

Описание, проверка значимости и валидности линейных моделей

Марина Варфоломеева, Вадим Хайтов

Описание и проверка значимости линейных моделей

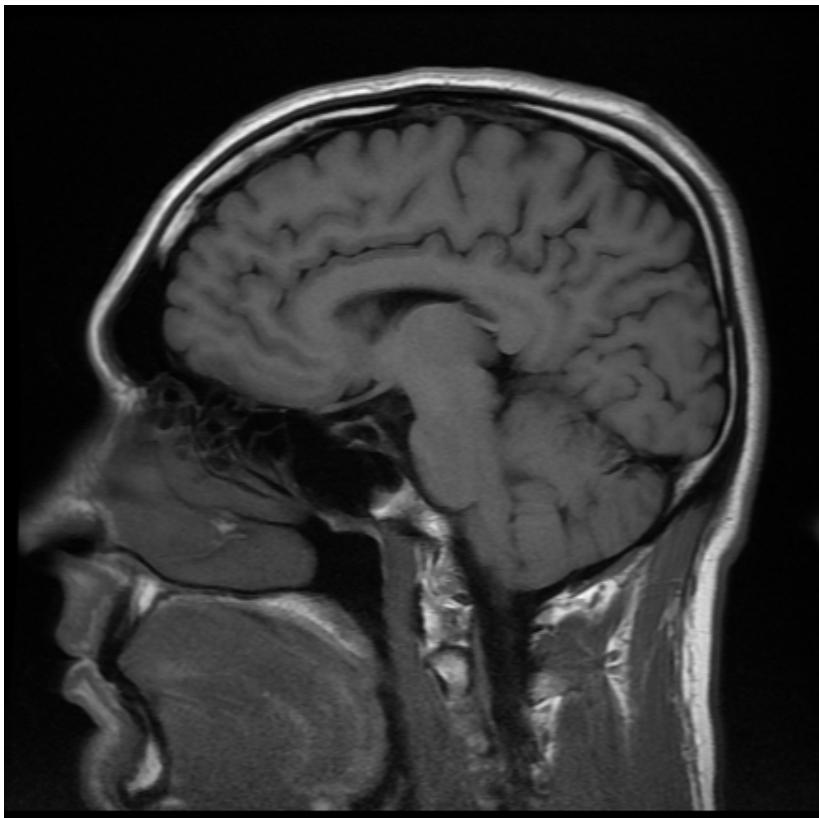
Вы сможете

- Подобрать линейную модель зависимости переменной-отклика от одного предиктора
- Протестировать значимость линейной модели в целом и значимость отдельных ее коэффициентов при помощи t и F критериев
- Проверить условия применимости линейной регрессии при помощи графиков

Вспомним пример из прошлой лекции

Пример: IQ и размеры мозга

Зависит ли уровень интеллекта от размера головного мозга? (Willerman et al. 1991)



Было исследовано 20 девушек и 20 молодых людей

У каждого индивида измеряли:

- вес
- рост
- размер головного мозга (количество пикселей на изображении ЯМР сканера)
- Уровень интеллекта измеряли с помощью IQ тестов

Пример: Willerman, L., Schultz, R., Rutledge, J. N., and Bigler, E. (1991), “In Vivo Brain Size and Intelligence”, *Intelligence*, 15, p.223–228.

Данные: “[The Data and Story Library](#)”

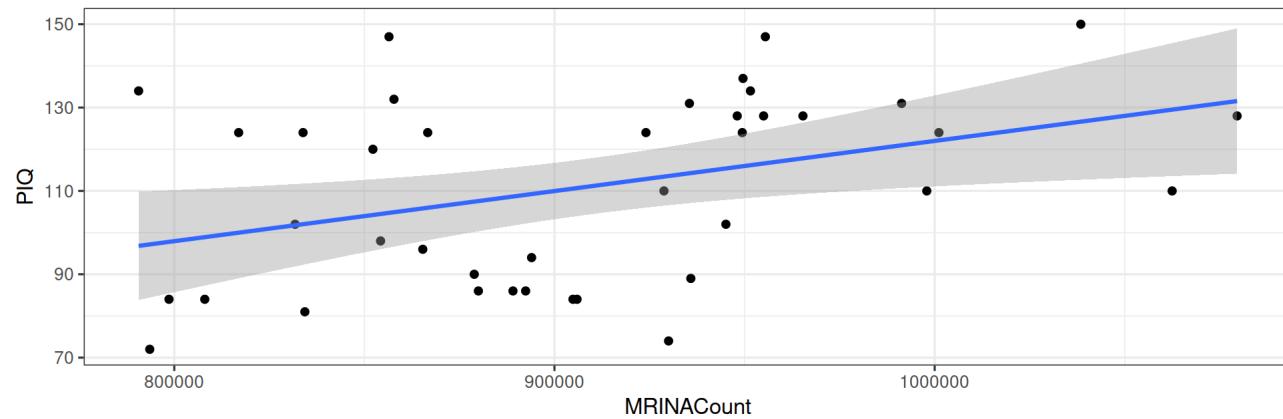
Фото: [Scan_03_11](#) by bucaorg (Paul Burnett) on Flickr

Вспомним, на чем мы остановились

```
##  
## Call:  
## lm(formula = PIQ ~ MRINACount, data = brain)  
##  
## Residuals:  
##     Min      1Q  Median      3Q     Max  
## -39.6   -17.9    -1.6    17.0    42.3  
##  
## Coefficients:  
##                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  1.7437570 42.3923825     0.04    0.967  
## MRINACount   0.0001203  0.0000465     2.59    0.014 *## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 21 on 38 degrees of freedom  
## Multiple R-squared:  0.15,    Adjusted R-squared:  0.127  
## F-statistic: 6.69 on 1 and 38 DF,  p-value: 0.0137
```

Уравнение и график зависимости

$$PIQ_i = 1.744 + 0.0001202 \cdot MRINACount_i$$



Тестирование гипотез о линейных моделях

Способы проверки значимости модели и ее коэффициентов

Два равноправных способа

Способы проверки значимости модели и ее коэффициентов

Два равноправных способа

Значима ли модель целиком?

- F критерий

Способы проверки значимости модели и ее коэффициентов

Два равноправных способа

Значима ли модель целиком?

- F критерий

Значима ли связь между предиктором и откликом?

- t-критерий
- F-критерий

Тестирование гипотез с помощью t -критерия

Тестирование гипотез с помощью t -критерия

Гипотезы

Зависимость есть, если $\beta_k \neq 0$

Нулевая гипотеза $H_0 : \beta_k = 0$

Тестирование гипотез с помощью t-критерия

Гипотезы

Зависимость есть, если $\beta_k \neq 0$

Нулевая гипотеза $H_0 : \beta_k = 0$

Тестовая статистика

$$t = \frac{b_k - \beta_k}{SE_{b_k}} = \frac{b_k - 0}{SE_{b_k}} = \frac{b_k}{SE_{b_k}}$$

Число степеней свободы: $df = n - p$, где n — объем выборки, p — число параметров модели, k — конкретный коэффициент регрессии.

Для простой линейной регрессии с одним предиктором $df = n - 2$.

Зависит ли IQ от размера головного мозга?

$$PIQ_i = 1.744 + 0.0001202 \cdot MRINACount_i$$

```
summary(brain_model)

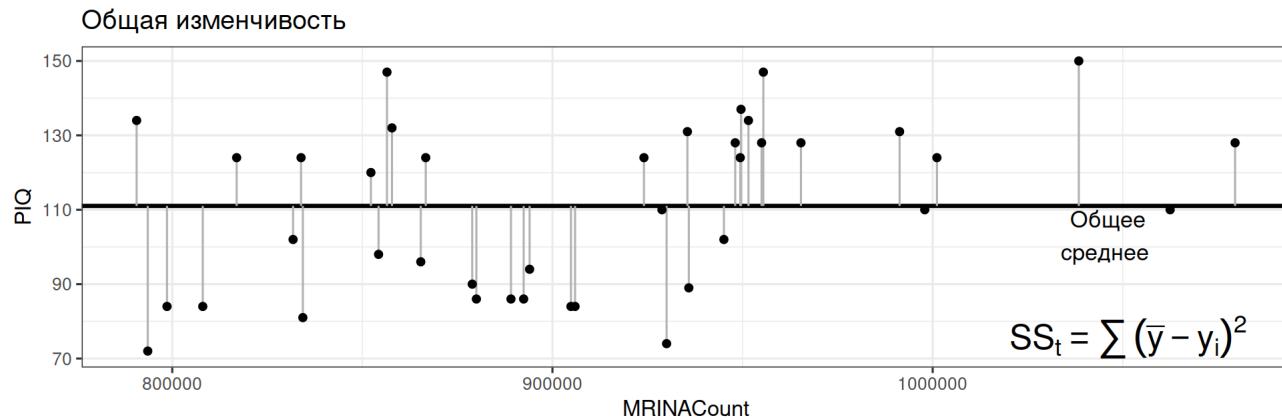
##
## Call:
## lm(formula = PIQ ~ MRINACount, data = brain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -39.6   -17.9   -1.6    17.0    42.3 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.7437570 42.3923825   0.04    0.967    
## MRINACount  0.0001203  0.0000465    2.59    0.014 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21 on 38 degrees of freedom
## Multiple R-squared:  0.15,    Adjusted R-squared:  0.127 
## F-statistic: 6.69 on 1 and 38 DF,  p-value: 0.0137
```

Результаты теста на IQ статистически значимо связаны с размерами мозга на МРТ ($t_{0.05,38} = 2.59$, $p = 0.01$).

Тестирование гипотез при помощи F-критерия

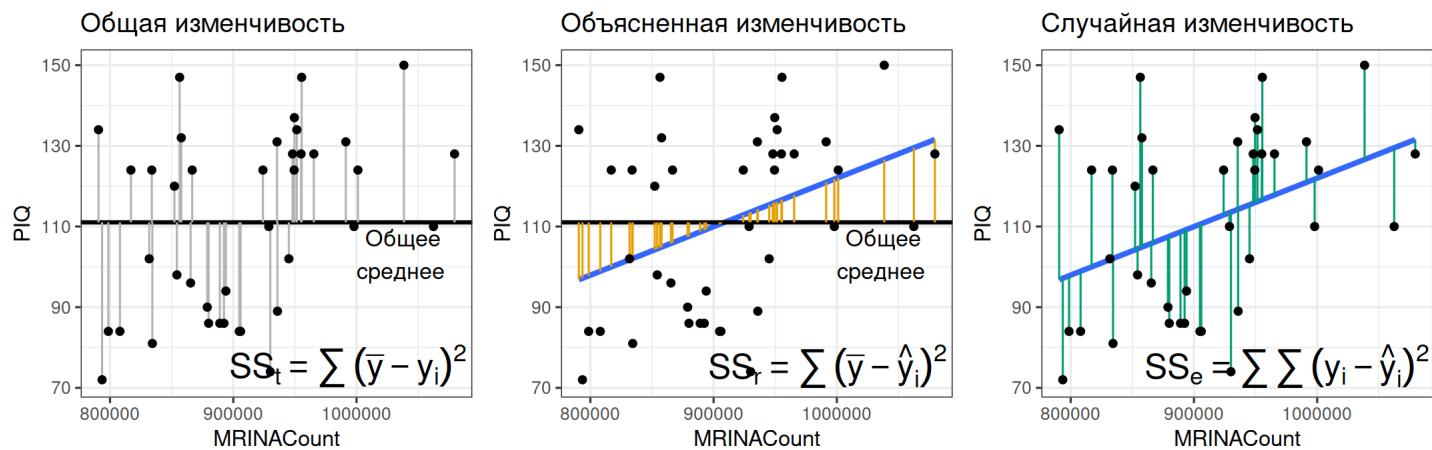
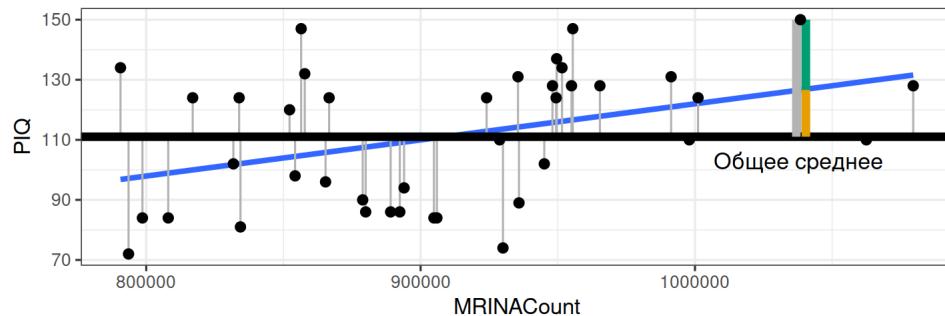
Общая изменчивость

Общая изменчивость SS_t — это сумма квадратов отклонений наблюдаемых значений y_i от общего среднего \bar{y}

