Регрессионный анализ, часть 2

Математические методы в зоологии с использованием R

Марина Варфоломеева

Вы сможете

- Подобрать модель множественной линейной регрессии
- Протестировать значимость модели и ее коэффициентов
- Интерпретировать коэффициенты множественной регрессии при разных предикторах
- Проверить условия применимости простой и множественной линейной регрессии при помощи анализа остатков

Условия применимости линейной регрессии

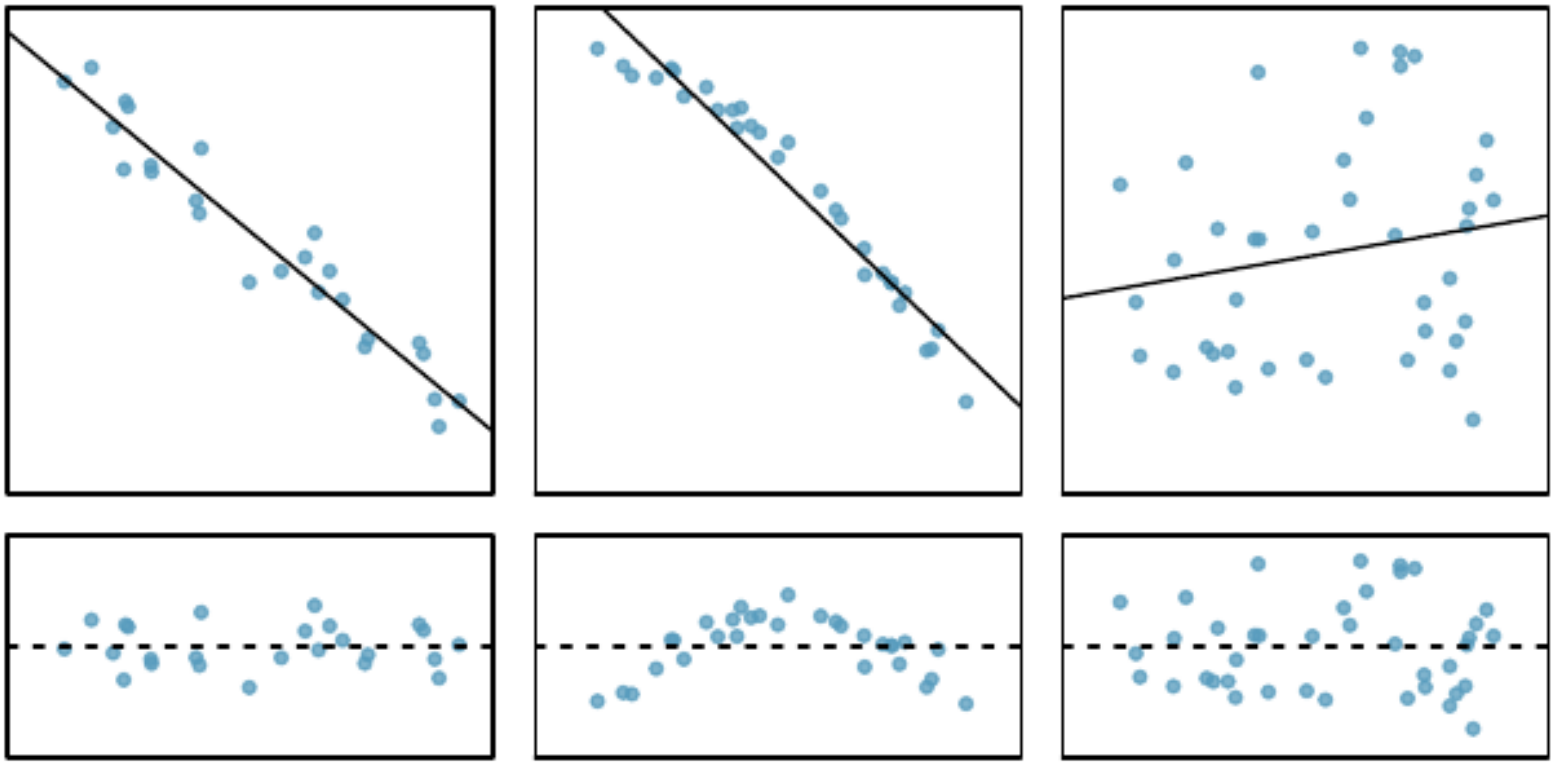
Условия применимости линейной регрессии

Условия применимости линейной регрессии должны выполняться, чтобы тестировать гипотезы

1. Независимость
2. Линейность
3. Нормальное распределение
4. Гомогенность дисперсий
5. Отсутствие коллинеарности предикторов (для множественной регрессии с этого условия нужно начинать!)

1. Независимость

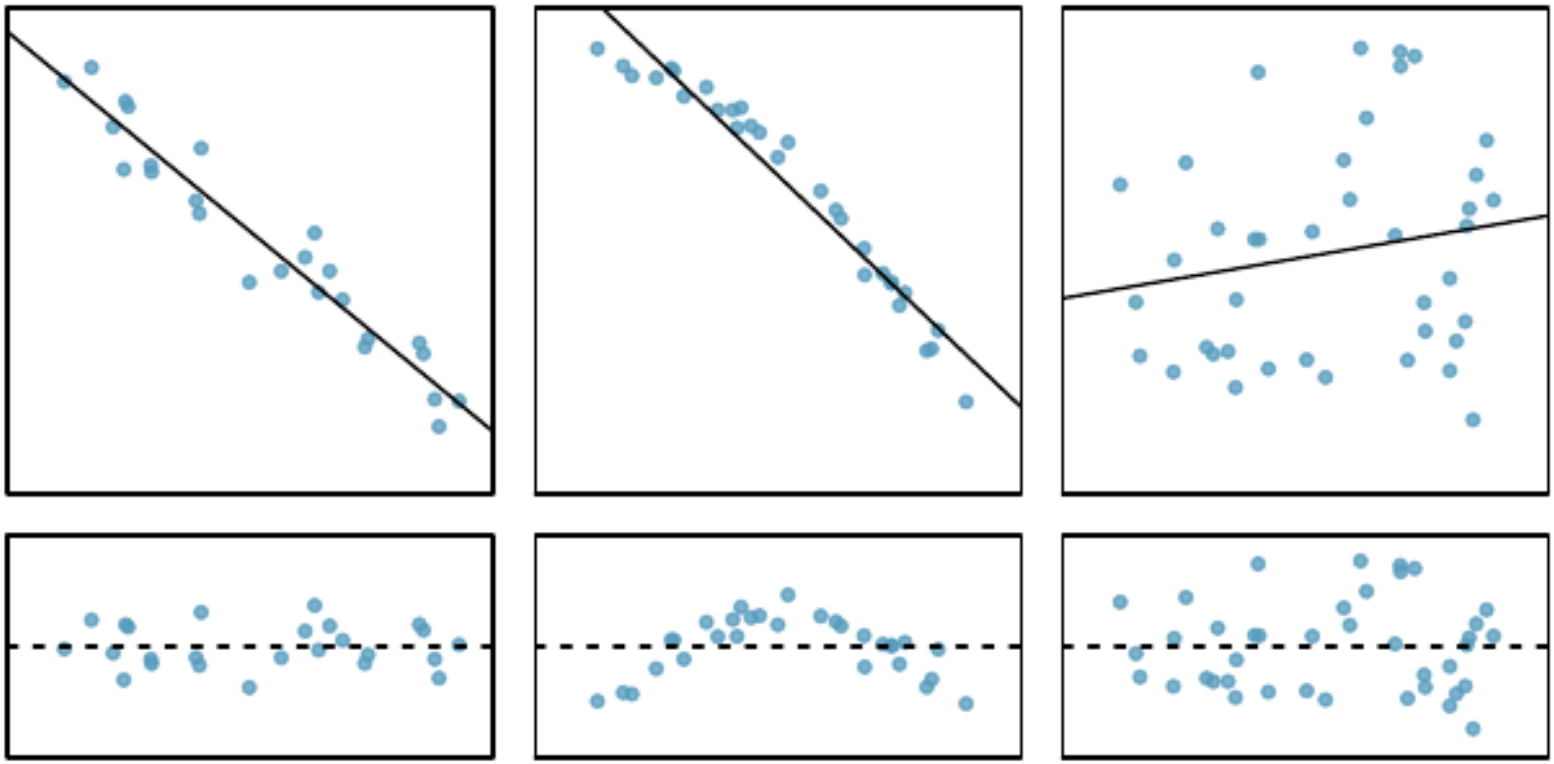
- Значения y_i должны быть независимы друг от друга
- Берегитесь псевдоповторностей и автокорреляций (например, временных)
- Контролируется на этапе планирования
- Проверяем на графике остатков



Из кн. Diez et al., 2010, стр. 332, рис. 7.8

2. Линейность связи

- Проверяем на графике рассеяния исходных данных
- Проверяем на графике остатков



Из кн. Diez et al., 2010, стр. 332, рис. 7.8

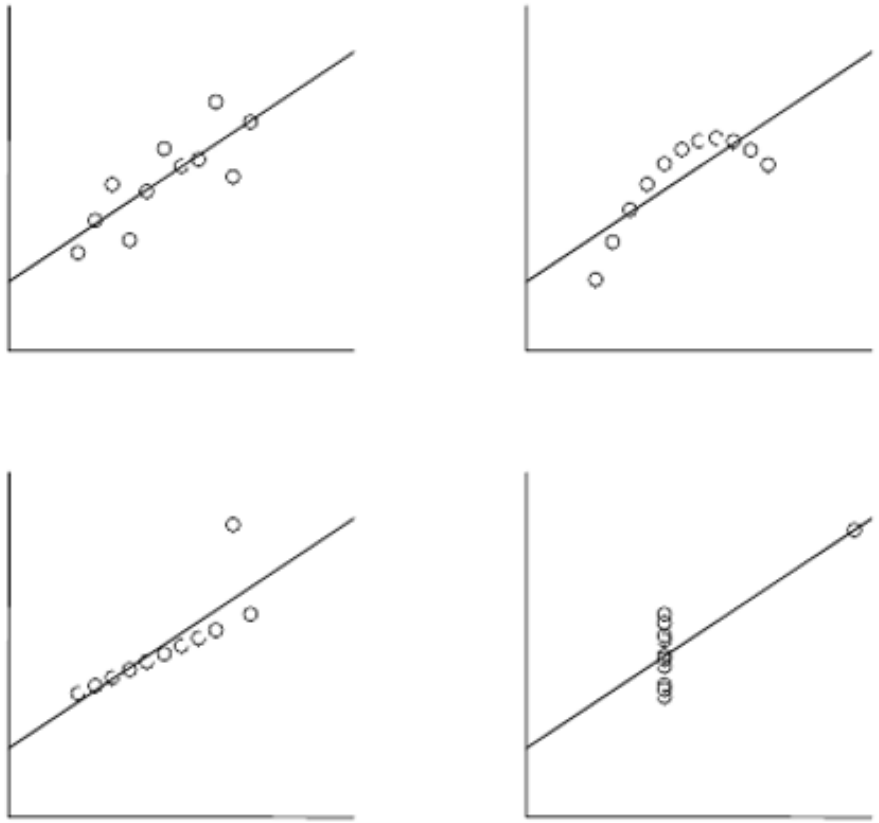
Что бывает, если не глядя применять линейную регрессию

Квартет Энскомба - примеры данных, где регрессии одинаковы во всех случаях (Anscombe, 1973)

$$y_i = 3.0 + 0.5x_i$$

$$r^2 = 0.68$$

$$H_0 : \beta_1 = 0, t = 4.24, p = 0.002$$

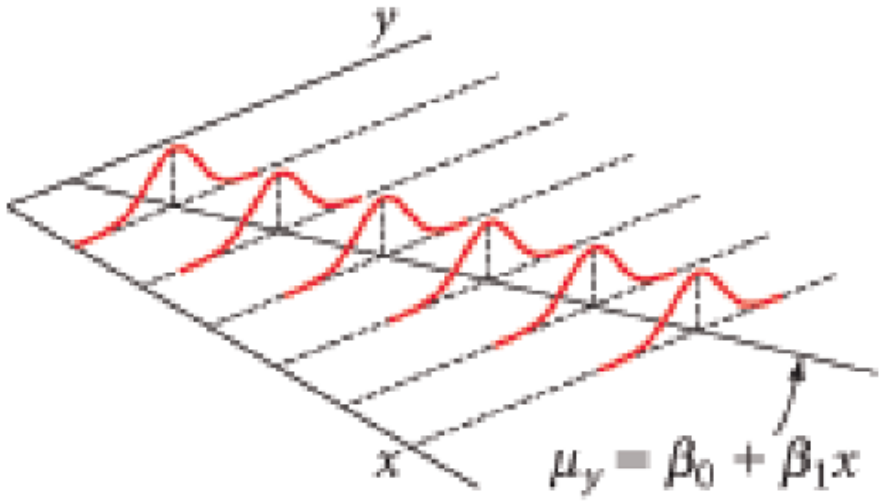


Из кн. Quinn, Keough, 2002, стр. 97, рис. 5.9

3. Нормальное распределение остатков

Нужно, т.к. в модели $Y_i = \beta_0 + \beta x_i + \epsilon_i$ зависимая переменная $Y \sim N(0, \sigma^2)$, а значит $\epsilon_i \sim N(0, \sigma^2)$

