Диагностика линейных моделей

Линейные модели…

Марина Варфоломеева, Вадим Хайтов

Кафедра Зоологии беспозвоночных, Биологический факультет, СПбГУ

Диагностика линейных моделей

- Анализ остатков
  - Проверка на наличие влиятельных наблюдений
  - Проверка условий применимости линейных моделей

Вы сможете

- перечислить условия применимости линейной регрессии
- идентифицировать основные нарушения условий применимости линейной регрессии по паттернам на графиках остатков
- написать код на языке R, который позволяет нарисовать диагностические графики остатков для линейной регрессии

Зачем нужна диагностика модели? Разве тестов было недостаточно?

```
dat <- read.table('http://www4.stat.ncsu.edu/~stefanski/NSF_Supported/
Hidden_Images/orly_owl_files/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 достоверно? Пишем статью?

Зачем нужна диагностика модели? Разве тестов было недостаточно?

```
dat <- read.table('http://www4.stat.ncsu.edu/~stefanski/NSF_Supported/
Hidden_Images/orly_owl_files/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
```

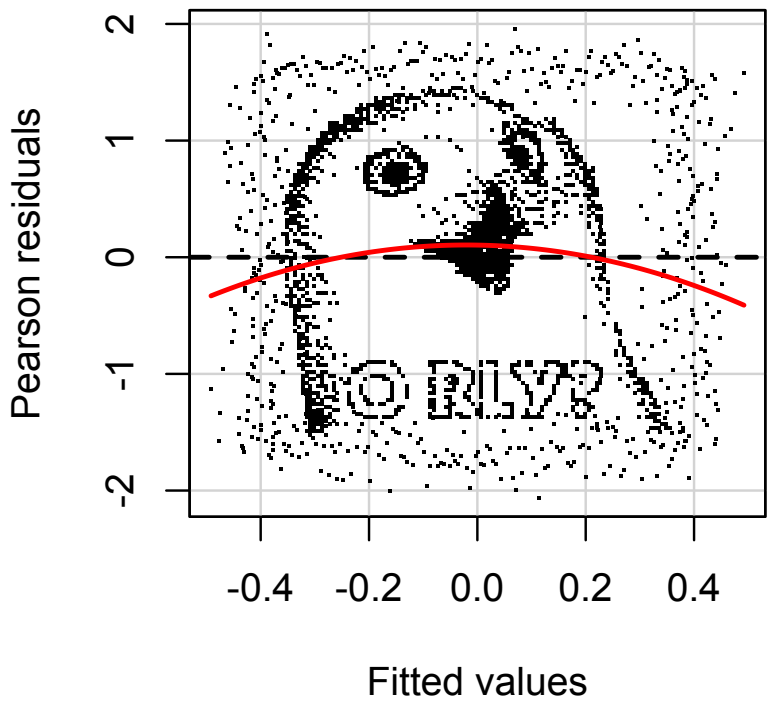
Все достоверно? Пишем статью?

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

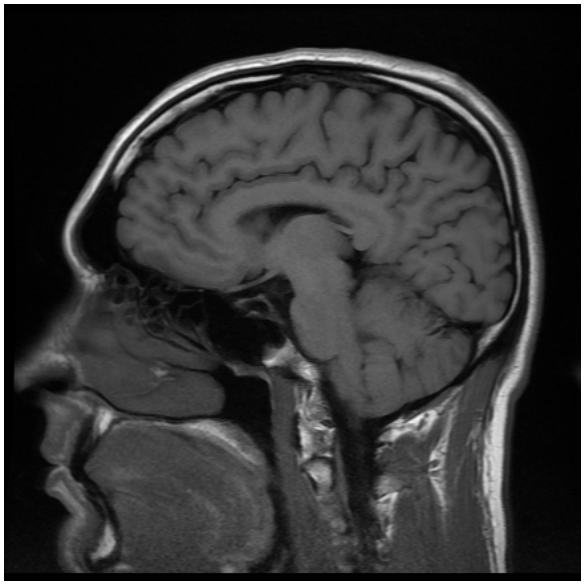
Проверка на наличие влиятельных наблюдений

Проверка условий применимости линейных моделей

- Линейная связь между зависимой перменной (Y) и предикторами (X)
- Независимость значений Y друг от друга
- Нормальное распределение Y для каждого уровня значений X
- Гомогенность дисерсий Y для каждого уровня значений X
- Отсутствие коллинеарности предикторов (для можественной регрессии)

С этим примером мы познакомились в прошлый раз: 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 (различные IQ тесты)

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

Не забудьте войти в вашу директорию для матметодов при помощи `setwd()`

```
# Данные можно загрузить с сайта
library(downloader)
# в рабочем каталоге создаем суб-директорию для данных
if(!dir.exists("data")) dir.create("data")
# скачиваем файл
download(
  url = "https://varmara.github.io/linmodr-course/data/IQ_brain.csv",
  destfile = "data/IQ_brain.csv")
```

Подберем модель

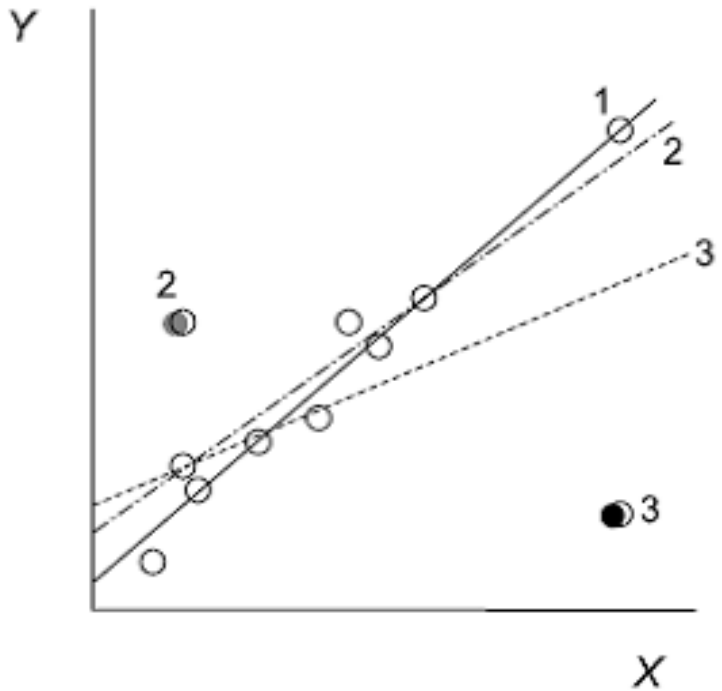
```
brain <- read.csv("data/IQ_brain.csv", header = TRUE)
brain_model <- lm(PIQ ~ MRINACount, data = brain)
coef(summary(brain_model))
```

```
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)  1.74376 42.3923825  0.0411   0.9674
# MRINACount   0.00012  0.0000465  2.5858   0.0137
```

Проверка на наличие влиятельных наблюдений

Влиятельные наблюдения — это…

наблюдения, которые вносят слишком большой вклад в оценку парметров (коэффициентов) модели.

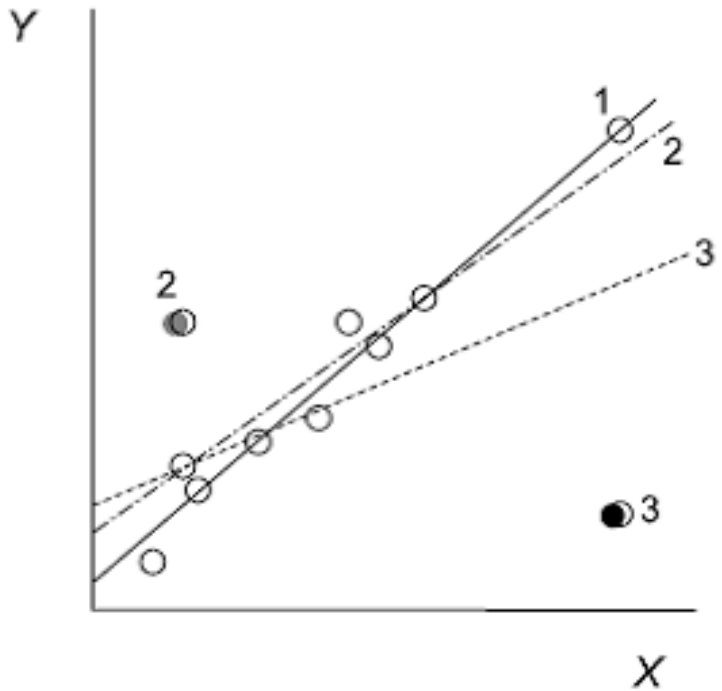


Из кн. Quinn, Keugh, 2002

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

Влиятельные наблюдения — это…

наблюдения, которые вносят слишком большой вклад в оценку парметров (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*

Типы остатков

“Сырые” остатки

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

Типы остатков

“Сырые” остатки

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

Пирсоновские остатки

$p_i = \frac{\varepsilon_i}{\sqrt{Var(\hat{y}_i)}}$, где $\sqrt{Var(\hat{y}_i)}$ — это стандартное отклонение предсказанных значений

легко сравнивать, т.к. стандартизованы

Типы остатков

“Сырые” остатки

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

Пирсоновские остатки

$p_i = \frac{\varepsilon_i}{\sqrt{Var(\hat{y}_i)}}$, где $\sqrt{Var(\hat{y}_i)}$ — это стандартное отклонение предсказанных значений

легко сравнивать, т.к. стандартизованы

Стьюдентовские остатки

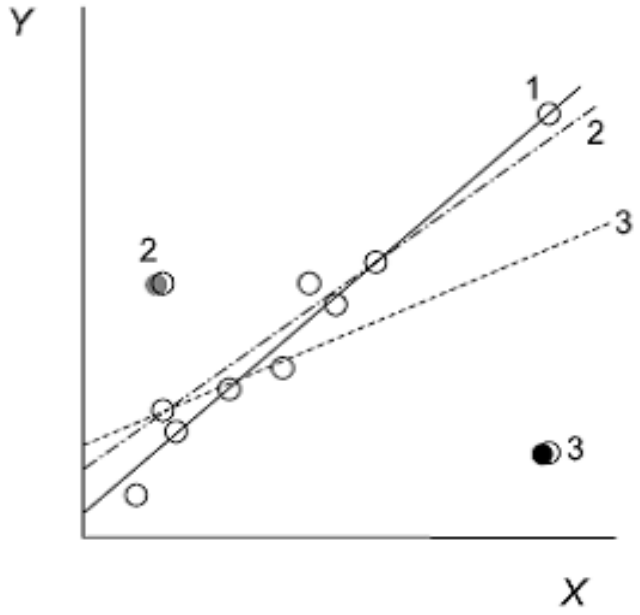
$s_i = \frac{p_i}{\sqrt{1-h_{ii}}} = \frac{\varepsilon_i}{\sqrt{Var(\hat{y}_i)(1-h_{ii})}}$,

где h_{ii} — “сила воздействия” отдельных наблюдений (leverage)

легко сравнивать, т.к. стандартизованы и учитывают влияние наблюдений

Воздействие точек h_{ii} (leverage)

показывает, насколько каждое значение x_i влияет на ход линии регрессии, то есть на $\hat{y}_i$



Из кн. Quinn, Keough, 2002

Weighing Machine by neys fadzil on Flickr

- Точки, располагающиеся дальше от $\bar{x}$, оказывают более сильное влияние на $\hat{y}_i$
- Эта величина, в норме, 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

$$D_i = \frac{\sum \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p \cdot MSE} \left(\frac{h_{ii}}{1 - h_{ii}}\right)$$

- $\hat{y}_j$ - значение предсказанное полной моделью
- $\hat{y}_{j(i)}$ - значение, предказанное моделью, построенной без учета i-го значения предиктора
- p - количество параметров в модели
- MSE - 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

$$D_i = \frac{\sum \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p \cdot MSE} \left(\frac{h_{ii}}{1 - h_{ii}}\right)$$

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

Расстояние Кука зависит одновременно от величины остатков и "силы воздействия" наблюдений.

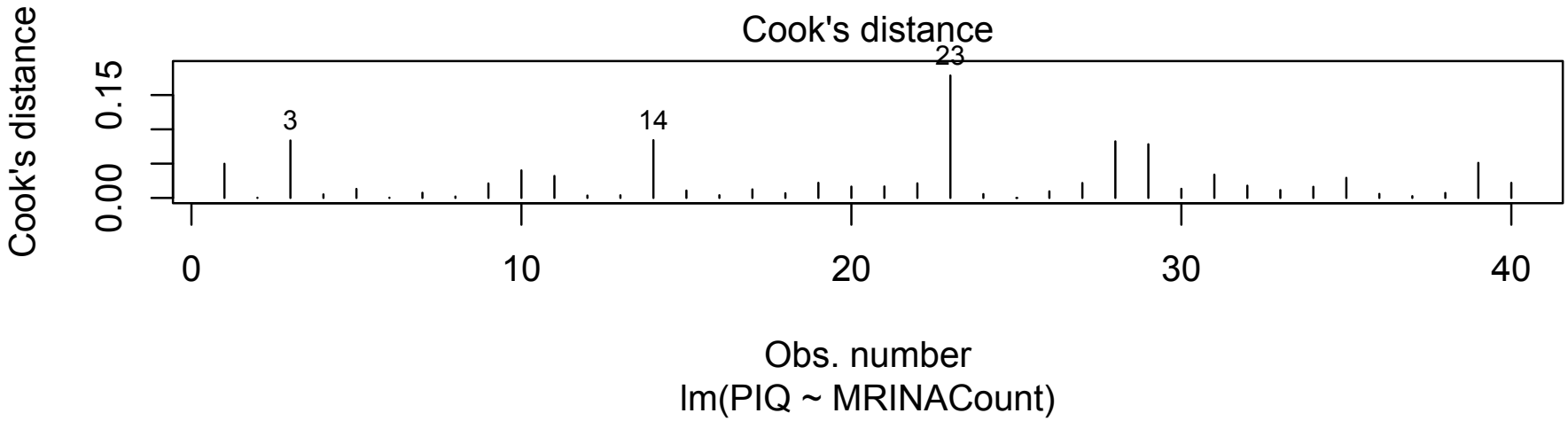
Статистических тестов для D_i нет, но можно использовать один из двух условных порогов. Наблюдение является выбросом (outlier), если:

- $D_i > 1$
- $D_i > 4/\left(Nk1\right)$ (N - объем выборки, k - число предикторов)

Проверяем наличие влиятельных наблюдений в brain_model

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

