

# **Регрессионный анализ, часть 2**

## **Математические методы в зоологии с использованием R**

Марина Варфоломеева

- 1 Множественная линейная регрессия**
- 2 Условия применимости линейной регрессии**
- 3 Проверка условий применимости линейной регрессии**

## **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

# Множественная линейная регрессия

## Пример: птицы Австралии

Зависит ли обилие птиц в лесах Австралии от характеристик леса? (Loyn, 1987, пример из кн. Quinn, Keough, 2002)

56 лесных участков в юго-восточной Виктории, Австралия

- `l10area` - Площадь леса, га
- `l10dist` - Расстояние до ближайшего леса, км (логарифм)
- `l10ldist` - Расстояние до ближайшего леса большего размера, км (логарифм)
- `yr.isol` - Год начала изоляции
- `abund` - Обилие птиц

## Скачиваем данные 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 в xlsx, либо в текстовом формате
```

```
if (!file.exists("data/loyn.xlsx")) {
```

```
  download(
```

```
    url = "https://varmara.github.io/mathmethr/data/loyn.xlsx",
```

```
    destfile = "data/loyn.xlsx")
```

```
}
```

```
if (!file.exists("data/loyn.csv")) {
```

```
  download(
```

```
    url = "https://varmara.github.io/mathmethr/data/loyn.xls",
```

```
    destfile = "data/loyn.csv")
```

```
}
```

# Читаем данные из файла одним из способов

## Чтение из xlsx

```
library(readxl)
bird <- read_excel(path = "data/loyn.xlsx", sheet = 1)
```

## Чтение из csv

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

## Все ли правильно открылось?

```
str(bird) # Структура данных
```

```
# 'data.frame': 56 obs. of 21 variables:
# $ abund      : num  5.3 2 1.5 17.1 13.8 14.1 3.8 2.2 3.3 3 ...
# $ area       : num  0.1 0.5 0.5 1 1 1 1 1 1 1 ...
# $ yr.isol    : int   1968 1920 1900 1966 1918 1965 1955 1920 1965 1900 ...
# $ dist       : int   39 234 104 66 246 234 467 284 156 311 ...
# $ ldist      : int   39 234 311 66 246 285 467 1829 156 571 ...
# $ graze      : int    2 5 5 3 5 3 5 5 4 5 ...
# $ alt        : int   160 60 140 160 140 130 90 60 130 130 ...
# $ l10dist    : num    1.59 2.37 2.02 1.82 2.39 ...
# $ l10ldist   : num    1.59 2.37 2.49 1.82 2.39 ...
# $ l10area    : num    -1 -0.301 -0.301 0 0 ...
# $ cyr.isol   : num    18.2 -29.8 -49.8 16.2 -31.8 ...
# $ cl10area   : num    -1.932 -1.233 -1.233 -0.932 -0.932 ...
# $ cgraze     : num    -0.9821 2.0179 2.0179 0.0179 2.0179 ...
# $ resid1    : num    -4.22 -1.03 -1.86 2.28 7.14 ...
# $ predict1   : num     9.52 3.03 3.36 14.82 6.66 ...
# $ arearesy   : num    -16.49 -3.28 -6.69 -1.78 4.71 ...
# $ arearesx   : num    -1.642 -0.3 -0.647 -0.543 -0.326 ...
# $ grazresy   : num    -1.318 -0.805 -1.425 2.459 6.157 ...
# $ grazresx   : num    -1.741 -0.137 -0.258 -0.108 0.580
```



## Знакомимся с данными

Есть ли пропущенные значения?

```
sapply(bird, function(x)sum(is.na(x)))
```

```
#      abund      area yr.isol      dist      ldist      graze      alt
#         0         0         0         0         0         0         0
#  ll0dist ll0ldist ll0area  cyr.isol  cl10area  cgraze  residl
#         0         0         0         0         0         0         0
# predictl arearesy arearesx grazresy grazresx  yrresy  yrresx
#         0         0         0         0         0         0         0
```

Каков объем выборки?