- Нужно для тестов параметров, а не для подбора методом наименьших квадратов
- Нарушение не страшно — тесты устойчивы к небольшим отклонениям от нормального распределения
- Проверяем распределение остатков на нормально-вероятностном графике



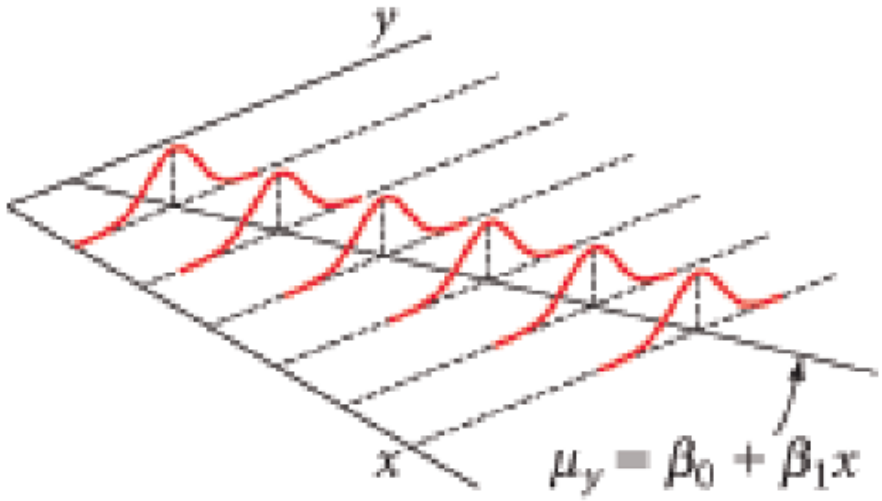
Из кн. Watkins et al., 2008, стр. 743, рис. 11.4

4. Гомогенность дисперсий

Нужно, т.к. в модели $Y_i = \beta_0 + \beta x_i + \epsilon_i$ зависимая переменная $Y \sim N(0, \sigma^2)$ и дисперсии $\sigma_1^2 = \sigma_2^2 = ... = \sigma_i^2$ для каждого Y_i

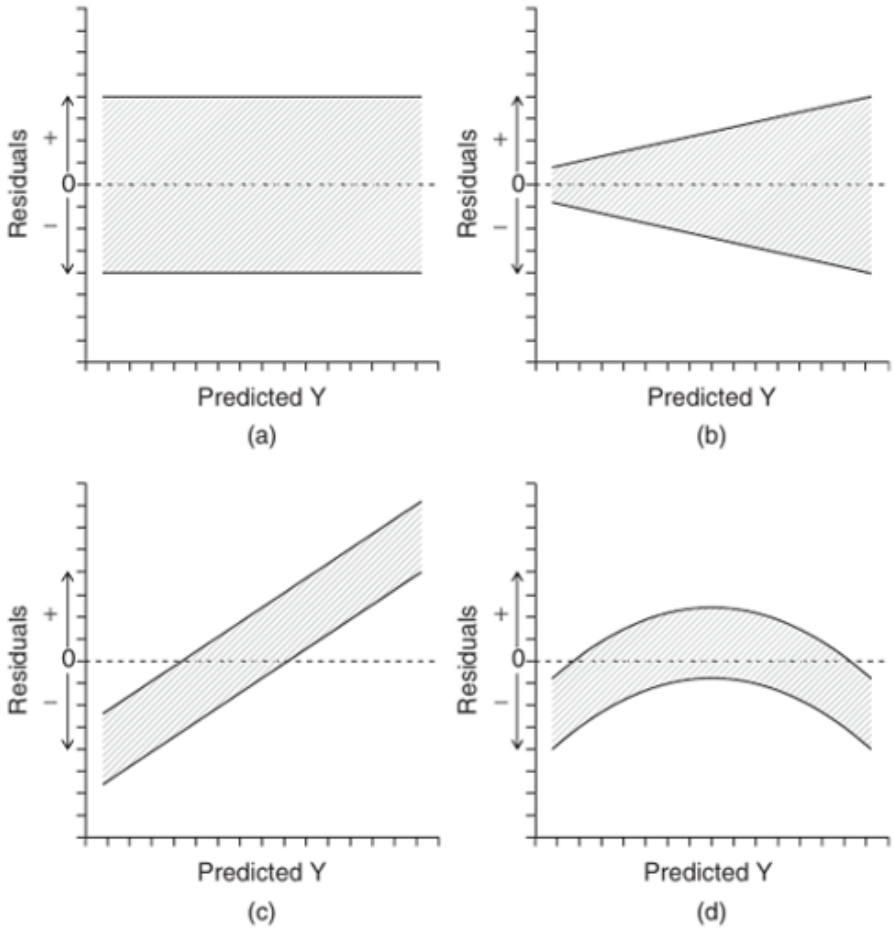
Но, поскольку $\epsilon_i \sim N(0, \sigma^2)$, можно проверить равенство дисперсий остатков ϵ_i

- Нужно и важно для тестов параметров
- Проверяем на графике остатков по отношению к предсказанным значениям
- Есть формальные тесты, но они очень чувствительны (тест Бройша-Пагана, тест Кокрана)



Из кн. Watkins et al., 2008, стр. 743, рис. 11.4

Диагностика регрессии по графикам остатков



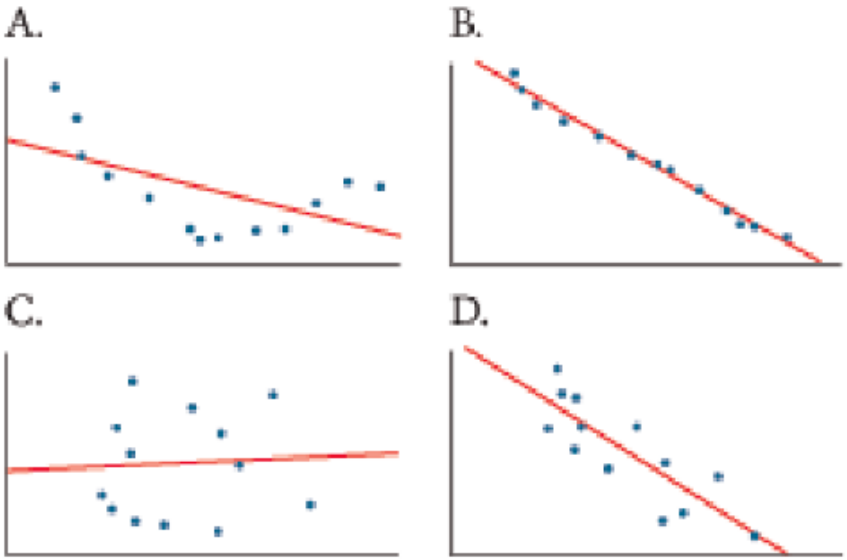
Из кн. Logan, 2010, стр. 174, рис. 8.5 d

- (a)все условия выполнены
- (b)разброс остатков разный (wedge-shaped pattern)
- (c)разброс остатков одинаковый, но нужны дополнительные предикторы
- (d)к нелинейной зависимости применили линейную регрессию

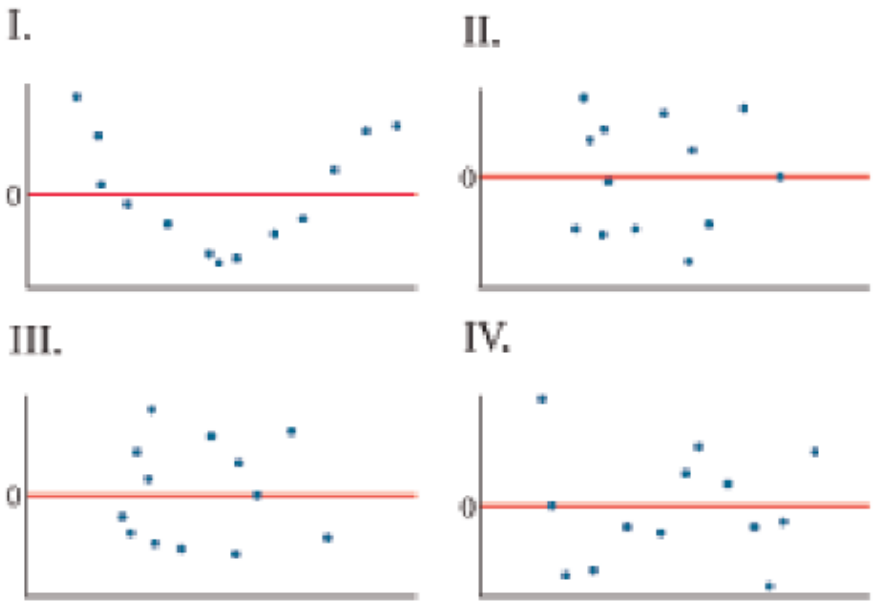
Задача: Проанализируйте графики остатков

Скажите пожалуйста

- какой регрессии соответствует какой график остатков?
- все ли условия применимости регрессии здесь выполняются?
- назовите случаи, в которых можно и нельзя применить линейную регрессию?



Display 3.84 Four scatterplots.

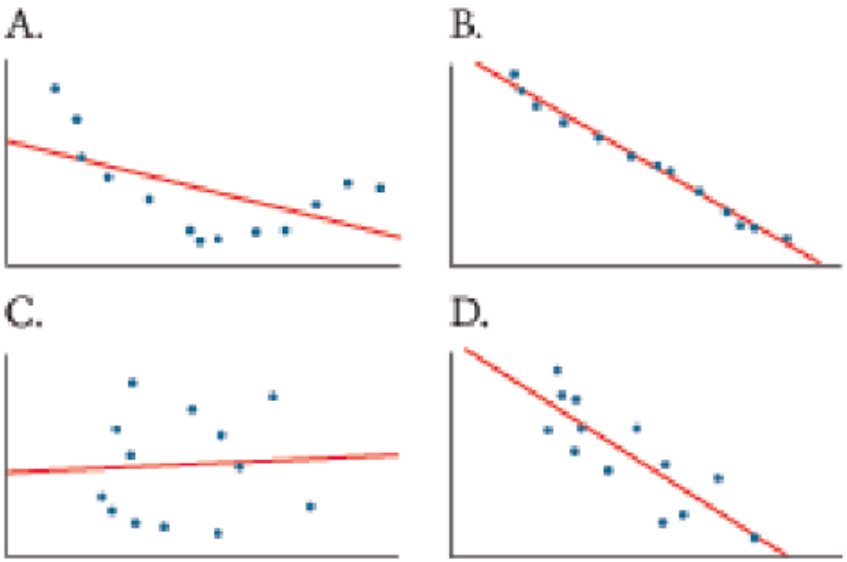


Display 3.85 Four residual plots.

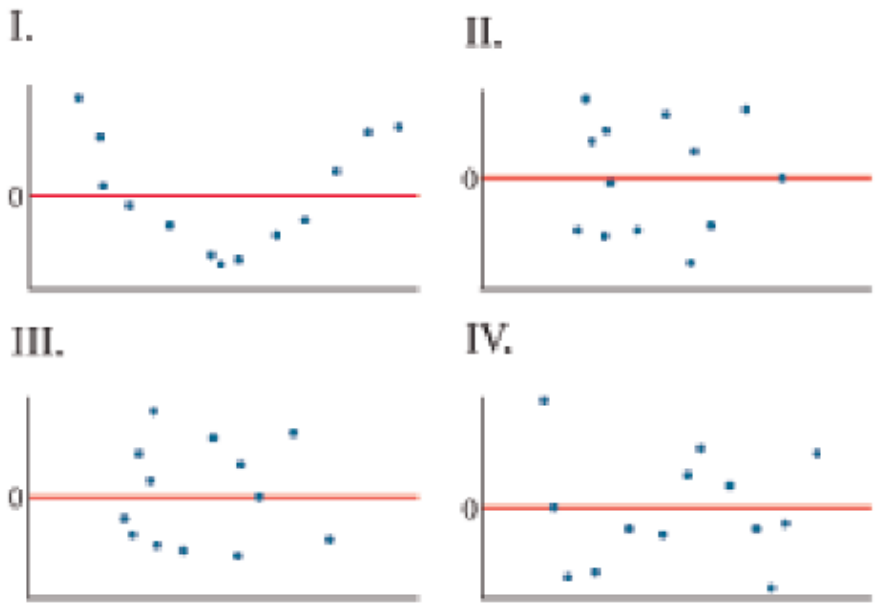
{Из кн. Watkins et al. 2008, стр. 177, рис. 3.84-3.85}

Решение

- A-I - нелинейная связь - нельзя;
- B-II - все в порядке, можно;
- C-III - все в порядке, можно;
- D-IV - синусоидальный паттерн в остатках, нарушено условие независимости или зависимость нелинейная - нельзя.