```
plot(brain_model, which = 4)
```



Второй вариант — построить обычный график остатков, и отметить на нем расстояния Кука.

Чтобы построить график остатков с расстояниями Кука нам понадобятся данные.

Извлечем из результатов сведения для анализа остатков при помощи функции `fortify()` из пакета `{ggplot2}`

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

```
#   PIQ MRINACount   .hat .sigma  .cooksd .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
```

Чтобы построить график остатков с расстояниями Кука нам понадобятся данные.

Извлечем из результатов сведения для анализа остатков при помощи функции `fortify()` из пакета `{ggplot2}`

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

```
#   PIQ MRINACount   .hat .sigma  .cooksd .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*)
- `.cooksd` — расстояние Кука
- `.fitted` — предсказанные значения
- `.resid` — остатки
- `.stdresid` — стандартизованные остатки

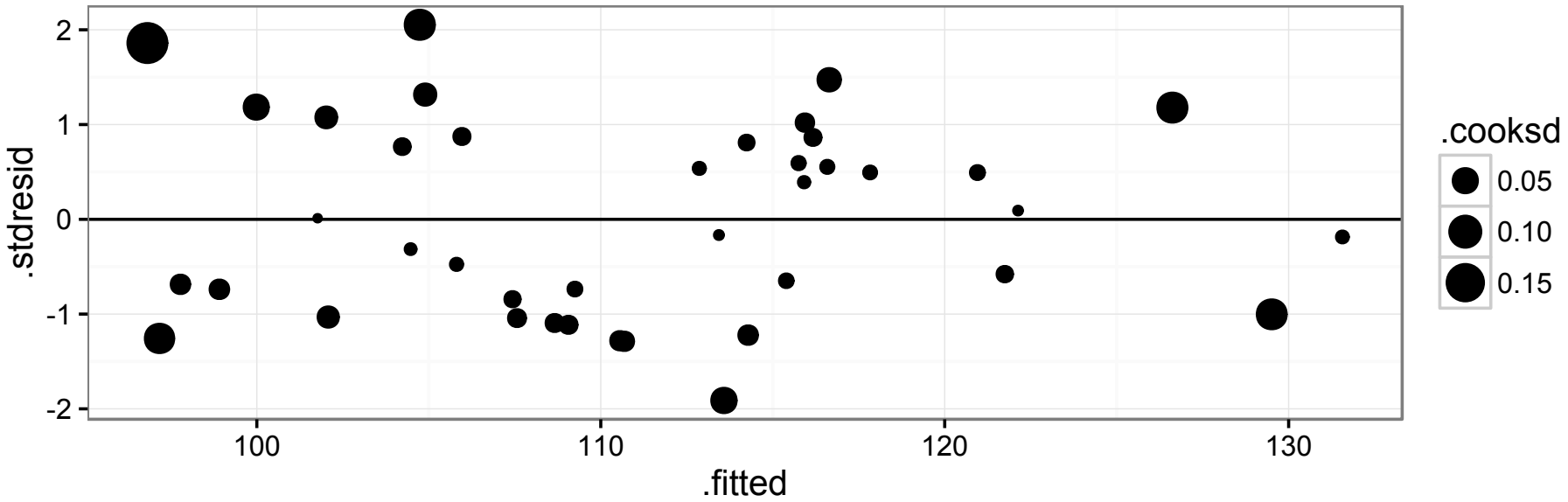
Задание

Используя данные из датафрейма `brain_diag`,
постройте график зависимости стандартизированных остатков модели `brain_model` от предсказанных значений.

Сделайте так, чтобы размер точек изменялся в зависимости от значения расстояния Кука.

Решение

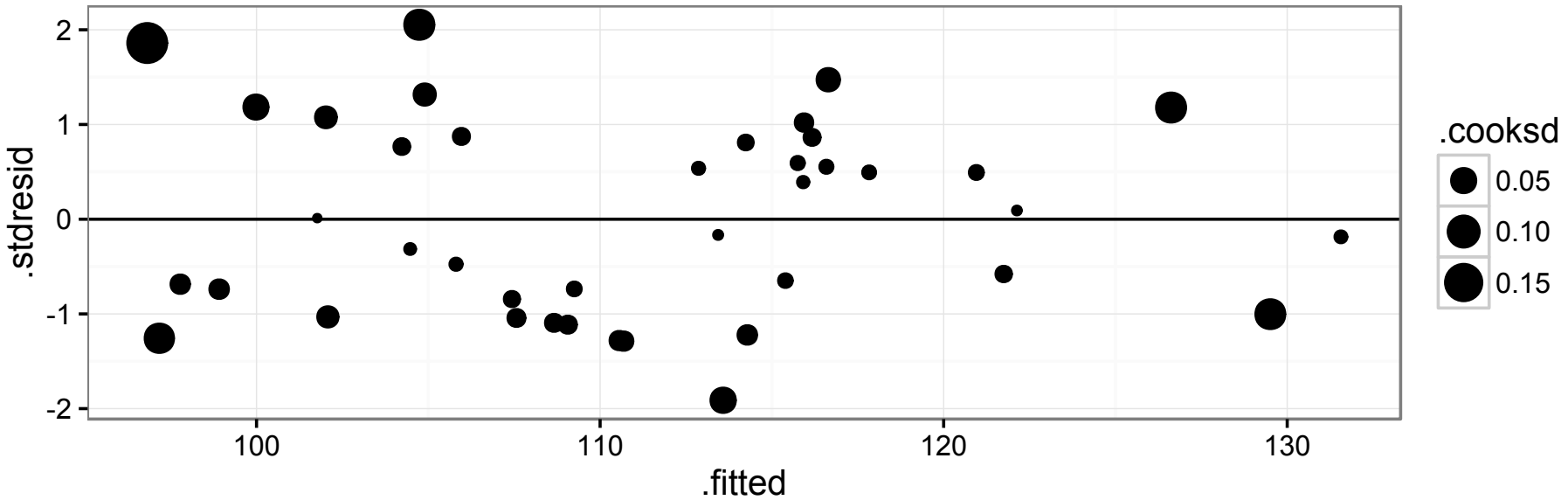
```
theme_set(theme_bw()) # устанавливаем тему (не обязательно)
ggplot(data = brain_diag, aes(x = .fitted, y = .stdresid, size = .cooksd)) +
  geom_point() + geom_hline(aes(yintercept = 0))
```



Что мы видим?

Решение

```
theme_set(theme_bw()) # устанавливаем тему (не обязательно)
ggplot(data = brain_diag, aes(x = .fitted, y = .stdresid, size = .cooksd)) +
  geom_point() + geom_hline(aes(yintercept = 0))
```

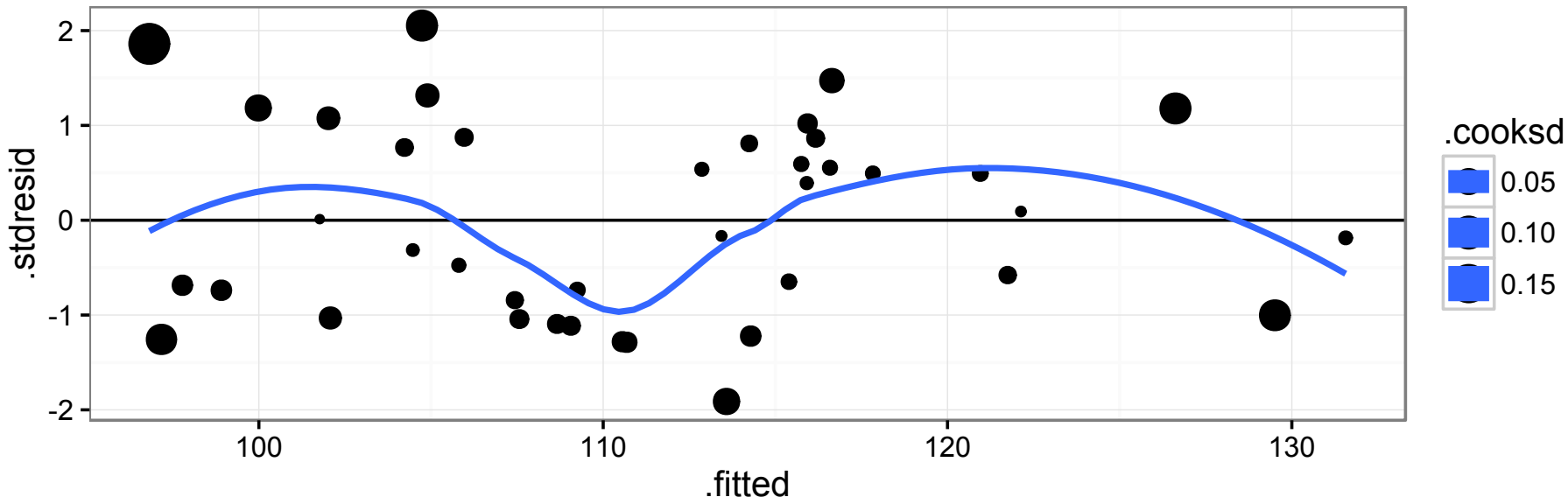


Что мы видим?

- Большая часть стандартизованных остатков в пределах двух стандартных отклонений
- Есть одно влиятельное наблюдение, которое нужно проверить, но сила его влияния невелика (расстояние Кука < 1)
- Среди остатков нет тренда, но, возможно, есть иной паттерн…

Добавим линию loess-сглаживания на график

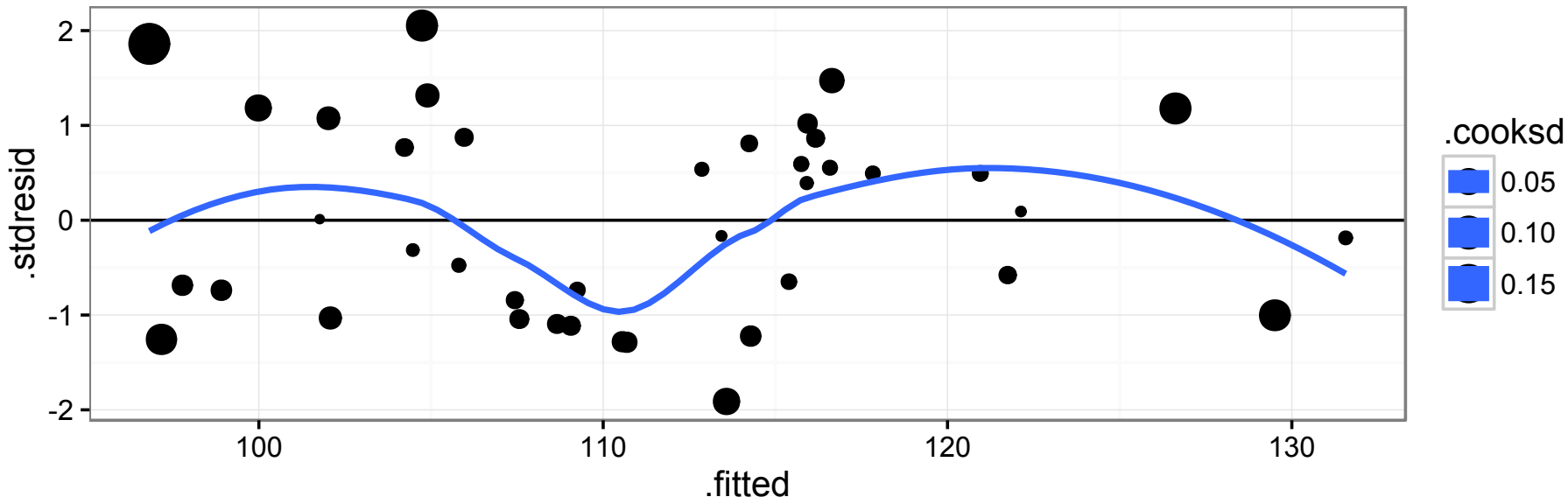
```
ggplot(data = brain_diag, aes(x = .fitted, y = .stdresid, size = .cooksd)) +
  geom_point() +  geom_hline(yintercept = 0) +
  geom_smooth(method="loess", se=FALSE)
```



Чем мог быть вызван такой странный паттерн?

