Диагностика линейных моделей

Линейные модели…

Марина Варфоломеева, Вадим Хайтов

СПбГУ

Мы рассмотрим

- Диагностика линейных моделей
-

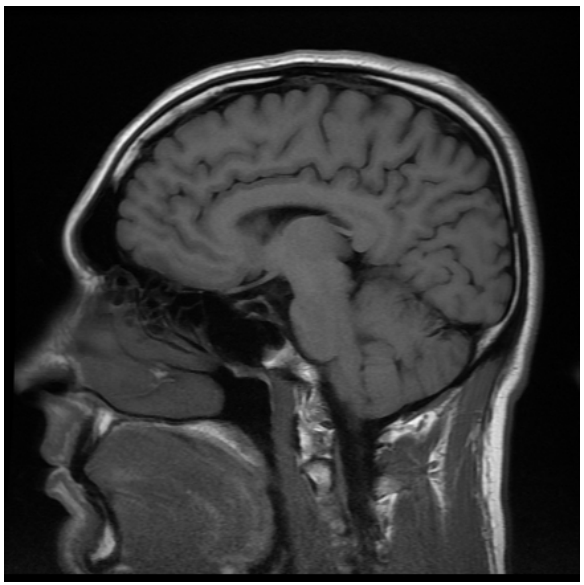
Вы сможете

-
-
-

Пример: IQ и размеры мозга

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

С этим примером мы познакомились в прошлый раз



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

Данные можно загрузить с сайта

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

```
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, описывающую зависимость результатов IQ-теста от размера головного мозга

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

```
#
# Call:
# lm(formula = PIQ ~ MRINACount, data = brain)
#
# Coefficients:
# (Intercept)   MRINACount
#     1.74376      0.00012
```

Пишем статью?

Подберем модель, описывающую зависимость результатов IQ-теста от размера головного мозга

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

```
#
# Call:
# lm(formula = PIQ ~ MRINACount, data = brain)
#
# Coefficients:
# (Intercept)   MRINACount
#     1.74376      0.00012
```

Пишем статью?

Нет, еще рано. Нужно кое-что проверить.

Анализ остатков линейных моделей

Проверка на наличие влиятельных наблюдений

Проверка условий применимости линейных моделей

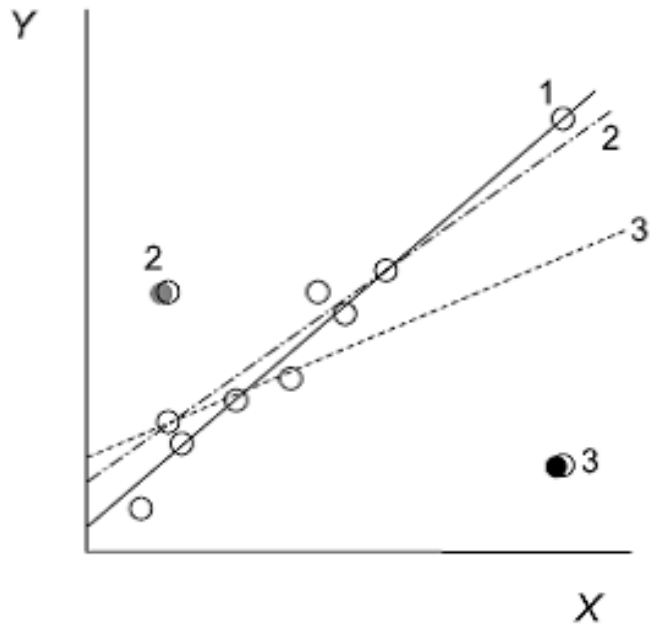
- Линейная связь между зависимой перменной (Y) и предикторами (X)
- Независимость значений Y друг от друга
- Нормальное распределение Y для каждого уровня значений X
- Гомогенность дисерсий Y для каждого уровня значений X
- Отсутствие коллинеарности предикторов (для можественной регрессии)

Проверка на наличие влиятельных наблюдений

Влиятельные наблюдения — это…

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

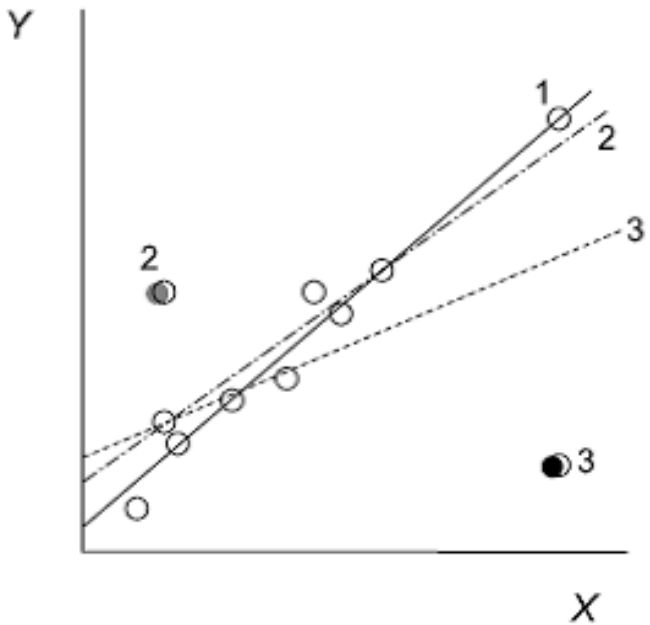
Учет каких из этих точек повлияет на ход регрессии и почему?



Из кн. Quinn, Keugh, 2002

Влиятельные наблюдения — это…

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



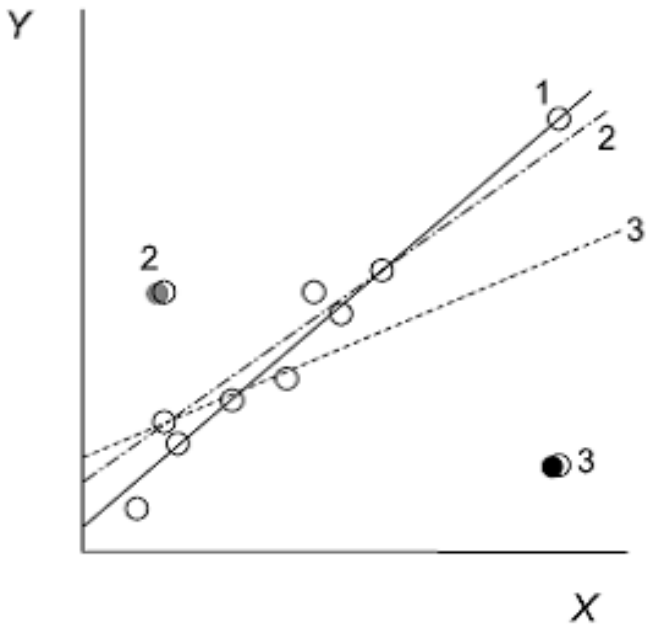
Из кн. Quinn, Keugh, 2002

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*

Влиятельные наблюдения — это…

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



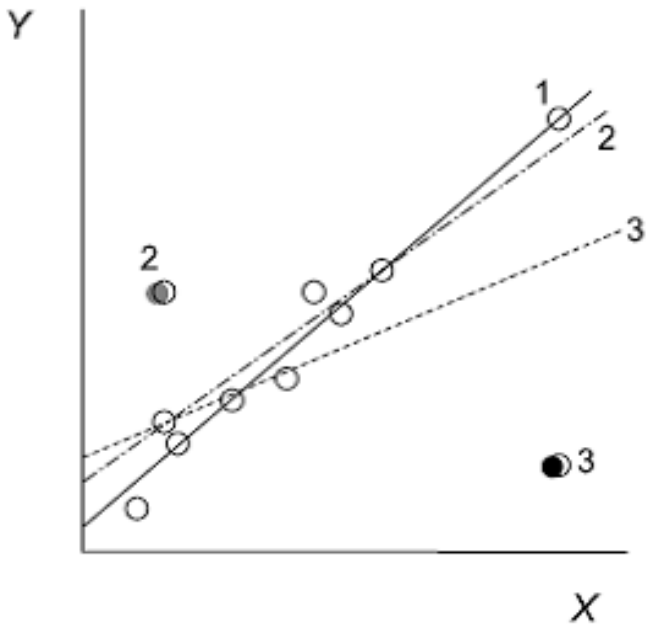
Из кн. Quinn, Keugh, 2002

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

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. выражены в стандартных отклонениях

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. выражены в стандартных отклонениях

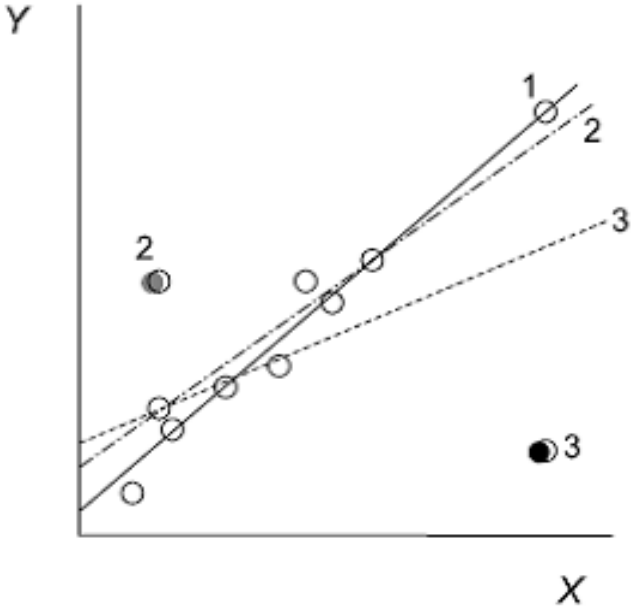
Стьюдентовские остатки

$s_i = \frac{p_i}{\sqrt{1-h_{ii}}} = \frac{\varepsilon_i}{\sqrt{Var(\hat{y}_i)(1-h_{ii})}}$,
где h_{ii} — “сила воздействия” отдельных наблюдений (leverage)

легко сравнивать, т.к. выражены в стандартных отклонениях 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 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)

Извлечем из результатов сведения для анализа остатков

Функция `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
```