Display 3.84 Four scatterplots.



Display 3.85 Four residual plots.

Рис. из кн. Watkins et al. 2008, стр. 177, рис. 3.84-3.85

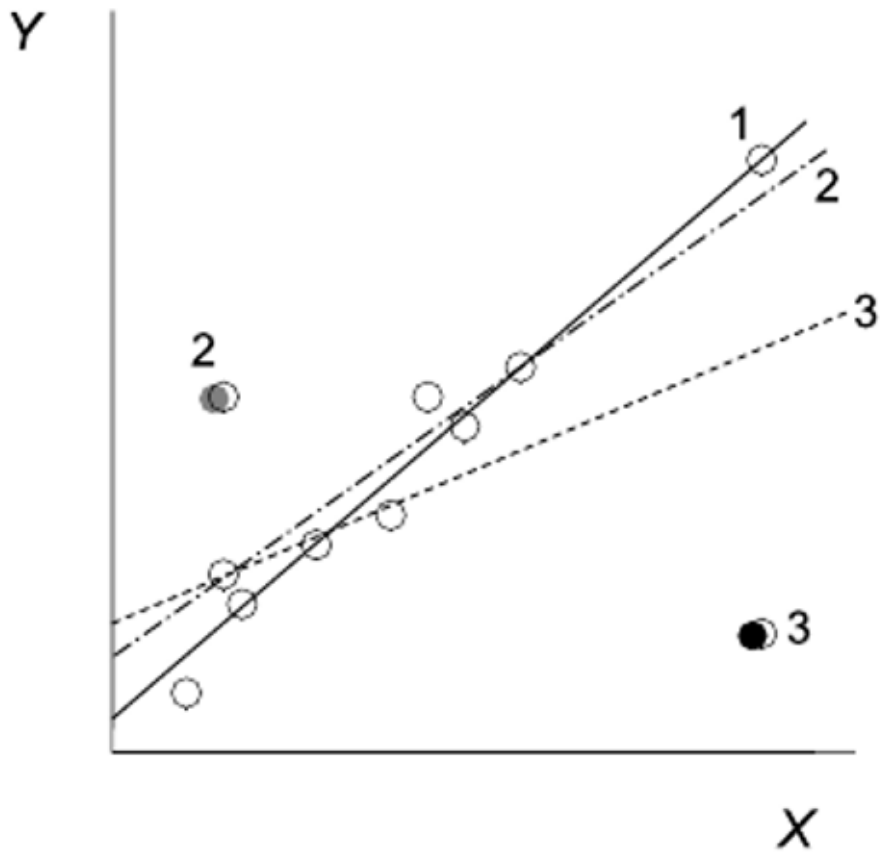
Какие наблюдения влияют на ход регрессии больше других?

Влиятельные наблюдения, выбросы, outliers

- большая абсолютная величина остатка
- близость к краям области определения (leverage - рычаг, сила; иногда называют hat)

На графике точки и линии регрессии построенные с их включением:

- 1 - не влияет на ход регрессии, т.к. лежит на прямой
- 2 - умеренно влияет (большой остаток, малая сила влияния)
- 3 - очень сильно влияет (большой остаток, большая сила влияния)

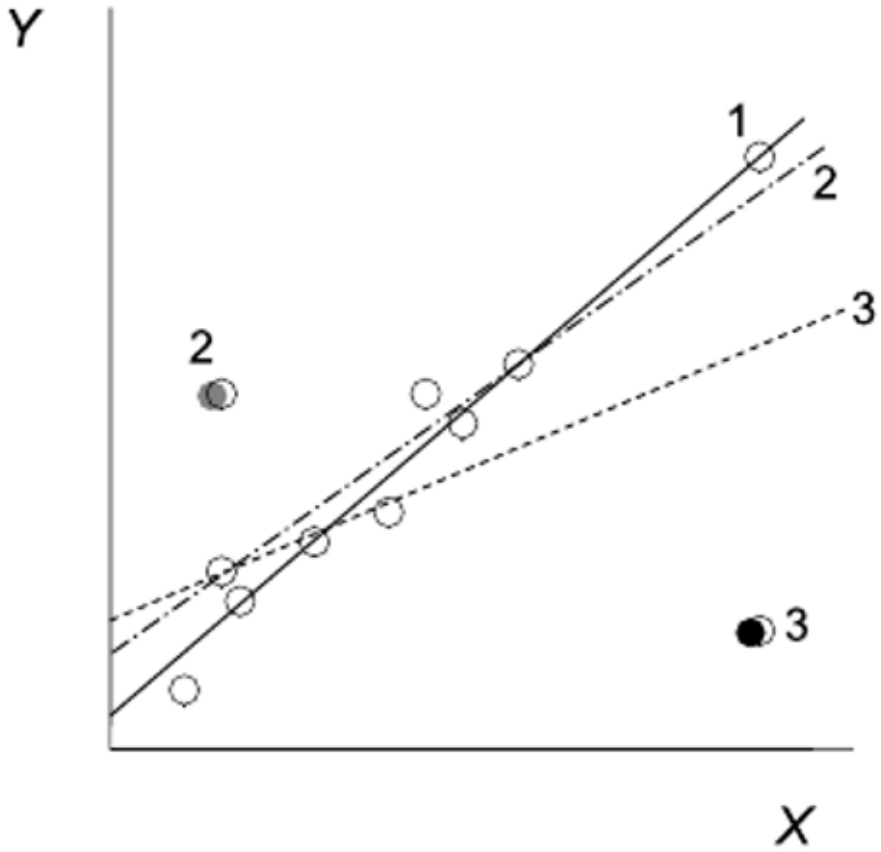


Из кн. Quinn, Keough, 2002, стр. 96, рис. 5.8

Как оценить влиятельность наблюдений?

Расстояние Кука (Cook's d, Cook, 1977)

- Учитывает одновременно величину остатка и близость к краям области определения (leverage)
- Условное пороговое значение: выброс, если $d \geq 4/(n-p)$, где n - объем выборки, p - число параметров модели. Иногда используют более мягкий порог $d \geq 1$

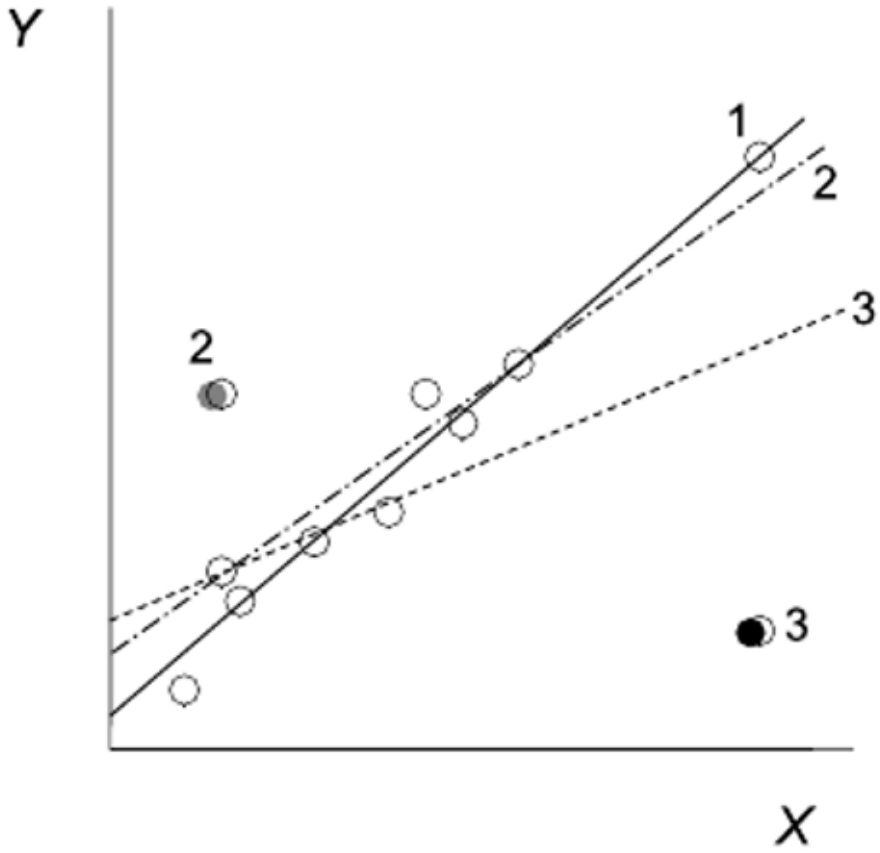


Из кн. Quinn, Keough, 2002, стр. 96, рис. 5.8

Как оценить влиятельность наблюдений?

Расстояние Кука (Cook's d, Cook, 1977)

- Учитывает одновременно величину остатка и близость к краям области определения (leverage)
- Условное пороговое значение: выброс, если $d \geq 4/(n-p)$, где n - объем выборки, p - число параметров модели. Иногда используют более мягкий порог $d \geq 1$

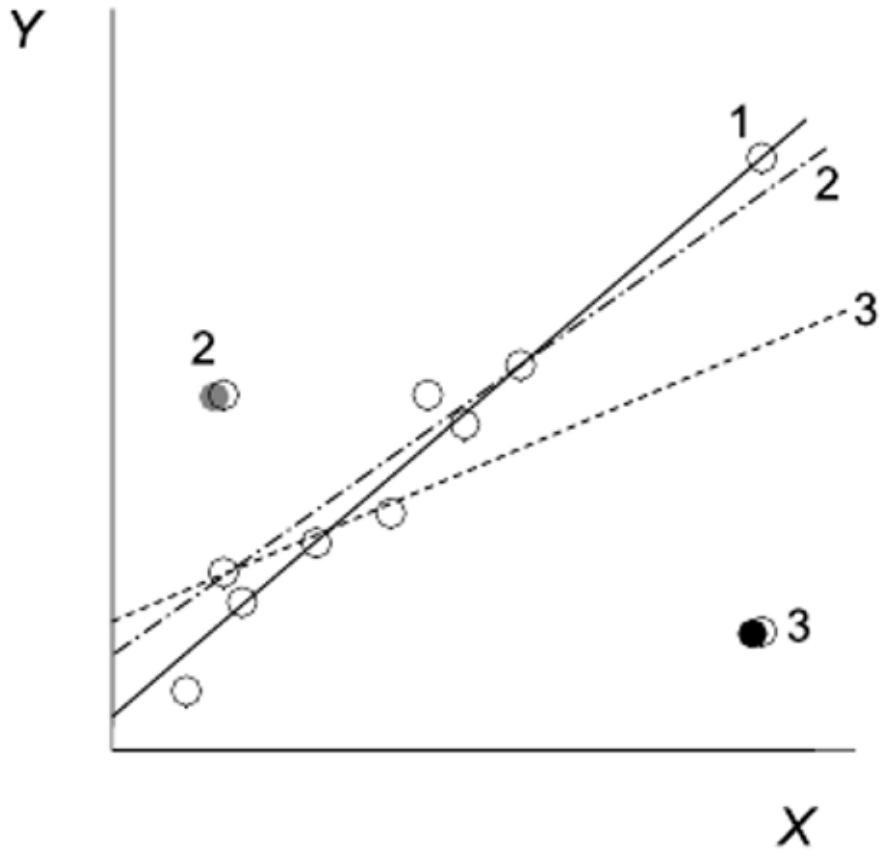


Из кн. Quinn, Keough, 2002, стр. 96, рис. 5.8

- Дж. Фокс советует не обращать внимания на пороговые значения (Fox, 1991)

Что делать с влиятельными точками и с выбросами?