Добавим линию loess-сглаживания на график

```
ggplot(data = brain_diag, aes(x = .fitted, y = .stdresid, size = .cooksd)) +
  geom_point() +  geom_hline(yintercept = 0) +
  geom_smooth(method="loess", se=FALSE)
```



Чем мог быть вызван такой странный паттерн?

- Неучтенная переменная — добавляем в модель
- Нелинейная зависимость — используем GAM, нелинейную регрессию и т.д.

Что делать с наблюдениями-выбросами?

Удалить?

Осторожно! Нельзя удалять выбросы только на основе такого диагноза. Задача диагностики — заставить вас искать причины такого поведения данных. Удалять следует только очевидные ошибки в наблюдениях.

Трансформировать?

Некоторые виды трансформаций

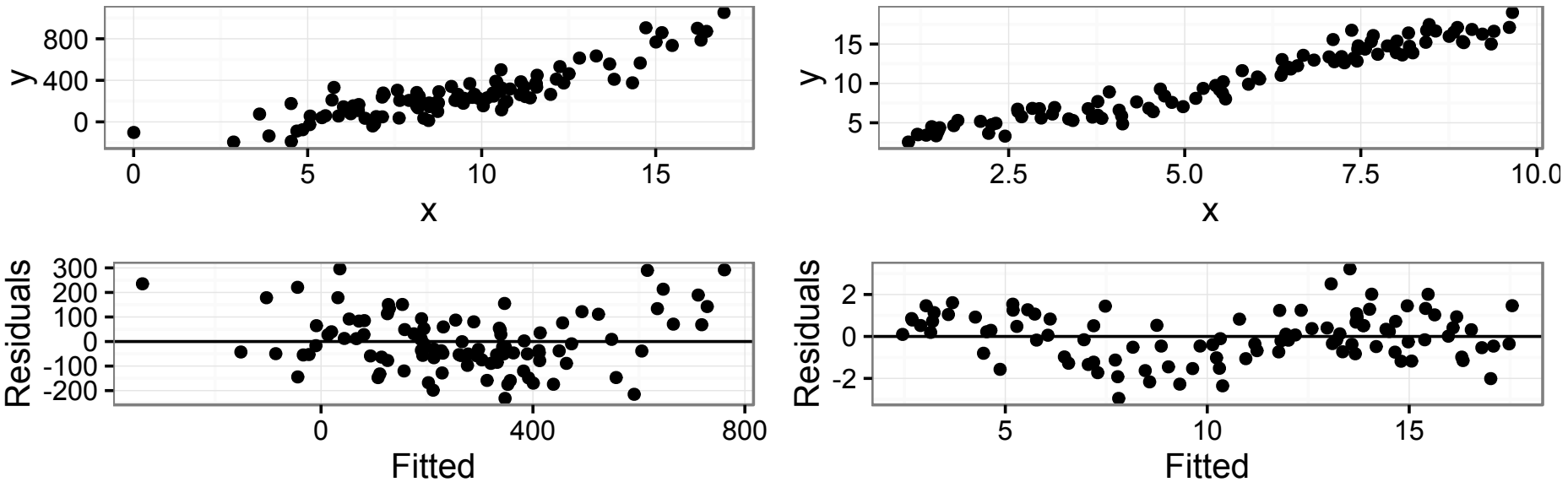
Трансформация	Формула
степень -2	$1/x^2$
степень -1	$1/x$
степень -0.5	$1/\sqrt{x}$
степень 0.5	$\sqrt{x}$
логарифмирование	$log(x)$

Условия применимости линейных моделей (Assumptions)

1. Линейность связи

Нелинейные зависимости не всегда видны на исходных графиках в осях Y vs X

Они становятся лучше заметны на графиках рассеяния остатков (Residual plots)



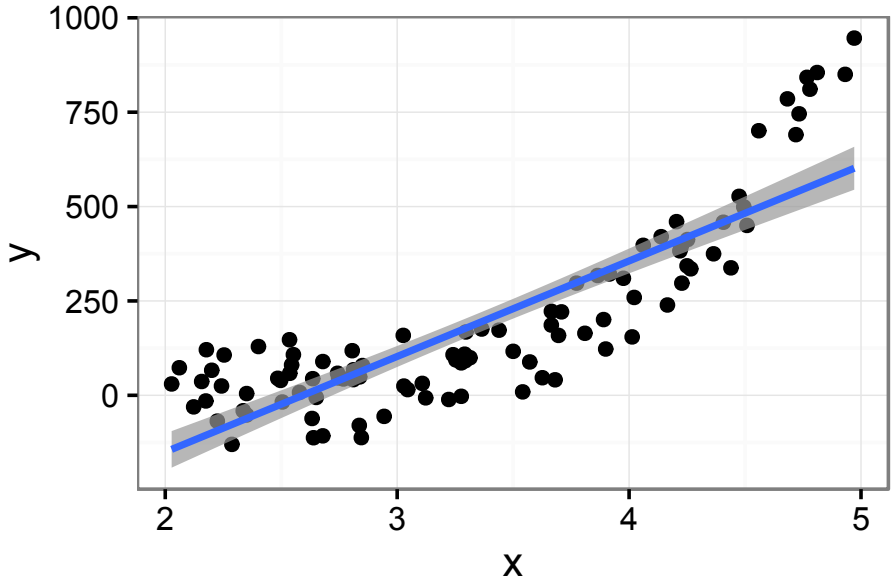
Проверка на линейность связи

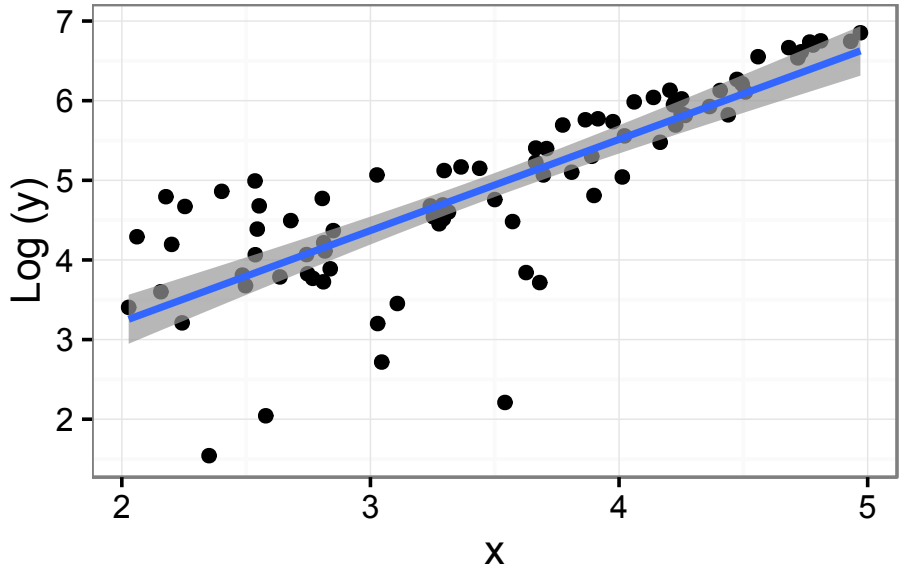
- График зависимости *Y* от *x* (для множественной регрессии - от всех *x*)
- График остатков от предсказанных значений

Что делать, если связь нелинейна?

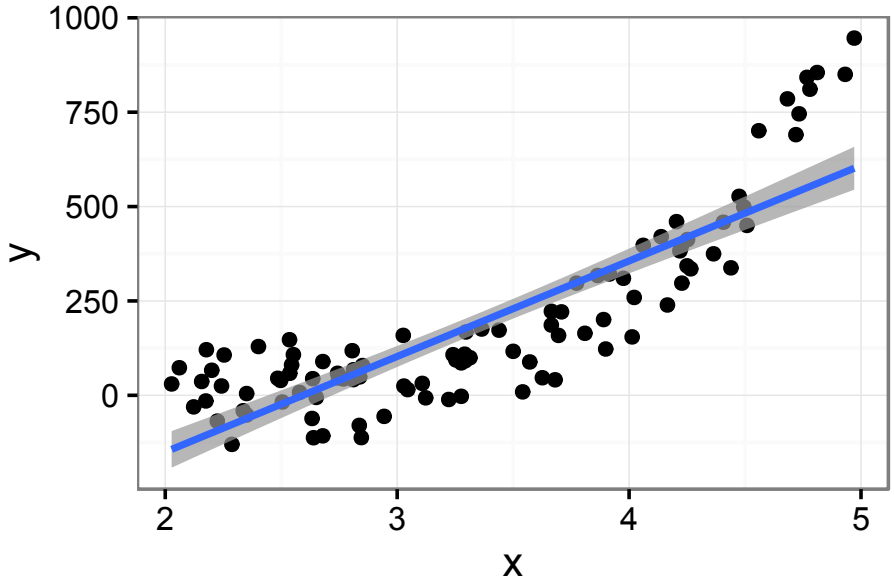
- Добавить неучтенные переменные
- Добавить взаимодействие переменных
- Применить линеаризующее преобразование (Осторожно!)
- Применить обобщенную линейную модель с другой функцией связи (об этом позже)
- Построить аддитивную модель (если достаточно наблюдений по x)
- Построить нелинейную модель (если известна форма зависимости)

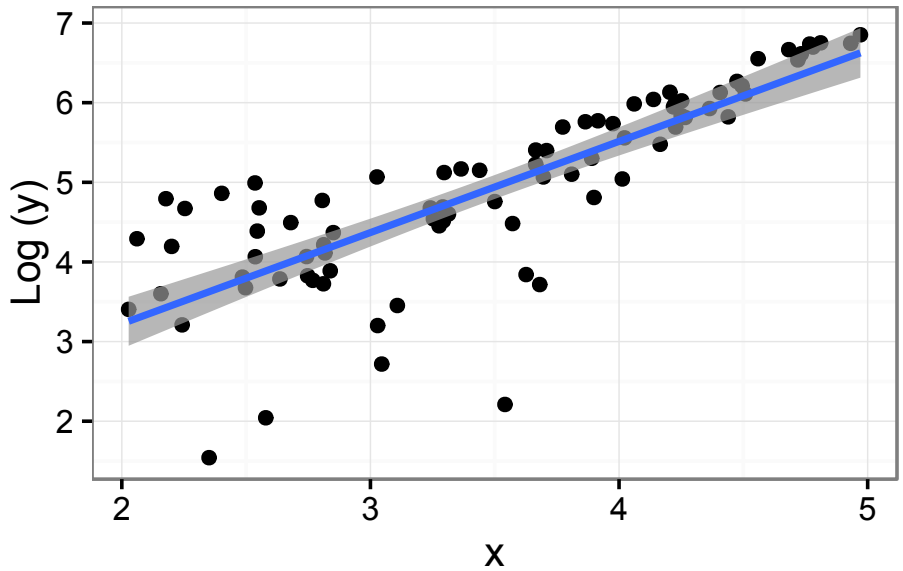
Пример линеаризующего преобразования





Пример линеаризующего преобразования





Осторожно! При таком преобразовании вы рискуете изучить не то, что хотели. Матожидание логарифма величины (как при трансформации) не то же самое, что логарифм матожидания величины (как при использовании обобщенной линейной модели с логарифмической функцией связи). Но об этом — позже.

2. Независимость Y друг от друга

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.