```
nrow(bird)
```

```
# [1] 56
```

## Задача

- Подберите модель множественной линейной регрессии, чтобы описать, как зависит обилие птиц от характеристик леса
  - Проверьте значимость ее коэффициентов при помощи t-критерия
- 
- abund - Обилие птиц
  - l10area - Площадь леса, га
  - l10dist - Расстояние до ближайшего леса, км (логарифм)
  - l10ldist - Расстояние до ближайшего леса большего размера, км (логарифм)
  - yr.isol - Год изоляции лесного массива

## Решение

```
bird_lm <- lm(abund ~ l10area + l10dist + l10ldist + yr.isol, data = bird)
summary(bird_lm)
```

```
#
# Call:
# lm(formula = abund ~ l10area + l10dist + l10ldist + yr.isol,
#     data = bird)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -16.6635  -3.5460   0.0859   2.8838  16.5300
#
# Coefficients:
#              Estimate Std. Error t value    Pr(>|t|)
# (Intercept) -224.42456    74.85040  -2.998    0.00419 **
# l10area       9.23476     1.27597   7.237 0.0000000023 ***
# l10dist      -0.70464     2.70766  -0.260   0.79573
# l10ldist     -1.59350     2.09538  -0.760   0.45047
# yr.isol       0.12358     0.03794   3.257   0.00201 **
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 6.577 on 51 degrees of freedom
# Multiple R-squared:  0.6519, Adjusted R-squared:  0.6246
# F-statistic: 23.88 on 4 and 51 DF,  p-value: 3.622e-11
```

## Можно привести в виде таблицы результаты t-теста для коэффициентов

- Обилие птиц увеличивалось с увеличением площади леса, и с уменьшением продолжительности изоляции (Табл. 1).

**Table 1:** Коэффициенты линейной регрессии обилия птиц от различных характеристик леса:  $\ln 10 \text{area}$  - логарифм площади леса,  $\ln 10 \text{dist}$  — логарифм расстояния до ближайшего леса,  $\ln 10 \text{ldist}$  — логарифм расстояния до ближайшего большого леса,  $\text{yr.isol}$  — год изоляции лесного массива.  $t$  — значение t-критерия,  $P$  — доверительная вероятность.

	Оценка	Ст.ошибка	t	P
Отрезок	-224.42	74.85	-3.00	<0.01
$\ln 10 \text{area}$	9.23	1.28	7.24	<0.01
$\ln 10 \text{dist}$	-0.70	2.71	-0.26	0.80
$\ln 10 \text{ldist}$	-1.59	2.10	-0.76	0.45

## Задача

Запишите уравнение множественной линейной регрессии

В качестве подсказки:

```
coef(bird_lm)
```

```
# (Intercept)      ll0area      ll0dist      ll0ldist      yr.isol  
# -224.4245557      9.2347571     -0.7046391     -1.5934969      0.1235795
```

```
bird_lm$call
```

```
# lm(formula = abund ~ ll0area + ll0dist + ll0ldist + yr.isol,  
#      data = bird)
```

## Решение

Коэффициенты модели:

```
coef(bird_lm)
```

# (Intercept)	l10area	l10dist	l10ldist	yr.isol
# -224.4245557	9.2347571	-0.7046391	-1.5934969	0.1235795

Уравнение регрессии:

$$\text{abund} = -224.42 + 9.23 \text{ l10area} - 0.70 \text{ l10dist} - 1.59 \text{ l10ldist} + 0.12 \text{ yr.isol}$$

Более формальная запись:

$$Y = -224.42 + 9.23 X_1 - 0.70 X_2 - 1.59 X_3 + 0.12 X_4$$

# Интерпретация коэффициентов регрессии

```
coef(bird_lm)
```

#	(Intercept)	l10area	l10dist	l10ldist	yr.isol
#	-224.4245557	9.2347571	-0.7046391	-1.5934969	0.1235795

# Интерпретация коэффициентов регрессии

```
coef(bird_lm)
```

```
# (Intercept)      ll0area      ll0dist      ll0ldist      yr.isol
# -224.4245557      9.2347571     -0.7046391     -1.5934969      0.1235795
```

## 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  $X_1 = \dots = X_p = 0$ .
- Коэффициенты при  $X_p$  показывают, на сколько изменяется  $Y$ , когда предиктор  $X_p$  меняется на единицу, при условии, что остальные предикторы не меняют своих значений.



## Для сравнения влияния разных факторов — стандартизованные коэффициенты

```
scaled_bird_lm <- lm(abund ~ scale(l10area) + scale(l10dist) +  
                      scale(l10ldist) + scale(yr.isol), data = bird)  
coef(scaled_bird_lm)
```