- Проверить, не ошибка ли это. Если нет, не удалять - обсуждать!
- Проверить, что будет, если их исключить из модели



Из кн. Quinn, Keough, 2002, стр. 96, рис. 5.8

Коллинеарность предикторов

Коллинеарность

Коллинеарные предикторы коррелируют друг с другом, т.е. не являются взаимно независимыми

Последствия

- Модель неустойчива к изменению данных
- При добавлении или исключении наблюдений может меняться оценка и знак коэффициентов

Что делать с коллинеарностью?

- Удалить из модели избыточные предикторы
- Получить вместо скоррелированных предикторов один новый комбинированный при помощи метода главных компонент

Проверка на коллинеарность

Показатель инфляции для дисперсии

(коэффициент распространения дисперсии, Variance inflation factor, VIF)
VIF оценивает степень избыточности каждого из предикторов модели:

$$VIF = 1/(1 - R'^2)$$

Здесь в знаменателе используется R^2 регрессии данного предиктора от всех других предикторов в модели.
Хорошо, если $VIF < 10$ (по Marquardt, 1970), но лучше $VIF < 3$, а иногда и $VIF < 2$. Если больше — есть коллинеарность.
Предикторы с VIF больше порогового значения нужно последовательно удалить из модели (по-одному, проверяя, как изменился VIF после каждого этапа удаления).

Множественная линейная регрессия

Пример: реки штата Нью-Йорк

В 70-е годы в штате Нью Йорк обследовали 20 речных бассейнов (Haith, 1976), чтобы оценить качество воды. Как влияют особенности землепользования на среднюю концентрацию азота (мг/л) в воде? (Датасет river из пакета bstats, источник Chatterjee & Hadi, 2006)

20 рек в штате Нью Йорк

- `River` - название реки
- `Agr` - процент сельскохозяйственных земель
- `Forest` - процент земли, занятой лесом
- `Rsdntial` - процент земель, занятых поселениями
- `ComIndl` - процент земель, занятых коммерцией и промышленностью
- `Nitrogen` - средняя концентрация азота в воде, мг/л

Т.е. мы хотим подобрать модель вида:

$$Nitrogen_i = b_0 + b_1 Agr_i + b_2 Forest_i + b_3 Rsdntial_i + b_4 ComIndl_i + e_i$$

Читаем данные из файла одним из способов

Чтение из xlsx

```
library(readxl)
river <- read_excel(path = "data/river.xlsx", sheet = "river-data")
```

Чтение из csv

```
river <- read.table("data/river.csv", header = TRUE, sep = "\t")
```

Все ли правильно открылось?

```
str(river)      # Структура данных
```

```
# 'data.frame': 20 obs. of  6 variables:
#  $ River   : chr  "Olean" "Cassadaga" "Oatka" "Neversink" ...
#  $ Agr     : int  26 29 54 2 3 19 16 40 28 26 ...
#  $ Forest  : int  63 57 26 84 27 61 60 43 62 60 ...
#  $ Rsdntial: num  1.2 0.7 1.8 1.9 29.4 3.4 5.6 1.3 1.1 0.9 ...
#  $ ComIndl : num  0.29 0.09 0.58 1.98 3.11 0.56 1.11 0.24 0.15 0.23 ...
#  $ Nitrogen: num  1.1 1.01 1.9 1 1.99 1.42 2.04 1.65 1.01 1.21 ...
```

```
head(river)     # Первые несколько строк файла
```

```
#         River Agr Forest Rsdntial ComIndl Nitrogen
# 1       Olean  26     63      1.2    0.29     1.10
# 2  Cassadaga  29     57      0.7    0.09     1.01
# 3       Oatka  54     26      1.8    0.58     1.90
# 4  Neversink   2     84      1.9    1.98     1.00
# 5 Hackensack   3     27     29.4    3.11     1.99
# 6  Wappinger  19     61      3.4    0.56     1.42
```

Знакомимся с данными

Есть ли пропущенные значения?

```
colSums(is.na(river))
```

```
#    River      Agr   Forest Rsdntial  ComIndl Nitrogen
#        0        0        0        0        0        0
```

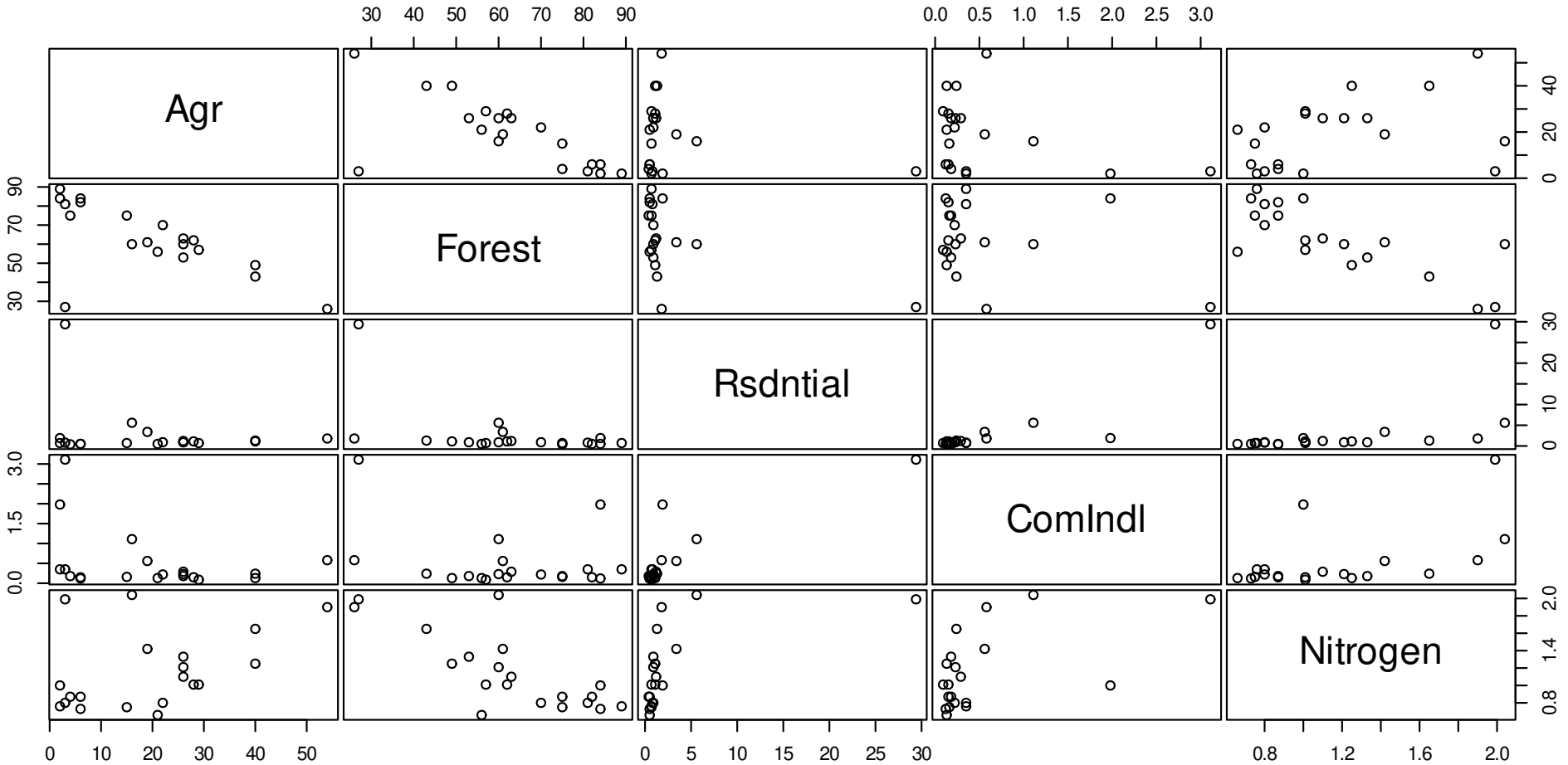
Каков объем выборки?

```
nrow(river)
```

```
# [1] 20
```

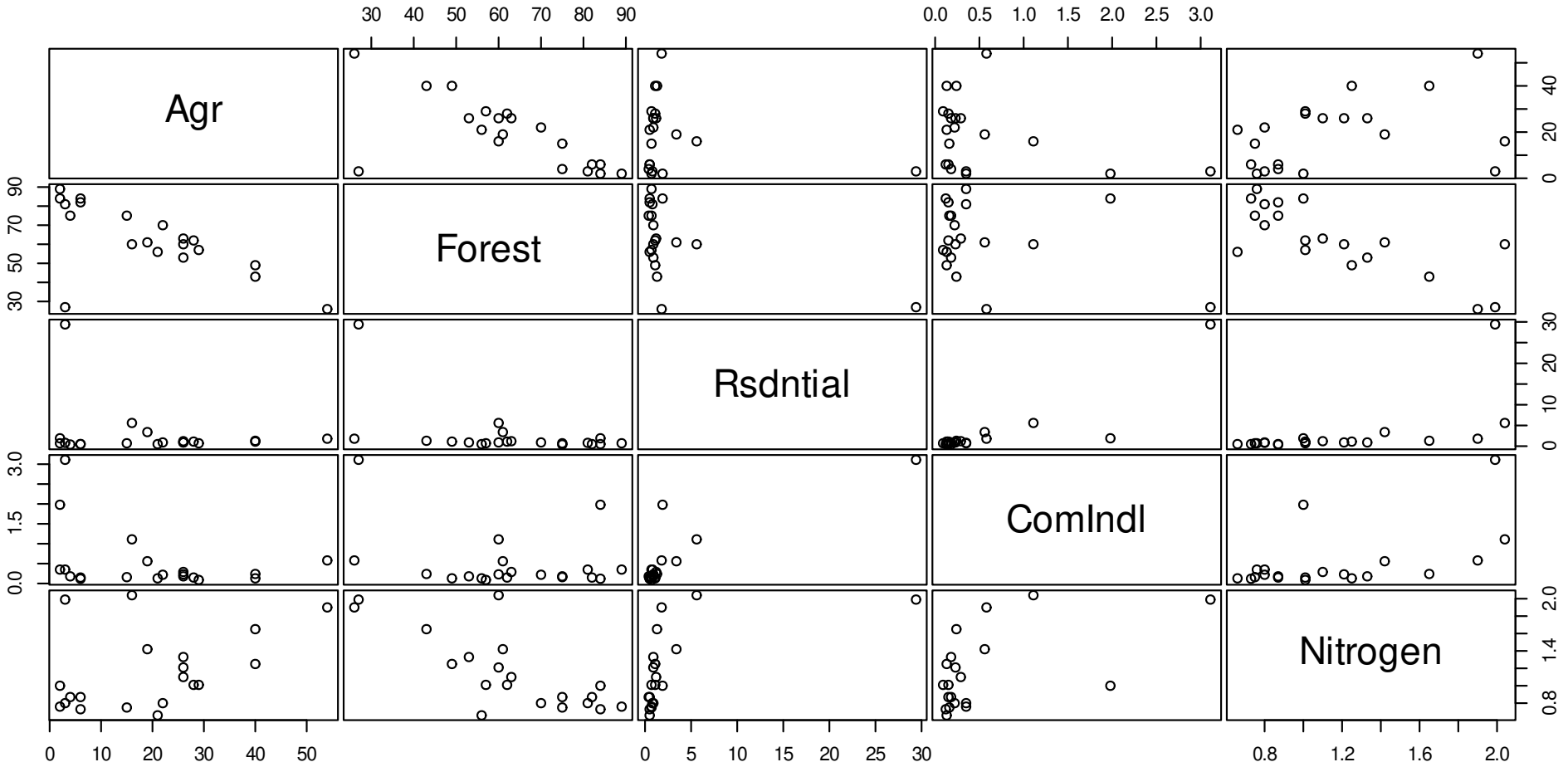
Парные графики для всех числовых переменных

```
pairs(river[, -1])
```



Парные графики для всех числовых переменных

```
pairs(river[, -1])
```



Выброс — сильно отскакивающее значение.

Похоже, что в этих данных есть выброс в столбце `Rsdntial`.

Варианты действий с выбросами

- удалить это наблюдение, т.к. рек с таким большим уровнем застройки территории больше нет в датасете.
- трансформировать `Rsdntial` (извлечь логарифм), чтобы “растянуть” начало шкалы и “сплющить” ее конец.

Варианты действий с выбросами

- удалить это наблюдение, т.к. рек с таким большим уровнем застройки территории больше нет в датасете.
- трансформировать `Rsdntial` (извлечь логарифм), чтобы “растянуть” начало шкалы и “сплющить” ее конец.

Мы пока продолжим, чтобы посмотреть как будет выглядеть это значение при анализе остатков.