2. Независимость Y друг от друга

Каждое значение Y_i должно быть независимо от любого другого Y_j

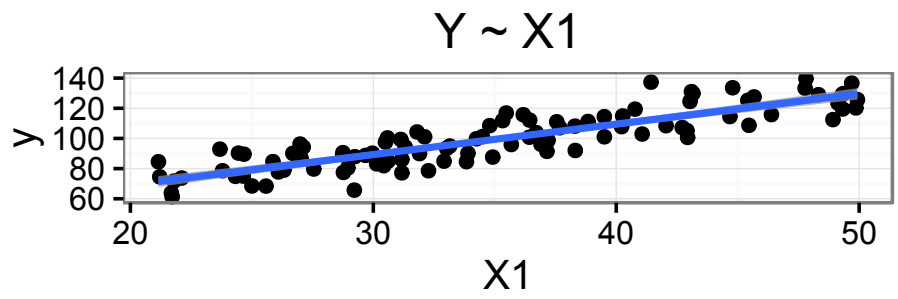
Это нужно контролировать на этапе планирования сбора материала

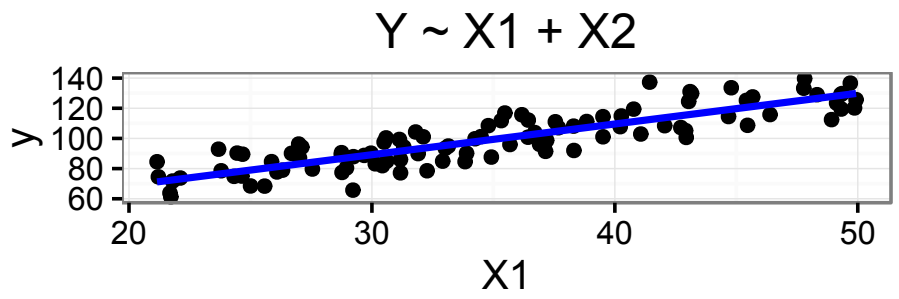
- Наиболее частые источники зависимостей:
  - псевдоповторности (повторно измеренные объекты)
  - неучтенные переменные
  - временные автокорреляции (если данные - временной ряд)
  - пространственные автокорреляции (если пробы взяты в разных местах)
  - и т.п.

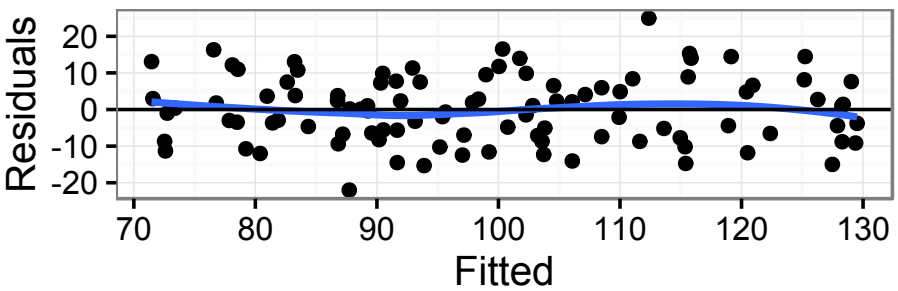
Взаимозависимости можно заметить на графиках остатков:

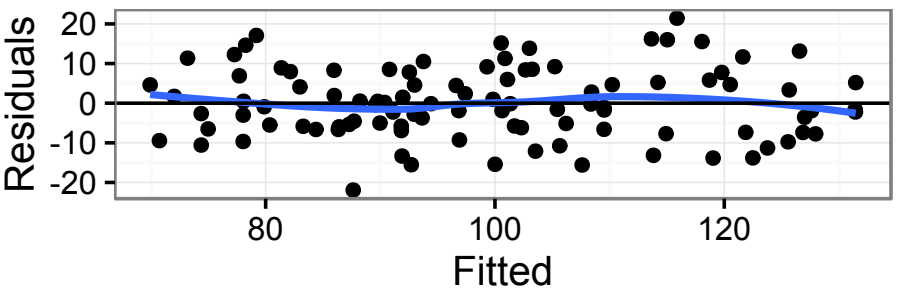
- остатки vs. предсказанные значения
- остатки vs. переменные в модели
- остатки vs. переменные не в модели

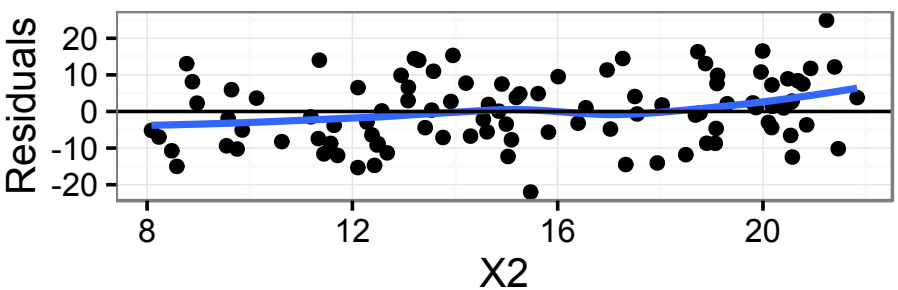
Нарушение условия независимости: Неучтенная переменная

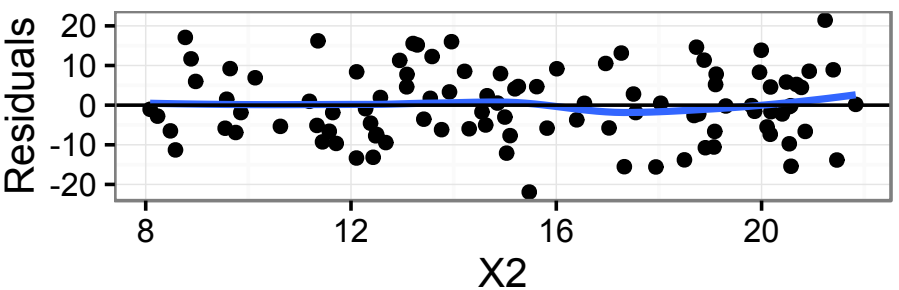










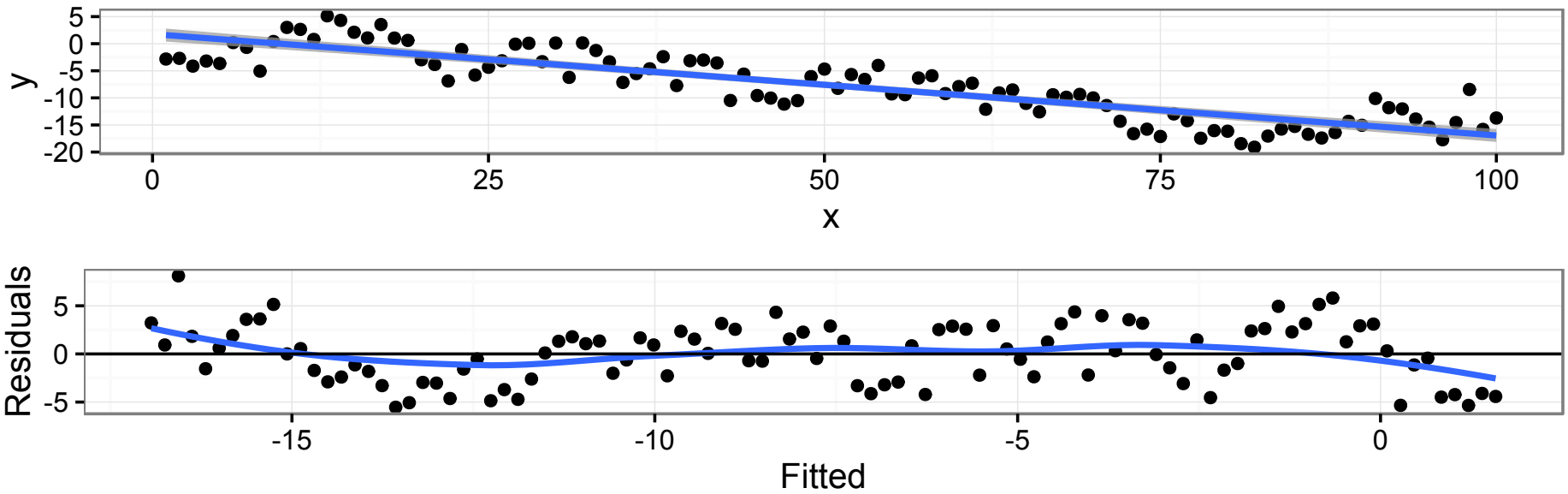


Если в модели не учтена переменная *X*2 (слева), внешне все нормально (только остатки большие), но если построить график зависимости остатков от *X*2.

Если *X*2 учесть (справа) — остатки становятся меньше, 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 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

3. Нормальное распределение Y (для каждого уровня значений X)

Это условие невозможно проверить “влоб”, т.к. обычно каждому X сообветствует лишь небольшое число Y

Если Y это нормально распределенная случайная величина

$$Y_i \in N(\mu_{y_i}, \sigma^2)$$

и мы моделируем ее как

$$Y_i \sim b_0 + b_1 x_{1i} + \cdots + \varepsilon_i$$

то остатки от этой модели — тоже нормально распределенная случайная величина

$$\varepsilon_i \in N(\mu_{y_i}, \sigma^2)$$

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

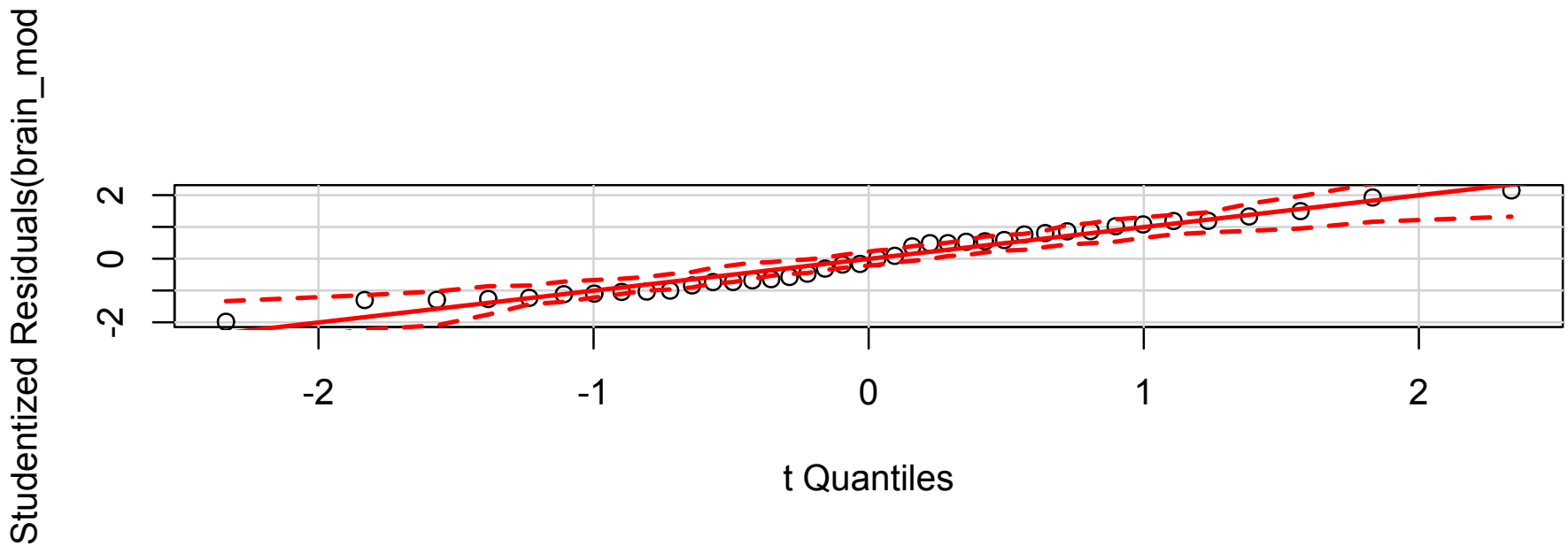
Лучший способ проверки — квантильный график остатков.

Квантильный график остатков

Квантиль - значение, которое заданная случайная величина не превышает с фиксированной вероятностью.