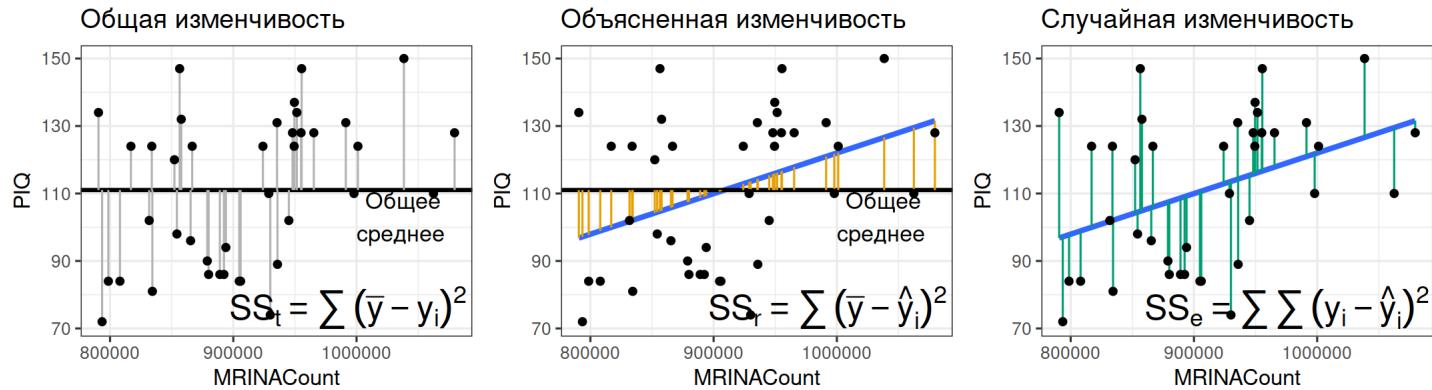


Структура общей изменчивости

$$SS_t = SS_r + SS_e$$



От изменчивостей к дисперсиям



MS_t , полная дисперсия

$$MS_t = \frac{SS_t}{df_t}$$

$$SS_t = \sum (\bar{y} - y_i)^2$$

$$df_t = n - 1$$

MS_r , дисперсия, объясненная регрессией

$$MS_r = \frac{SS_r}{df_r}$$

$$SS_r = \sum (\bar{y} - \hat{y})^2$$

$$df_r = 1$$

MS_e , остаточная дисперсия

$$MS_e = \frac{SS_e}{df_e}$$

$$SS_e = \sum (\hat{y} - y_i)^2$$

$$df_e = n - 2$$

С помощью MS_r и MS_e можно тестировать значимость коэффициентов

Если дисперсии остатков для всех значений X равны, то

$$E(MS_r) = \sigma^2 + \beta_1^2 \sum(x_i - \bar{x})^2$$

С помощью MS_r и MS_e можно тестировать значимость коэффициентов

Если дисперсии остатков для всех значений X равны, то

$$E(MS_r) = \sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2 = \sigma^2 + \sigma_x^2$$

$$E(MS_e) = \sigma^2$$

С помощью MS_r и MS_e можно тестировать значимость коэффициентов

Если дисперсии остатков для всех значений X равны, то

$$E(MS_r) = \sigma^2 + \beta_1^2 \sum(x_i - \bar{x})^2 = \sigma^2 + \sigma_x^2$$

$$E(MS_e) = \sigma^2$$

Если зависимости нет, то $\beta_1 = 0$, и тогда

С помощью MS_r и MS_e можно тестировать значимость коэффициентов

Если дисперсии остатков для всех значений X равны, то

$$E(MS_r) = \sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2 = \sigma^2 + \sigma_x^2$$

$$E(MS_e) = \sigma^2$$

Если зависимости нет, то $\beta_1 = 0$, и тогда $MS_r \approx MS_e$

С помощью MS_r и MS_e можно тестировать значимость коэффициентов

Если дисперсии остатков для всех значений X равны, то

$$E(MS_r) = \sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2 = \sigma^2 + \sigma_x^2$$

$$E(MS_e) = \sigma^2$$

Если зависимости нет, то $\beta_1 = 0$, и тогда $MS_r \approx MS_e$

Значит, $MS_r \approx MS_e$

С помощью MS_r и MS_e можно тестировать значимость коэффициентов

Если дисперсии остатков для всех значений X равны, то

$$E(MS_r) = \sigma^2 + \beta_1^2 \sum(x_i - \bar{x})^2 = \sigma^2 + \sigma_x^2$$

$$E(MS_e) = \sigma^2$$

Если зависимости нет, то $\beta_1 = 0$, и тогда $MS_r \approx MS_e$

Значит, $MS_r \approx MS_e$

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$

$$F_{df_r, df_e} = \frac{MS_r}{MS_e}$$

Тестирование значимости коэффициентов регрессии при помощи F-критерия

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$

$$F_{df_r, df_e} = \frac{MS_r}{MS_e}$$

Для простой линейной регрессии $df_r = 1$ и $df_e = n - 2$

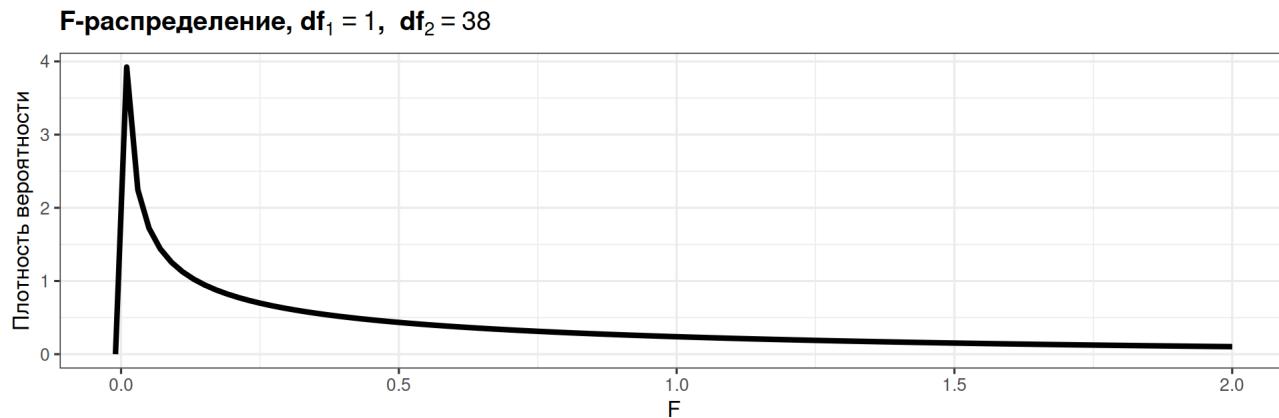


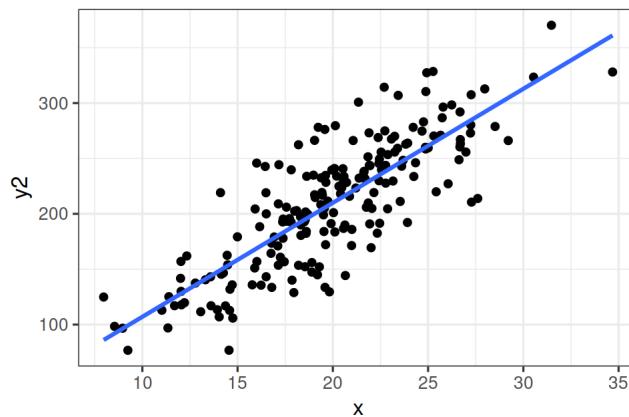
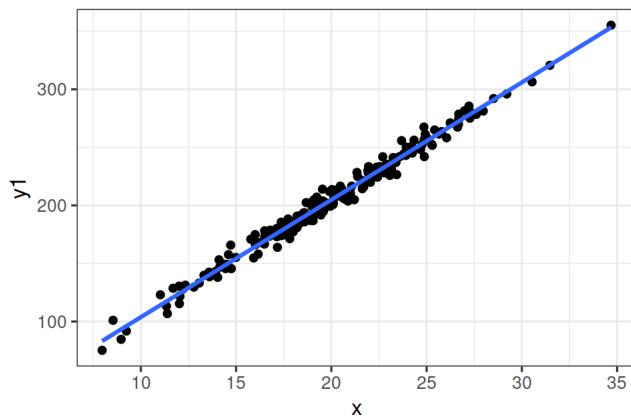
Таблица результатов дисперсионного анализа

Источник изменчивости	df	SS	MS	F
Регрессия	$df_r = 1$	$SS_r = \sum (\bar{y} - \hat{y}_i)^2$	$MS_r = \frac{SS_r}{df_r}$	$F_{df_r, df_e} = \frac{MS_r}{MS_e}$
Остаточная	$df_e = n - 2$	$SS_e = \sum (y_i - \hat{y}_i)^2$	$MS_e = \frac{SS_e}{df_e}$	
Общая	$df_t = n - 1$	$SS_t = \sum (\bar{y} - y_i)^2$		

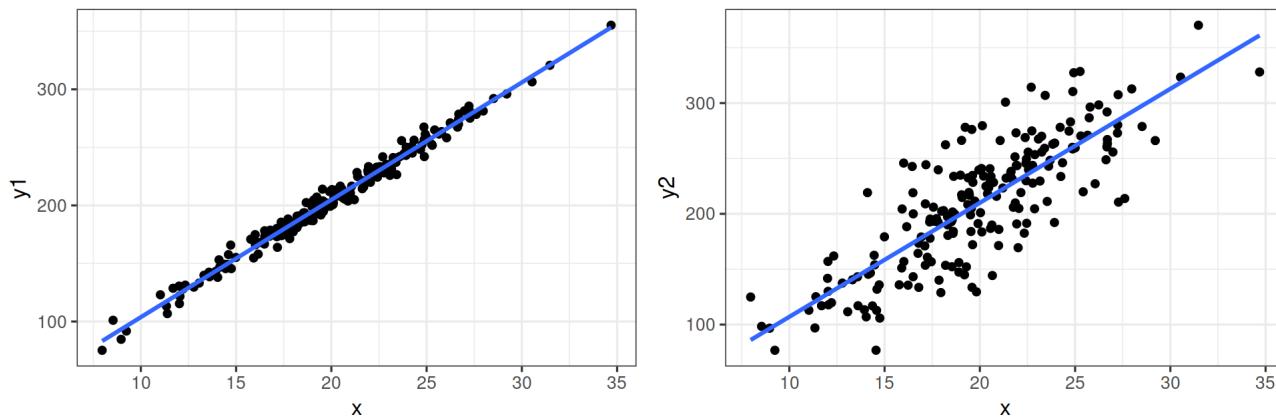
Минимальное упоминание результатов в тексте должно содержать F_{df_r, df_e} и p .

Оценка качества подгонки модели

В чем различие между этими двумя моделями?



В чем различие между этими двумя моделями?



У этих моделей разный разброс остатков:

- Модель слева объясняет практически всю изменчивость
- Модель справа объясняет не очень много изменчивости

Коэффициент детерминации — мера качества подгонки модели

Коэффициент детерминации

описывает какую долю дисперсии зависимой переменной объясняет модель

$$R^2 = \frac{SS_r}{SS_t}$$

- $0 < R^2 < 1$
- $R^2 = r^2$ — для простой линейной регрессии коэффициент детерминации равен квадрату коэффициента Пирсоновской корреляции

Если в модели много предикторов, нужно внести поправку

Скорректированный коэффициент детерминации (adjusted R-squared)

Применяется если необходимо сравнить две модели с разным количеством параметров

$$R_{adj}^2 = 1 - (1 - R^2) \frac{n - 1}{n - p}$$

p - количество параметров в модели

Вводится штраф за каждый новый параметр

Еще раз смотрим на результаты регрессионного анализа зависимости IQ от размеров мозга

```
summary(brain_model)

##
## Call:
## lm(formula = PIQ ~ MRINACount, data = brain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -39.6   -17.9   -1.6    17.0    42.3 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.7437570 42.3923825    0.04    0.967    
## MRINACount  0.0001203  0.0000465    2.59    0.014 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21 on 38 degrees of freedom
## Multiple R-squared:  0.15,    Adjusted R-squared:  0.127  
## F-statistic: 6.69 on 1 and 38 DF,  p-value: 0.0137
```

Как записываются результаты регрессионного анализа в тексте статьи?

Мы показали, что связь между результатами теста на IQ и размером головного мозга на МРТ описывается моделью вида

$$IQ = 1.74 + 0.00012 \text{ MRINACount} \quad (F_{1,38} = 6.686, p = 0.0136, R^2 = 0.149)$$

Как записываются результаты регрессионного анализа в тексте статьи?

Мы показали, что связь между результатами теста на IQ и размером головного мозга на МРТ описывается моделью вида

$$IQ = 1.74 + 0.00012 \text{ MRINACount} \quad (F_{1,38} = 6.686, p = 0.0136, R^2 = 0.149)$$

Неужели уже пора писать статью?

Задание 1 & 2

Выполните задания 1 и 2 в одном из этих файлов:

- 07_task_assumptions_catsM.R
- 07_task_assumptions_GAG.R

Зачем нужна диагностика линейных моделей

Зачем нужна диагностика модели? Разве тестов было недостаточно?

```
dat <- read.table('data/orly_owl_Lin_4p_5_flat.txt')
fit <- lm(V1 ~ V2 + V3 + V4 + V5 - 1, data = dat)
coef(summary(fit))

##      Estimate Std. Error t value Pr(>|t|)
## V2      0.986     0.1280    7.70 1.99e-14
## V3      0.971     0.1266    7.67 2.50e-14
## V4      0.861     0.1196    7.20 8.30e-13
## V5      0.927     0.0833   11.13 4.78e-28
```

Все значимо? Пишем статью?

