

# Регрессионный анализ, часть 2

## Математические методы в зоологии - на R

Марина Варфоломеева

осень 2015

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

## Вы сможете

- Подобрать модель множественной линейной регрессии, проверить ее валидность и интерпретировать коэффициенты при разных предикторах.
- Проверить условия применимости 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 - Продолжительности изоляции, лет
- abund - Обилие птиц

# Открываем данные

```
# установите рабочую директорию
# birds <- read.delim(file = "loyn.csv") # из .csv
library(readxl)
birds <- read_excel("loyn.xls", sheet = 1)
str(birds)
```

```
# Classes 'tbl_df', 'tbl' and '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 : num 1968 1920 1900 1966 1918 ...
# $ dist : num 39 234 104 66 246 234 467 284 156 311 ...
# $ ldist : num 39 234 311 66 246 ...
# $ graze : num 2 5 5 3 5 3 5 5 4 5 ...
# $ alt : num 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 ...
```

## Задача: запишите формулу модели регрессии

Как зависит обилие птиц от характеристик леса? Запишите в обозначениях R модель множественной линейной регрессии

$$Y_i = b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + b_4x_{4i}$$

Используйте названия переменных вместо  $x_{1i}$  —  $x_{4i}$

- 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 - Продолжительности изоляции, лет

# Решение

В обозначениях R модель множественной линейной регрессии

$$abund \sim l10area + l10dist + l10ldist + yr.isol$$

Названия переменных:

- 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 - Продолжительности изоляции, лет



# Подбираем параметры модели и проверяем валидность с помощью t-критерия

$$H_0 : \beta_i = 0$$

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

```
#
# Call:
# lm(formula = abund ~ l10area + l10dist + l10ldist + yr.isol,
#     data = birds)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -16.663  -3.546   0.086   2.884  16.530
#
# Coefficients:
#              Estimate Std. Error t value    Pr(>|t|)
# (Intercept) -224.4246    74.8504  -3.00    0.0042 **
# l10area       9.2348     1.2760   7.24 0.0000000023 ***
# l10dist      -0.7046     2.7077  -0.26    0.7957
# l10ldist     -1.5935     2.0954  -0.76    0.4505
# yr.isol       0.1236     0.0379   3.26    0.0020 **
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Задача: Запишите уравнение множественной линейной регрессии

Запишите уравнение множественной линейной регрессии

В качестве подсказки:

```
coef(bird_lm)
```

# (Intercept)	l10area	l10dist	l10ldist	yr.isol
# -224.425	9.235	-0.705	-1.593	0.124

```
bird_lm$call
```

```
# lm(formula = abund ~ l10area + l10dist + l10ldist + yr.isol,
#     data = birds)
```

# Уравнение множественной линейной регрессии

```
coef(bird_lm)
```

# (Intercept)	l10area	l10dist	l10ldist	yr.isol
# -224.425	9.235	-0.705	-1.593	0.124

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
# -224.425	9.235	-0.705	-1.593	0.124

## Обычные коэффициенты

- величина зависит от единиц измерения

## Сравнение влияния разных факторов

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

```
#      (Intercept)  scale(l10area)  scale(l10dist)  scale(l10ldist)
#           19.514           7.502          -0.292          -0.916
#  scale(yr.isol)
#           3.161
```

### Бета-коэффициенты

- измерены в стандартных отклонениях
- относительная оценка влияния фактора
- можно сравнивать

## Задача: Сравните влияние разных факторов

Определите по значениям beta-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 = birds)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -16.663  -3.546   0.086   2.884  16.530
#
# Coefficients:
#              Estimate Std. Error t value    Pr(>|t|)
# (Intercept)    19.514     0.879   22.20 < 2e-16 ***
# scale(l10area)    7.502     1.037    7.24 0.0000000023 ***
# scale(l10dist)   -0.292     1.120   -0.26   0.796
# scale(l10ldist)  -0.916     1.205   -0.76   0.450
# scale(yr.isol)    3.161     0.971    3.26   0.002 **
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 6.58 on 51 degrees of freedom
# Multiple R-squared:  0.652    Adjusted R-squared:  0.625
```