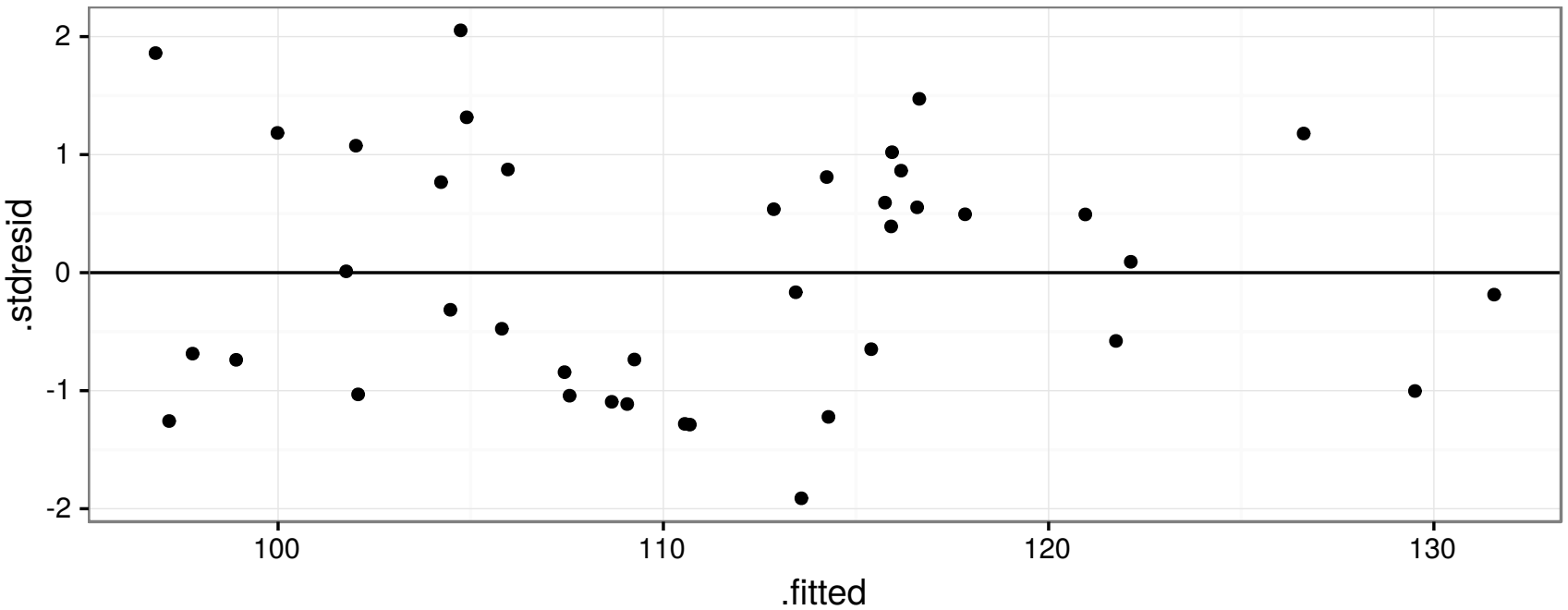
- `.hat` — “сила воздействия” данного наблюдения (*leverage*)
- `.cooksd` — расстояние Кука
- `.fitted` — предсказанные значения
- `.resid` — остатки
- `.stdresid` — стандартизованные остатки

Задание

Для модели `brain_model` постройте график рассеяния стандартизированных остатков в зависимости от предсказанных значений, используя данные из датафрейма `brain_diag`

Решение

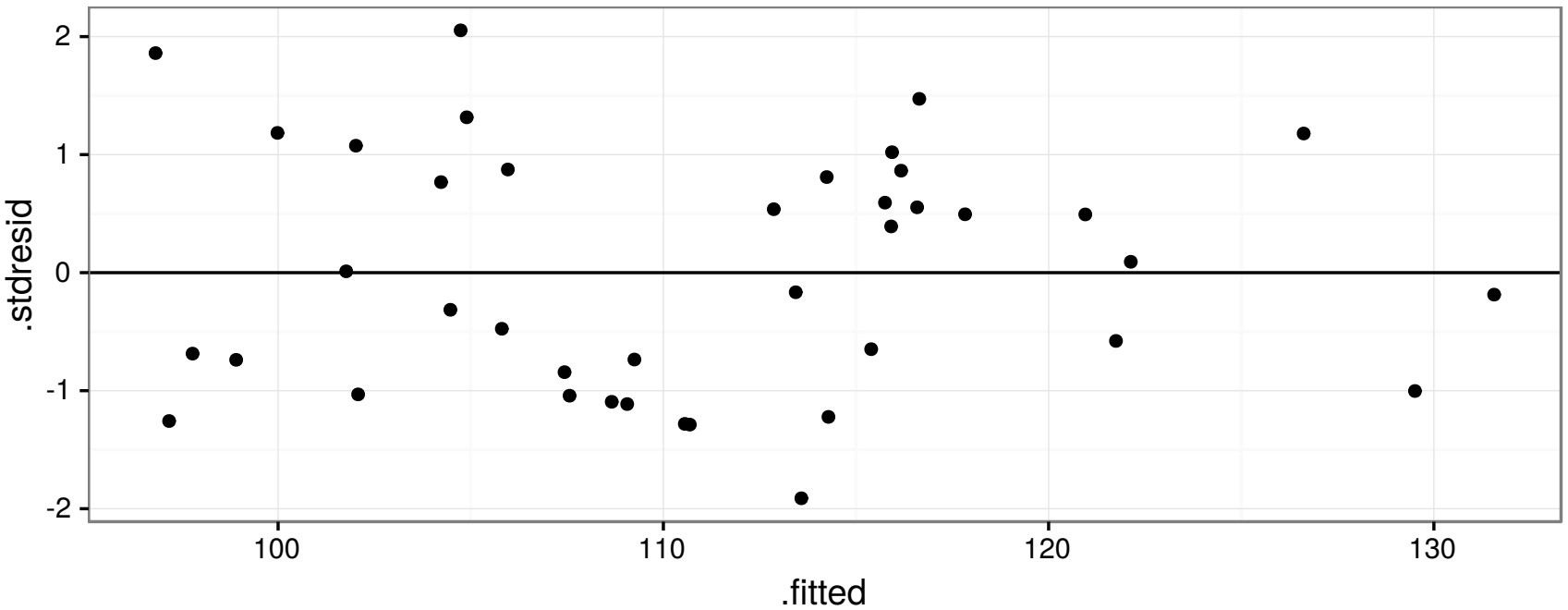
```
theme_set(theme_bw()) # устанавливаем тему (не обязательно)
ggplot(data = brain_diag, aes(x = .fitted, y = .stdresid)) +
  geom_point() + geom_hline(aes(yintercept = 0))
```



Что мы видим?