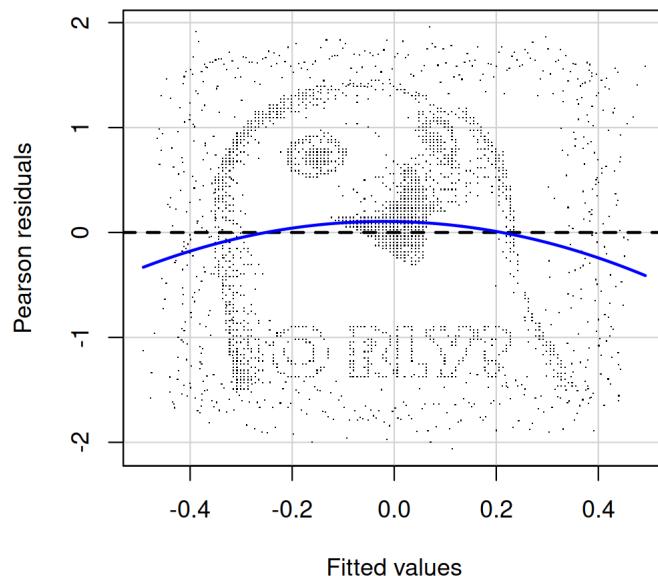
Задание 3

Постройте график зависимости остатков от предсказанных значений при помощи этого кода

```
library(car)
residualPlot(fit, pch = ".")
```

Oh, really?

```
library(car)
residualPlot(fit, pch = ".")
```



http://www4.stat.ncsu.edu/~stefanski/NSF_Supported/Hidden_Images/stat_res_plots.html

Анализ остатков линейных моделей

- 1) Проверка на наличие влиятельных наблюдений
- 2) Проверка условий применимости линейных моделей

1. Линейная связь
2. Независимость
3. Нормальное распределение
4. Гомогенность дисперсий
5. Отсутствие коллинеарности предикторов (для множественной регрессии)

Анализ остатков

Какие бывают остатки?

Какие бывают остатки?

“Сырые” остатки

$$e_i = y_i - \hat{y}_i$$

Какие бывают остатки?

“Сырые” остатки

$$e_i = y_i - \hat{y}_i$$

Стандартизованные (стьюдентизированные) остатки

$$s_i = \frac{e_i}{\sqrt{MS_e(1 - h_{ii})}}$$

- легко сравнивать (стандартизованы), учитывают силу влияния наблюдений
- $\sqrt{MS_e}$ — стандартное отклонение остатков
- h_{ii} — “сила воздействия” отдельных наблюдений (leverage, рычаг проекционной матрицы)

Что такое проекционная матрица?

По определению, остатки $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$.

Тогда $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b} = \mathbf{Y} - \mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Y}$.

Обозначим $\mathbf{H} \equiv \mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'$.

Матрица \mathbf{H} — называется “хэт”-матрица (hat-matrix) или проекционная матрица, т.к. она позволяет получить предсказанные значения из наблюдаемых.

$$\hat{\mathbf{Y}} = \mathbf{HY}$$

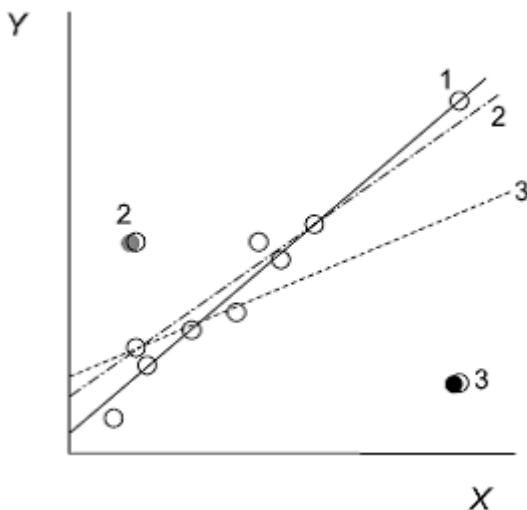
Тогда остатки можно получить как $\mathbf{e} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$.

Диагональные элементы проекционной матрицы — это мера воздействия точек на ход линии регрессии.

Влиятельные наблюдения

Влияние наблюдения

Влияние наблюдения — это наблюдения, которые вносят слишком большой вклад в оценку параметров (коэффициентов) модели.

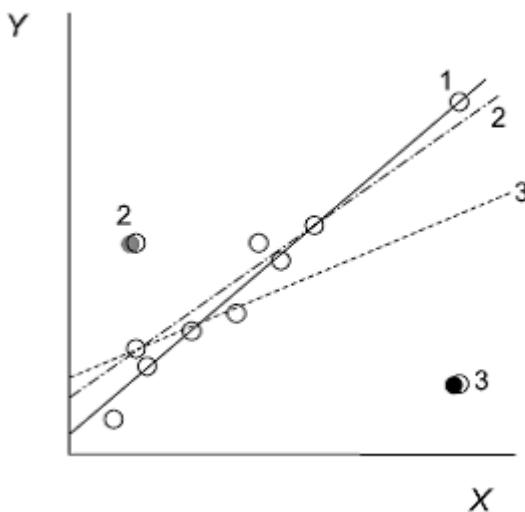


Из кн. Quinn, Keugh, 2002

Учет каких из этих точек повлияет на ход регрессии и почему?

Влияние наблюдения

Влияние наблюдения — это наблюдения, которые вносят слишком большой вклад в оценку параметров (коэффициентов) модели.



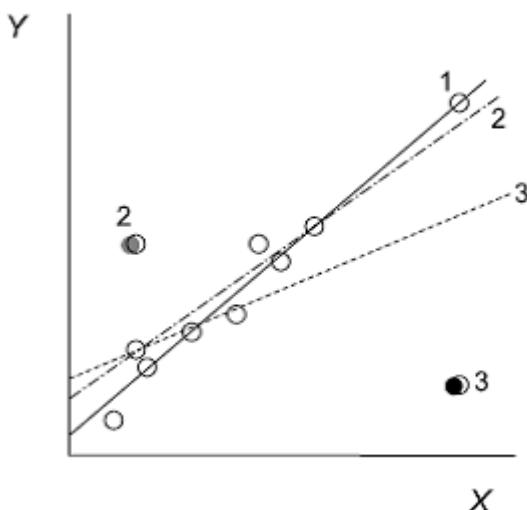
Из кн. Quinn, Keugh, 2002

Учет каких из этих точек повлияет на ход регрессии и почему?

- Точка 1 почти не повлияет, т.к. у нее маленький остаток, хоть и большой X
- Точка 2 почти не повлияет, т.к. ее X близок к среднему, хоть и большой остаток
- Точка 3 повлияет сильно, т.к. у нее не только большой остаток, но и большой X

Воздействие точек h_{ii} (leverage)

показывает силу влияния значений x_i на ход линии регрессии, то есть на \hat{y}_i

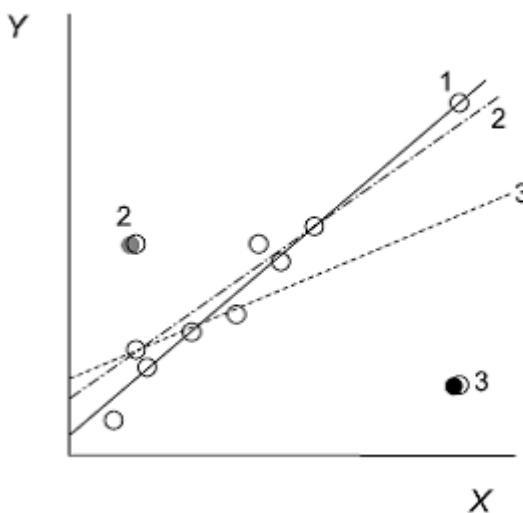


Weighing Machine by neys fadzil on Flickr

Из кн. Quinn, Keough, 2002

Воздействие точек h_{ii} (leverage)

показывает силу влияния значений x_i на ход линии регрессии, то есть на \hat{y}_i



Weighing Machine by neys fadzil on Flickr

Из кн. Quinn, Keough, 2002

Точки, располагающиеся дальше от \bar{x} , оказывают более сильное влияние на \hat{y}_i

- h_{ii} варьирует в промежутке от $1/n$ до 1
- Если $h_{ii} > 2(p/n)$, то надо внимательно посмотреть на данное значение (p — число параметров, n — объем выборки)

Расстояние Кука (Cook's distance)

описывает, как повлияет на модель удаление данного наблюдения

$$D_i = \frac{\sum (\hat{y}_j - \hat{y}_{j(i)})^2}{p MS_e} = \frac{e_i^2}{p MS_e} \frac{h_{ii}}{(1 - h_{ii})^2}$$

- \hat{y}_j - значение предсказанное полной моделью
- $\hat{y}_{j(i)}$ - значение, предсказанное моделью, построенной без учета i -го значения предиктора
- p - количество параметров в модели
- MS_e - среднеквадратичная ошибка модели ($\hat{\sigma}^2$)
- h_{ii} — “сила воздействия” отдельных наблюдений (leverage)

Расстояние Кука (Cook's distance)

описывает, как повлияет на модель удаление данного наблюдения

$$D_i = \frac{\sum (\hat{y}_j - \hat{y}_{j(i)})^2}{p MS_e} = \frac{e_i^2}{p MS_e} \frac{h_{ii}}{(1 - h_{ii})^2}$$

- \hat{y}_j - значение предсказанное полной моделью
 - $\hat{y}_{j(i)}$ - значение, предсказанное моделью, построенной без учета i -го значения предиктора
 - p - количество параметров в модели
 - MS_e - среднеквадратичная ошибка модели ($\hat{\sigma}^2$)
 - h_{ii} — “сила воздействия” отдельных наблюдений (leverage)
-
- Зависит одновременно от величины остатков и “силы воздействия” наблюдений.
 - Условное пороговое значение. Наблюдение является выбросом (outlier), если:
 - $D_i > 1$ — это “мягкий” порог
 - $D_i > 4/(n - p)$ (n — объем выборки, p — число параметров) — это “жесткий” порог

Что делать с наблюдениями-выбросами?

Что делать с наблюдениями-выбросами?

- Удалить?

Что делать с наблюдениями-выбросами?

- Удалить?

Осторожно! Только очевидные ошибки в наблюдениях можно удалять. Лучше найти причины.

Что делать с наблюдениями-выбросами?

- Удалить?

Осторожно! Только очевидные ошибки в наблюдениях можно удалять. Лучше найти причины.

- Трансформировать? Это не всегда поможет.
- Иногда можно переформулировать модель.

Некоторые виды трансформаций

Трансформация	Формула
степень -2	$1/x^2$
степень -1	$1/x$
степень -0.5	$1/\sqrt{x}$
степень 0.5	\sqrt{x}
логарифмирование	$\log(x)$

Данные для анализа остатков

```
library(ggplot2)
brain_diag <- fortify(brain_model)
head(brain_diag, 2)

##   PIQ MRINACount   .hat .sigma .cooksdi .fitted .resid .stdresid
## 1 124     816932 0.0664    20.9 0.049838      100  24.02    1.1840
## 2 124     1001121 0.0669    21.3 0.000304      122   1.87    0.0921
```

- `.hat` — “сила воздействия” данного наблюдения (leverage)
- `.cooksdi` — расстояние Кука
- `.fitted` — предсказанные значения
- `.resid` — остатки
- `.stdresid` — стандартизованные остатки

График расстояния Кука

Проверяем наличие влиятельных наблюдений в `brain_model`.

Значения на графике расстояния Кука приведены в том же порядке, что и в исходных данных.

```
# График расстояния Кука
ggplot(brain_diag, aes(x = 1:nrow(brain_diag), y = .cooksdi)) +
  geom_bar(stat = "identity")
```

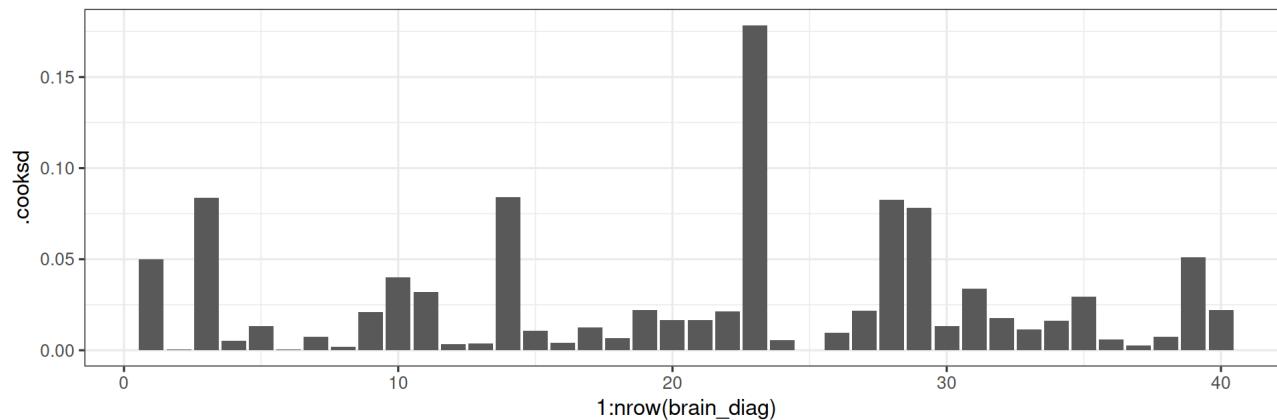
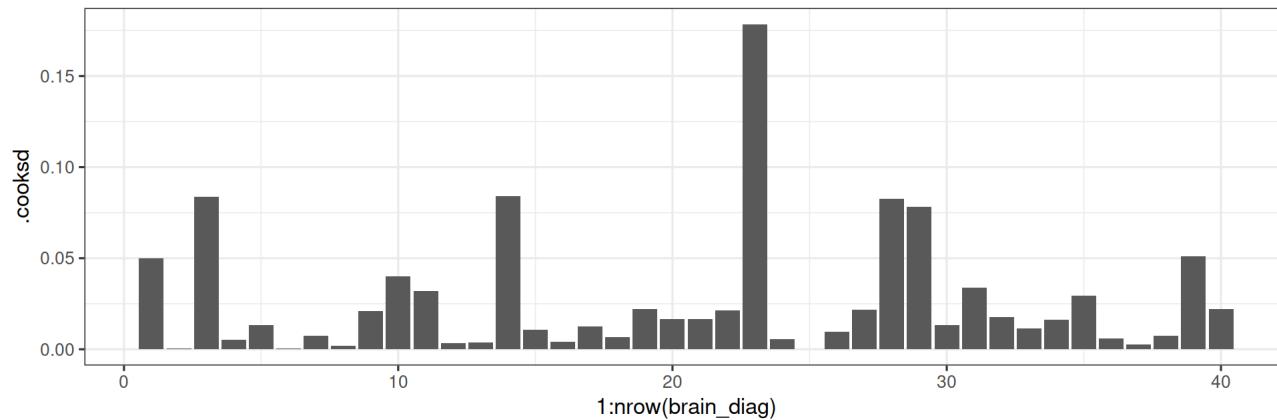


График расстояния Кука

Проверяем наличие влиятельных наблюдений в `brain_model`.