```
#      (Intercept)  scale(l10area)  scale(l10dist) scale(l10ldist)  
#      19.5142857      7.5024269      -0.2915814      -0.9160679  
# scale(yr.isol)  
#      3.1613396
```

## Для сравнения влияния разных факторов — стандартизованные коэффициенты

```
scaled_bird_lm <- lm(abund ~ scale(l10area) + scale(l10dist) +
                     scale(l10ldist) + scale(yr.isol), data = bird)
coef(scaled_bird_lm)
```

```
#      (Intercept)  scale(l10area)  scale(l10dist) scale(l10ldist)
#      19.5142857      7.5024269      -0.2915814      -0.9160679
#  scale(yr.isol)
#      3.1613396
```

### 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  $X_1 = \dots = X_p = 0$ . Для стандартизованных величин среднее значение равно нулю, поэтому  $b_0$  — это значение зависимой переменной при средних значениях всех предикторов.
- Коэффициенты при  $X_p$  показывают, на сколько изменяется  $Y$ , когда предиктор  $X_p$  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 факторы сильнее всего влияют на обилие птиц

```
summary(scaled_bird_lm)
```

```
#
# Call:
# lm(formula = abund ~ scale(l10area) + scale(l10dist) + scale(l10ldist) +
#     scale(yr.isol), data = bird)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -16.6635  -3.5460   0.0859   2.8838  16.5300
#
# Coefficients:
#              Estimate Std. Error t value    Pr(>|t|)
# (Intercept)   19.5143    0.8789  22.203 < 2e-16 ***
# scale(l10area)  7.5024    1.0366   7.237 0.0000000023 ***
# scale(l10dist) -0.2916    1.1204  -0.260   0.79573
# scale(l10ldist) -0.9161    1.2046  -0.760   0.45047
# scale(yr.isol)  3.1613    0.9707   3.257   0.00201 **
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 6.577 on 51 degrees of freedom
# Multiple R-squared:  0.6519, Adjusted R-squared:  0.6246
# F-statistic: 23.88 on 4 and 51 DF, p-value: 3.622e-11
```

## Оценка качества подгонки модели

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