Задача

1. Подберите модель множественной линейной регрессии, чтобы описать, как зависит концентрация азота от особенностей землепользования.

$$Nitrogen_i = b_0 + b_1 Agr_i + b_2 Forest_i + b_3 Rsdntial_i + b_4 ComIndl_i + e_i$$

1. Запишите уравнение этой линейной модели с коэффициентами.

Решение

```
river_lm1 <- lm(Nitrogen ~ Agr + Forest + Rsdntial + ComIndl, data = river)
# summary(river_lm1)
```

Коэффициенты модели:

```
coef(river_lm1)
```

```
#  (Intercept)          Agr       Forest     Rsdntial      ComIndl
#  1.722213529  0.005809126 -0.012967887 -0.007226768  0.305027765
```

Уравнение регрессии:

$$Nitrogen_i = 1.722 + 0.006Agr_i - 0.013Forest_i - 0.007Rsdntial_i + 0.305ComIndl_i$$

Решение

```
river_lm1 <- lm(Nitrogen ~ Agr + Forest + Rsdntial + ComIndl, data = river)
# summary(river_lm1)
```

Коэффициенты модели:

```
coef(river_lm1)
```

```
#  (Intercept)          Agr       Forest     Rsdntial      ComIndl
#  1.722213529  0.005809126 -0.012967887 -0.007226768  0.305027765
```

Уравнение регрессии:

$$Nitrogen_i = 1.722 + 0.006 Agr_i - 0.013 Forest_i - 0.007 Rsdntial_i + 0.305 ComIndl_i$$

Более формальная запись
(и та и другая запись требует расшифровки обозначений):

$$Y_i = 1.722 + 0.006 X_{1i} - 0.013 X_{2i} - 0.007 X_{3i} + 0.305 X_{4i}$$

Решение

```
river_lm1 <- lm(Nitrogen ~ Agr + Forest + Rsdntial + ComIndl, data = river)
# summary(river_lm1)
```

Коэффициенты модели:

```
coef(river_lm1)
```

```
#  (Intercept)          Agr       Forest     Rsdntial      ComIndl
#  1.722213529  0.005809126 -0.012967887 -0.007226768  0.305027765
```

Уравнение регрессии:

$$Nitrogen_i = 1.722 + 0.006Agr_i - 0.013Forest_i - 0.007Rsdntial_i + 0.305ComIndl_i$$

Более формальная запись
(и та и другая запись требует расшифровки обозначений):

$$Y_i = 1.722 + 0.006X_{1i} - 0.013X_{2i} - 0.007X_{3i} + 0.305X_{4i}$$

Важно! Прежде чем интерпретировать результаты нужно обязательно проверить, выполняются ли условия применимости линейной регрессии.

Проверка условий применимости линейной регрессии

Как проверить условия применимости?

1. Вычисляем VIF — коллинеарность предикторов (для множественной регрессии с этого всегда нужно начинать)
2. График расстояния Кука для разных наблюдений — проверка на наличие выбросов
3. График остатков от предсказанных значений — величина остатков, влиятельность наблюдений, отсутствие паттернов, гомогенность дисперсий.
4. График квантилей остатков — распределение остатков

1. Проверка на коллинеарность предикторов

```
library(car)
vif(river_lm1) # variance inflation factors

#       Agr    Forest  Rsdntial   ComIndl
# 13.277430 16.727089 12.682219  4.144766
```

1. Проверка на коллинеарность предикторов

```
library(car)
vif(river_lm1) # variance inflation factors
```

```
#       Agr    Forest  Rsdntial   ComIndl
# 13.277430 16.727089 12.682219  4.144766
```

Самое большое значение vif для предиктора `Forest`. Удалим его из модели и пересчитаем vif.

1. Проверка на коллинеарность предикторов

```
library(car)
vif(river_lm1) # variance inflation factors
```

```
#       Agr    Forest  Rsdntial   ComIndl
# 13.277430 16.727089 12.682219  4.144766
```

Самое большое значение vif для предиктора `Forest`. Удалим его из модели и пересчитаем vif.

```
river_lm2 <- update(river_lm1, .~.-Forest)
vif(river_lm2) # variance inflation factors
```

```
#      Agr Rsdntial  ComIndl
# 1.151355 3.868585 4.137905
```

1. Проверка на коллинеарность предикторов

```
library(car)
vif(river_lm1) # variance inflation factors
```

```
#       Agr    Forest  Rsdntial   ComIndl
# 13.277430 16.727089 12.682219  4.144766
```

Самое большое значение vif для предиктора `Forest`. Удалим его из модели и пересчитаем vif.

```
river_lm2 <- update(river_lm1, .~.-Forest)
vif(river_lm2) # variance inflation factors
```

```
#      Agr Rsdntial  ComIndl
# 1.151355 3.868585 4.137905
```

Самое большое значение vif для предиктора `ComIndl`. Аналогично.

1. Проверка на коллинеарность предикторов, продолжение

Удаляем ComIndl

```
river_lm3 <- update(river_lm2, .~.-ComIndl)
vif(river_lm3) # variance inflation factors

#      Agr Rsdntial
# 1.062021 1.062021
```

1. Проверка на коллинеарность предикторов, продолжение

Удаляем ComIndl

```
river_lm3 <- update(river_lm2, .~.-ComIndl)
vif(river_lm3) # variance inflation factors
```

```
#      Agr Rsdntial
# 1.062021 1.062021
```

Все в порядке. Судя по значениям vif после пошагового удаления всех коллинеарных предикторов оставшиеся предикторы независимы.

Теперь наша модель river_lm3 выглядит так:

$$Nitrogen_i = b_0 + b_1 Agr_i + b_3 Rsdntial_i + e_i$$

Для анализа остатков создадим диангостический датафрейм

```
library(ggplot2) # там есть функция fortify()
river_diag3 <- fortify(river_lm3)
# вот, что записано в диагностическом датафрейме
head(river_diag3, 2)
```

```
#   Nitrogen Agr Rsdntial       .hat    .sigma     .cooksd  .fitted
# 1     1.10  26       1.2 0.06148896 0.2891313 0.002728688 1.196408
# 2     1.01  29       0.7 0.07373962 0.2848682 0.016423620 1.223162
#         .resid  .stdresid
# 1 -0.09640751 -0.3534750
# 2 -0.21316218 -0.7867036
```

Для анализа остатков создадим диангостический датафрейм

```
library(ggplot2) # там есть функция fortify()
river_diag3 <- fortify(river_lm3)
# вот, что записано в диагностическом датафрейме
head(river_diag3, 2)
```

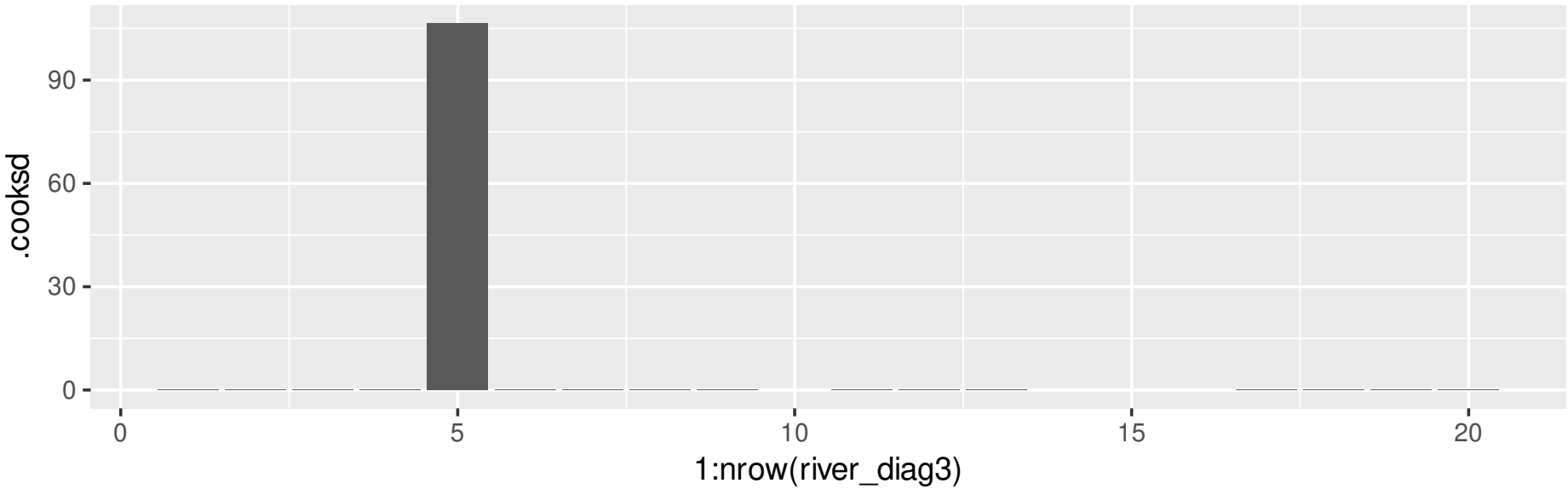
```
#   Nitrogen Agr Rsdntial       .hat    .sigma     .cooksd  .fitted
# 1     1.10  26       1.2 0.06148896 0.2891313 0.002728688 1.196408
# 2     1.01  29       0.7 0.07373962 0.2848682 0.016423620 1.223162
#         .resid  .stdresid
# 1 -0.09640751 -0.3534750
# 2 -0.21316218 -0.7867036
```

- .cooksd - расстояние Кука
- .fitted - предсказанные значения
- .resid - остатки
- .stdresid - стандартизованные остатки

2. Проверка на наличие влиятельных наблюдений

График расстояния Кука для всех наблюдений

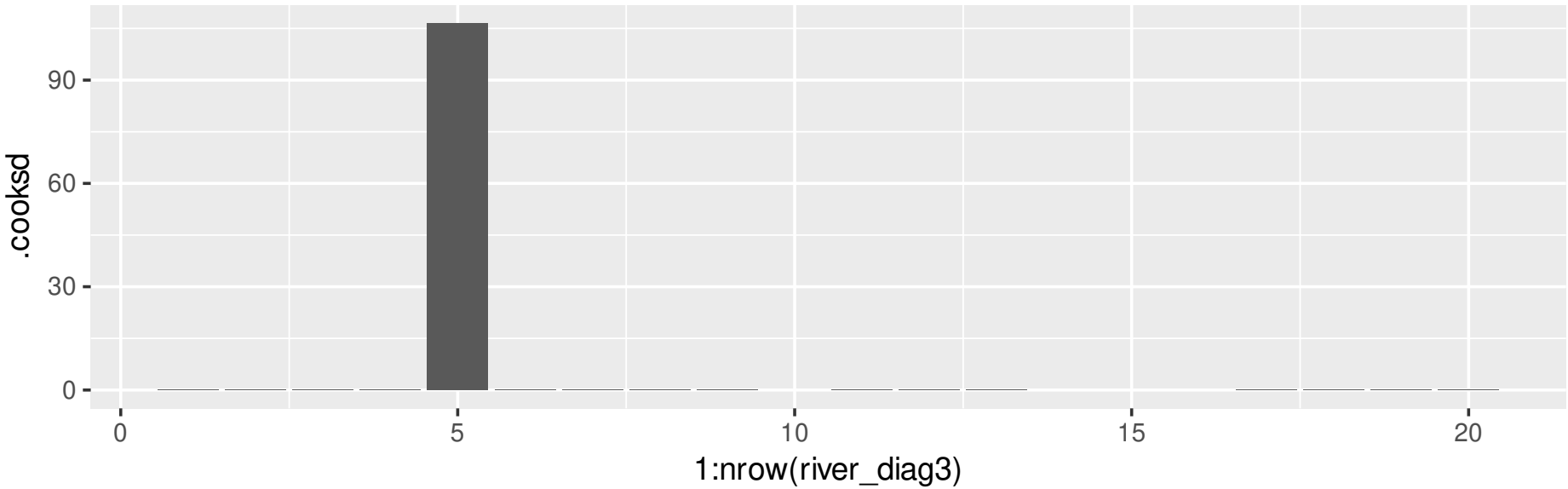
```
ggplot(data = river_diag3, aes(x = 1:nrow(river_diag3), y = .cooksd)) +
  geom_bar(stat = "identity")
```