Значения на графике расстояния Кука приведены в том же порядке, что и в исходных данных.

```
# График расстояния Кука
ggplot(brain_diag, aes(x = 1:nrow(brain_diag), y = .cooksdi)) +
  geom_bar(stat = "identity")
```

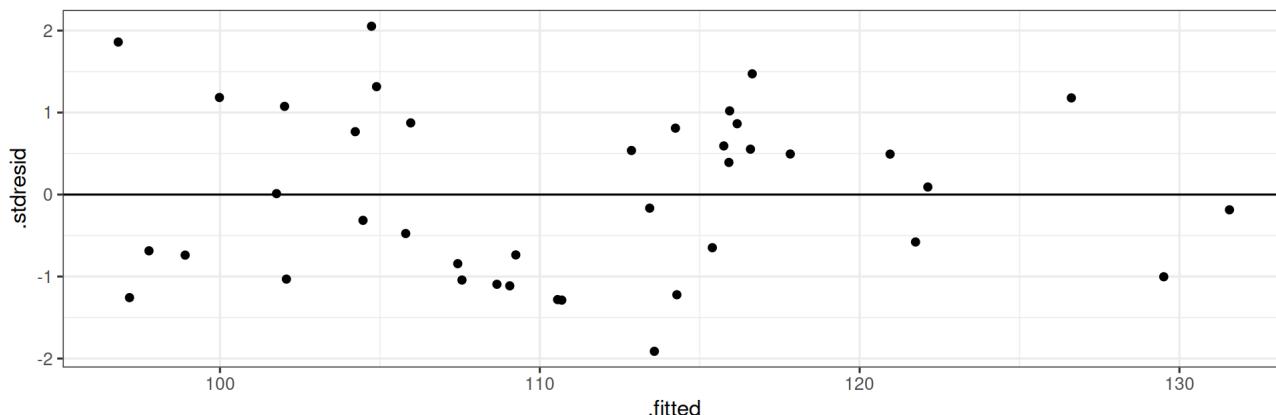


- Есть одно влиятельное наблюдение, которое нужно проверить, но сила его влияния невелика (расстояние Кука < 1 , и только одно наблюдение больше $4/(n - p) = 0.11$)

График остатков от предсказанных значений

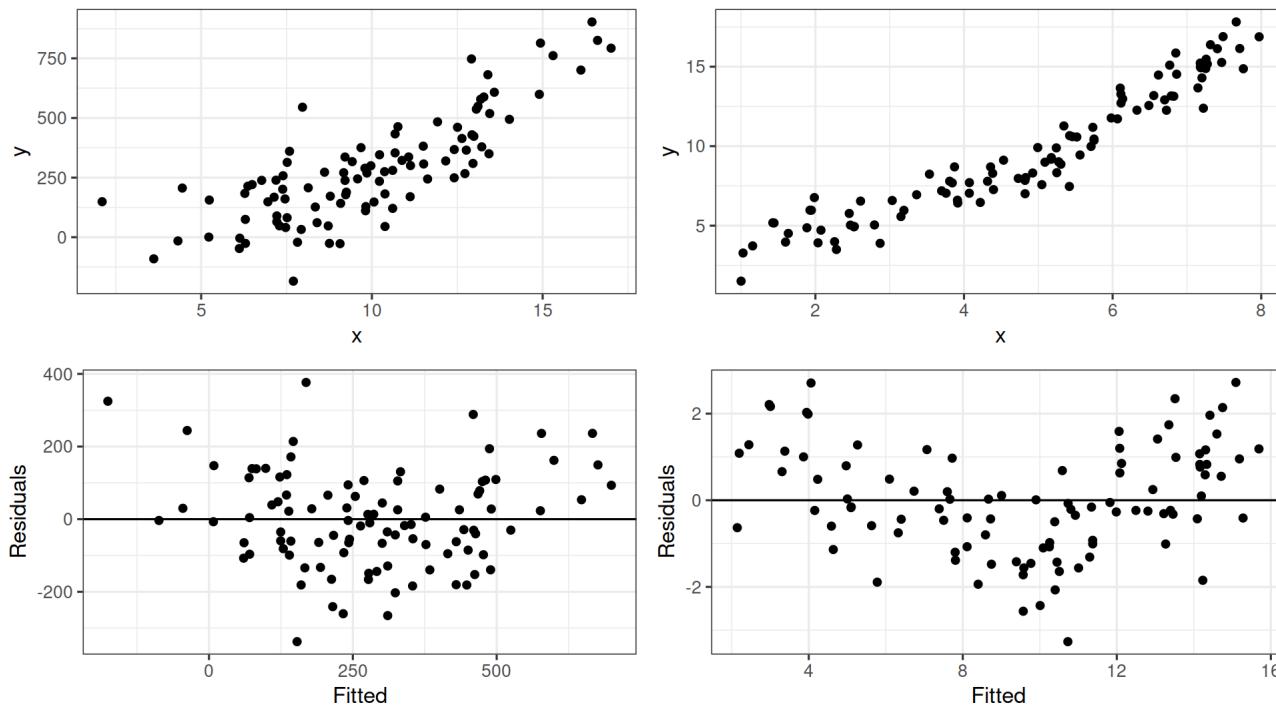
Большую часть того, что нужно знать про остатки вы увидите на этом графике. А сейчас давайте научимся читать такой график.

```
gg_resid <- ggplot(data = brain_diag, aes(x = .fitted, y = .stdresid)) +  
  geom_point() +  
  geom_hline(yintercept = 0)  
gg_resid
```



1. Линейность связи

Нелинейность связи видна на графиках остатков



Проверка на линейность связи

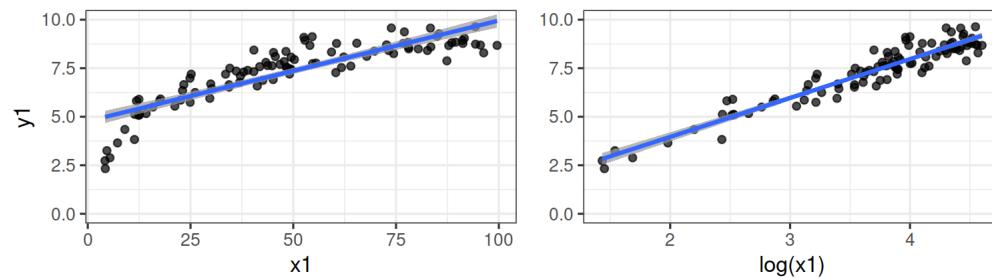
- График зависимости y от x (и от других переменных, не включенных в модель)
- График остатков от предсказанных значений

Что делать, если связь нелинейна?

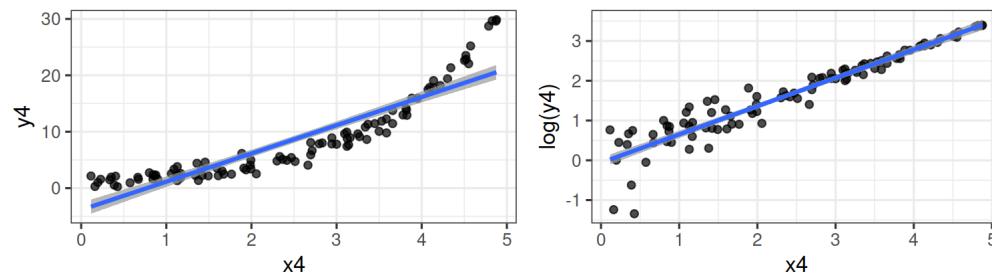
- Добавить неучтенные переменные или взаимодействия
- Применить линеаризующее преобразование (Осторожно!)
- Применить обобщенную линейную модель с другой функцией связи (GLM)
- Построить аддитивную модель (GAM), если достаточно наблюдений по x
- Построить нелинейную модель, если известна форма зависимости

Примеры линеаризующих преобразований

Логарифмирование

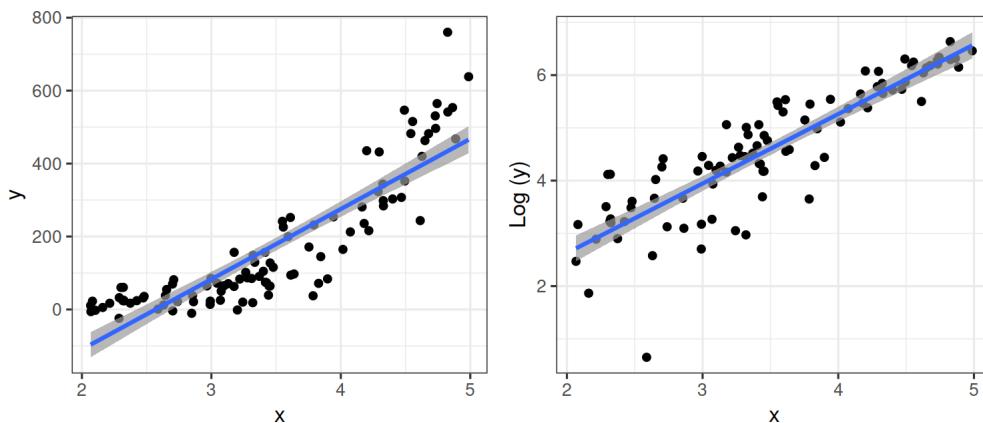


Возведение в степень



и т.д.

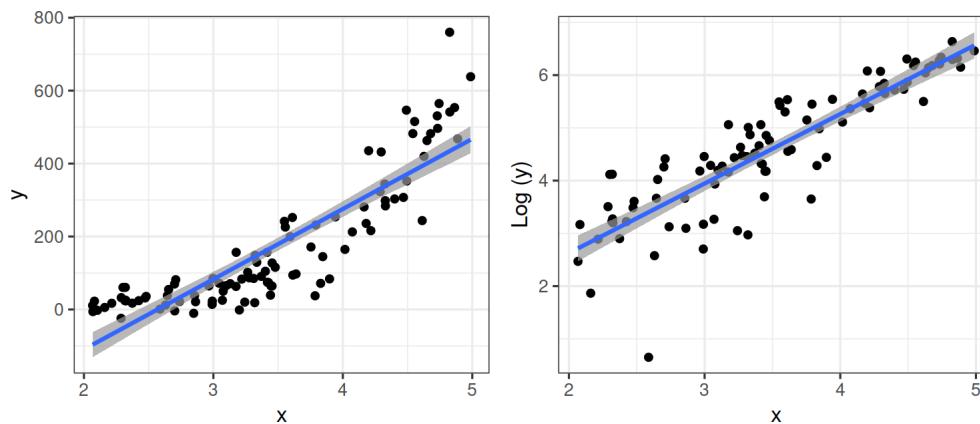
При прочих равных не стоит трансформировать отклик



Осторожно! Вы рискуете изучить не то, что хотели:

1. При логарифмировании отклика вы будете изучать поведение мат.ожидания логарифма $E(\log(y_i)) = b_0 + b_1 x_{1i} + \dots + e_i$.
2. Трансформация отклика не только линеаризует зависимость, но и затронет величину остатков e .

При прочих равных не стоит трансформировать отклик



Осторожно! Вы рискуете изучить не то, что хотели:

1. При логарифмировании отклика вы будете изучать поведение мат.ожидания логарифма $E(\log(y_i)) = b_0 + b_1 x_{1i} + \dots + e_i$.
2. Трансформация отклика не только линеаризует зависимость, но и затронет величину остатков e .

Вместо трансформации отклика лучше использовать обобщенную линейную модель с подходящей функцией связи, например:

$$\log(E(y_i)) = b_0 + b_1 x_{1i} + \dots + e_i$$

2. Независимость

Каждое значение y_i должно быть независимо от любого другого y_j

Это нужно контролировать на этапе планирования сбора материала

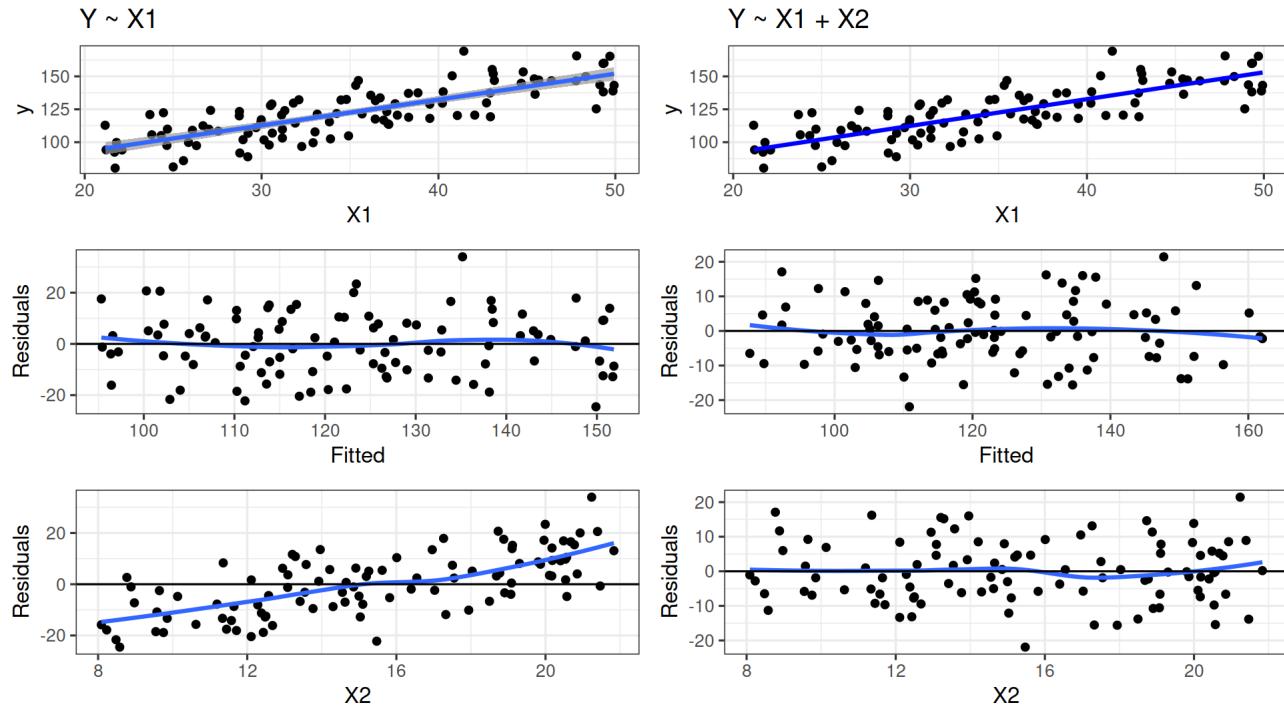
- Наиболее частые источники зависимостей:
 - псевдоповторности (повторно измеренные объекты)
 - неучтенные переменные
 - временные автокорреляции (если данные - временной ряд)
 - пространственные автокорреляции (если пробы взяты в разных местах)
 - и т.п.