```
# [1] 0.6246181
```

**Обычный  $R^2$  — доля объясненной изменчивости**

$$R^2 = \frac{SS_{model}}{SS_{total}} = 1 - \frac{SS_{error}}{SS_{total}}$$

**Не используйте обычный  $R^2$  для множественной регрессии!**

**$R^2_{adj}$  — скорректированный  $R^2$**

$$R^2_{adj} = 1 - \frac{SS_{error}/df_{error}}{SS_{total}/df_{total}}$$

где  $df_{error} = n - p - 1$ ,  $df_{total} = n - 1$

$R^2_{adj}$  учитывает число переменных в модели, вводится штраф за каждый новый параметр.

Используйте  $R^2_{adj}$  для сравнения моделей с разным числом параметров.

## Условия применимости линейной регрессии

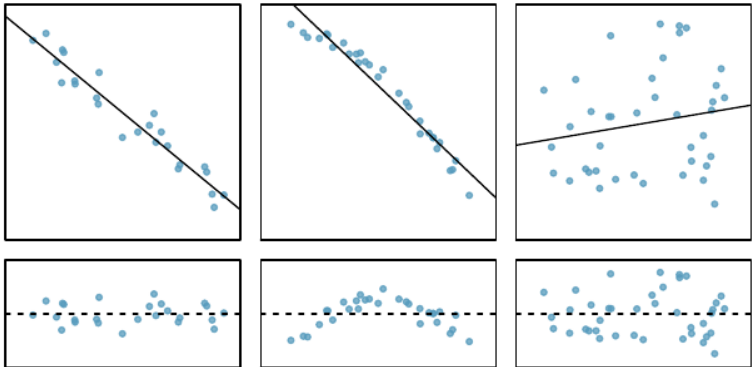
# Условия применимости линейной регрессии

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)

# 1. Независимость

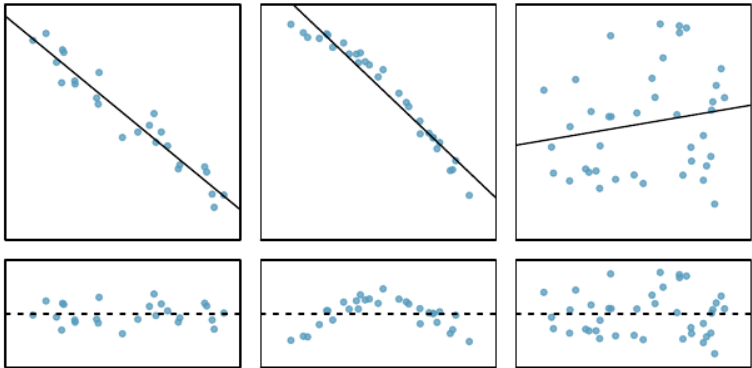
- Значения  $y_i$  должны быть независимы друг от друга
- берегитесь псевдоповторностей и автокорреляций (например, временных)
- Контролируется на этапе планирования
- Проверяем на графике остатков



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



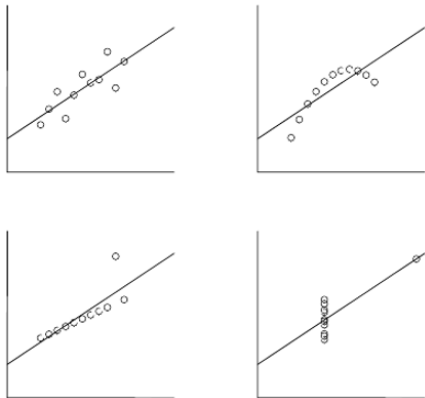
# Что бывает, если не глядя применять линейную регрессию

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

$$y_i = 3.0 + 0.5x_i$$

$$r^2 = 0.68$$

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



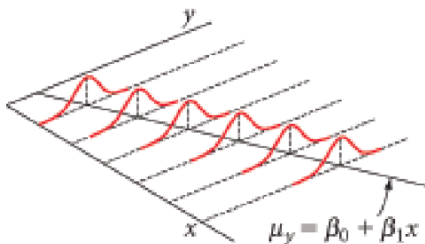
Из кн. Quinn, Keough, 2002, стр. 97, рис. 5.9

### 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

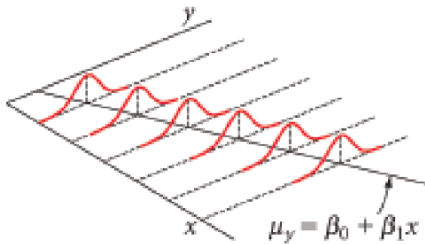
$Y_i = \beta_0 + \beta x_i + \epsilon_i$  зависимая

переменная  $Y \sim N(0, \sigma^2)$  и

дисперсии  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_i^2$  для каждого  $Y_i$

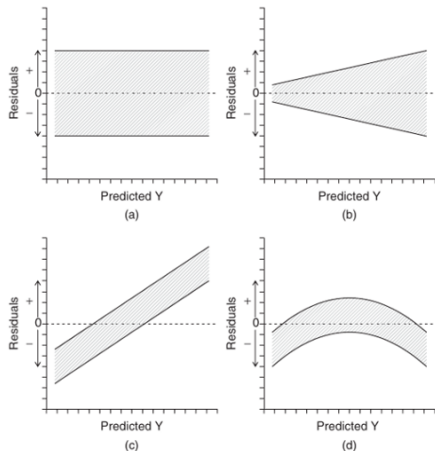
Но, поскольку  $\epsilon_i \sim N(0, \sigma^2)$ , можно проверить равенство дисперсий остатков  $\epsilon_i$

- Нужно и важно для тестов параметров
- Проверяем на графике остатков по отношению к предсказанным значениям
- Есть формальные тесты (Cochran's C), но только если несколько значений  $y$  для каждого  $x$



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

# Диагностика регрессии по графикам остатков



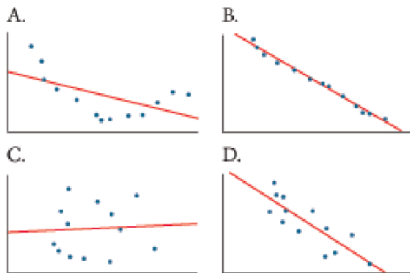
\begin{enumerate}[(a)] - все условия выполнены - разброс остатков разный (wedge-shaped pattern) - разброс остатков одинаковый, но нужны дополнительные предикторы - к нелинейной зависимости применили линейную регрессию \end{enumerate}

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