2. Проверка на наличие влиятельных наблюдений

График расстояния Кука для всех наблюдений

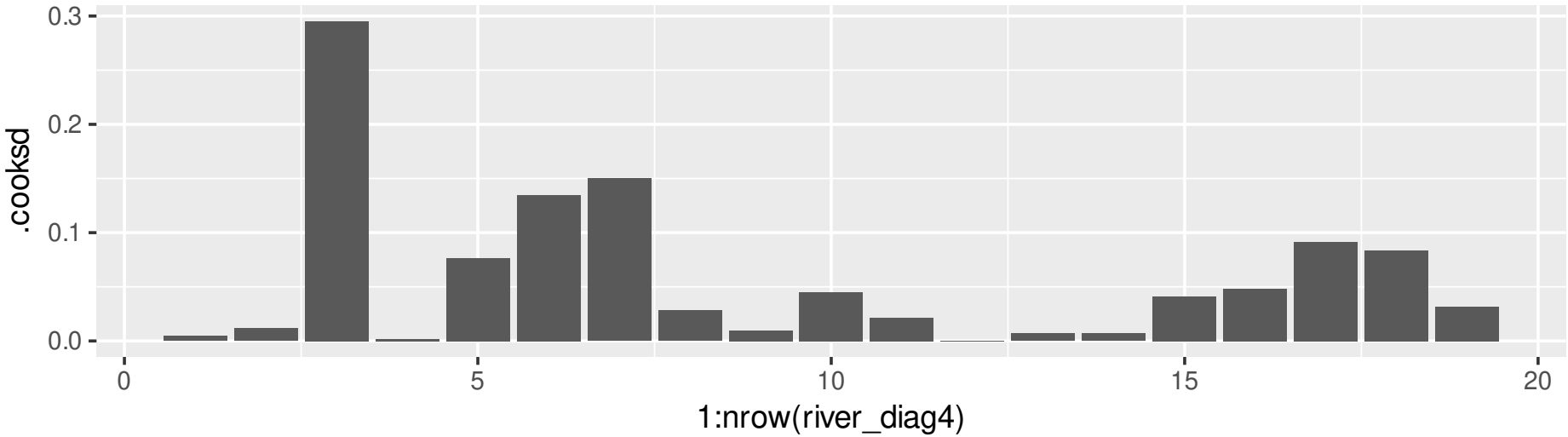
```
ggplot(data = river_diag3, aes(x = 1:nrow(river_diag3), y = .cooksd)) +
  geom_bar(stat = "identity")
```



Вот оно, то самое отскакивающее значение `Rsdntial` больше 25% застройки. Сейчас оно слишком сильно влияет на ход регрессии. Давайте попробуем его удалить и переподобрать модель.

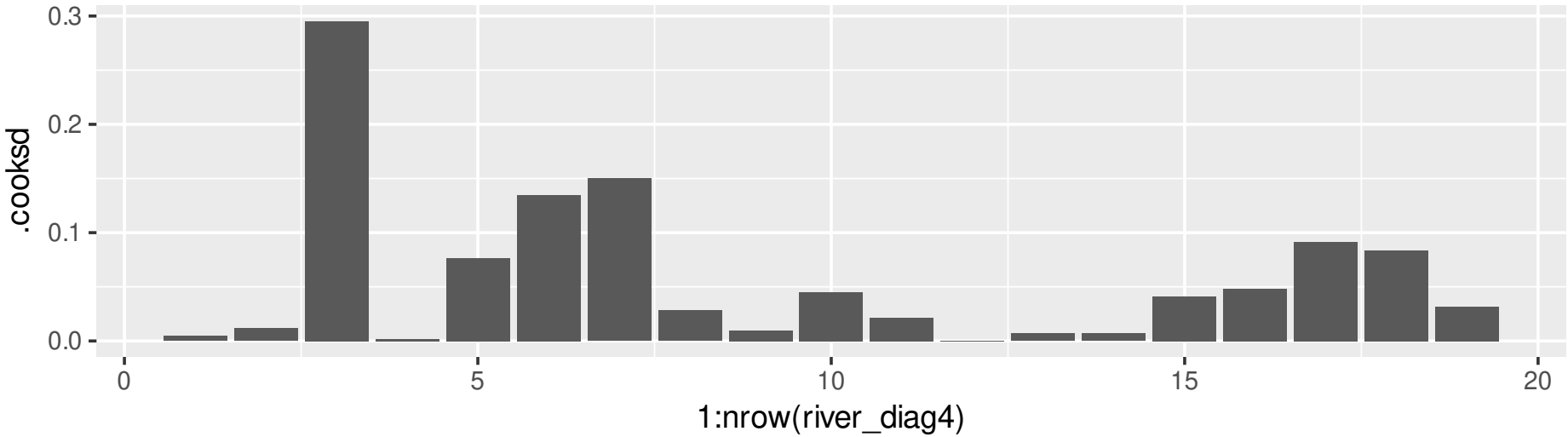
Новая модель, на очищенных данных

```
# данные без выброса
river_subset <- river[river$Rsdntial < 25, ]
# новая модель
river_lm4 <- lm(Nitrogen ~ Agr + Rsdntial, data = river_subset)
# диагностический датафрейм
river_diag4 <- fortify(river_lm4)
# график расстояния Кука
ggplot(data = river_diag4, aes(x = 1:nrow(river_diag4), y = .cooksd)) +
  geom_bar(stat = "identity")
```



Новая модель, на очищенных данных

```
# данные без выброса
river_subset <- river[river$Rsdntial < 25, ]
# новая модель
river_lm4 <- lm(Nitrogen ~ Agr + Rsdntial, data = river_subset)
# диагностический датафрейм
river_diag4 <- fortify(river_lm4)
# график расстояния Кука
ggplot(data = river_diag4, aes(x = 1:nrow(river_diag4), y = .cooksd)) +
  geom_bar(stat = "identity")
```

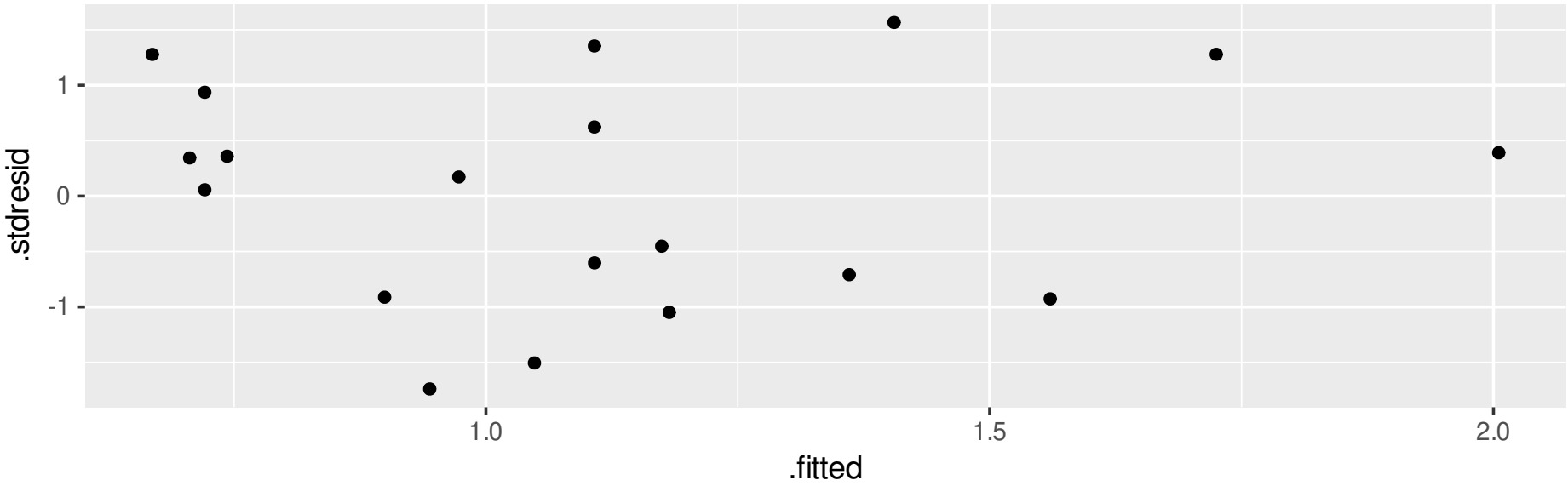


Отлично, больше нет чрезмерно влиятельных наблюдений с $d > 1$.

Задача

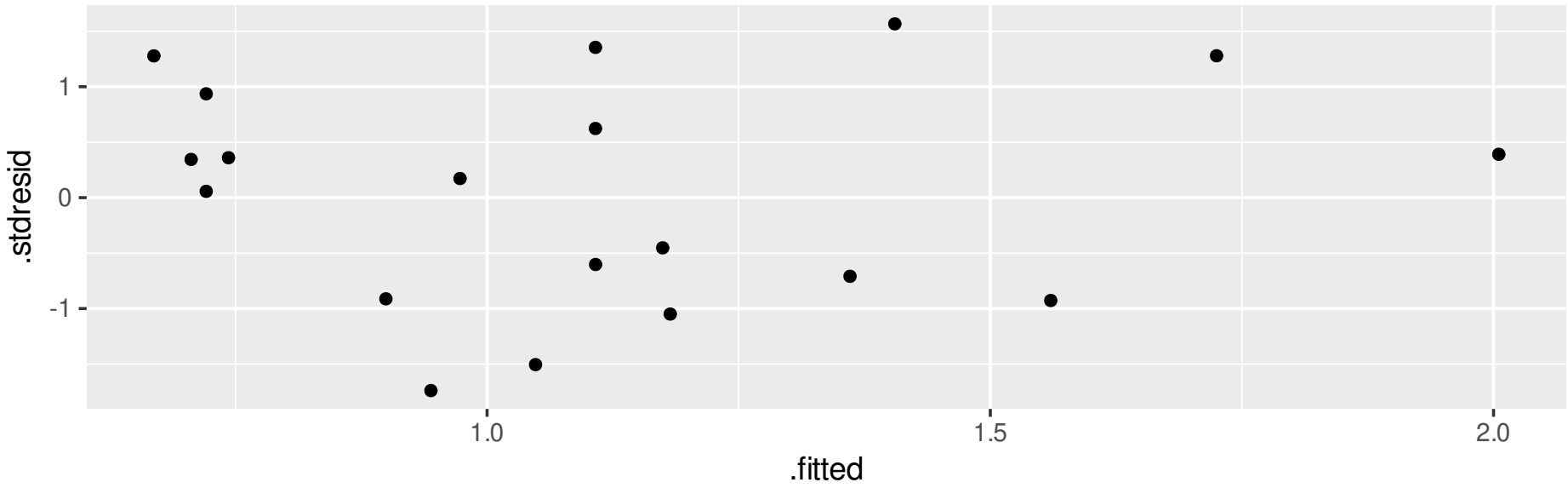
Постройте график зависимости стандартизованных остатков от предсказанных значений

Используйте данные из `river_diag4`



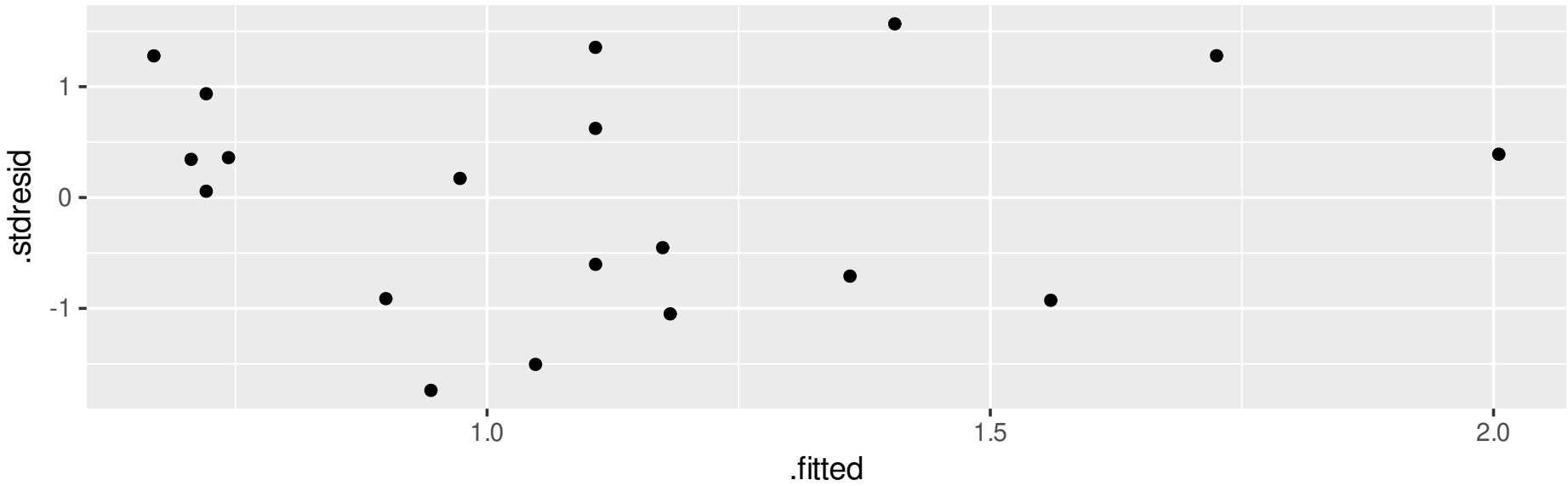
3. График зависимости стандартизованных остатков от предсказанных значений

```
gg_resid <- ggplot(data = river_diag4, aes(x = .fitted, y = .stdresid)) +
  geom_point()
gg_resid
```