Диагностика нарушений независимости

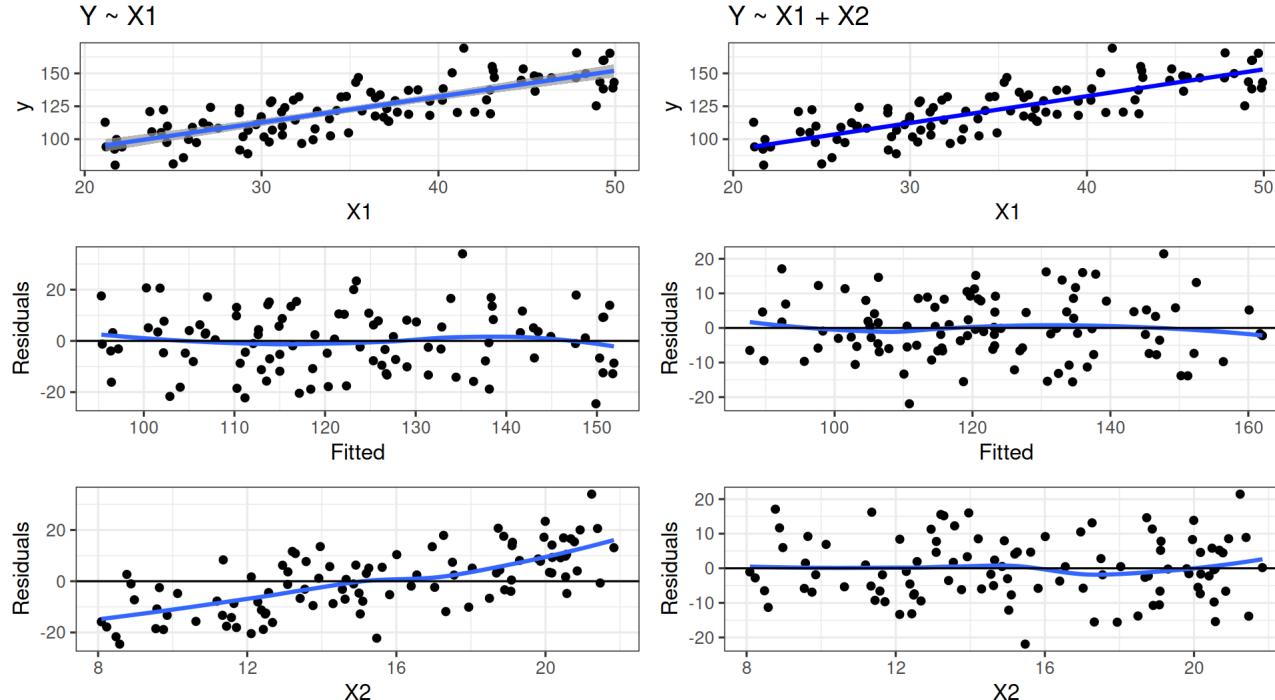
Взаимозависимости можно заметить на графиках остатков

- остатки VS. предсказанные значения
- остатки VS. переменные в модели
- остатки VS. переменные не в модели

Нарушение условия независимости: Неучтенная переменная



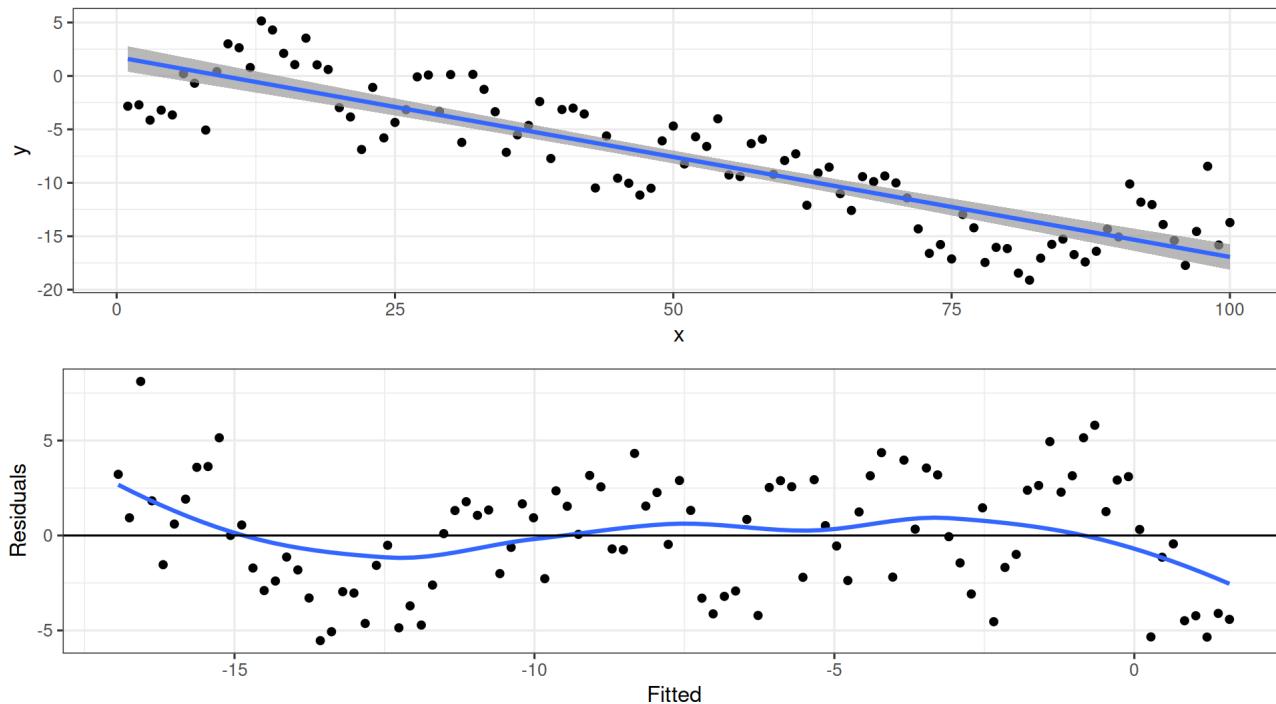
Нарушение условия независимости: Неучтенная переменная



- Слева: Если в модели не учтена переменная X_2 , внешне все нормально, но величина остатков зависит от X_2
- Справа: Если X_2 учесть, то зависимость остатков от X_2 исчезает

Нарушение условия независимости: Автокорреляция

В данном случае, наблюдения — это временной ряд.



На графиках остатков четко видно, что остатки не являются независимыми.

Проверка на автокорреляцию

Проверка на автокорреляцию нужна если данные это временной ряд, или если известны координаты или время сбора проб.

Способы проверки временной автокорреляции (годятся, если наблюдения в ряду расположены через равные интервалы):

- График автокорреляционной функции остатков (ACF-plot) покажет корреляции с разными лагами.
- Критерий Дарбина-Уотсона (значимость автокорреляции 1-го порядка).

Для проверки пространственных автокорреляций

- вариограмма
- I Морана (Moran's I)

Что делать, если у вас нарушено условие независимости значений?

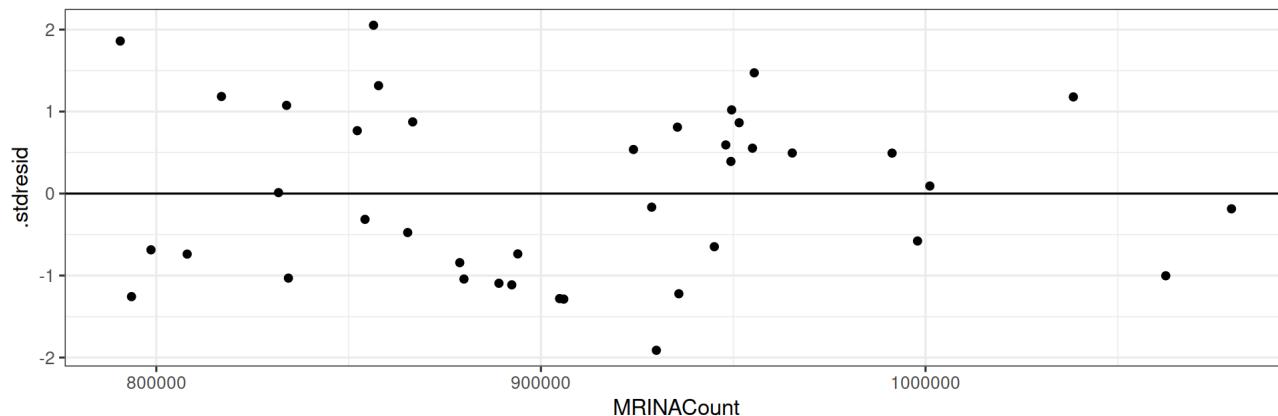
Выбор зависит от обстоятельств. Вот несколько возможных вариантов.

- псевдоповторности
 - избавляемся от псевдоповторностей, вычислив среднее
 - подбираем модель со случайным фактором
- неучтенные переменные
 - включаем в модель (если возможно)
- временные автокорреляции
 - моделируем автокорреляцию
 - подбираем модель со случайным фактором
- пространственные автокорреляции
 - моделируем пространственную автокорреляцию
 - делим на пространственные блоки и подбираем модель со случайным фактором (= random effects model, mixed model)

Проверка условия независимости

Графики зависимости остатков от предикторов в модели

```
# Полный код  
ggplot(data = brain_diag, aes(x = MRINACount, y = .stdresid)) +  
  geom_point() +  
  geom_hline(yintercept = 0)
```



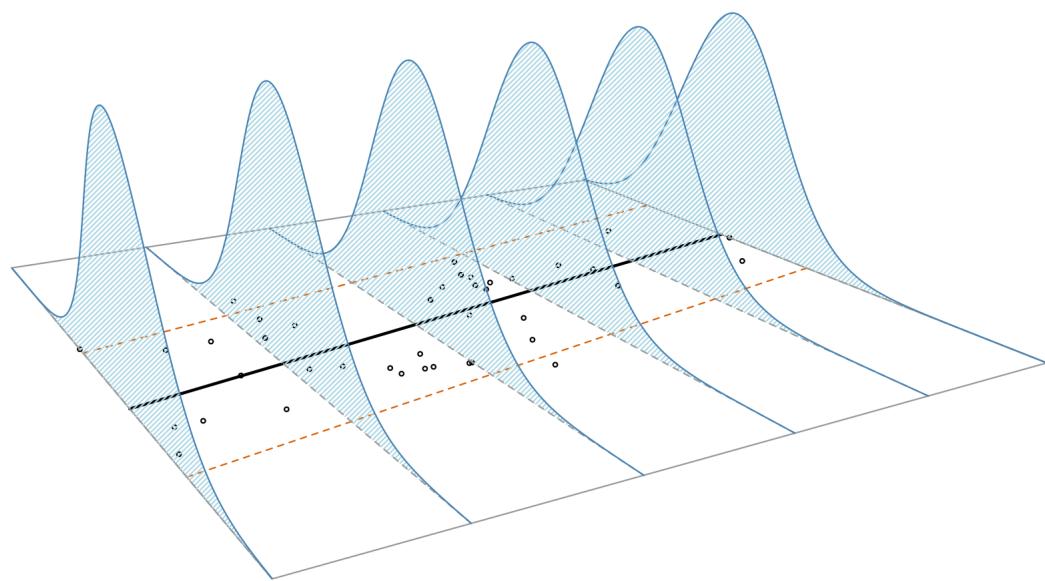
```
# То же самое с использованием ранее созданного gg_resid  
gg_resid + aes(x = MRINACount)
```

Графики зависимости остатков от предикторов не в модели

В данном случае их нет

3. Нормальное распределение

3. Нормальное распределение y (для каждого уровня значений x)



Проверка нормальности распределения остатков

Если y это нормально распределенная случайная величина

$$y_i \sim N(\mu_{y_i}, \sigma)$$

Проверка нормальности распределения остатков

Если y это нормально распределенная случайная величина

$$y_i \sim N(\mu_{y_i}, \sigma)$$

и мы моделируем ее как

$$y_i = b_0 + b_1 x_{1i} + \dots + e_i$$

Проверка нормальности распределения остатков

Если y это нормально распределенная случайная величина

$$y_i \sim N(\mu_{y_i}, \sigma)$$

и мы моделируем ее как

$$y_i = b_0 + b_1 x_{1i} + \dots + e_i$$

то остатки от этой модели — тоже нормально распределенная случайная величина

$$e_i \sim N(0, \sigma)$$

Проверка нормальности распределения остатков

Если y это нормально распределенная случайная величина

$$y_i \sim N(\mu_{y_i}, \sigma)$$

и мы моделируем ее как

$$y_i = b_0 + b_1 x_{1i} + \dots + e_i$$

то остатки от этой модели — тоже нормально распределенная случайная величина

$$e_i \sim N(0, \sigma)$$

Т.е. выполнение этого условия можно оценить по поведению случайной части модели.