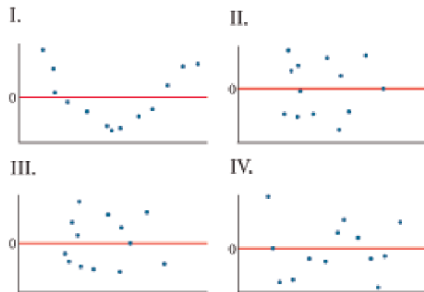
# Задача: Проанализируйте графики остатков

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.

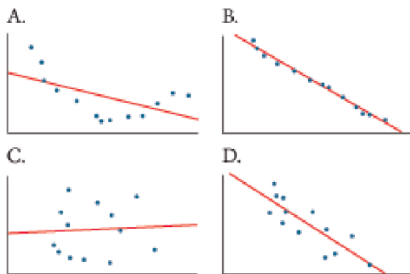


**Display 3.85** Four residual plots.

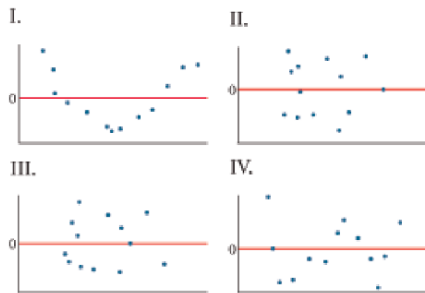
Из кн. Watkins et al. 2008, стр. 177, рис. 3.84-3.85

# 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.

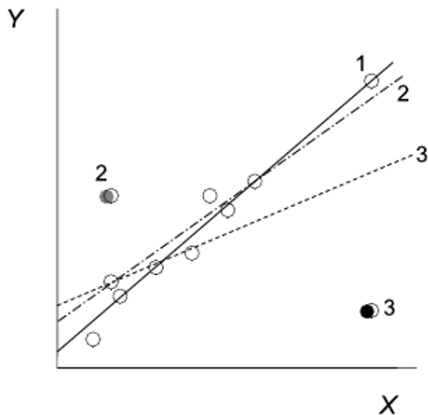
# Какие наблюдения влияют на ход регрессии больше других?

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)

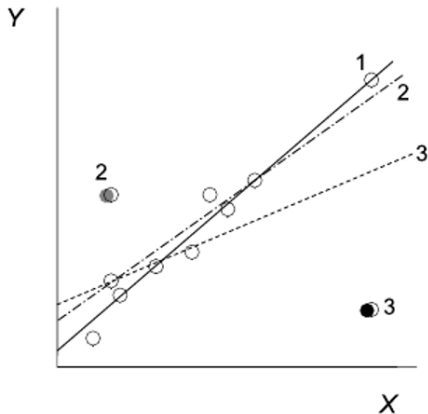


Из кн. Quinn, Keough, 2002, стр. 96, рис. 5.8

## Как оценить влияние наблюдений?

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

- Учитывает одновременно величину остатка и близость к краям области определения (leverage)
- Условное пороговое значение: выброс, если  $d \geq 4/(N - k - 1)$ , где  $N$  - объем выборки,  $k$  - число предикторов.



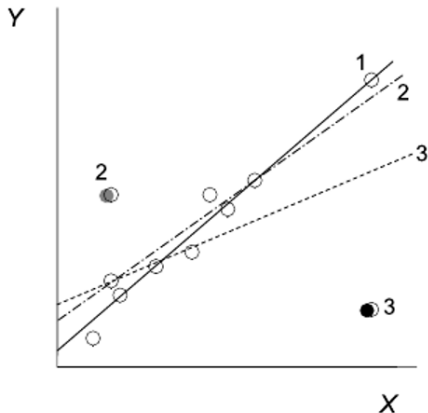
Из кн. Quinn, Keough, 2002, стр. 96, рис. 5.8



## Как оценить влияние наблюдений?

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

- Учитывает одновременно величину остатка и близость к краям области определения (leverage)
- Условное пороговое значение: выброс, если  $d \geq 4/(N - k - 1)$ , где  $N$  - объем выборки,  $k$  - число предикторов.

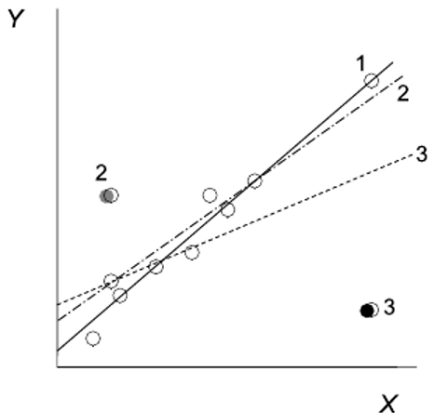


Из кн. Quinn, Keough, 2002, стр. 96, рис. 5.8

- 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

## Колинеарность предикторов

### Колинеарность

Колинеарные 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 колинеарностью?

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

## Проверка на коллинеарность

### Толерантность (tolerance)

$1 - R^2$  регрессии данного предиктора от всех других  
 $T \leq 0.25$  — коллинеарность

### Показатель инфляции для дисперсии

(коэффициент распространения дисперсии, Variance inflation factor, VIF)  
 $VIF = 1/T$   
 $VIF > 10$  (по Marquardt, 1970), но лучше  $VIF > 3$ , а иногда и  $VIF > 2$  — коллинеарность

## Проверка условий применимости линейной регрессии

## Как проверить условия применимости?

- ❶ VIF — коллинеарность предикторов (для множественной регрессии)
- ❷ График остатков от предсказанных значений — величина остатков, влияние наблюдений, отсутствие паттернов, гомогенность дисперсий.
- ❸ График квантилей остатков — распределение остатков

# 1. Проверим, есть ли в этих данных коллинеарность предикторов

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