Решение

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

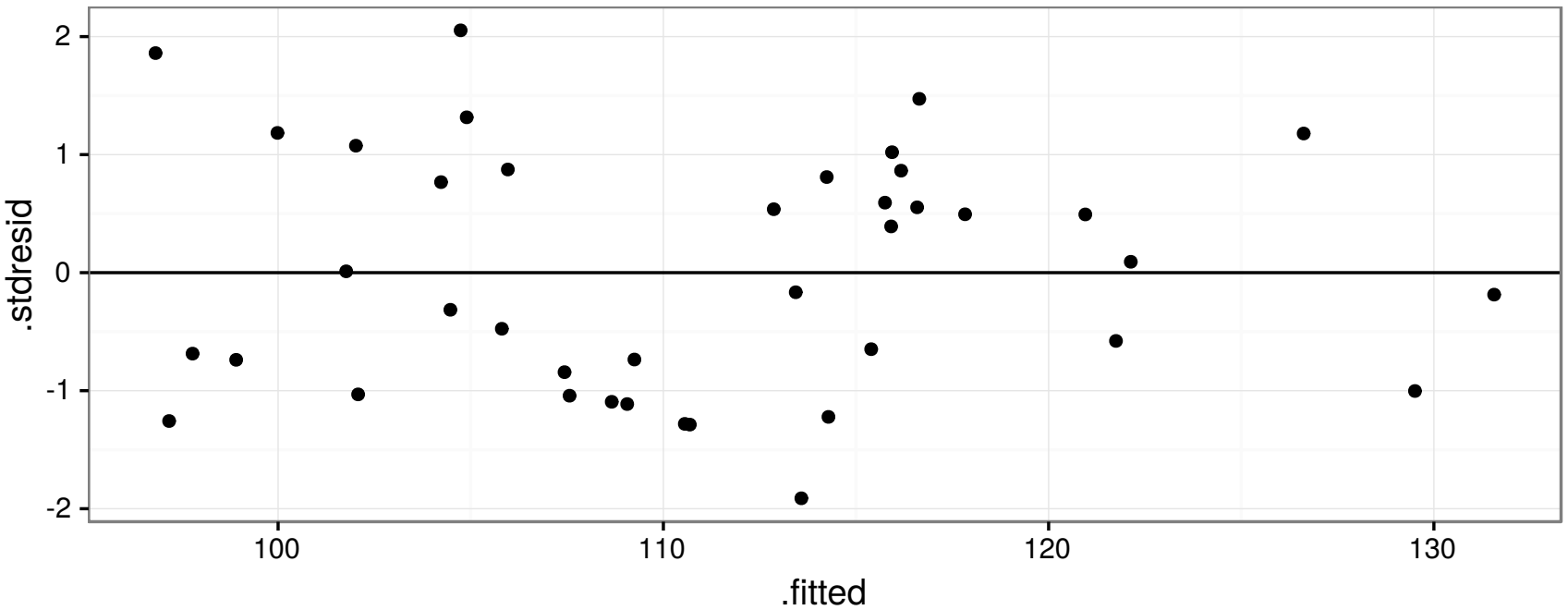


Что мы видим?

- Большая часть стандартизованных остатков в пределах двух стандартных отклонений

Решение

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

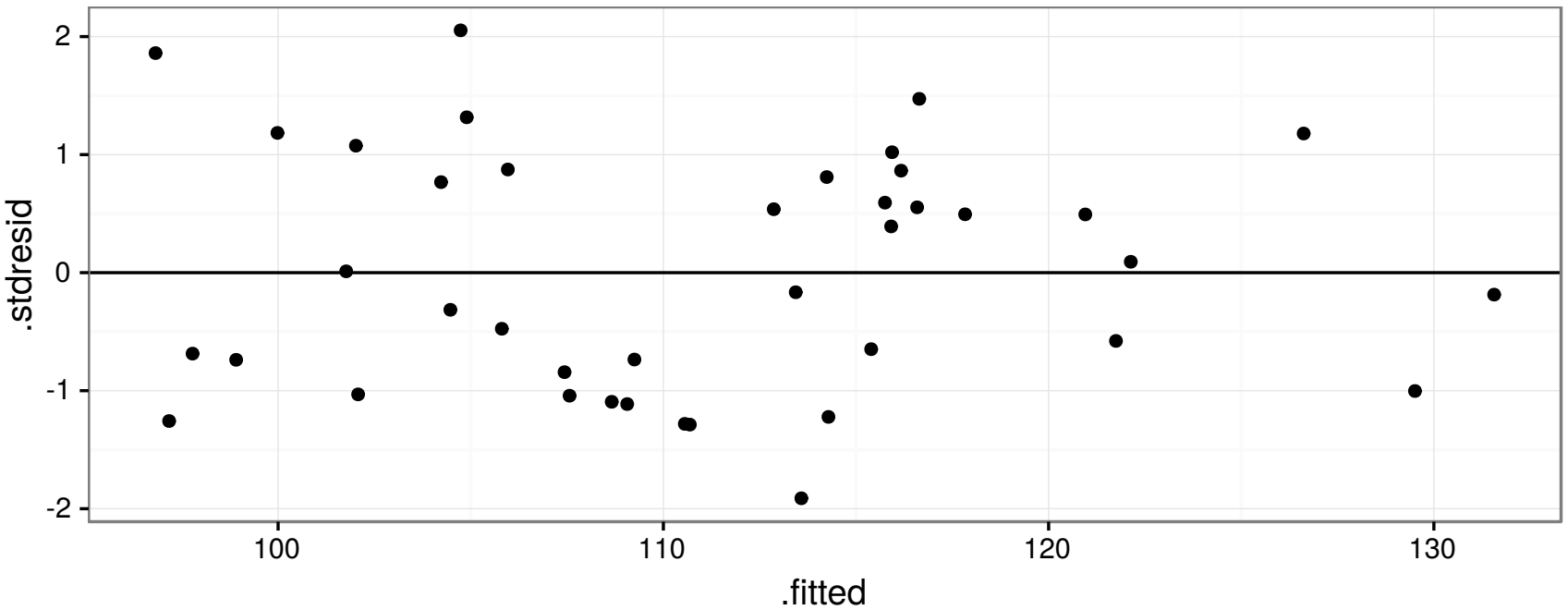


Что мы видим?

- Большая часть стандартизованных остатков в пределах двух стандартных отклонений
- Есть одно влиятельное наблюдение, которое нужно проверить, но сила его влияния невелика

Решение

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

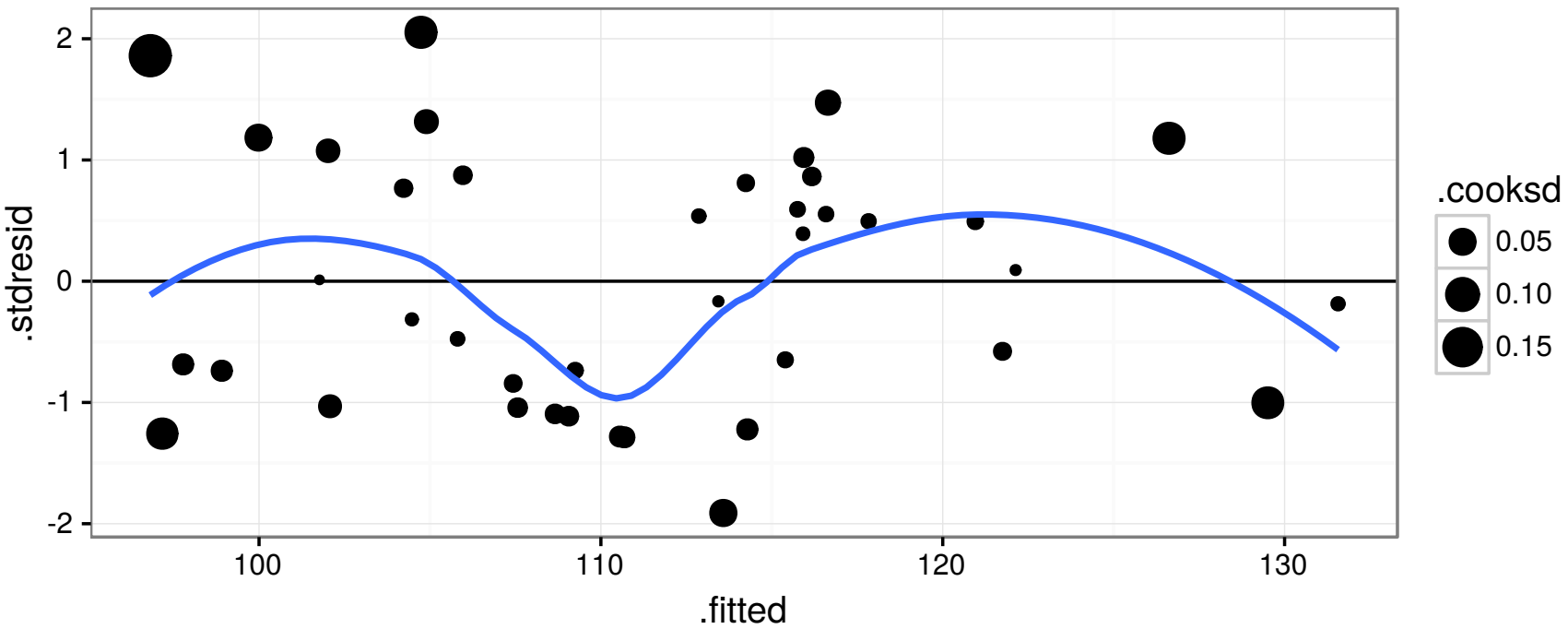


Что мы видим?

