

Bretz F., Hothorn T., Westfall P. *Multiple Comparisons Using R*, 2010.

Cook D.R., Weisberg S. *Residuals and influence in regression*, 1982.

Литература

Cribari-Neto F. (2004). *Asymptotic inference under heteroskedasticity of unknown form*. Computational Statistics & Data Analysis, 45(2), 215–233.

Davidson R., MacKinnon J. (1981). *Several Tests for Model Specification in the Presence of Alternative Hypotheses*. Econometrica, 49, 781-793.

Freedman D.A. *A Note on Screening Regression Equations*. The American Statistician, 37(2), 152-155.

MacKinnon J., White H. (1985). *Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties*. Journal of Econometrics, 29, 305–325.

White H. (1980). *A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity*. Econometrica: Journal of the Econometric Society, 48(4), 817–838.

Wooldridge J. *Introductory Econometrics: A Modern Approach*, 2016.