```
#  ll0area  ll0dist ll0ldist  yr.isol
# 1.366278 1.596165 1.844939 1.197991
```

```
vif(bird_lm) > 2 # есть ли проблемы?
```

```
#  ll0area  ll0dist ll0ldist  yr.isol
#   FALSE    FALSE    FALSE    FALSE
```

```
1/vif(bird_lm) # tolerance
```

```
#  ll0area  ll0dist ll0ldist  yr.isol
# 0.7319153 0.6265018 0.5420234 0.8347311
```

# 1. Проверим, есть ли в этих данных коллинеарность предикторов

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

```
#  ll0area  ll0dist ll0ldist  yr.isol
# 1.366278 1.596165 1.844939 1.197991
```

```
vif(bird_lm) > 2 # есть ли проблемы?
```

```
#  ll0area  ll0dist ll0ldist  yr.isol
#   FALSE    FALSE    FALSE    FALSE
```

```
1/vif(bird_lm) # tolerance
```

```
#  ll0area  ll0dist ll0ldist  yr.isol
# 0.7319153 0.6265018 0.5420234 0.8347311
```

Все в порядке, предикторы независимы



## Для анализа остатков выделим нужные данные в новый датафрейм

```
library(ggplot2) # там есть функция fortify()
bird_diag <- fortify(bird_lm)

head(bird_diag, 2)
```

```
#   abund  ll0area  ll0dist ll0ldist yr.isol      .hat   .sigma
# 1    5.3 -1.00000 1.591065 1.591065   1968 0.16621067 6.641837
# 2    2.0 -0.30103 2.369216 2.369216   1920 0.08525566 6.631126
#           .cooksd  .fitted      .resid   .stdresid
# 1 0.0003830847 5.888692 -0.5886922 -0.09802371
# 2 0.0032420786 4.623396 -2.6233963 -0.41704702
```

## Для анализа остатков выделим нужные данные в новый датафрейм

```
library(ggplot2) # там есть функция fortify()
bird_diag <- fortify(bird_lm)

head(bird_diag, 2)
```

```
#   abund  ll0area  ll0dist ll0ldist yr.isol      .hat   .sigma
# 1    5.3 -1.00000 1.591065 1.591065   1968 0.16621067 6.641837
# 2    2.0 -0.30103 2.369216 2.369216   1920 0.08525566 6.631126
#           .cooksd   .fitted     .resid   .stdresid
# 1 0.0003830847 5.888692 -0.5886922 -0.09802371
# 2 0.0032420786 4.623396 -2.6233963 -0.41704702
```

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

## Задача

Постройте график зависимости стандартизованных остатков от предсказанных значений

Используйте данные из `bird_diag`