Проверка нормальности распределения остатков

Есть формальные тесты, но:

- у формальных тестов тоже есть свои условия применимости
- при больших выборках формальные тесты покажут, что значимы даже небольшие отклонения от нормального распределения
- тесты, которые используются в линейной регрессии, устойчивы к небольшим отклонениям от нормального распределения

Проверка нормальности распределения остатков

Есть формальные тесты, но:

- у формальных тестов тоже есть свои условия применимости
- при больших выборках формальные тесты покажут, что значимы даже небольшие отклонения от нормального распределения
- тесты, которые используются в линейной регрессии, устойчивы к небольшим отклонениям от нормального распределения

Лучший способ проверки — квантильный график остатков.

Квантильный график остатков

По оси X — квантили теоретического распределения, по оси Y — квантили остатков модели. Если наблюдаемое распределение соответствует теоретическому, то точки должны лежать вдоль прямой по диагонали графика.

Квантильный график остатков

По оси X — квантили теоретического распределения, по оси Y — квантили остатков модели. Если наблюдаемое распределение соответствует теоретическому, то точки должны лежать вдоль прямой по диагонали графика.

Обычные остатки должны подчиняться нормальному распределению $e \sim N(0, \sigma)$.

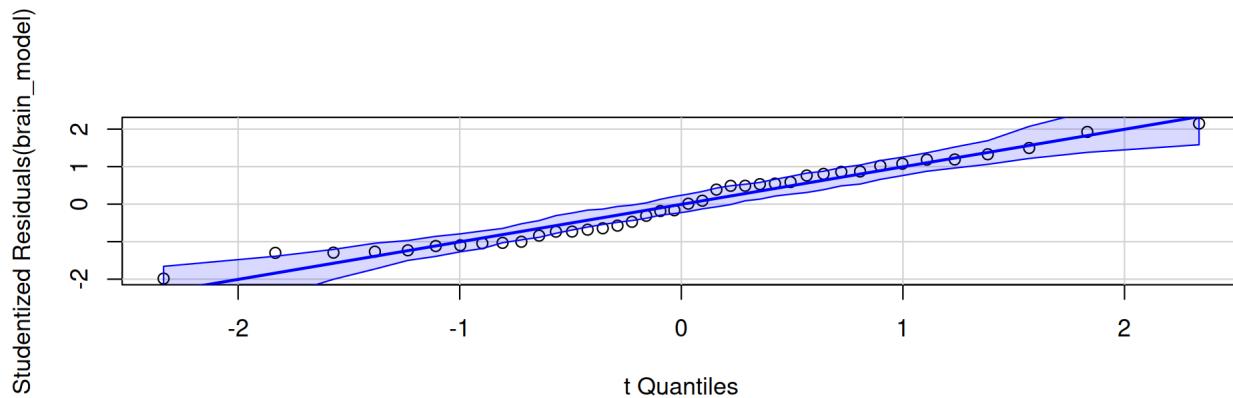
Квантильный график остатков

По оси X — квантили теоретического распределения, по оси Y — квантили остатков модели. Если наблюдаемое распределение соответствует теоретическому, то точки должны лежать вдоль прямой по диагонали графика.

Обычные остатки должны подчиняться нормальному распределению $e \sim N(0, \sigma)$.

Стьюдентизированные остатки — t-распределению.

```
# library(car)
qqPlot(brain_model, id = FALSE) # из пакета car
```



Что делать, если остатки распределены не нормально?

Зависит от причины

- Нелинейная связь?
 - Построить аддитивную модель (если достаточно наблюдений по x)
 - Построить нелинейную модель (если известна форма зависимости)
- Неучтенные переменные?
 - добавляем в модель
- Зависимая переменная распределена по-другому?
 - трансформируем данные (неудобно)
 - подбираем модель с другим распределением остатков (обобщенную линейную модель)

4. Постоянство дисперсии

4. Постоянство дисперсии (= гомогенность дисперсии, гомоскедастичность)

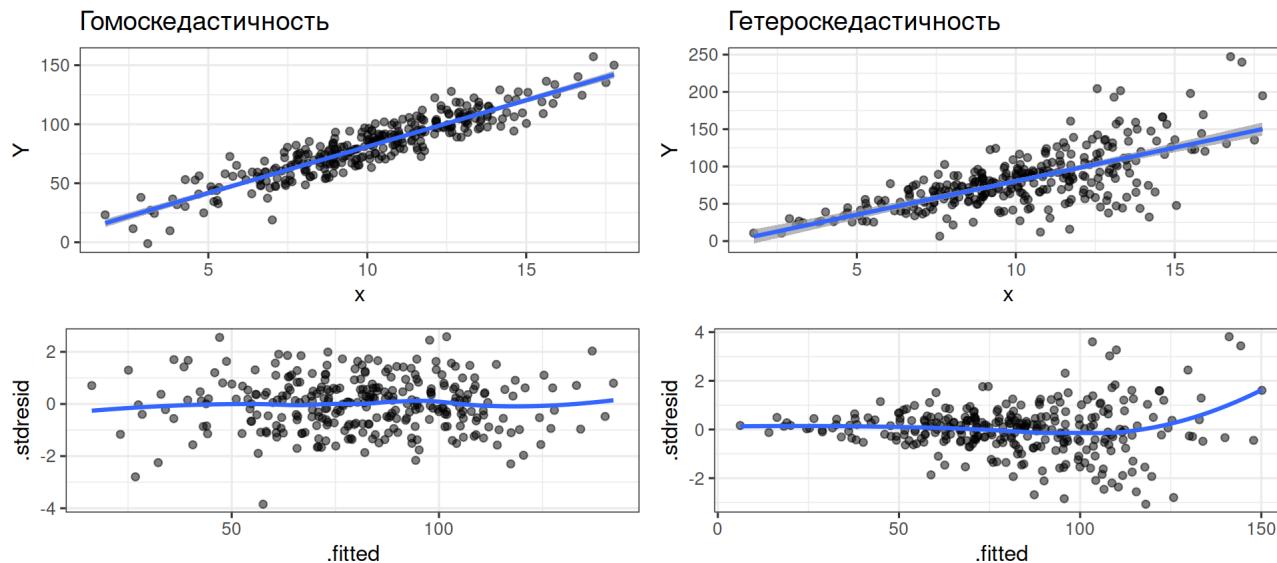
Это самое важное условие, поскольку многие тесты чувствительны к гетероскедастичности.

Проверка гомогенности дисперсий

Есть формальные тесты (тест Брайша-Пагана, тест Ко크рана), но:

- у формальных тестов тоже есть свои условия применимости, и многие сами неустойчивы к гетероскедастичности
- при больших выборках формальные тесты покажут, что значима даже небольшая гетероскедастичность

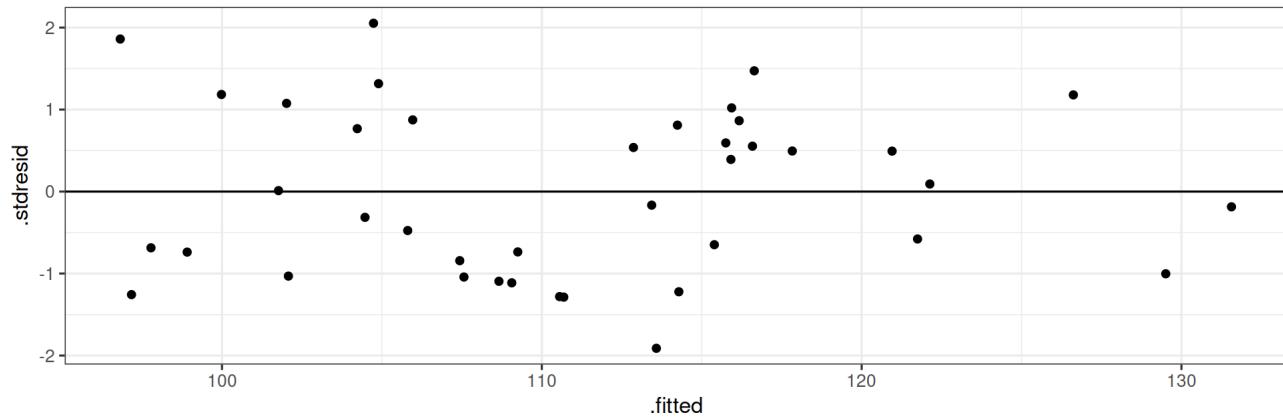
Лучший способ проверки на гомогенность дисперсий — график остатков.



Проверка на гетероскедастичность

Этот график у нас уже есть

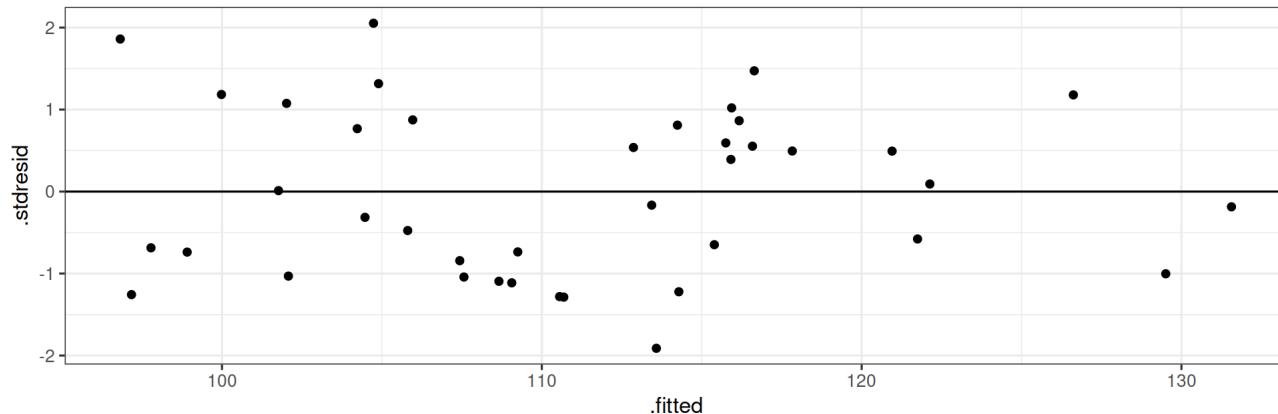
```
gg_resid
```



Проверка на гетероскедастичность

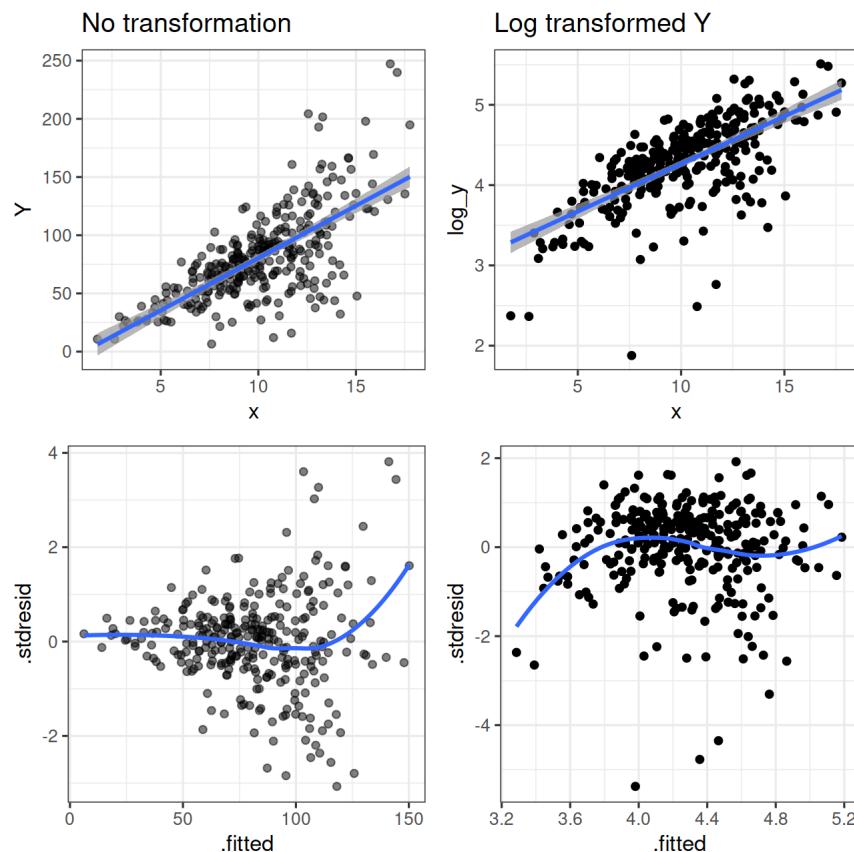
Этот график у нас уже есть

```
gg_resid
```



- Гетерогенность дисперсий не выражена.

Что делать если вы столкнулись с гетероскедастичностью?



Трансформация может помочь...