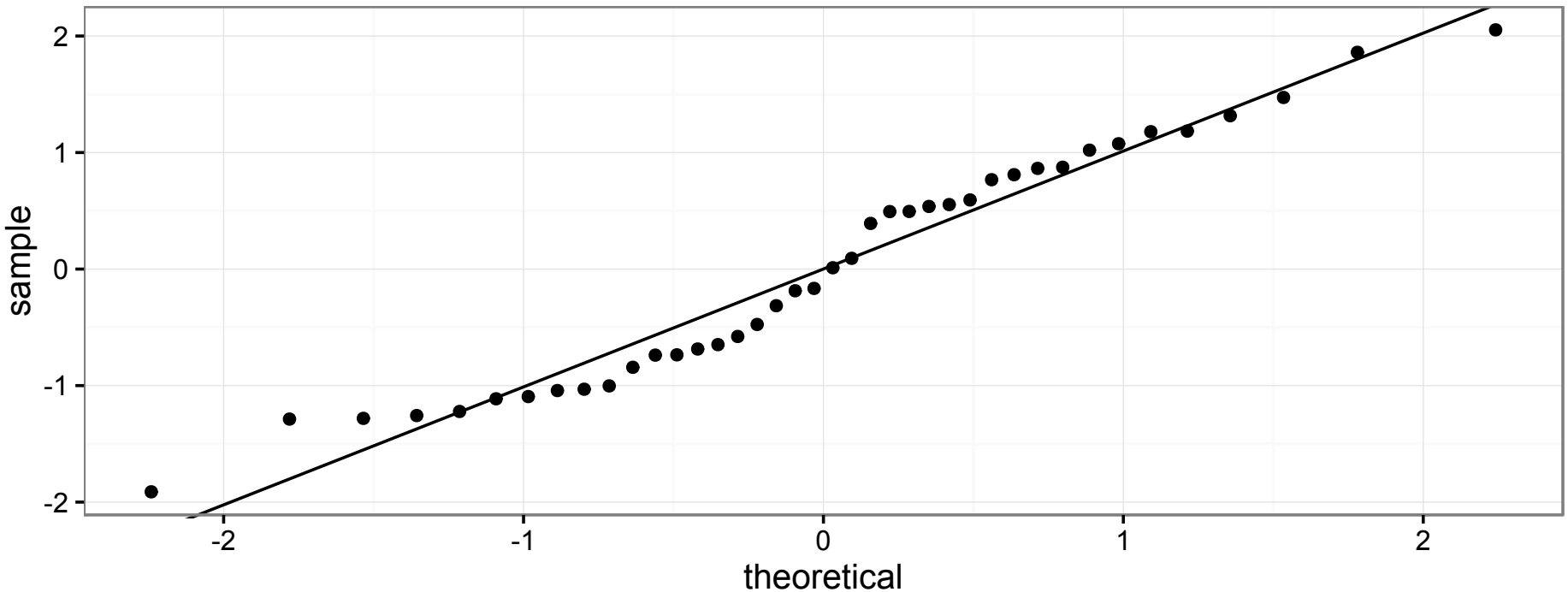
Если точки - это реализации случайной величины из $N(0, \sigma^2)$, то они должны лечь вдоль прямой $Y = X$. Если это стьюдентизированные остатки — то используются квантили t-распределения

```
qqPlot(brain_model) # из пакета car
```



Аналогичный график при помощи ggplot2

```
mean_val <- mean(brain_diag$.stdresid)
sd_val <- sd(brain_diag$.stdresid)
ggplot(brain_diag, aes(sample = .stdresid)) + geom_point(stat = "qq") +
    geom_abline(intercept = mean_val, slope = sd_val)
```



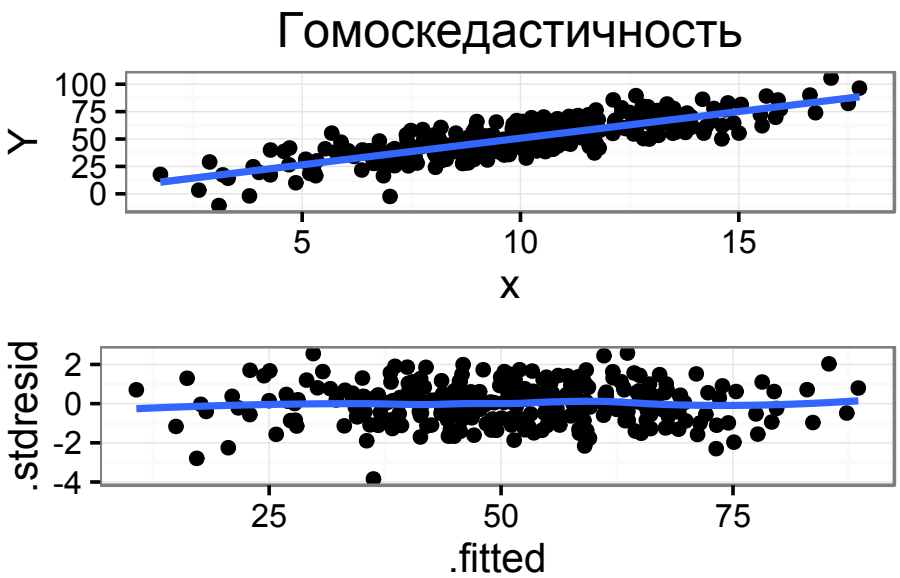
Что делать, если остатки распределены не нормально?

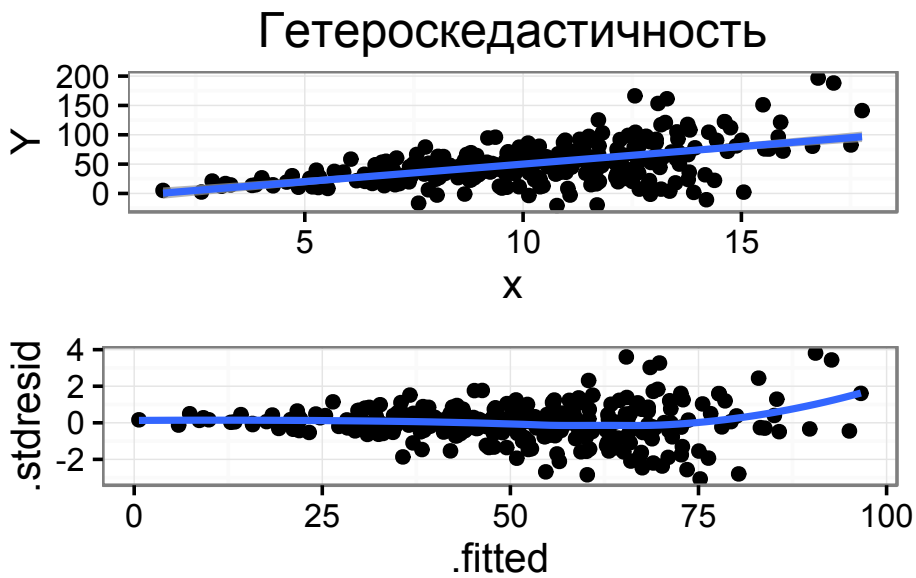
Зависит от причины

- 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 (гомоскедастичность)

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





Проверка постоянства дисперсий

Есть формальные тесты (тест Бройша-Пагана, тест Кокрана), но:

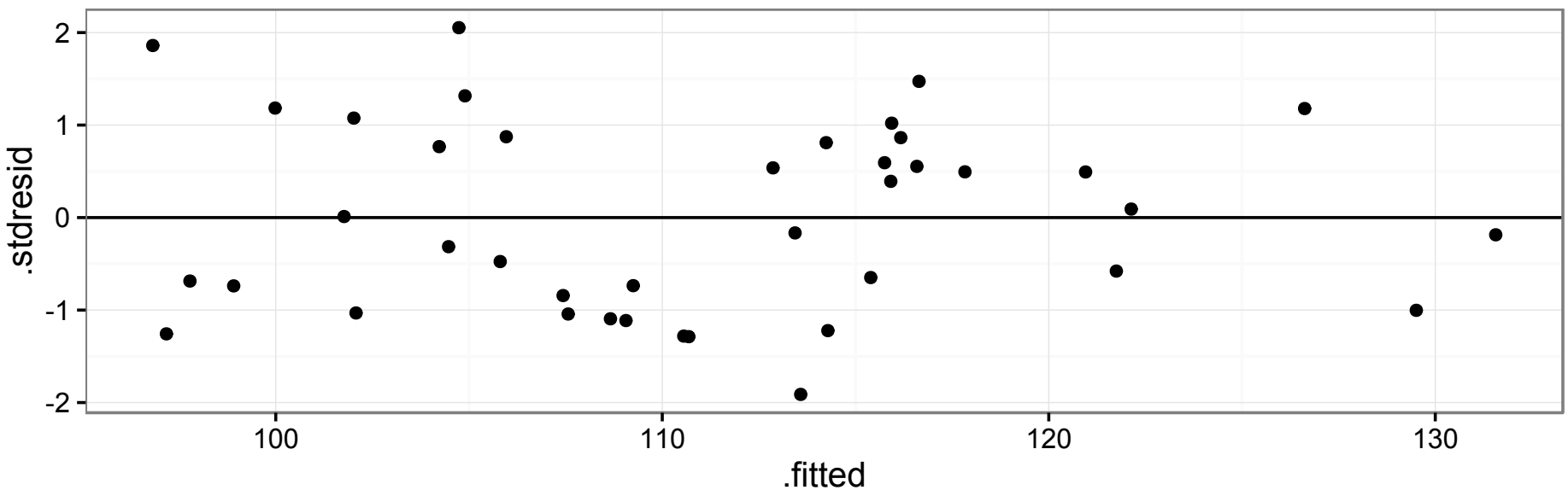
- у формальных тестов тоже есть свои условия применимости, и многие сами неустойчивы к гетероскедастичности
- при больших выборках формальные тесты покажут, что значима даже небольшая гетероскедастичность

Лучший способ проверки — график остатков.

Проверка на гетероскедастичность

Мы уже строили график остатков в ggplot2

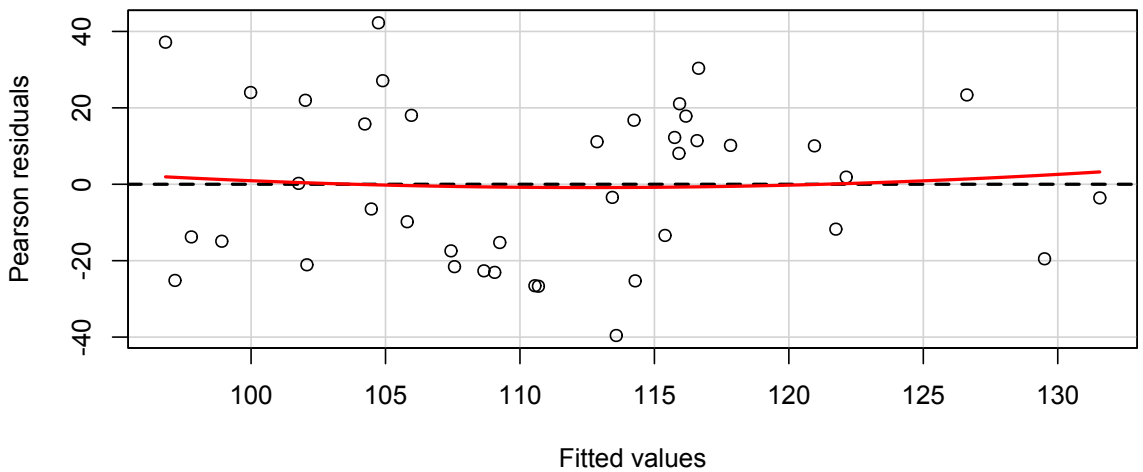
```
ggplot(data = brain_diag,
       aes(x = .fitted, y = .stdresid)) +
  geom_point() + geom_hline(yintercept = 0)
```