- Большая часть стандартизованных остатков в пределах двух стандартных отклонений
- Есть одно влиятельное наблюдение, которое нужно проверить, но сила его влияния невелика
- Среди остатков нет тренда, но, возможно, есть иной паттерн...

Добавим линию loess-сглаживания на график

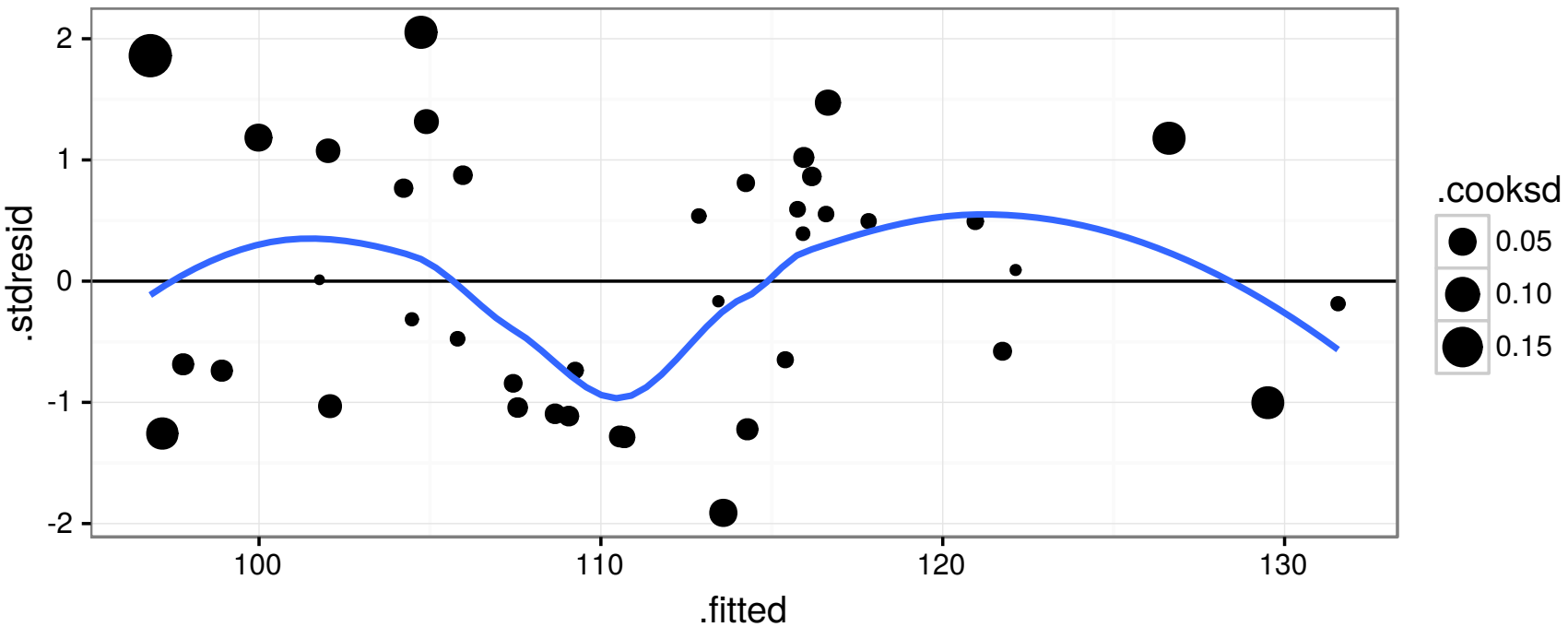
```
ggplot(data = brain_diag, aes(x = .fitted, y = .stdresid)) +
  geom_point(aes(size = .cooksd)) +
  geom_hline(yintercept = 0) +
  geom_smooth(method="loess", se=FALSE)
```



Чем мог быть вызван такой странный паттерн?

Добавим линию loess-сглаживания на график

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

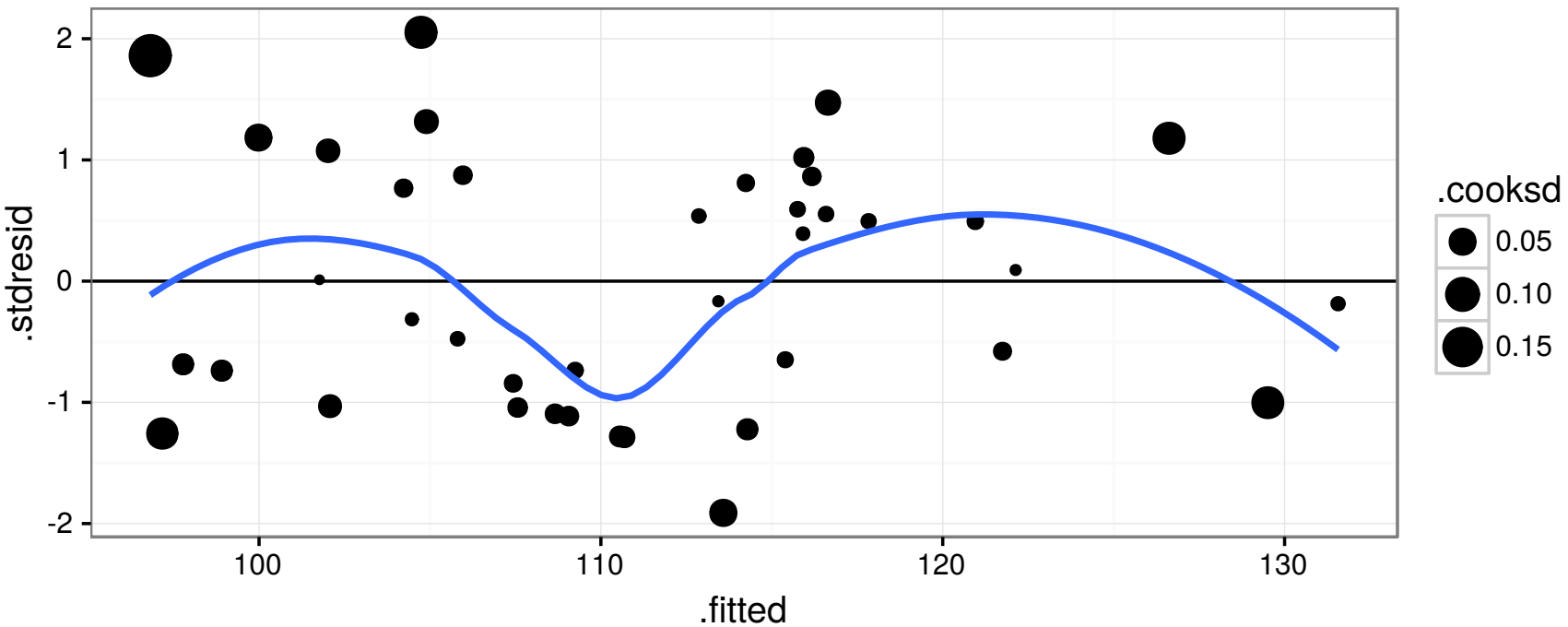


Чем мог быть вызван такой странный паттерн?