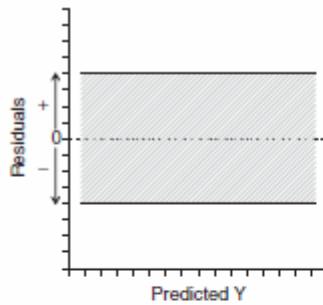
Возможные причины гетероскедастичности

Даже если трансформация может помочь, лучше поискать причину гетерогенности дисперсий

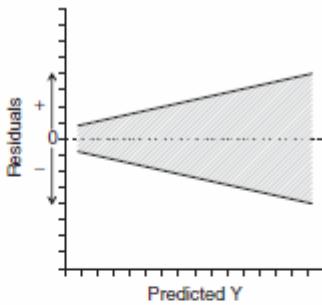
- Неучтенные переменные
 - добавляем в модель
- Зависимая переменная распределена по-другому
 - трансформируем данные (неудобно)
 - подбираем модель с другим распределением остатков (обобщенную линейную модель)
- Моделируем гетерогенность дисперсии.

Тренинг по анализу остатков

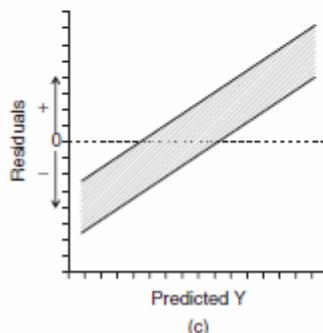
Некоторые частые паттерны на графиках остатков



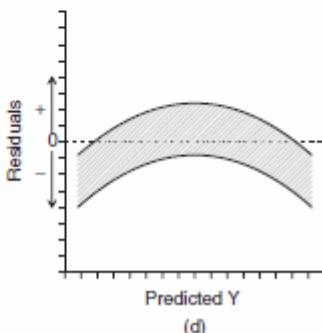
(a)



(b)



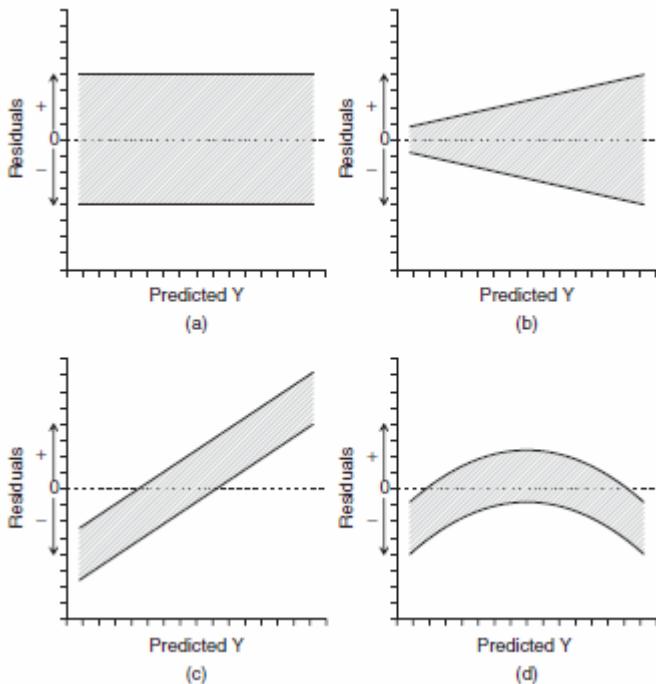
(c)



(d)

Из кн. Logan, 2010, стр. 174

Некоторые частые паттерны на графиках остатков



- Рис. а — Условия применимости соблюдаются, модель хорошая
- Рис. б — Клиновидный паттерн. Есть гетероскедастичность. Модель плохая
- Рис. с — Остатки рассеяны равномерно, но нужны дополнительные предикторы
- Рис. д — Нелинейный паттерн. Линейная модель использована некорректно

Из кн. Logan, 2010, стр. 174

Задание 4

Выполните три блока кода

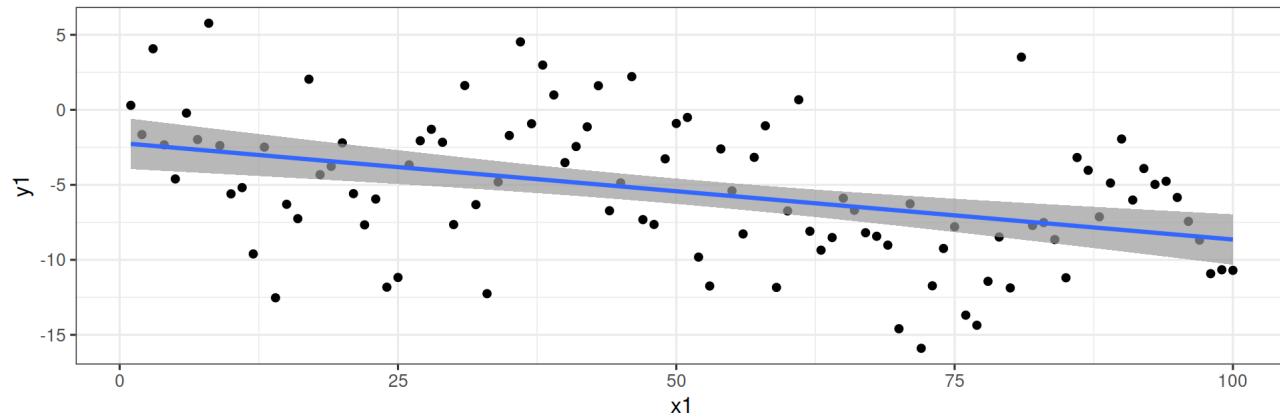
Какие нарушения условий применимости линейных моделей здесь наблюдаются?

Вам понадобятся

1. График расстояния Кука
2. График остатков от предсказанных значений
3. Графики остатков от предикторов в модели и не в модели
4. Квантильный график остатков

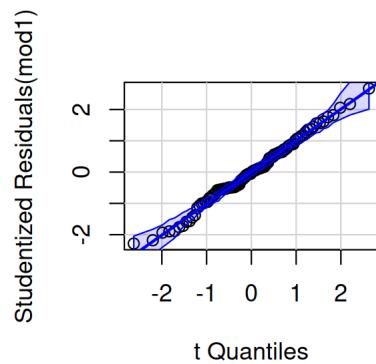
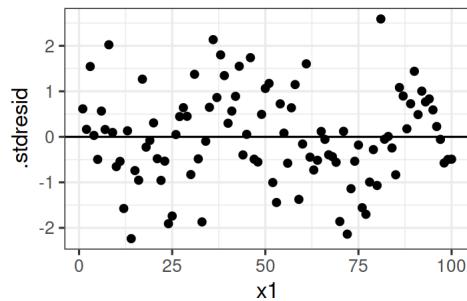
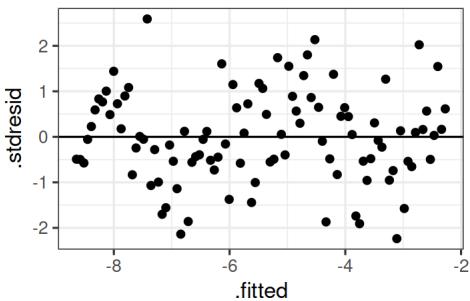
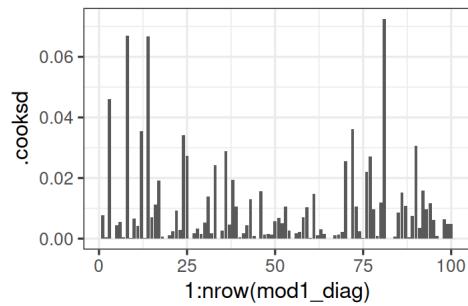
Задание 4, блок 1

```
set.seed(90829)
x1 <- seq(1, 100, 1)
y1 <- diffinv(rnorm(99)) + rnorm(100, 0.2, 4)
dat1 = data.frame(x1, y1)
ggplot(dat1, aes(x = x1, y = y1)) + geom_point() +
  geom_smooth(method="lm", alpha = 0.7)
```



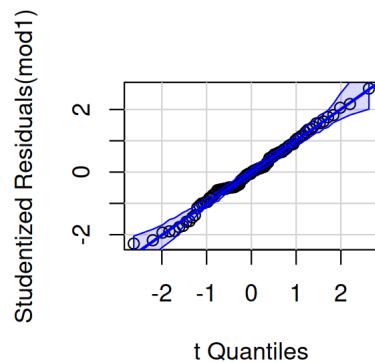
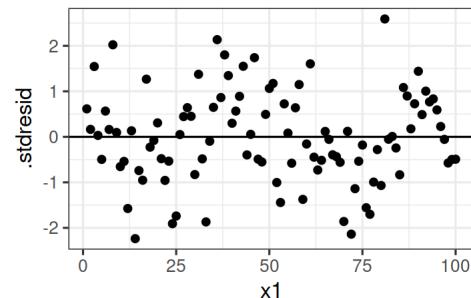
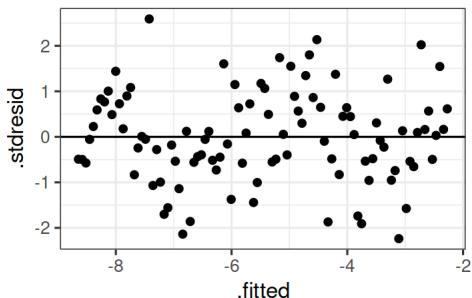
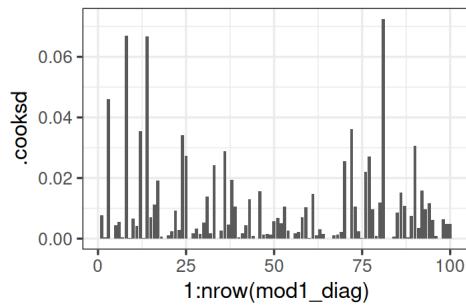
Решение, блок 1

Графики



Решение, блок 1

Графики



- Выбросов нет, зависимость нелинейна
- Небольшие отклонения от нормального распределения

Решение, блок 1

```
mod1 <- lm(y1 ~ x1, data = dat1)

# Данные для графиков остатков
mod1_diag <- fortify(mod1)

# 1) График расстояния Кука
ggplot(mod1_diag, aes(x = 1:nrow(mod1_diag), y = .cooksdi)) +
  geom_bar(stat = "identity")

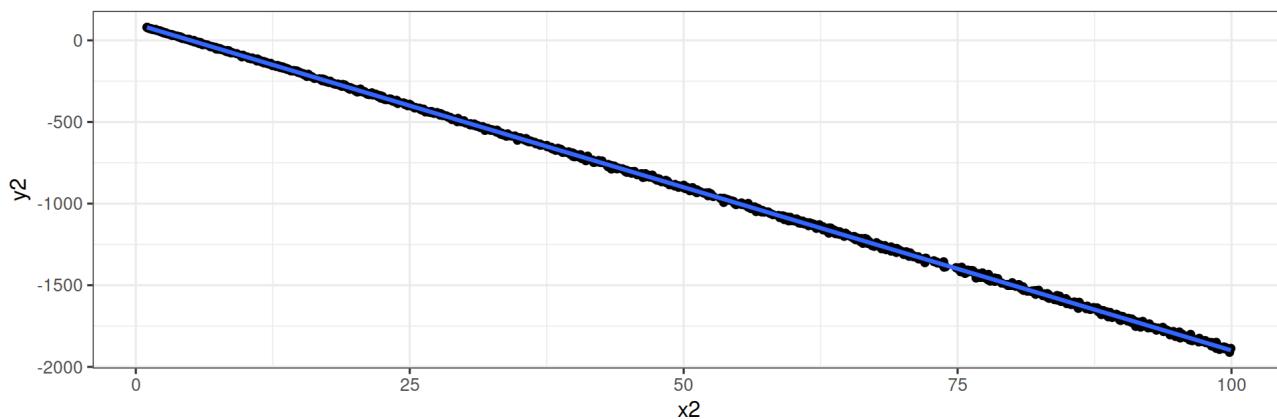
# 2) График остатков от предсказанных значений
gg_resid <- ggplot(data = mod1_diag, aes(x = .fitted, y = .stdresid)) +
  geom_point() + geom_hline(yintercept = 0)
gg_resid

# 3) Графики остатков от предикторов в модели и не в модели
gg_resid + aes(x = x1)

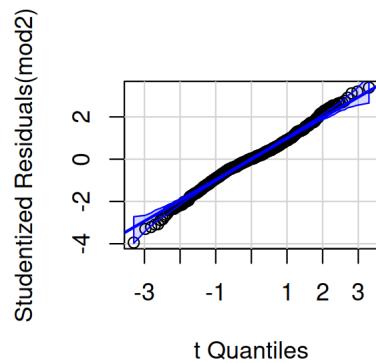
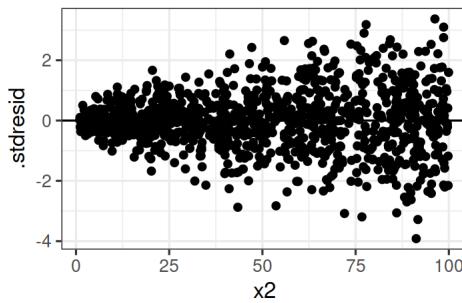
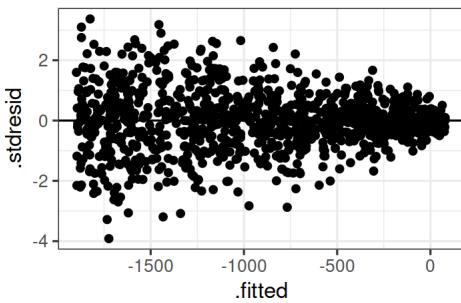
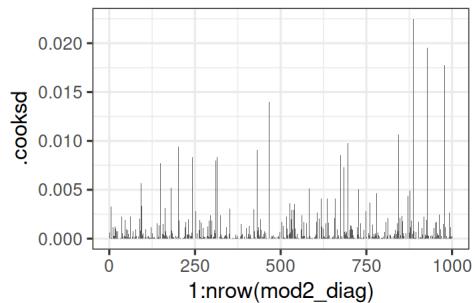
# 4) Квантильный график остатков
qqPlot(mod1, id = FALSE)
```