## Оценка качества подгонки модели

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

```
# [1] 0.625
```

### Скорректированный $R^2$

- Учитывает число переменных в модели

## Условия применимости линейной регрессии



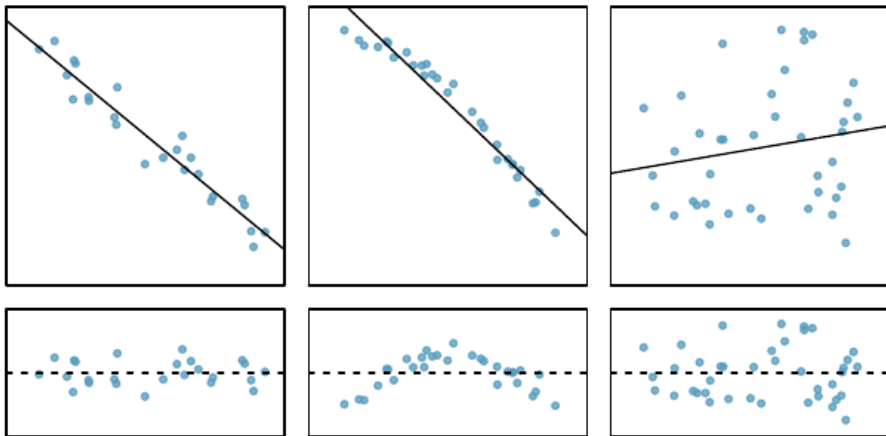
# Условия применимости линейной регрессии

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

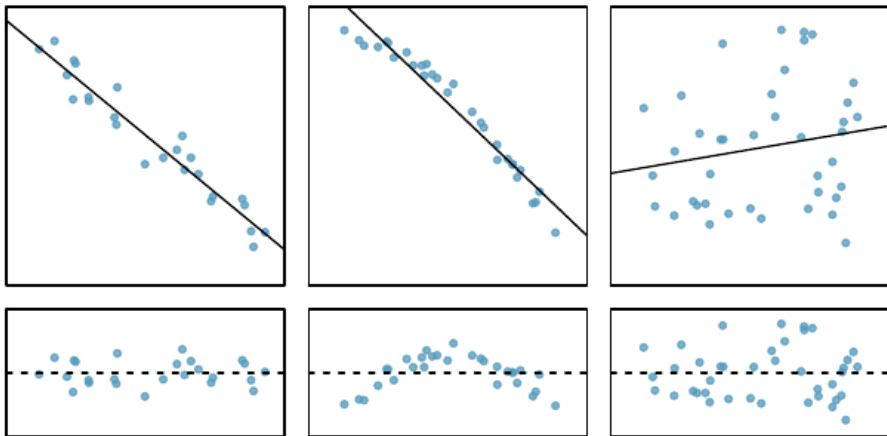
- Значения  $y_i$  должны быть независимы друг от друга
  - берегитесь псевдповторностей и автокорреляций (например, временных)
- Контролируется на этапе планирования
- Проверяем на графике остатков



Остаточная изменчивость (Рис. из кн. Diez et al., 2010, стр. 332, рис. 7.8)

## 2. Линейность связи

- проверяем на графике рассеяния исходных данных
- проверяем на графике остатков



Остаточная изменчивость (Рис. из кн. Diez et al., 2010, стр. 332, рис. 7.8)

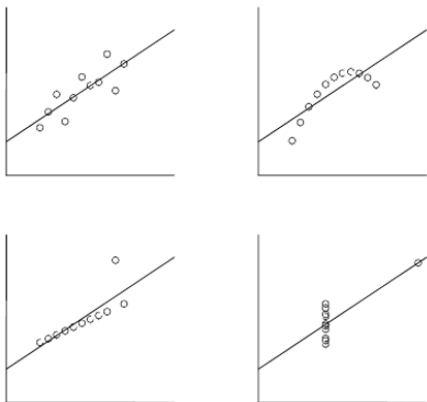
# Что бывает, если не глядя применять линейную регрессию

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

$$y_i = 3.0 + 0.5x_i$$

$$r^2 = 0.68$$

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



Квартет Энскомба (рис. из кн. Quinn, Keough, 2002, стр. 97, рис. 5.9)

### 3. Нормальное распределение остатков

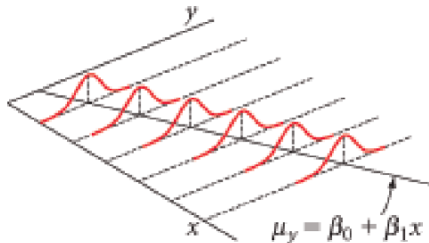
Нужно, т.к. в модели

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



Условие нормальности и гомогенность дисперсий (рис. 11.4 из кн. Watkins et al., 2008, стр. 743)