Проверка на гетероскедастичность

Можем построить аналогичный график остатков средствами пакета `car`

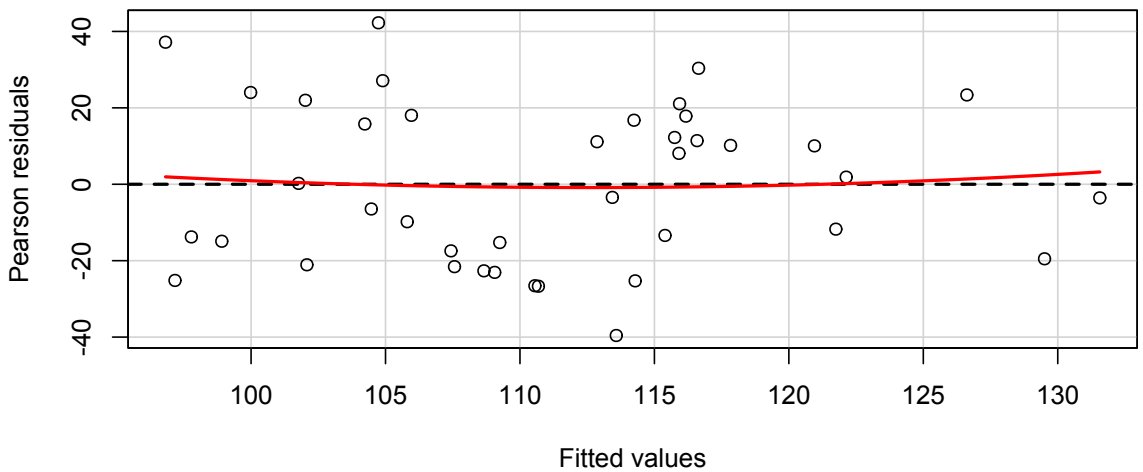
```
residualPlot(brain_model)
```



Проверка на гетероскедастичность

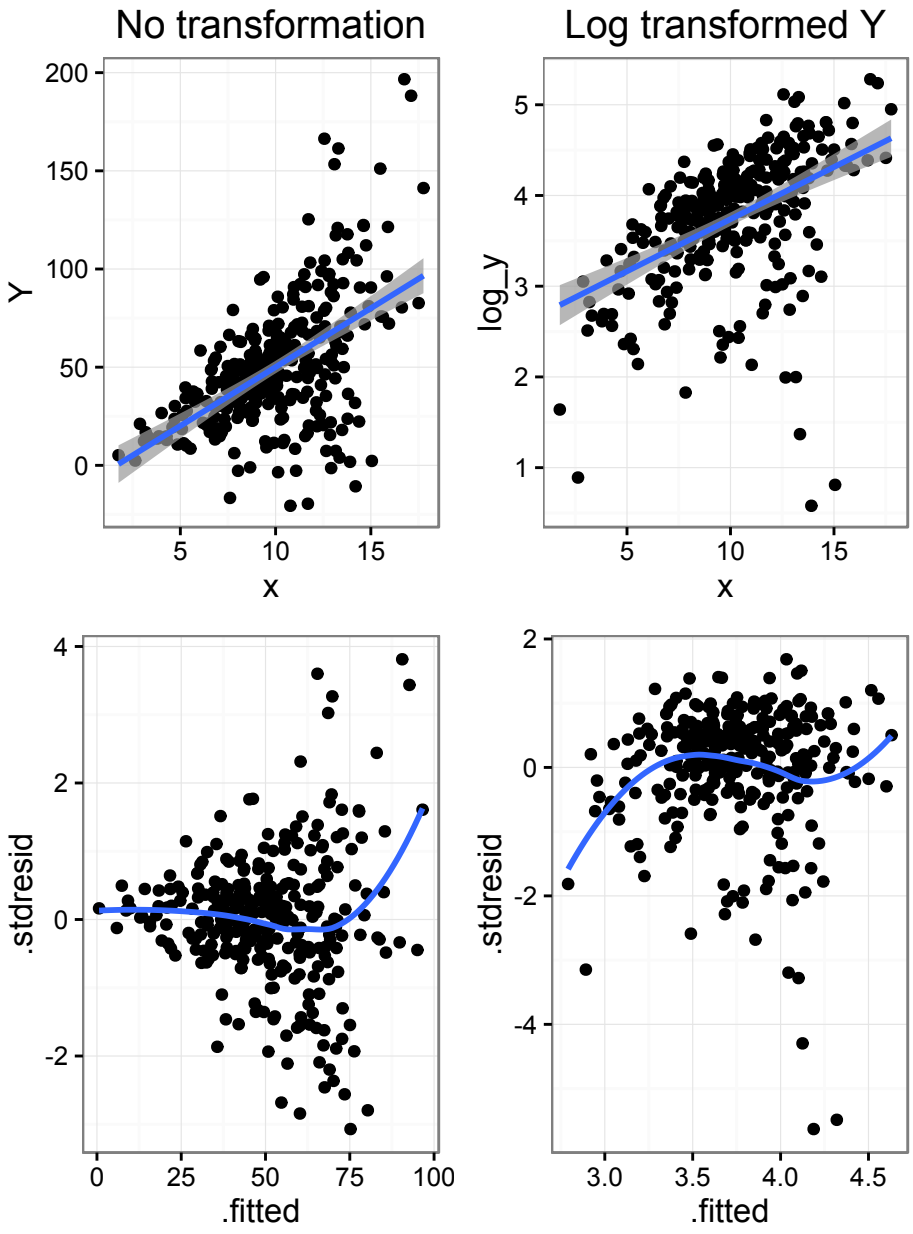
Можем построить аналогичный график остатков средствами пакета `car`

```
residualPlot(brain_model)
```



- Гетерогенность дисперсий не выражена.

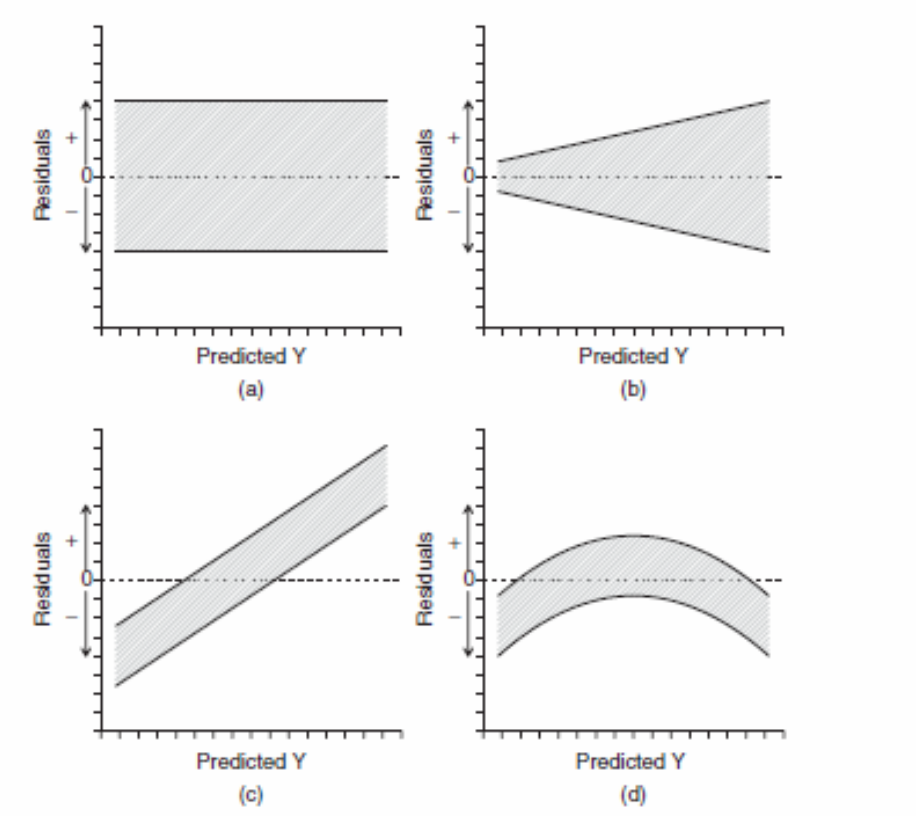
Что делать если вы столкнулись с гетероскедастичностью?



Трансформация может помочь…
Но на самом деле, нужно смотреть на причину гетерогенности

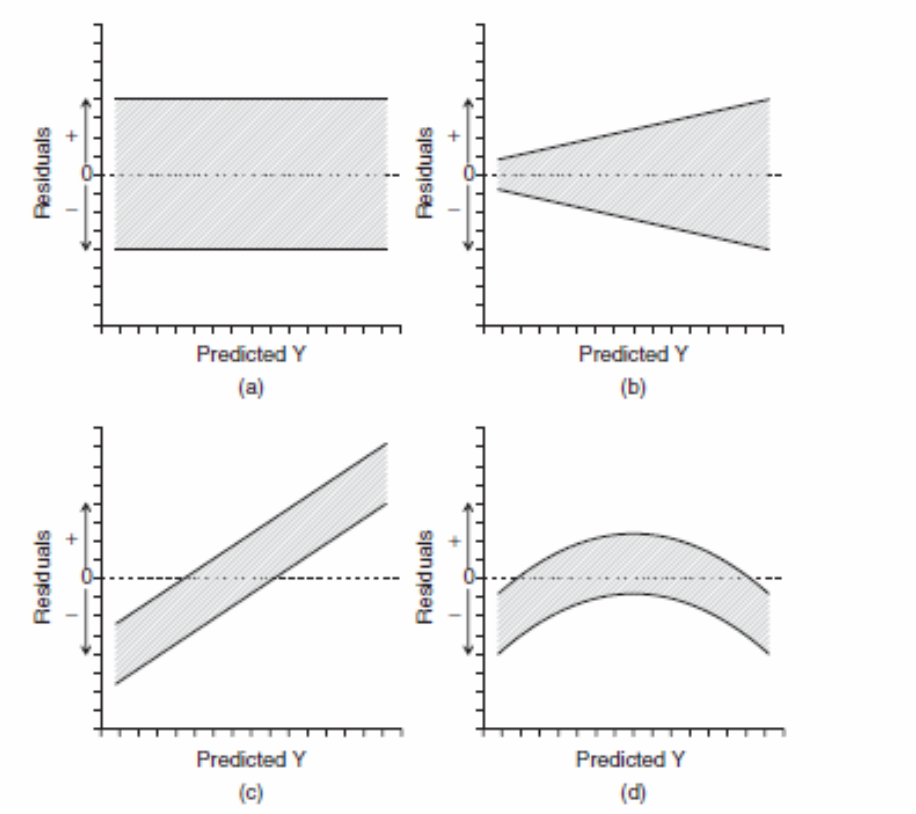
- Неучтенные переменные
  - добавляем в модель
- Зависимая переменная распределена по-другому
  - трансформируем данные (неудобно)
  - подбираем модель с другим распределением остатков (обобщенную линейную модель)
- Моделируем гетерогенность дисперсии.

Некоторые распространенные паттерны на графиках остатков



Вз кн. Logan, 2010, стр. 174

Некоторые распространенные паттерны на графиках остатков



Вз кн. Logan, 2010, стр. 174

1. Условия применимости соблюдаются. Модель хорошая
2. Клиновидный паттерн. Есть гетероскедастичность. Модель плохая
3. Остатки рассеяны равномерно, но модель неполна. Нужны дополнительные предикторы. Модель можно улучшить
4. Нелинейный паттерн сохранился. Линейная модель использована некорректно. Модель плохая

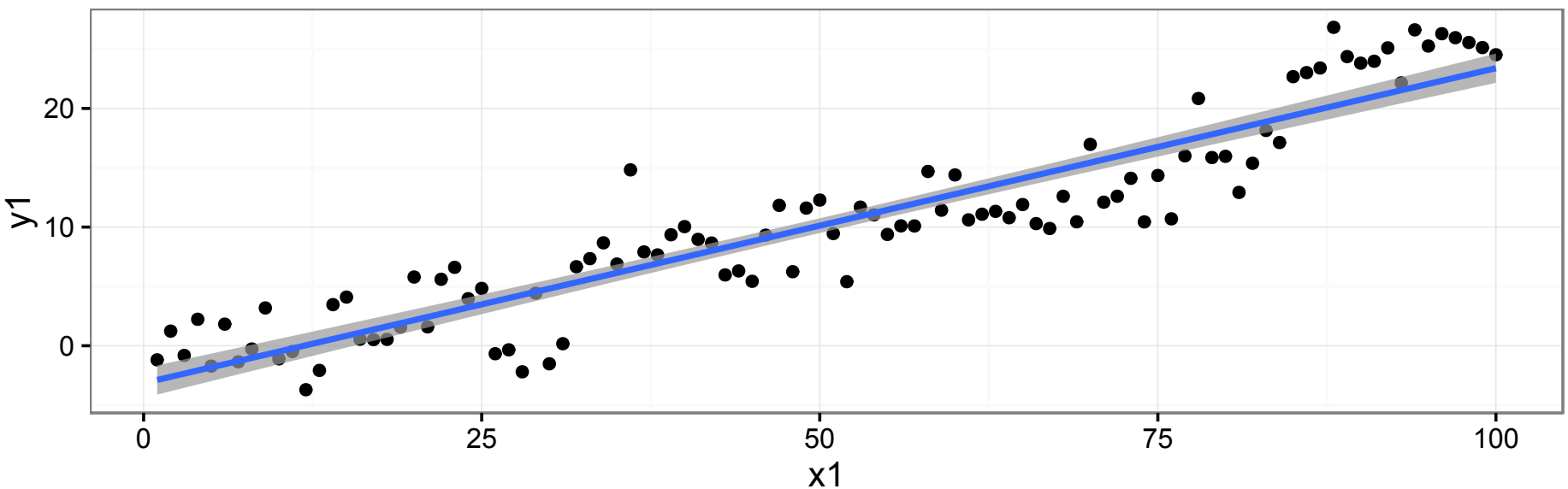
Задание

Выполните три блока кода (см. код лекции).

Какие нарушения условий применимости линейных моделей здесь наблюдаются?