Задание 4, блок 2

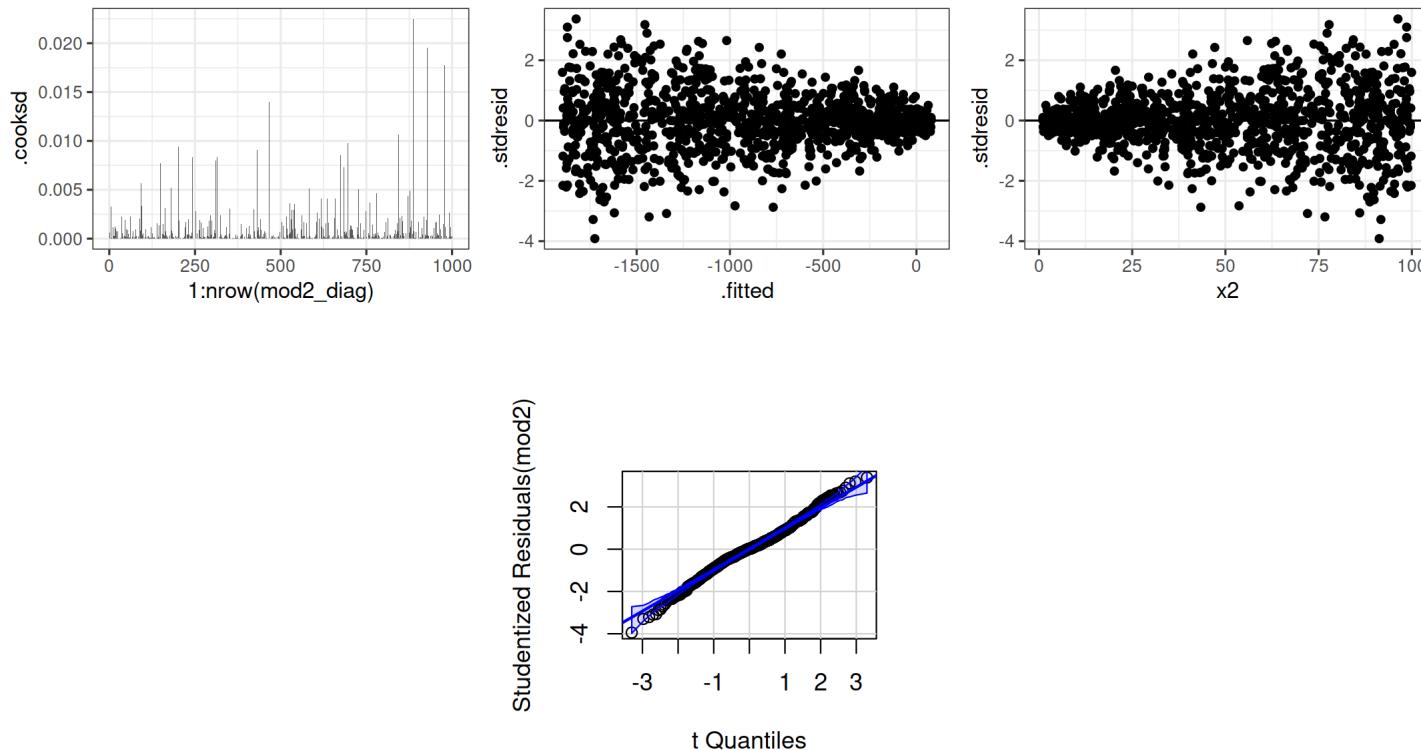
```
set.seed(7657674)
x2 <- runif(1000, 1, 100)
b_0 <- 100; b_1 <- -20
h <- function(x) x^0.5
eps <- rnorm(1000, 0, h(x2))
y2 <- b_0 + b_1 * x2 + eps
dat2 <- data.frame(x2, y2)
ggplot(dat2, aes(x = x2, y = y2)) +
  geom_point() +
  geom_smooth(method = "lm")
```



Решение, блок 2



Решение, блок 2



- Выбросов нет
- Гетерогенность дисперсий, остатки не подчиняются нормальному распределению

Решение, блок 2

```
mod2 <- lm(y2 ~ x2, data = dat2)

# Данные для графиков остатков
mod2_diag <- fortify(mod2)

# 1) График расстояния Кука
ggplot(mod2_diag, aes(x = 1:nrow(mod2_diag), y = .cooksdi)) +
  geom_bar(stat = "identity")

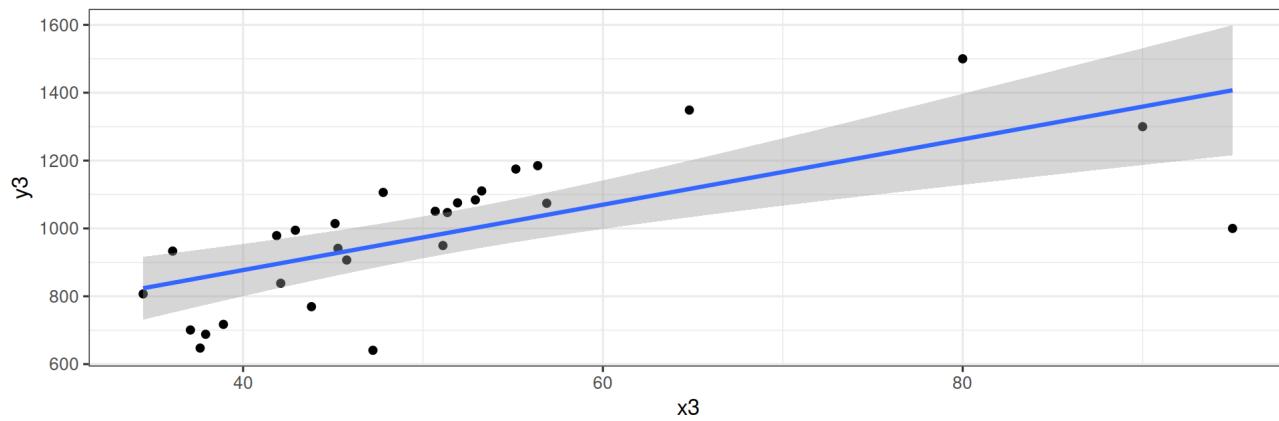
# 2) График остатков от предсказанных значений
gg_resid <- ggplot(data = mod2_diag, aes(x = .fitted, y = .stdresid)) +
  geom_point() + geom_hline(yintercept = 0)
gg_resid

# 3) Графики остатков от предикторов в модели и не в модели
gg_resid + aes(x = x2)

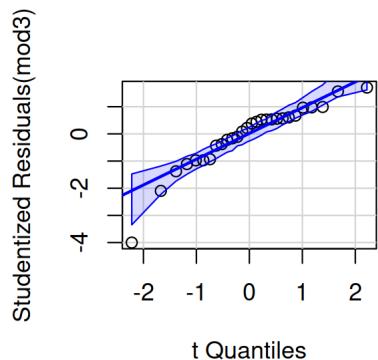
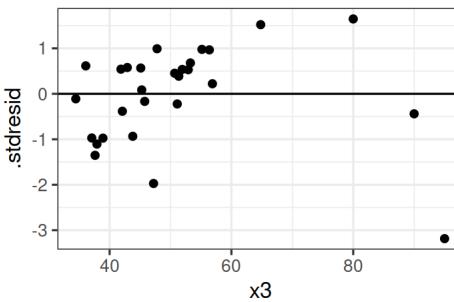
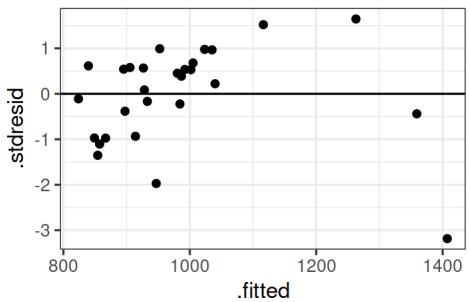
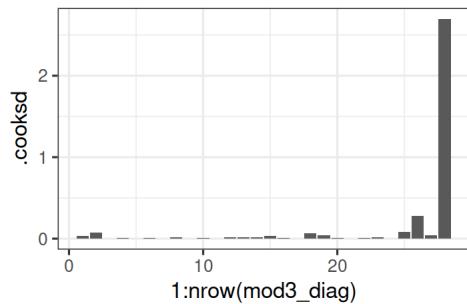
# 4) Квантильный график остатков
qqPlot(mod2, id = FALSE)
```

Задание 4, блок 3

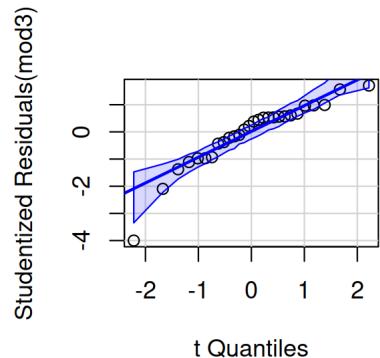
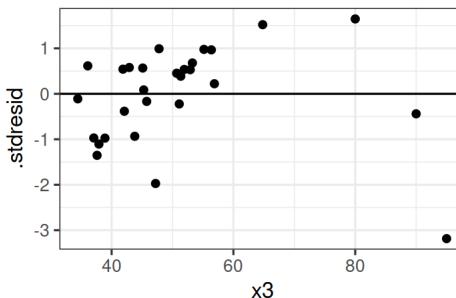
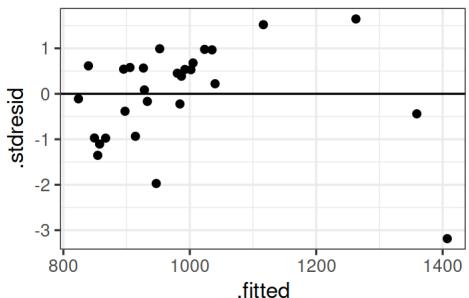
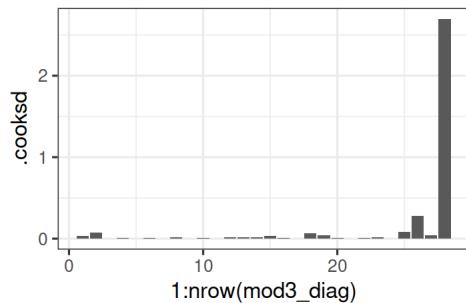
```
set.seed(9283)
x3 <- rnorm(25, 50, 10)
b_0 <- 20; b_1 <- 20; eps <- rnorm(50, 0, 100)
y3 <- b_0 + b_1*x3 + eps
y3[100] <- 1000; x3[100] <- 95; y3[99] <- 1300; x3[99] <- 90; y3[98] <- 1500; x3[98] <
dat3 <- data.frame(x3, y3)
ggplot(dat3, aes(x=x3, y=y3)) + geom_point() + geom_smooth(method="lm")
```



Решение, блок 3



Решение, блок 3



- 100-е наблюдение сильно влияет на ход регрессии
- Зависимость нелинейна

Решение, блок 3

```
mod3 <- lm(y3 ~ x3, data = dat3)

# Данные для графиков остатков
mod3_diag <- fortify(mod3)

# 1) График расстояния Кука
ggplot(mod3_diag, aes(x = 1:nrow(mod3_diag), y = .cooksdi)) +
  geom_bar(stat = "identity")

# 2) График остатков от предсказанных значений
gg_resid <- ggplot(data = mod3_diag, aes(x = .fitted, y = .stdresid)) +
  geom_point() + geom_hline(yintercept = 0)
gg_resid

# 3) Графики остатков от предикторов в модели и не в модели
gg_resid + aes(x = x3)

# 4) Квантильный график остатков
qqPlot(mod3, id = FALSE)
```

Задание 5

Выполните задание 3 в одном из этих файлов:

- 07_task_assumptions_catsM.R
- 07_task_assumptions_GAG.R

Take-home messages

- Гипотезы о наличии зависимости можно тестировать при помощи t - или F -теста.
- Качество подгонки модели можно оценить при помощи коэффициента детерминации R^2 .

Take-home messages

- Гипотезы о наличии зависимости можно тестировать при помощи t - или F -теста.
- Качество подгонки модели можно оценить при помощи коэффициента детерминации R^2 .
- У линейных моделей есть условия применимости, поэтому не спешите описывать результаты — сначала проверьте.

Take-home messages

- Гипотезы о наличии зависимости можно тестировать при помощи t - или F -теста.
- Качество подгонки модели можно оценить при помощи коэффициента детерминации R^2 .
- У линейных моделей есть условия применимости, поэтому не спешите описывать результаты — сначала проверьте.
- Если условия применимости нарушены, то результатам тестов для этой модели нельзя верить (получаются заниженные доверительные вероятности, возрастает вероятность ошибок I рода).
- Анализ остатков дает разностороннюю информацию о валидности моделей.

ЧТО ПОЧИТАТЬ

- Гланц, С., 1998. Медико-биологическая статистика. М., Практика
- Кабаков Р.И. R в действии. Анализ и визуализация данных на языке R. М.: ДМК Пресс, 2014
- Diez, D.M., Barr, C.D. and Çetinkaya-Rundel, M., 2015. OpenIntro Statistics. OpenIntro.
- Zuur, A., Ieno, E.N. and Smith, G.M., 2007. Analyzing ecological data. Springer Science & Business Media.
- Quinn G.P., Keough M.J. 2002. Experimental design and data analysis for biologists
- Logan M. 2010. Biostatistical Design and Analysis Using R. A Practical Guide