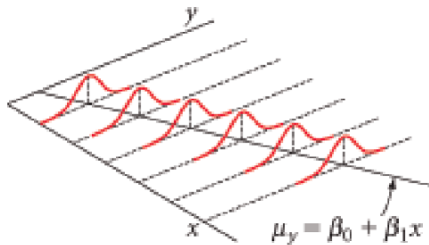
## 4. Гомогенность дисперсий

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  $Y_i$

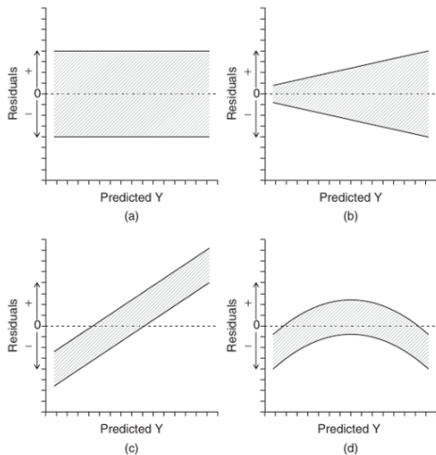
Но, поскольку  $\epsilon_i \sim N(0, \sigma^2)$ , можно проверить равенство дисперсий остатков  $\epsilon_i$

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



Условие нормальности и гомогенность дисперсий (рис. 11.4 из кн. Watkins et al., 2008, стр. 743)

# Диагностика регрессии по графикам остатков



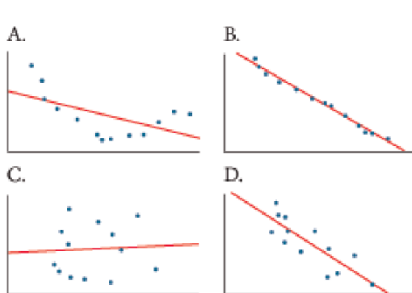
- (a)** все условия выполнены
- (b)** разброс остатков разный (wedge-shaped pattern)
- (c)** разброс остатков одинаковый, но нужны дополнительные предикторы
- (d)** к нелинейной зависимости применили линейную регрессию

Диагностика регрессии по графикам остатков (рис. 8.5 d из кн. Logan, 2010, стр. 174)

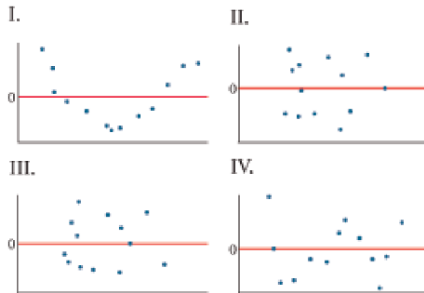
## Задача: Проанализируйте графики остатков

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.



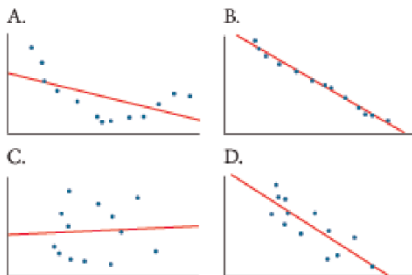
**Display 3.85** Four residual plots.

Графики регрессий и остатков (рис. 3.84-3.85 из кн. Watkins et al. 2008, стр. 177)

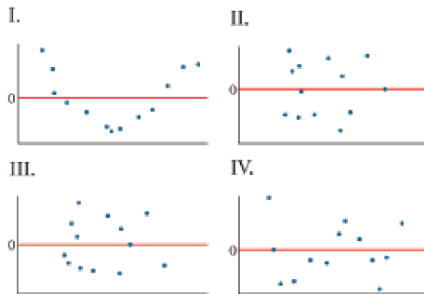


# Решение

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



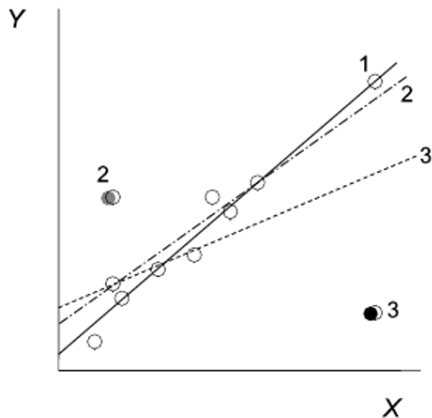
**Display 3.84** Four scatterplots.