- Неучтенная переменная — добавляем в модель

Добавим линию loess-сглаживания на график

```
ggplot(data = brain_diag, aes(x = .fitted, y = .stdresid)) +
  geom_point(aes(size = .cooksd)) +
  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

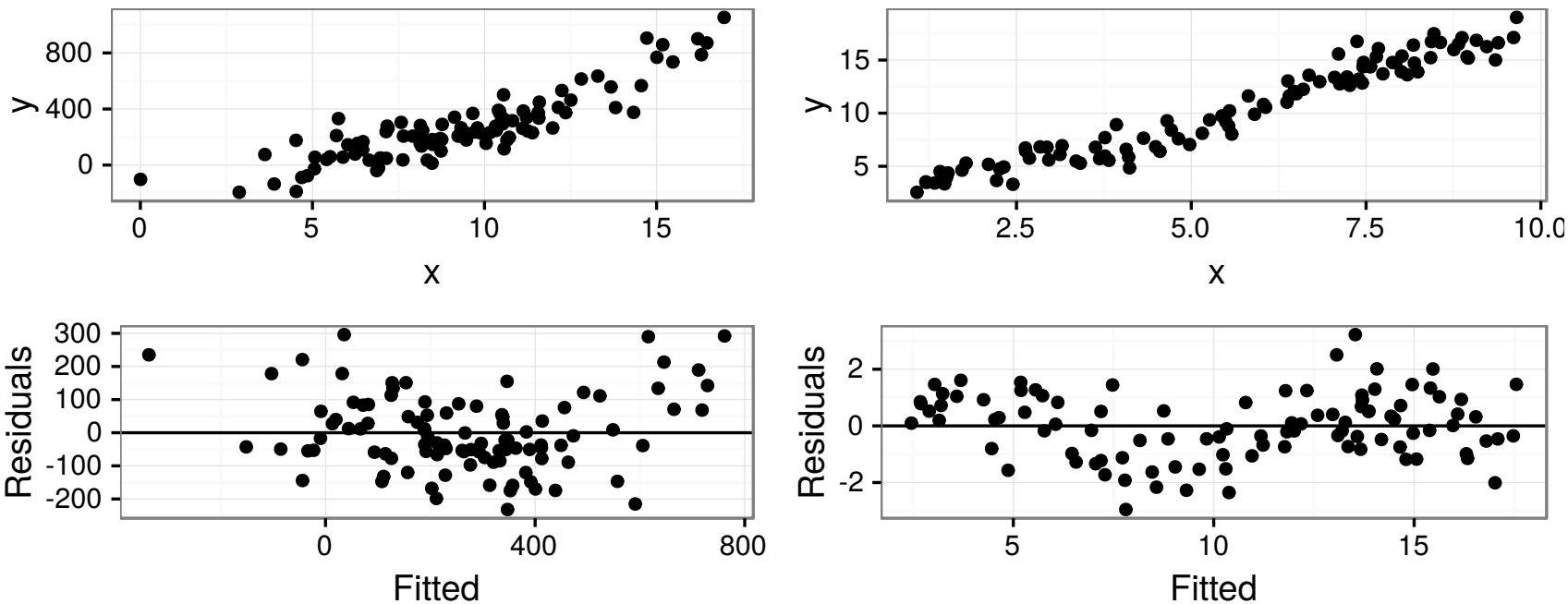
Трансформация	Формула
степень -2	$1/x^2$
степень -1	$1/x$
степень -0.5	$1/\sqrt{x}$
степень 0.5	$\sqrt{x}$
логарифмирование	$log(x)$

Условия применимости линейных моделей (Assumptions)

1. Линейность связи

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

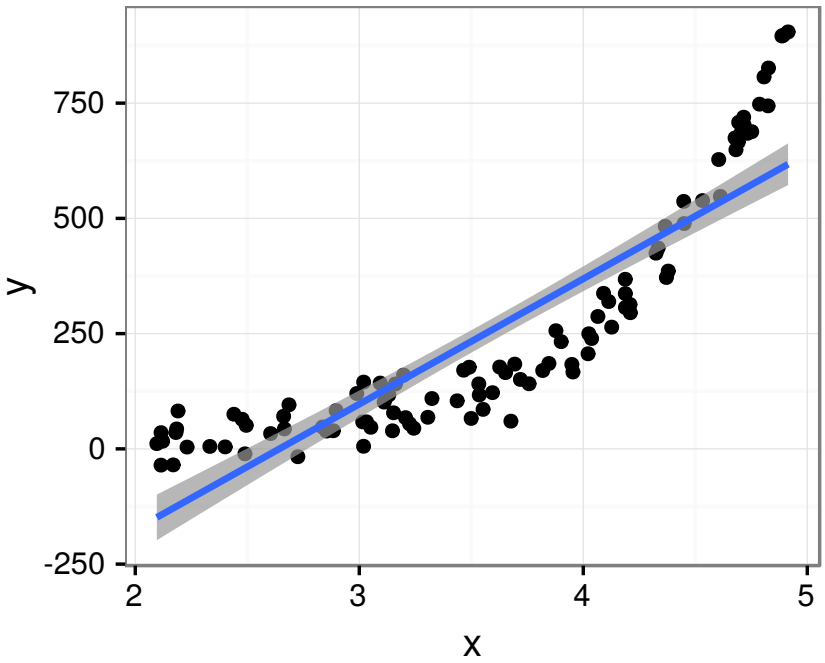
Они становятся лучше заметны на графиках рассеяния остатков (Residual plots)

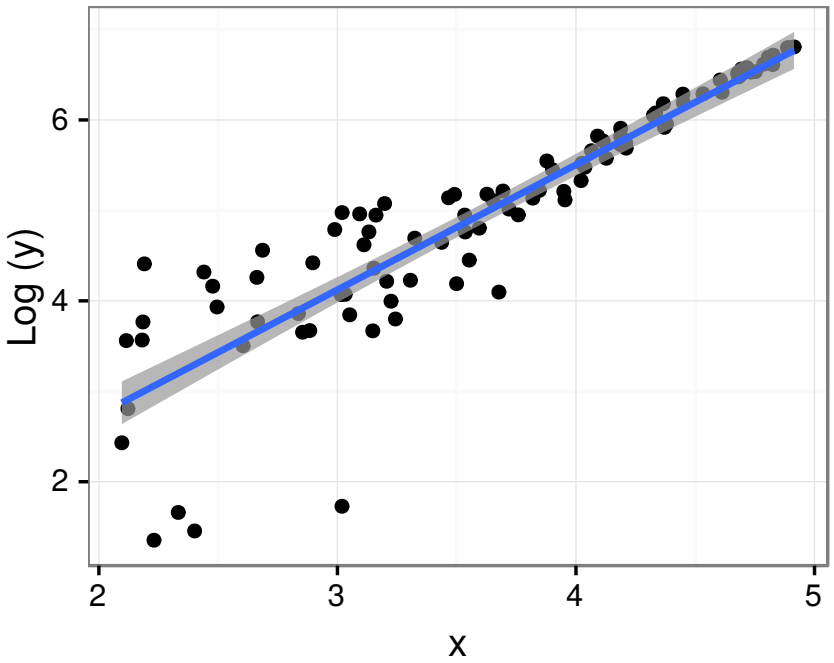


Что делать, если связь нелинейна?

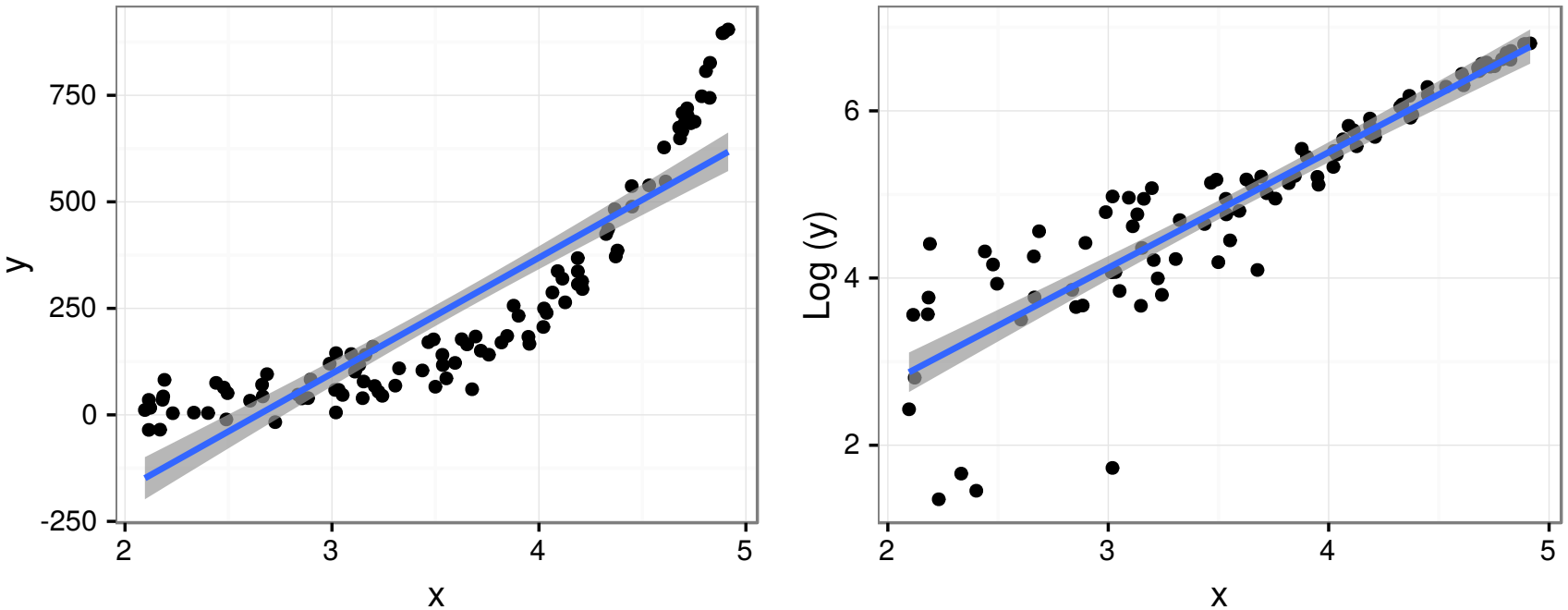
- Построить аддитивную модель (если достаточно наблюдений по x)
- Построить нелинейную модель (если известна форма зависимости)
- Применить линеаризующее преобразование (Осторожно!)
- Применить обобщенную линейную модель с другой функцией связи (об этом позже)

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 матриала

- Наиболее частые источники зависимостей:
  - псевдоповторности (повторно измеренные объекты — чтобы исправить, используем случайные факторы в модели)
  - неучтенные переменные (чтобы исправить, включаем переменные)
  - временные автокорреляции (если данные - временной ряд)
  - пространственные автокорреляции (если пробы взяты в разных местах)