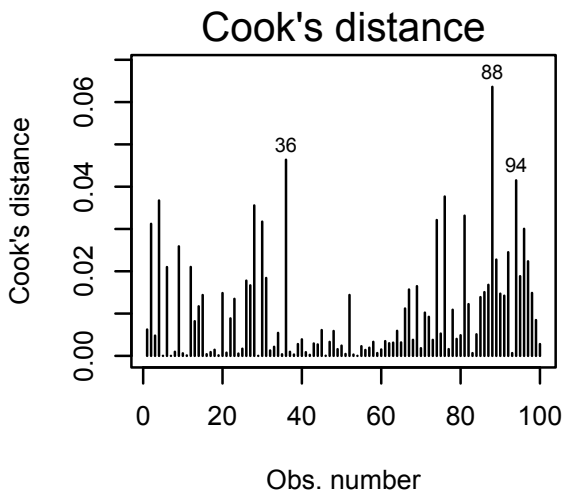
Задание, блок 1

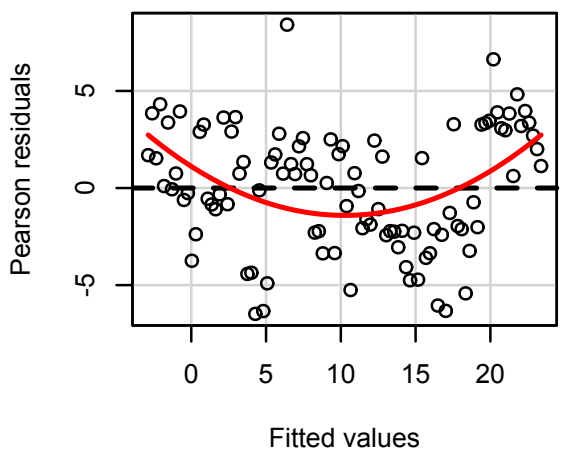
```
set.seed(12345)
x1 <- seq(1, 100, 1)
y1 <-  diffinv(rnorm(99)) + rnorm(100, 0.2, 2)
dat1 = data.frame(x1, y1)
ggplot(dat1, aes(x = x1, y = y1)) + geom_point()+
  geom_smooth(method="lm", alpha = 0.7)
```

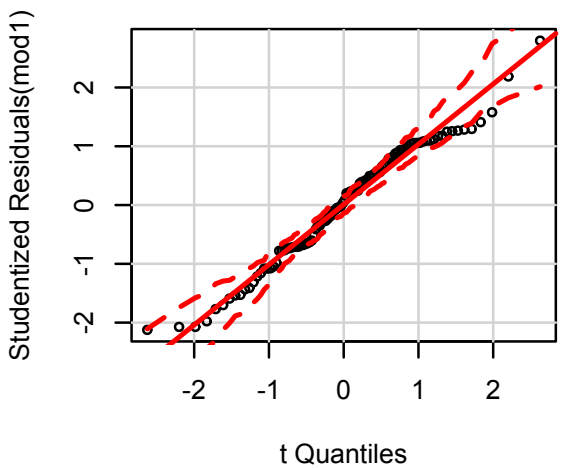


Решение, блок 1

```
mod1 <- lm(y1 ~ x1, data = dat1)
op <- par(mfrow = c(1, 3)) # располагаем картинки в 3 колонки
plot(mod1, which = 4)          # Расстояние Кука
residualPlot(mod1)             # График остатков
qqPlot(mod1)                   # Квантильный график остатков
par(op)                      # возвращаем старые графические параметры
```

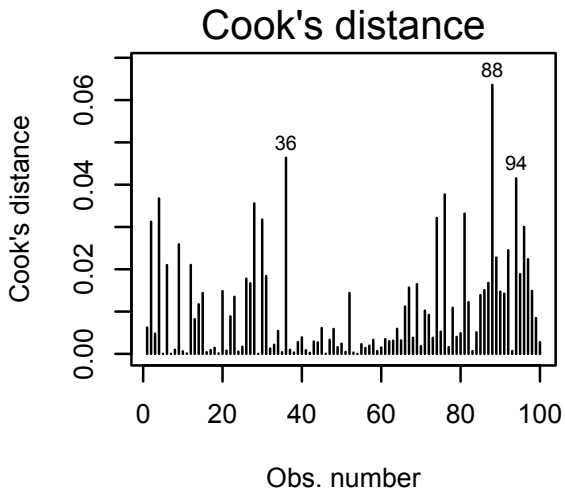


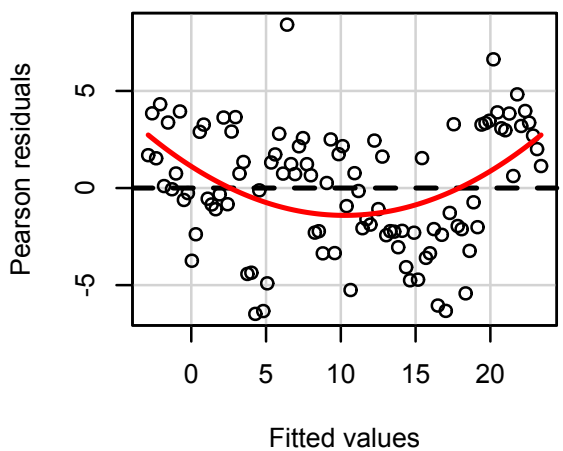


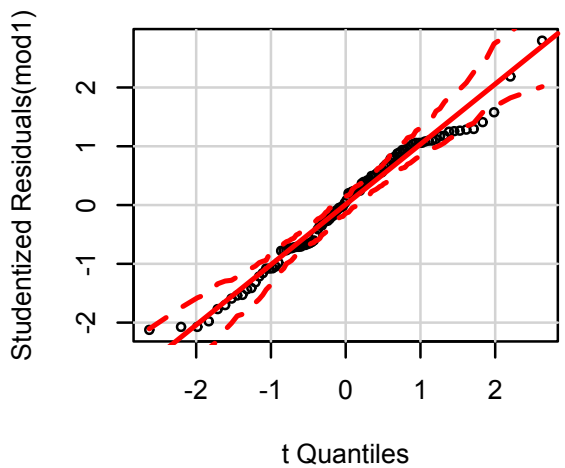


Решение, блок 1

```
mod1 <- lm(y1 ~ x1, data = dat1)
op <- par(mfrow = c(1, 3)) # располагаем картинки в 3 колонки
plot(mod1, which = 4)          # Расстояние Кука
residualPlot(mod1)             # График остатков
qqPlot(mod1)                   # Квантильный график остатков
par(op)                      # возвращаем старые графические параметры
```



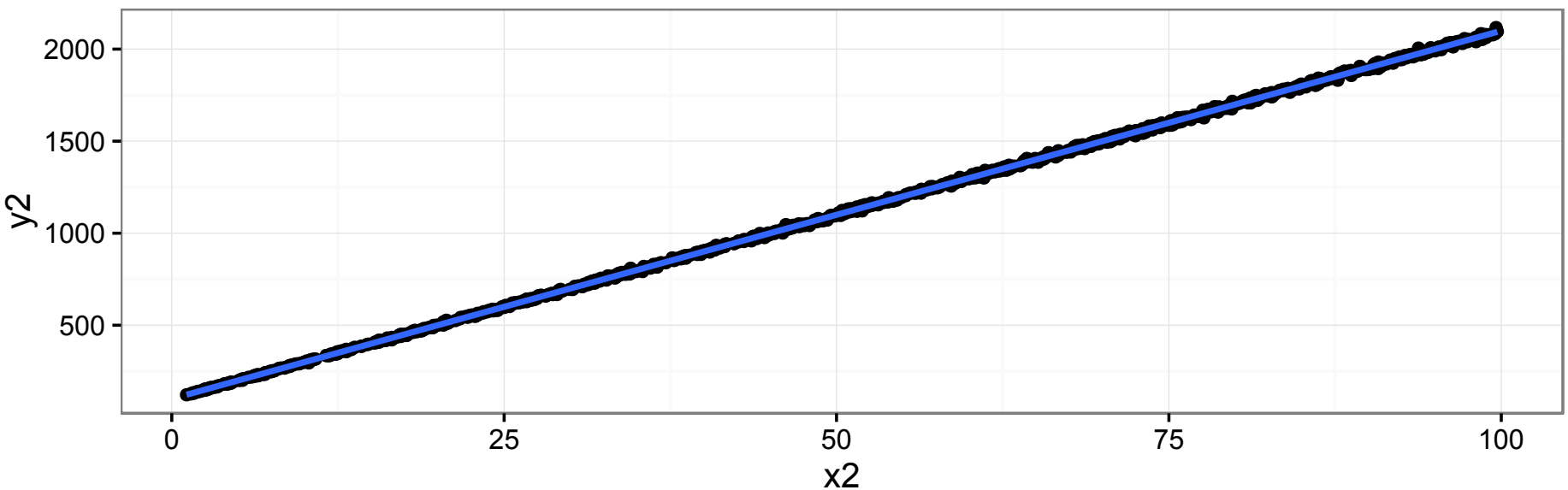




- Выбросов нет
- Зависимость нелинейна
- Остатки не подчиняются нормальному распределению

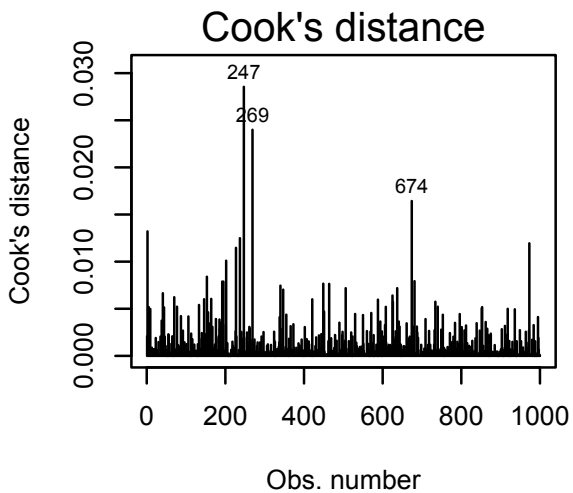
Задание, блок 2

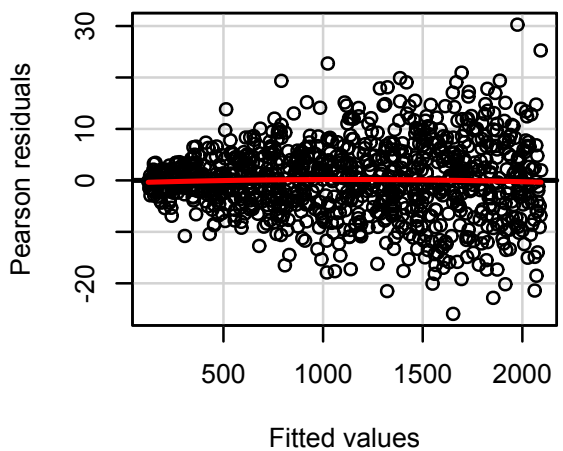
```
set.seed(12345)
x2 <- runif(1000, 1, 100)
b_0 <- 100;  b_1 <- 20
h <- function(x) x^0.5
eps <- rnorm(1000, 0, h(x2))
y2 <- b_0 + b_1 * x2 + eps
dat2 <- data.frame(x2, y2)
ggplot(dat2, aes(x = x2, y = y2)) + geom_point() + geom_smooth(method = "lm
```

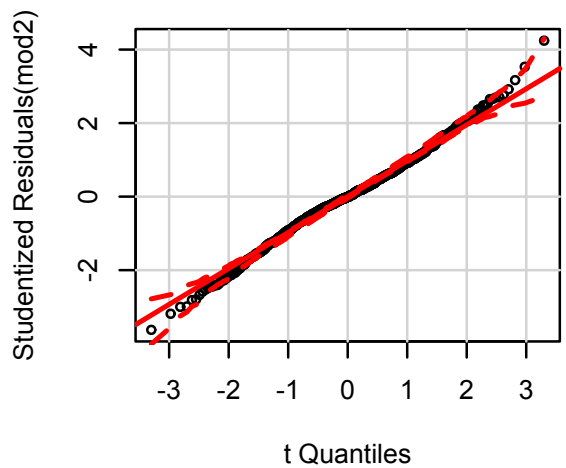


Решение, блок 2

```
mod2 <- lm(y2 ~ x2, data = dat2)
op <- par(mfrow = c(1, 3))
plot(mod2, which = 4)
residualPlot(mod2)
qqPlot(mod2)
par(op)
```