```
ggplot()  
aes()  
geom_point()
```

### Стандартизованные остатки

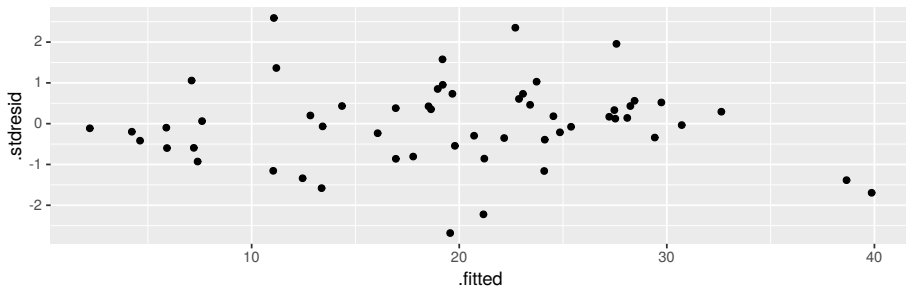
$$\frac{y_i - \hat{y}_i}{\sqrt{MS_e}}$$

- можно сравнивать между регрессиями
- можно сказать, какие остатки большие, какие нет
  - $\leq 2SD$  - обычные
  - $> 3SD$  - редкие

## Решение

График зависимости стандартизованных остатков от предсказанных значений

```
ggplot(data = bird_diag, aes(x = .fitted, y = .stdresid)) + geom_point()
```



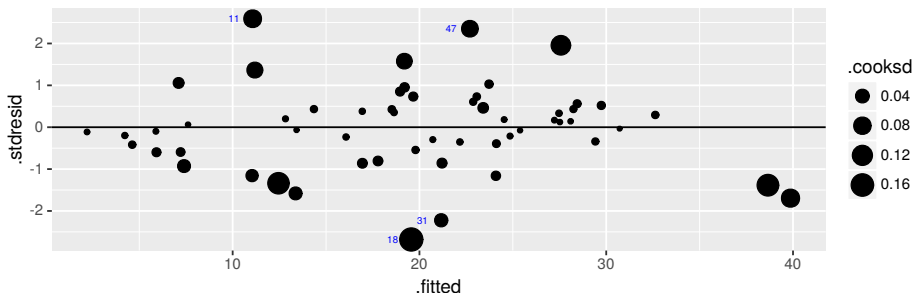
## График станет информативнее, если кое-что добавить

```
# Создаем логический вектор, где TRUE,
# если стандартизованный остаток больше 2
f_outlier <- abs(bird_diag$.stdresid) > 2
# Создаем будущие ярлыки
labs <- ifelse(test = f_outlier,
               yes = row(bird_diag), # Если test == TRUE
               no = "") # Если test == FALSE