2. Независимость Y друг от друга

Каждое значение Y_i должно быть независимо от любого другого Y_j

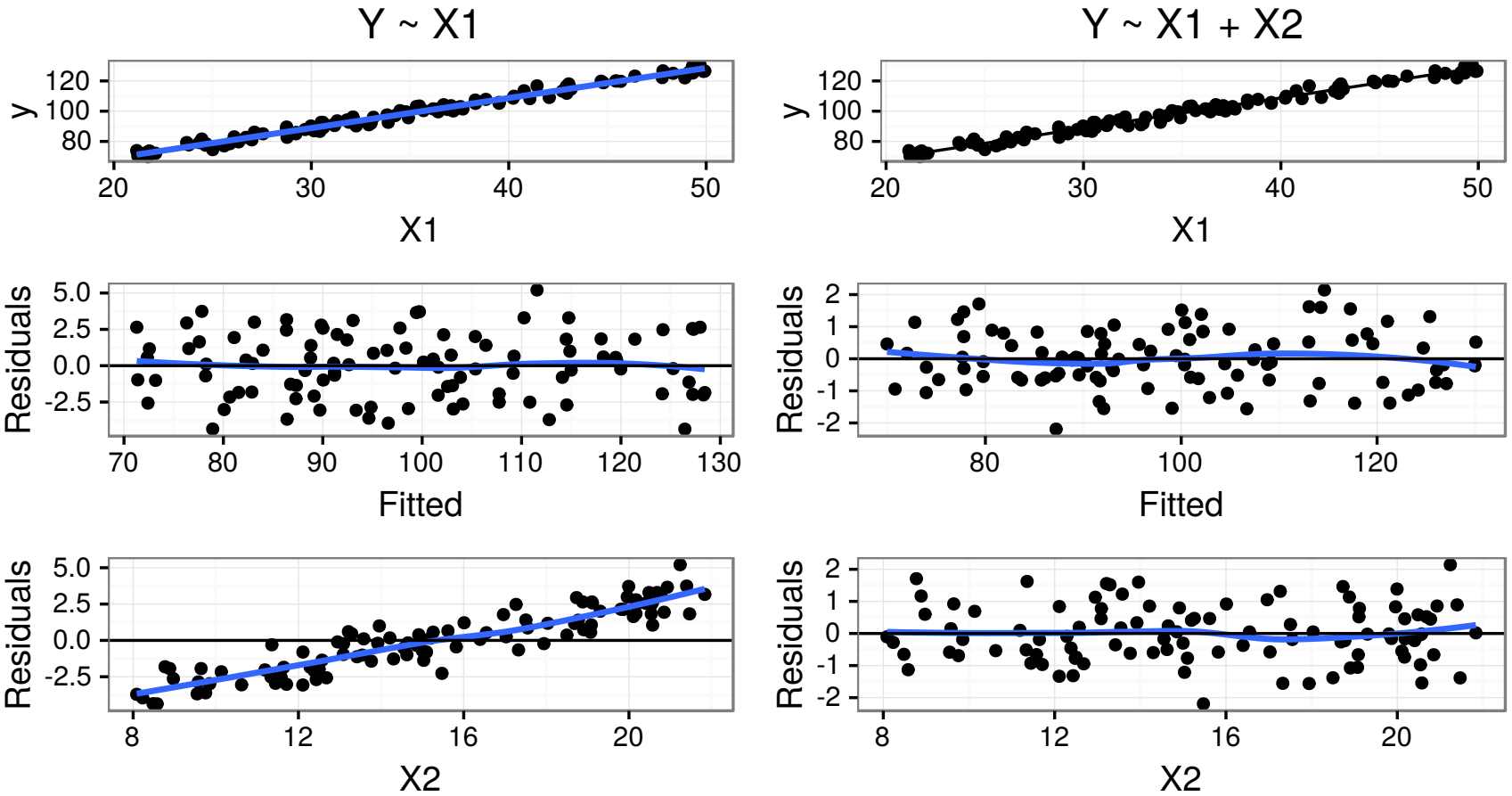
Это нужно контролировать на этапе планирования сбора матриала

- Наиболее частые источники зависимостей:
  - псевдоповторности (повторно измеренные объекты — чтобы исправить, используем случайные факторы в модели)
  - неучтенные переменные (чтобы исправить, включаем переменные)
  - временные автокорреляции (если данные - временной ряд)
  - пространственные автокорреляции (если пробы взяты в разных местах)

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

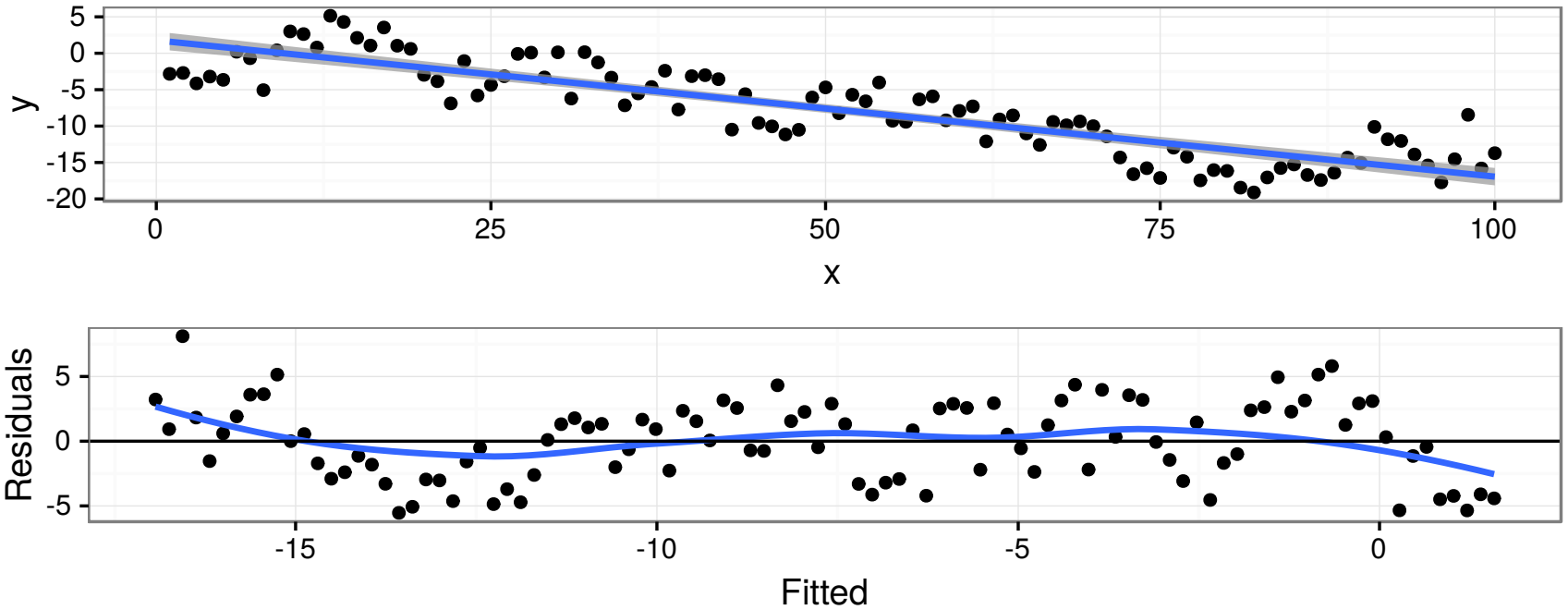


Если в модели не учтена переменная *X*2 (слева), внешне все нормально (только остатки большие), но если построить график зависимости остатков от *X*2.

Если *X*2 учесть (справа) — остатки становятся меньше, зависимость остатков от *X*2 исчезает.

Нарушение условия независимости: Автокоррелированные данные

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)

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

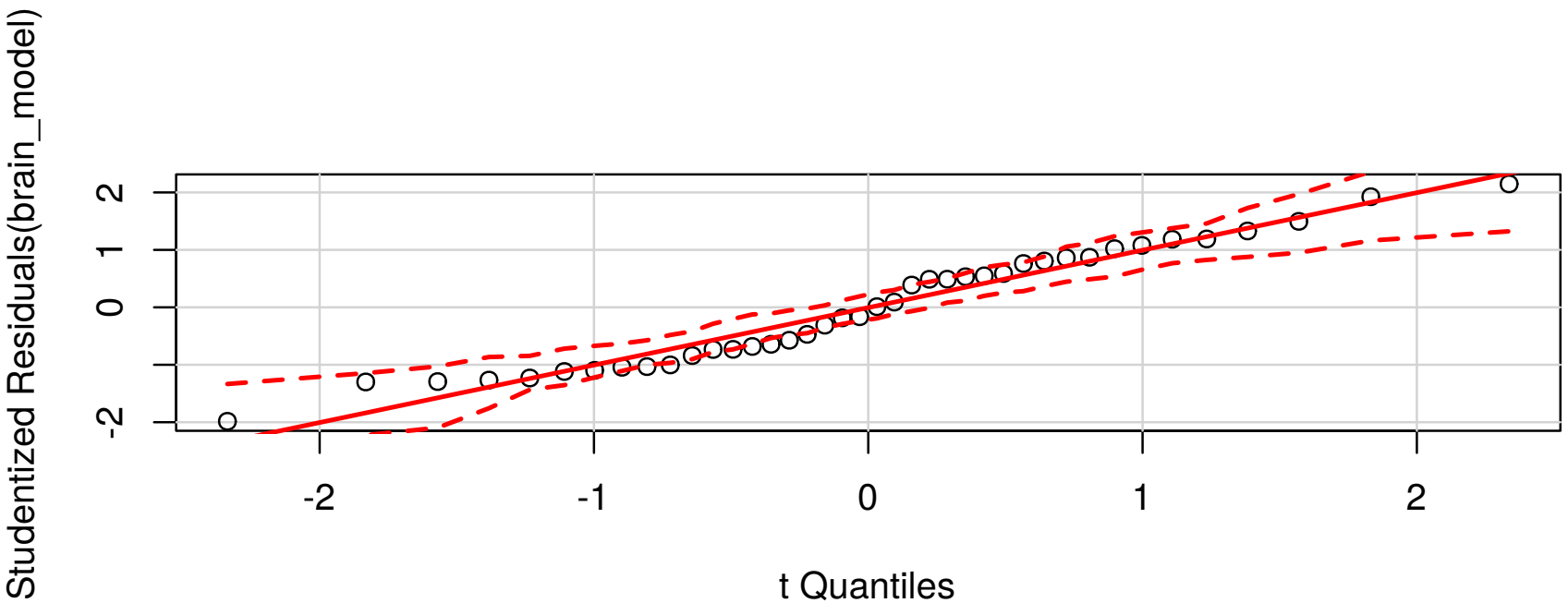
Лучший способ проверки — квантильный график остатков.

Квантильный график остатков

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

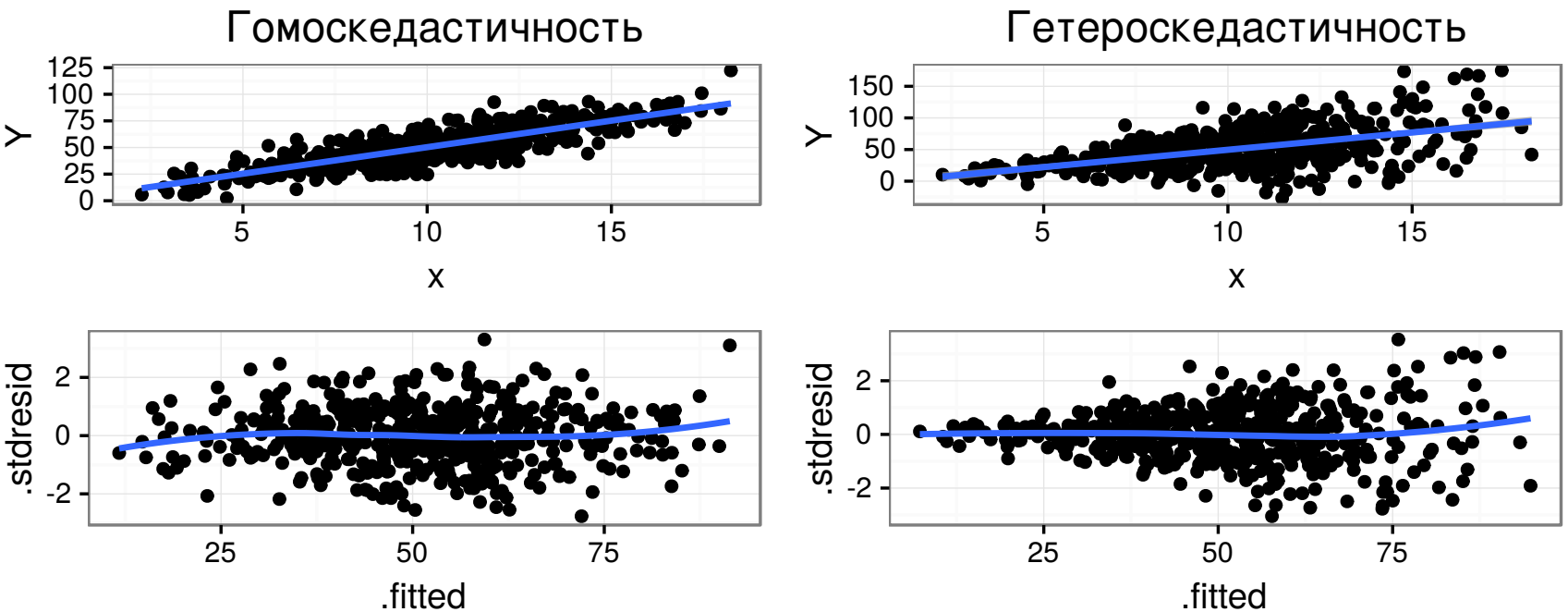
Если точки - это реализации случайной величины из $N(0, \sigma^2)$, то они должны лечь вдоль прямой $Y = X$. Если это стьюдентизированные остатки — то используются квантили t-распределения