3. График зависимости стандартизованных остатков от предсказанных значений

```
gg_resid <- ggplot(data = river_diag4, aes(x = .fitted, y = .stdresid)) +
  geom_point()
gg_resid
```



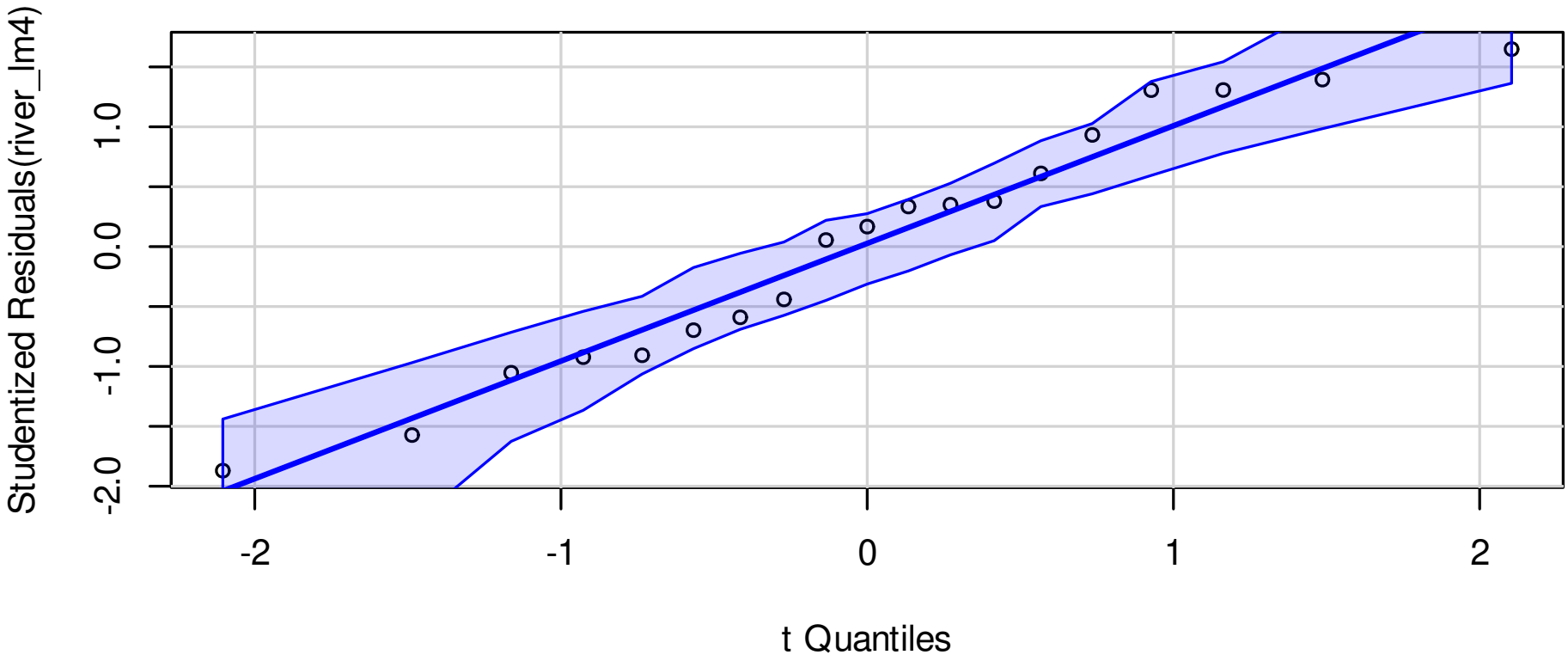
Большая часть стандартизованных остатков в пределах двух стандартных отклонений. В правой части графика мало наблюдений (с большими предсказанными значениями концентрации азота) - с этим ничего не поделаешь… Тренда среди остатков нет.

4. Квантильный график стандартизованных остатков

Используется, чтобы оценить форму распределения. По оси X — квантили теоретического распределения, по оси Y — квантили остатков модели.

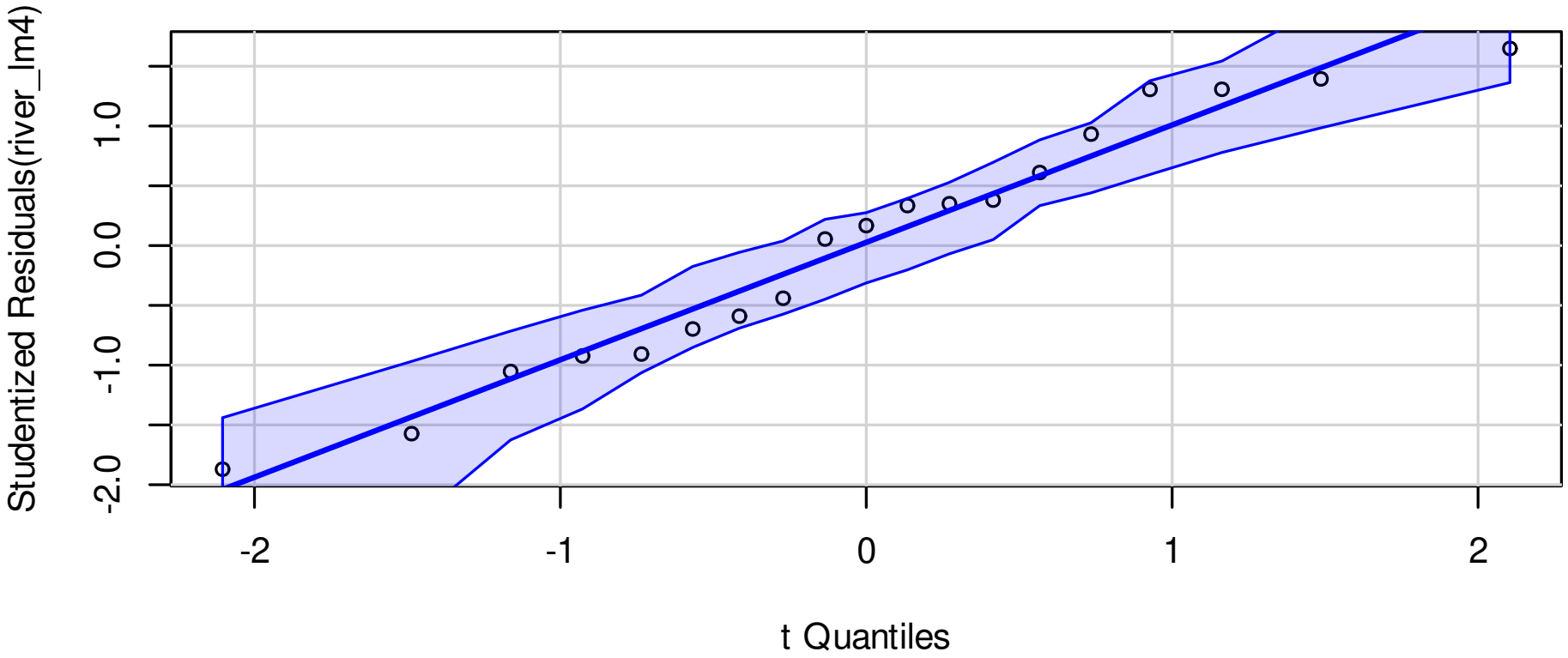
Если точки лежат на одной прямой — все в порядке.

```
library(car)
qqPlot(river_lm4, id = FALSE) # из пакета car
```



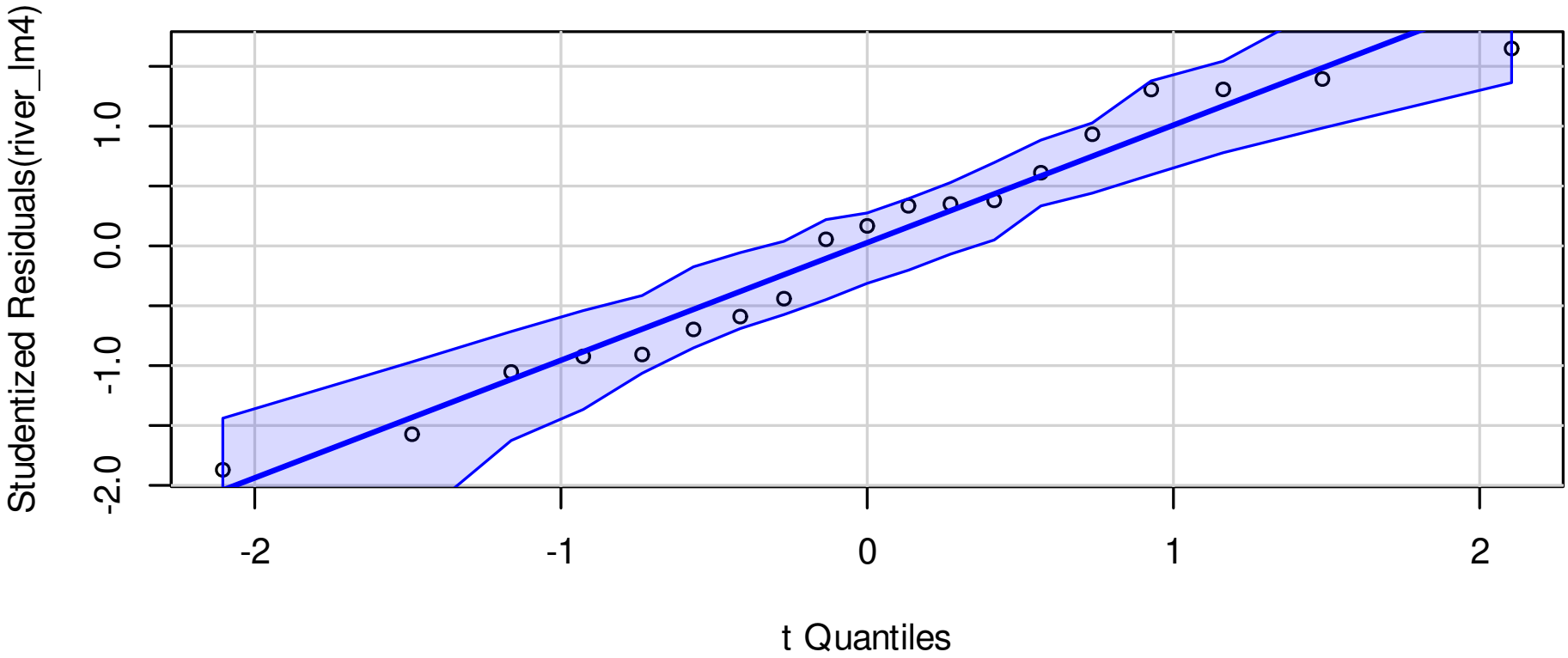
Интерпретируем квантильный график

Какие выводы можно сделать по квантильному графику?



Интерпретируем квантильный график

Какие выводы можно сделать по квантильному графику?



Отклонений от нормального распределения нет

Внимание!

Только если все условия выполняются, можно приступить к интерпретации результатов тестов значимости коэффициентов регрессии.