gg_resid <- ggplot(data = bird_diag,
                  aes(x = .fitted, y = .stdresid)) +
  geom_point(aes(size = .cooksdi)) + # расстояние Кука
  geom_hline(yintercept = 0) + # горизонтальная линия y = 0
  geom_text(aes(label = labs), hjust = 2, colour = "blue",
            size = 2) # номера наблюдений с остатками больше 2SD
```

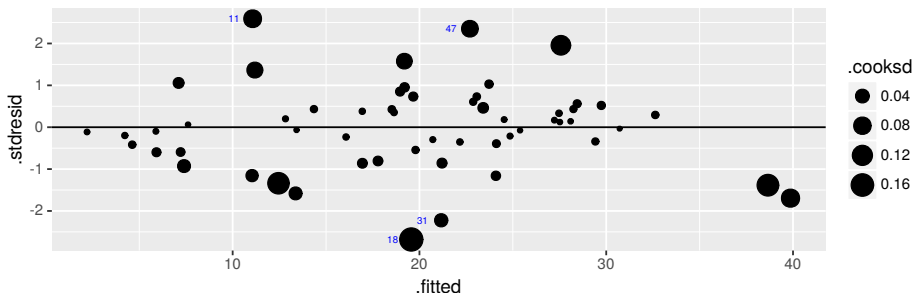
# Интерпретируем график стандартизованных остатков от предсказанных значений

Какие выводы можно сделать по графику остатков?



## Интерпретируем график стандартизованных остатков от предсказанных значений

Какие выводы можно сделать по графику остатков?

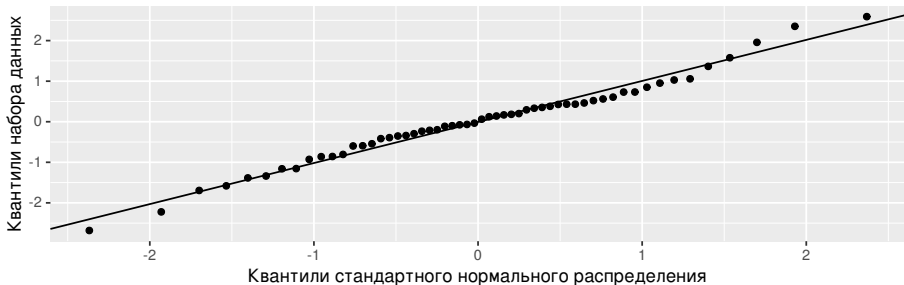


- Большая часть стандартизованных остатков в пределах двух стандартных отклонений. Есть отдельные влиятельные наблюдения, которые нужно проверить
- Разброс остатков не совсем одинаков. Похоже на гетерогенность дисперсий
- Тренда среди остатков нет

### 3. Квантильный график стандартизованных остатков

Используется, чтобы оценить форму распределения. Если точки лежат на одной прямой — все в порядке.

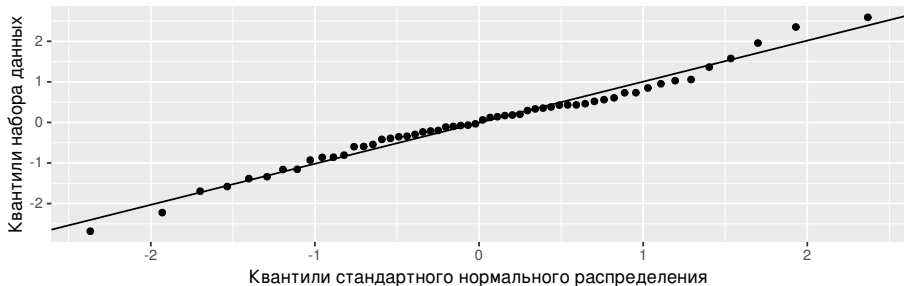
```
mean_val <- mean(bird_diag$.stdresid)
sd_val <- sd(bird_diag$.stdresid)
ggplot(bird_diag, aes(sample = .stdresid)) + geom_point(stat = "qq") +
  geom_abline(intercept = mean_val, slope = sd_val) + # точки должны быть здесь
  labs(x = "Квантили стандартного нормального распределения", y = "Квантили н
```





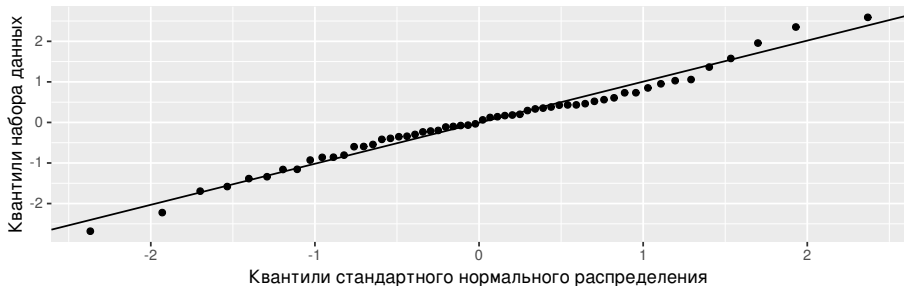
# Интерпретируем квантильный график

Какие выводы можно сделать по квантильному графику?



# Интерпретируем квантильный график

Какие выводы можно сделать по квантильному графику?



- Отклонений от нормального распределения нет

## Take home messages

- Для сравнения влияния разных предикторов можно использовать бета-коэффициенты
- 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)

## Дополнительные ресурсы

### Учебники

- Quinn, Keough, 2002, pp. 92-98, 111-130
- [Open Intro to Statistics: Chapter 8. Multiple and logistic regression](#), pp. 354-367.
- Logan, 2010, pp. 170-173, 208-211
- Sokal, Rohlf, 1995, pp. 451-491, 609-653
- Zar, 2010, pp. 328-355, 419-439

### Упражнения для тренировки

- OpenIntro Labs, Lab 7: Introduction to linear regression (Осторожно, они используют базовую графику а не ggplot)
  - [Обычный вариант](#), после упражнения 4
  - [Интерактивный вариант на Data Camp](#), после вопроса 4
- OpenIntro Labs, Lab 8: Multiple linear regression
  - [Обычный вариант](#), до упражнения 11
  - [Интерактивный вариант на Data Camp](#), до вопроса 8