```
library(car)
qqPlot(brain_model)
```



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:

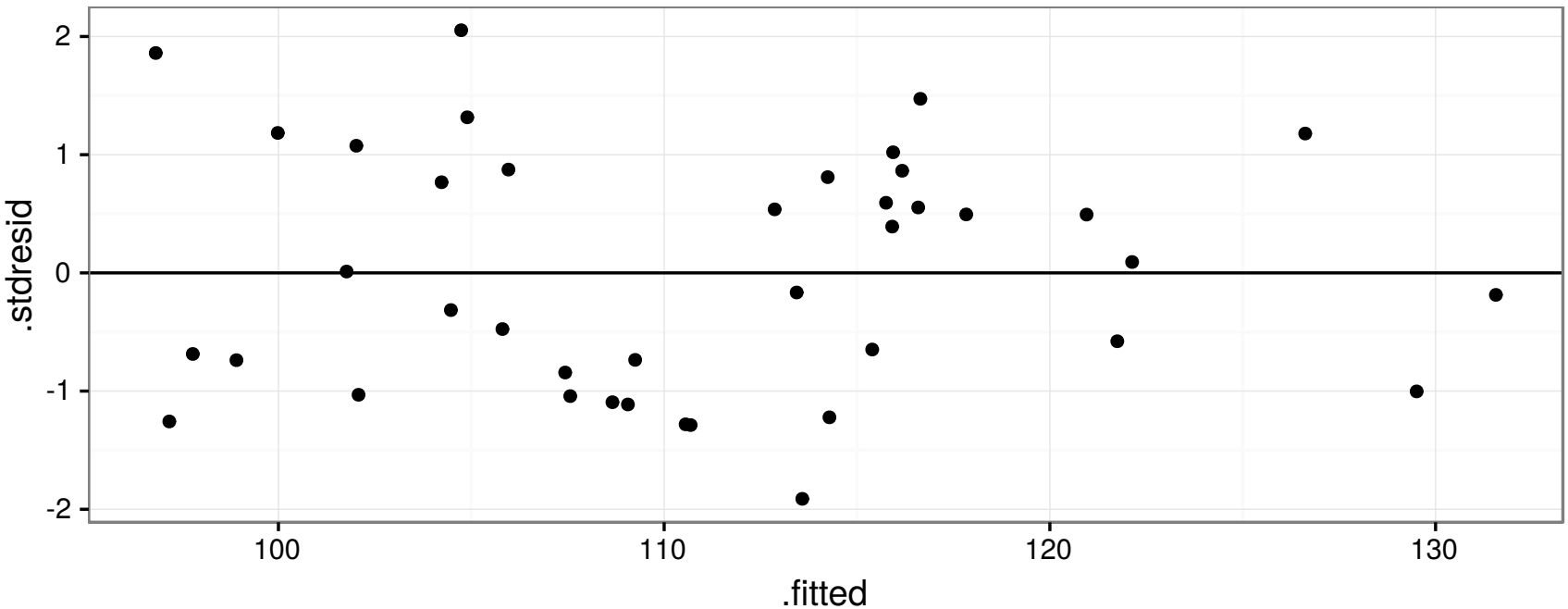
- у формальных тестов тоже есть свои условия применимости, и многие сами неустойчивы к гетероскедастичности
- при больших выборках формальные тесты покажут, что значима даже небольшая гетероскедастичность

Лучший способ проверки — график остатков.

Проверка на гетероскедастичность

Мы уже строили график остатков в `ggplot2`

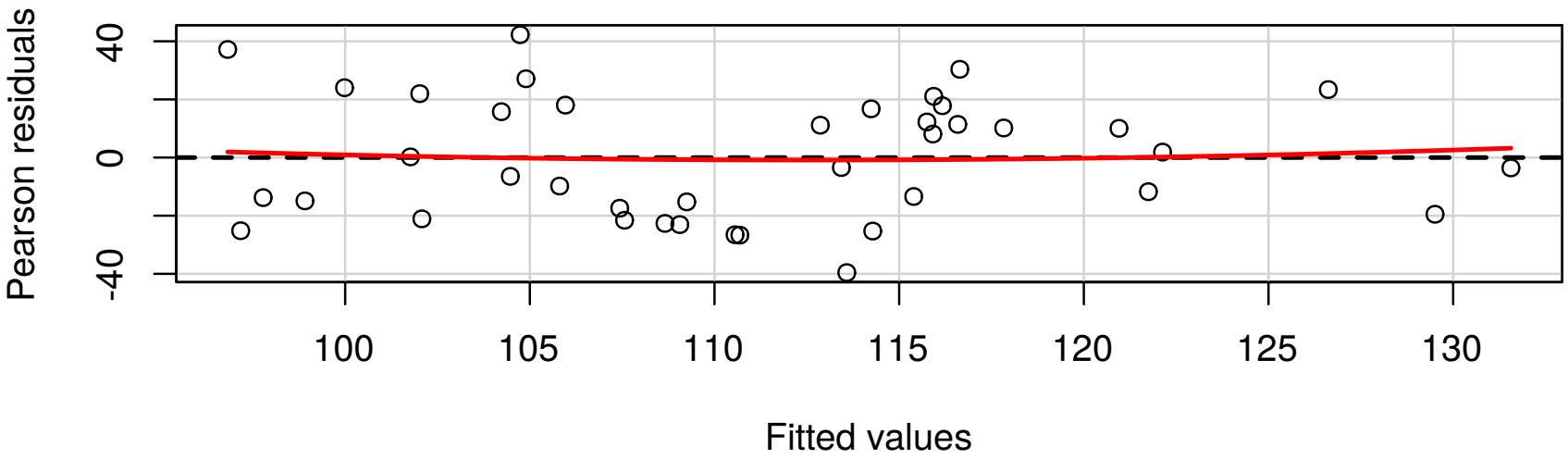
```
ggplot(data = brain_diag,
       aes(x = .fitted, y = .stdresid)) +
  geom_point() + geom_hline(yintercept = 0)
```



Проверка на гетероскедастичность

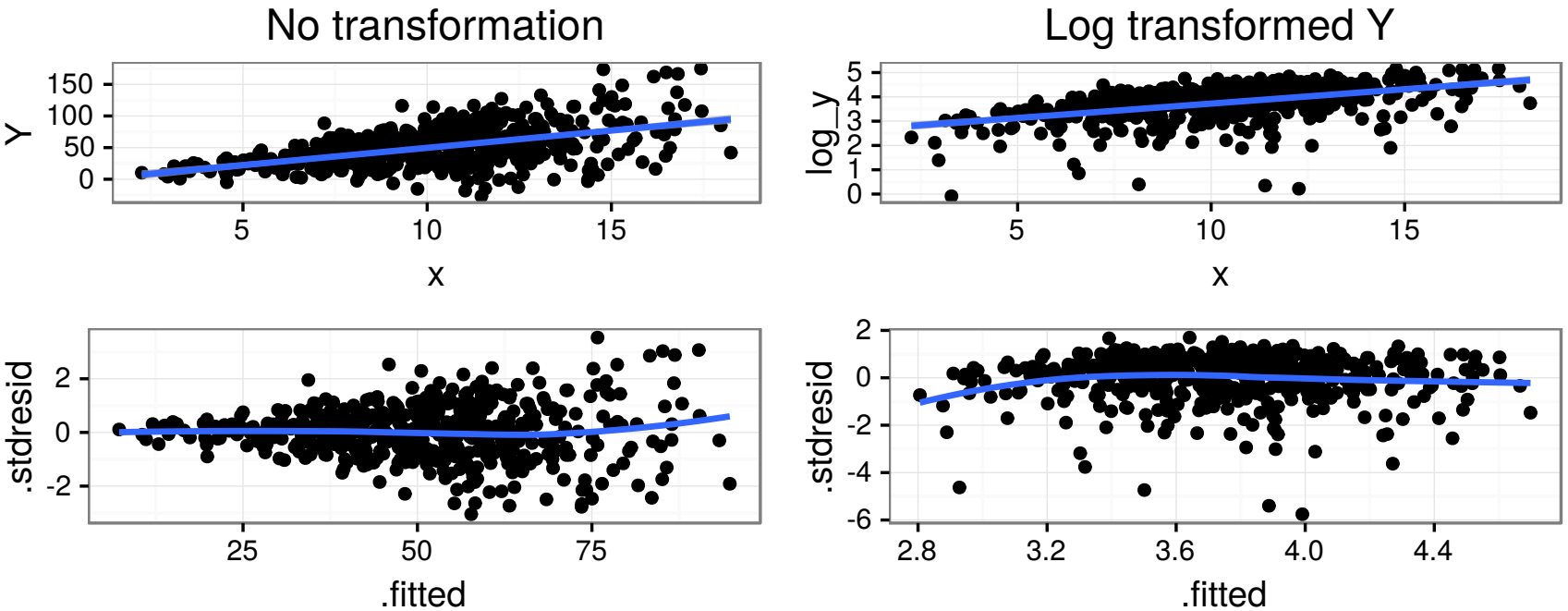
Можем построить аналогичный график остатков средствами пакета `car`