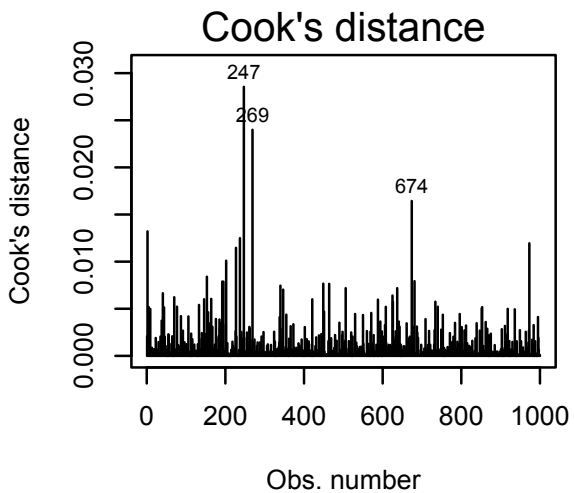


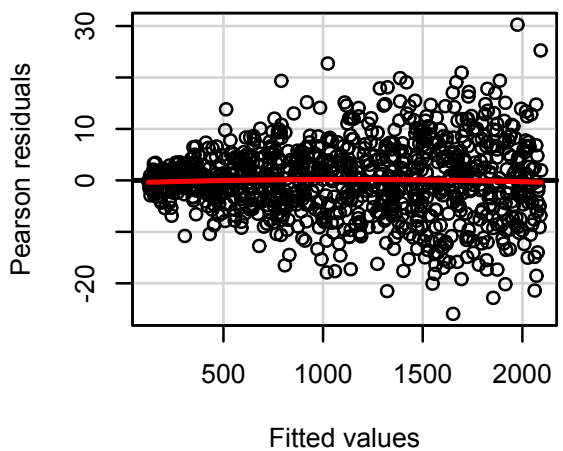


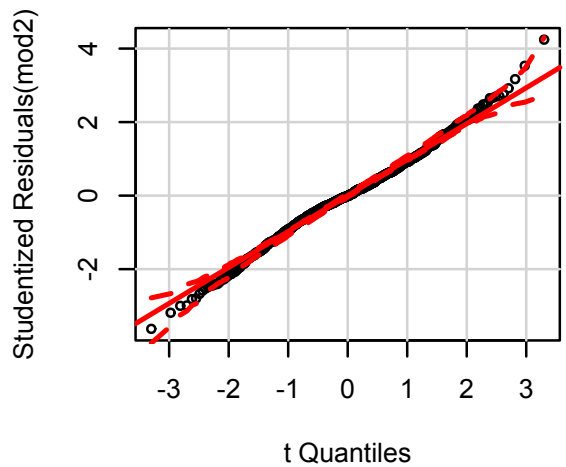


Решение, блок 2

```
mod2 <- lm(y2 ~ x2, data = dat2)
op <- par(mfrow = c(1, 3))
plot(mod2, which = 4)
residualPlot(mod2)
qqPlot(mod2)
par(op)
```



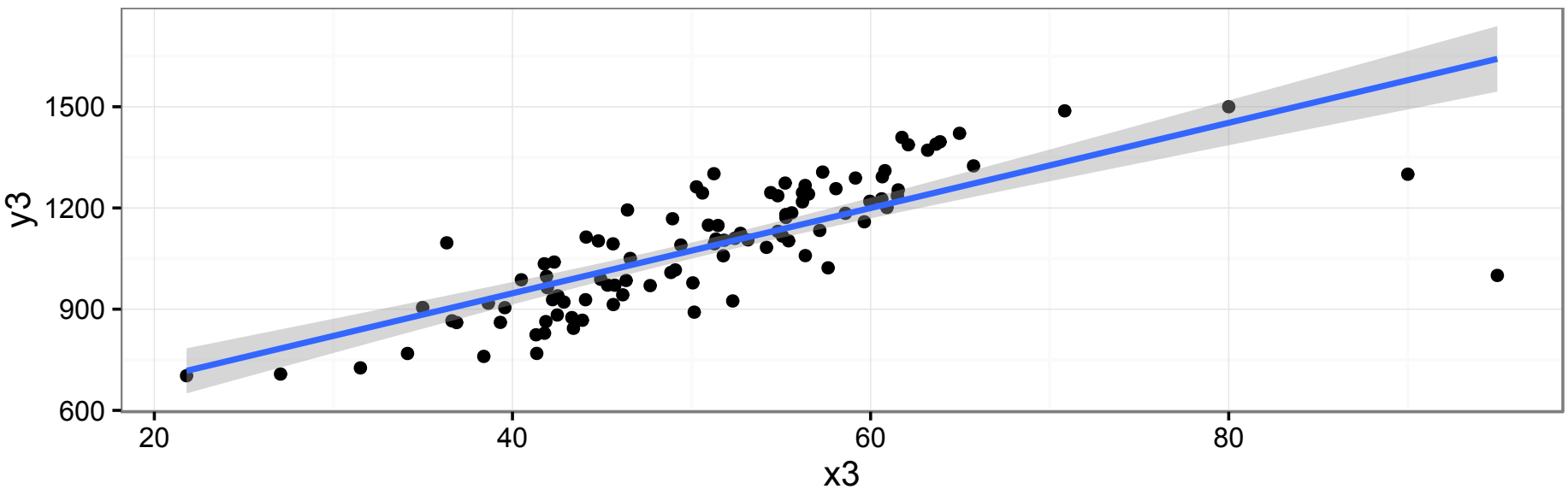




- Выбросов нет
- Гетерогенность дисперсий
- Остатки не подчиняются нормальному распределению

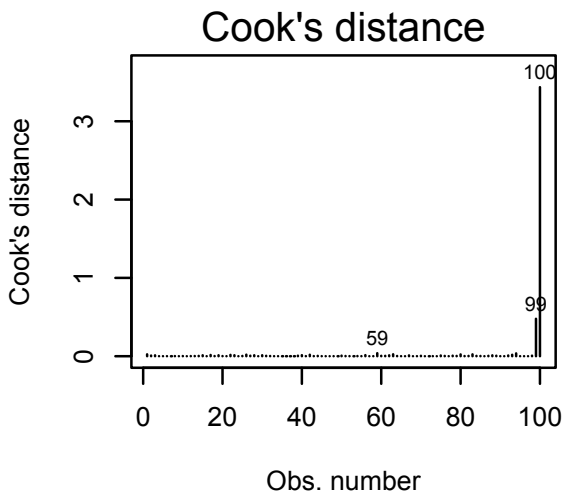
Задание, блок 3

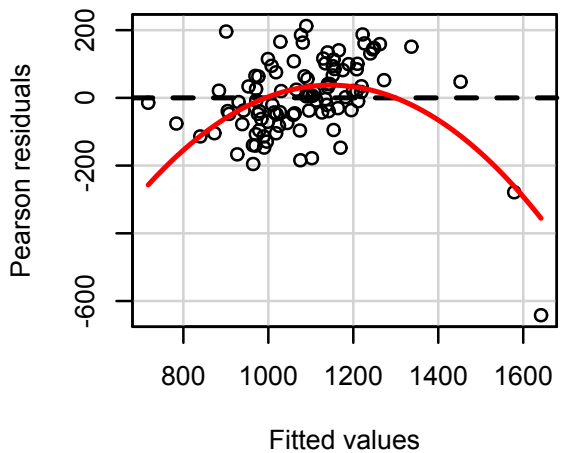
```
set.seed(2309587)
x3 <- rnorm(100, 50, 10)
b_0 <- 100; b_1 <- 20; eps <- rnorm(100, 0, 100)
y3 <- b_0 + b_1*x3 + eps
y3[100] <- 1000; x3[100] <- 95; y3[99] <- 1300; x3[99] <- 90; y3[98] <- 1500;
dat3 <- data.frame(x3, y3)
ggplot(dat3, aes(x=x3, y=y3)) + geom_point() + geom_smooth(method="lm")
```

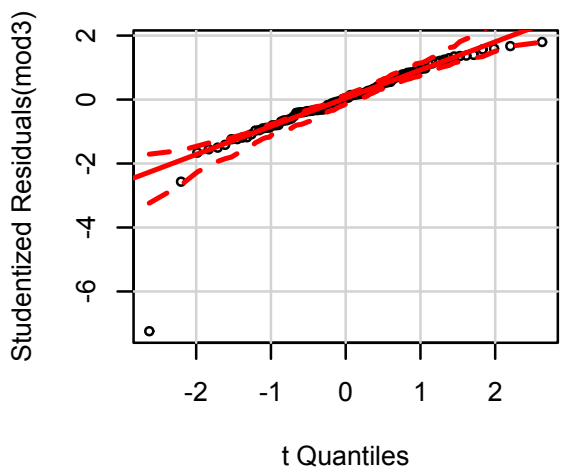


Решение, блок 3

```
mod3 <- lm(y3 ~ x3, data = dat3)
op <- par(mfrow = c(1, 3))
plot(mod3, which = 4)
residualPlot(mod3)
qqPlot(mod3)
par(op)
```

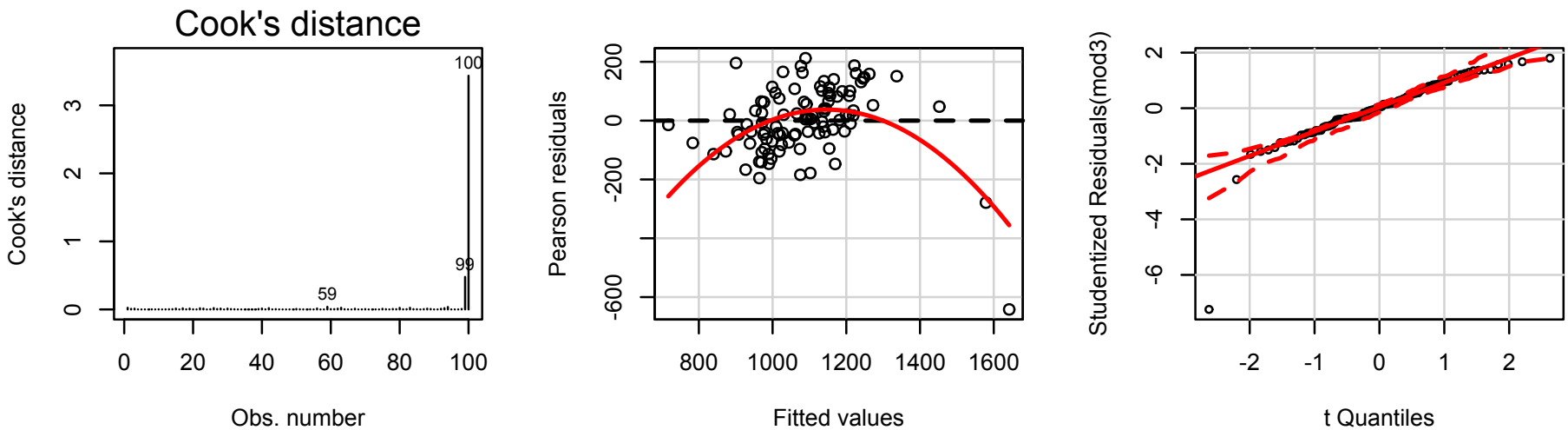






Решение, блок 3

```
mod3 <- lm(y3 ~ x3, data = dat3)
op <- par(mfrow = c(1, 3))
plot(mod3, which = 4)
residualPlot(mod3)
qqPlot(mod3)
par(op)
```



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

Что нужно писать в тексте статьи по поводу проверки валидности моделей?

Вариант 1

Привести необходимые графики в электронных приложениях.

Что нужно писать в тексте статьи по поводу проверки валидности моделей?

Вариант 1

Привести необходимые графики в электронных приложениях.

Вариант 2

Привести в тексте работы результаты тестов на гомогеность дисперсии, автокорреляцию (если используются пространственые или временные предикторы) и нормальность распределиня остатков.

Что нужно писать в тексте статьи по поводу проверки валидности моделей?

Вариант 1

Привести необходимые графики в электронных приложениях.

Вариант 2

Привести в тексте работы результаты тестов на гомогеность дисперсии, автокорреляцию (если используются пространственые или временные предикторы) и нормальность распределиня остатков.

Вариант3

Написать в главе *“Материал и методика”* фразу вроде такой: “Визуальная проверка графиков рассяния остатков не выявила заметных отклонений от условий гомогенности дисперсий и нормальности”.

Summary

- Не всякая модель, в которой коэффициенты достоверно отличаются от нуля, может считаться валидной
- Обязательный этап работы с моделями - проверка условий применимости
- Наиболее важную информацию о валидности модели дает анализ остатков

Что почитать

- Кабаков Р.И. R в действии. Анализ и визуализация данных на языке R. М.: ДМК Пресс, 2014.
- Quinn G.P., Keough M.J. (2002) Experimental design and data analysis for biologists, pp. 92-98, 111-130
- Diez D. M., Barr C. D., Cetinkaya-Rundel M. (2014) Open Intro to Statistics., pp. 354-367.
- Logan M. (2010) Biostatistical Design and Analysis Using R. A Practical Guide, pp. 170-173, 208-211
- Legendre P., Legendre L. (2012) Numerical ecology. Second english edition. Elsevier, Amsterdam.