Интерпретация коэффициентов регрессии

```
coef(river_lm4)

# (Intercept)         Agr    Rsdntial
#  0.52032778  0.01488396  0.22262844
```

Интерпретация коэффициентов регрессии

```
coef(river_lm4)
```

```
# (Intercept)         Agr    Rsdntial
#  0.52032778  0.01488396  0.22262844
```

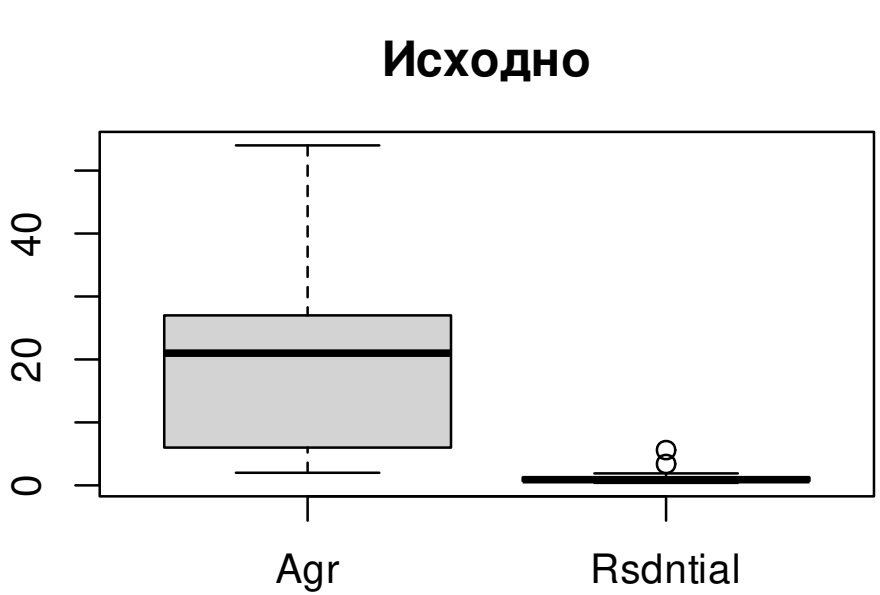
Обычные коэффициенты

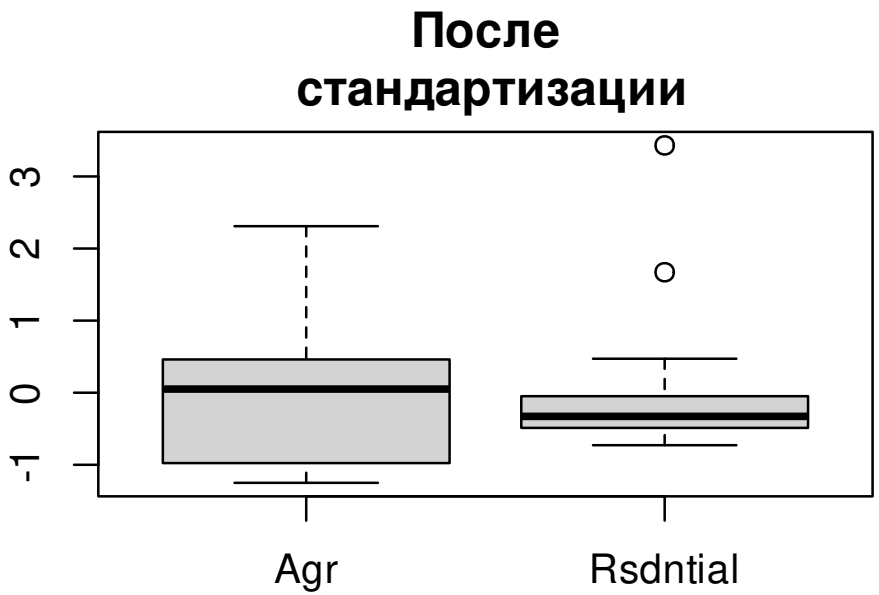
- Величина обычных коэффициентов зависит от единиц измерения
- b_0 — Отрезок (Intercept), отсекаемый регрессионной прямой на оси y. Значение зависимой переменной Y, если предикторы равны нулю.
- Коэффициенты при предикторах показывают, на сколько изменяется Y, когда данный предиктор меняется на единицу, при условии, что остальные предикторы не меняют своих значений.

Если предикторы измерены в разных единицах

Обычные коэффициенты отражают силу влияния предикторов, но не учитывают масштаб их варьирования.

Если стандартизовать переменные ($x_{std} = \frac{x_i - \bar{x}}{SD_x}$), то масштабы их изменений выровняются: они будут измеряться в одних и тех же единицах — в стандартных отклонениях.

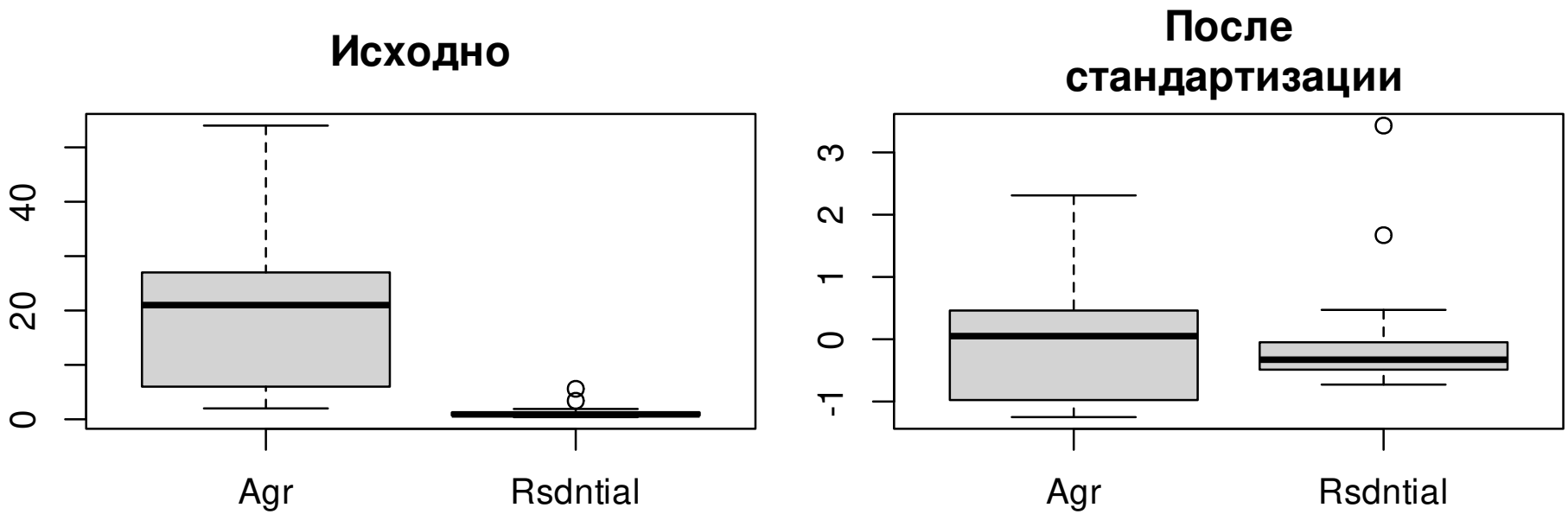




Если предикторы измерены в разных единицах

Обычные коэффициенты отражают силу влияния предикторов, но не учитывают масштаб их варьирования.

Если стандартизовать переменные ($x_{std} = \frac{x_i - \bar{x}}{SD_x}$), то масштабы их изменений выровняются: они будут измеряться в одних и тех же единицах — в стандартных отклонениях.



Если подобрать по линейную регрессию по стандартизованным значениям предикторов, то можно будет сравнивать силу их влияния с учетом масштаба их варьирования.

Для сравнения влияния разных предикторов — стандартизованные коэффициенты

```
scaled_river_lm4 <- lm(Nitrogen ~ scale(Agr) + scale(Rsdntial), data = river_subset)
coef(scaled_river_lm4)

#     (Intercept)      scale(Agr) scale(Rsdntial)
#       1.1136842       0.2173852       0.2783493
```

Для сравнения влияния разных предикторов — стандартизованные коэффициенты

```
scaled_river_lm4 <- lm(Nitrogen ~ scale(Agr) + scale(Rsdntial), data = river_subset)
coef(scaled_river_lm4)

#     (Intercept)      scale(Agr) scale(Rsdntial)
#       1.1136842       0.2173852       0.2783493
```

Стандартизованные коэффициенты

- Стандартизованные коэффициенты измерены в стандартных отклонениях. Их можно сравнивать друг с другом, поскольку они дают относительную оценку влияния фактора.
- b_0 — Отрезок (Intercept), отсекаемый регрессионной прямой на оси y. Значение зависимой переменной Y, если предикторы равны нулю. Для стандартизованных величин среднее значение равно нулю, поэтому b_0 — это значение зависимой переменной при средних значениях всех предикторов.
- Коэффициенты при предикторах показывают, на сколько изменяется Y, когда предиктор меняется на одно стандартное отклонение, при условии, что остальные предикторы не меняют своих значений. Это относительная оценка влияния фактора.

Задача

Определите по значениям стандартизованных коэффициентов, какие предикторы сильнее всего влияют на концентрацию азота в воде?

```
summary(scaled_river_lm4)
```

```
#
# Call:
# lm(formula = Nitrogen ~ scale(Agr) + scale(Rsdntial), data = river_subset)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -0.28420 -0.12532  0.02691  0.12569  0.24490
#
# Coefficients:
#                 Estimate Std. Error t value Pr(>|t|)
# (Intercept)      1.11368    0.03901  28.549 3.74e-15 ***
# scale(Agr)       0.21739    0.04022   5.404 5.85e-05 ***
# scale(Rsdntial)  0.27835    0.04022   6.920 3.45e-06 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.17 on 16 degrees of freedom
# Multiple R-squared:  0.8401,  Adjusted R-squared:  0.8201
# F-statistic: 42.03 on 2 and 16 DF,  p-value: 0.0000004274
```

Задача

Определите по значениям стандартизованных коэффициентов, какие предикторы сильнее всего влияют на концентрацию азота в воде?

```
summary(scaled_river_lm4)
```

```
#
# Call:
# lm(formula = Nitrogen ~ scale(Agr) + scale(Rsdntial), data = river_subset)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -0.28420 -0.12532  0.02691  0.12569  0.24490
#
# Coefficients:
#                 Estimate Std. Error t value Pr(>|t|)
# (Intercept)      1.11368    0.03901  28.549 3.74e-15 ***
# scale(Agr)       0.21739    0.04022   5.404 5.85e-05 ***
# scale(Rsdntial)  0.27835    0.04022   6.920 3.45e-06 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.17 on 16 degrees of freedom
# Multiple R-squared:  0.8401,  Adjusted R-squared:  0.8201
# F-statistic: 42.03 on 2 and 16 DF,  p-value: 0.0000004274
```

Влияние обоих предикторов сопоставимо по силе, но сильнее всего все же влияет процент застройки `Rsdntial`.

Оценка качества подгонки модели

```
summary(river_lm1)$adj.r.squared
```

```
# [1] 0.6319037
```

Обычный R^2 — доля объясненной изменчивости

$$R^2 = \frac{SS_r}{SS_t} = 1 - \frac{SS_e}{SS_t}$$

Не используйте обычный R^2 для множественной регрессии!

Оценка качества подгонки модели

```
summary(river_lm1)$adj.r.squared
```

```
# [1] 0.6319037
```

Обычный R^2 — доля объясненной изменчивости

$$R^2 = \frac{SS_r}{SS_t} = 1 - \frac{SS_e}{SS_t}$$

Не используйте обычный R^2 для множественной регрессии!

R^2_{adj} — скорректированный R^2

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

где $n - p = df_e$, $n - 1 = df_t$
R^2_{adj} учитывает число переменных в модели, вводится штраф за каждый новый параметр.
Используйте R^2_{adj} для сравнения моделей с разным числом параметров.

Описание результатов

Для описания зависимости концентрации азота в речной воде от особенностей землепользования была подобрана линейная модель: $Nitrogen_i = 0.52 + 0.01Agr_i + 0.22Rsdntial_i$, где Agr — процент сельскохозяйственных земель, $Rsdntial$ — процент земель, занятых поселениями. Эта модель объяснила 63.2% общей изменчивости концентрации азота в речной воде. С увеличением процента застройки и процента сельскохозяйственных земель в бассейнах рек концентрация азота статистически значимо увеличивалась (Табл. 1).

Table 1: Коэффициенты линейной регрессии, описывающей зависимость средней концентрации азота в воде (мг/л) от характеристик землепользования: Agr — процент сельскохозяйственных земель, Rsdntial — процент земель, занятых поселениями. t — значение t-критерия, P — уровень значимости.

	Оценка	Ст.ошибка	t	P
Отрезок	0.52	0.078	6.71	< 0.01
Agr	0.01	0.003	5.40	< 0.01
Rsdntial	0.22	0.032	6.92	< 0.01

Что еще можно сделать?

Можно было бы нарисовать график предсказаний модели. Например, это мог бы быть график зависимости предсказанной концентрации азота от площади застройки, если процент сельскохозяйственных земель зафиксирован на среднем уровне. Однако в этом курсе мы не будем разбирать, как можно построить такой график.

Take-home messages

- Для сравнения влияния разных предикторов можно использовать бета-коэффициенты
- Условия применимости линейной регрессии должны выполняться, чтобы можно было тестировать гипотезы
  1. Независимость
  2. Линейность
  3. Нормальное распределение
  4. Гомогенность дисперсий
  5. Отсутствие коллинеарности предикторов (для множественной регрессии)

Дополнительные ресурсы

- 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