```
residualPlot(brain_model)
```



Что делать если вы столкнулись с гетероскедастичностью?

Решение 1. Применить преобразование зависимой переменной (в некоторых случаях и предиктора).



Недостатки:

- Не всегда спасает.
- Модель описывает поведение не исходной, а преобразованной величины.

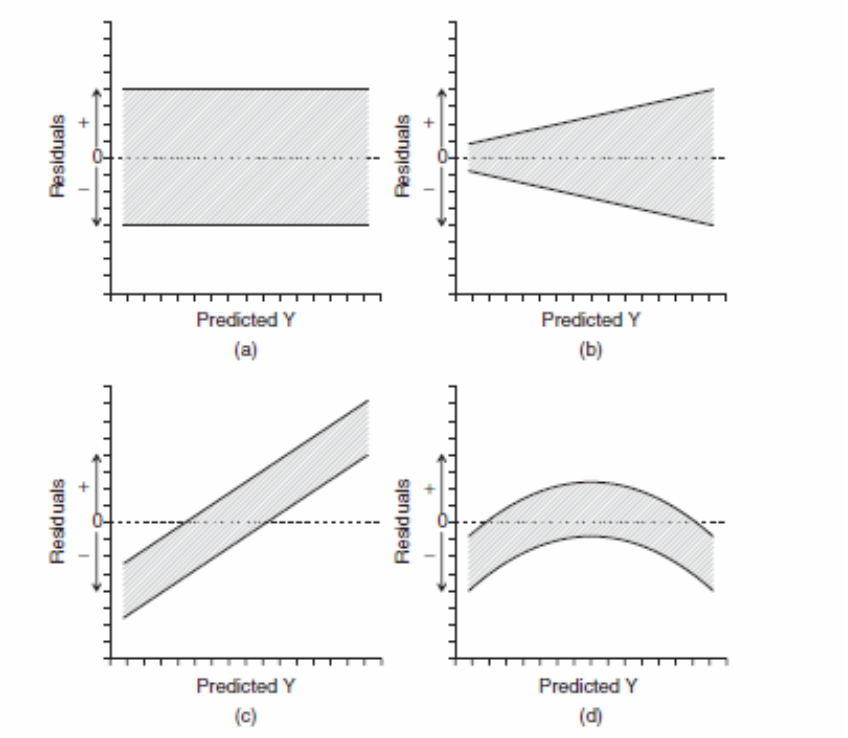
Что делать если вы столкнулись с гетероскедастичностью?

Решение 2. Построить более сложную модель, которая учитывала бы гетерогенность дисперсии зависимой перменной.

"Welcome to our world, the world of *mixed effects modelling*."(Zuur et al., 2009)

Об этом речь впереди!

Некоторые распространенные паттерны на графиках остатков



Вз кн. Logan, 2010, стр. 174

- a) Условия применимости соблюдаются. Модель хорошая
- b) Клиновидный паттерн. Есть гетероскедастичность. Модель плохая
- c) Остатки рассеяны равномерно, но модель неполна. Нужны дополнительные предикторы. Модель можно улучшить
- d) 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?

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 любая модель с достверными результатами проверки H_0 валидна.
- 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.