**Display 3.85** Four residual plots.

Графики регрессий и остатков (рис. 3.84-3.85 из кн. Watkins et al. 2008, стр. 177)

# Какие наблюдения влияют на ход регрессии больше других?

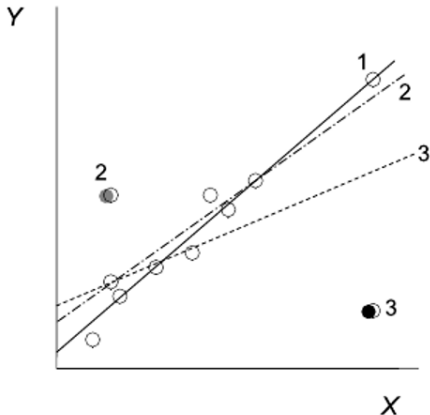


Влиятельные наблюдения (рис. 5.8 из кн. Quinn, Keough, 2002, стр. 96)

# 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)



Влиятельные наблюдения (рис. 5.8 из кн. Quinn, Keough, 2002, стр. 96)

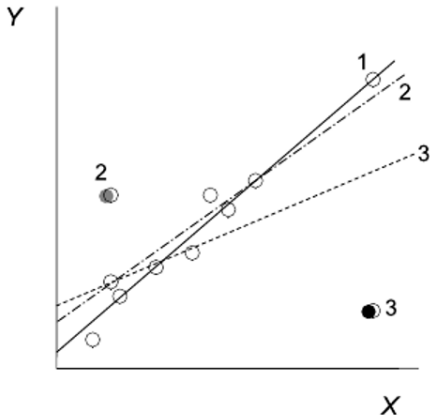
# Какие наблюдения влияют на ход регрессии больше других?

## 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
- 2 - умеренно влияет (большой остаток, малая сила влияния)
- 3 - очень сильно влияет (большой остаток, большая сила влияния)

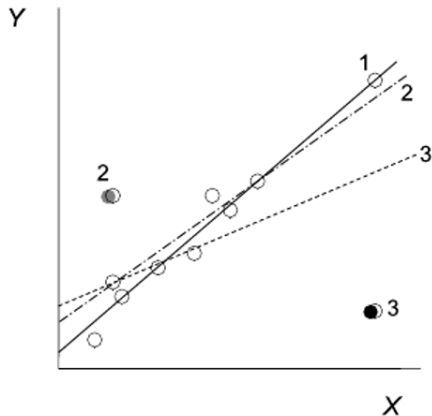


Влиятельные наблюдения (рис. 5.8 из кн. Quinn, Keough, 2002, стр. 96)

# 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.



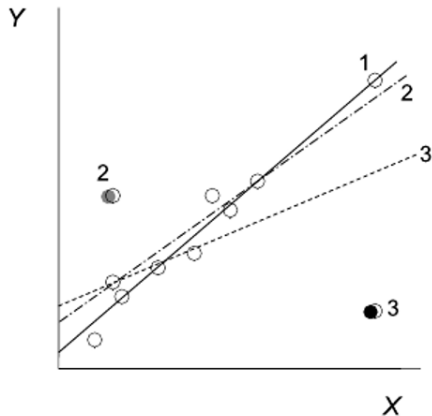
Влиятельные наблюдения (рис. 5.8 из кн. Quinn, Keough, 2002, стр. 96)

# 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.

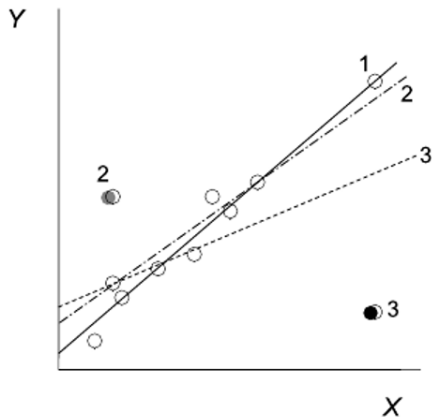
Дж. Фокс советует не обращать внимания на пороговые значения (Fox, 1991)



Влиятельные наблюдения (рис. 5.8 из кн. Quinn, Keough, 2002, стр. 96)

# 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



Влиятельные наблюдения (рис. 5.8 из кн. Quinn, Keough, 2002, стр. 96)

# Колинеарность предикторов

## Колинеарность

Когда 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)

$$VIF = 1/T$$

$\sqrt{VIF} > 2$  - коллинеарность

# Проверка условий применимости линейной регрессии

## Как проверить условия применимости?

- Величина остатков, влияние наблюдений, тренды - на графике остатков от предсказанных значений
- Форма распределения остатков - нормальное вероятностный график
- Коллинеарность предикторов - толерантность и показатель инфляции для дисперсии

## 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 l10area l10dist l10ldist yr.isol   .hat .sigma .cooksd
# 1    5.3  -1.000    1.59     1.59    1968 0.1662   6.64 0.000383
# 2     2.0  -0.301    2.37     2.37    1920 0.0853   6.63 0.003242
#   .fitted .resid .stdresid
# 1     5.89 -0.589   -0.098
# 2     4.62 -2.623   -0.417
```

Кроме abund, l10area, l10dist, l10ldist и yr.isol нам понадобятся

- .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

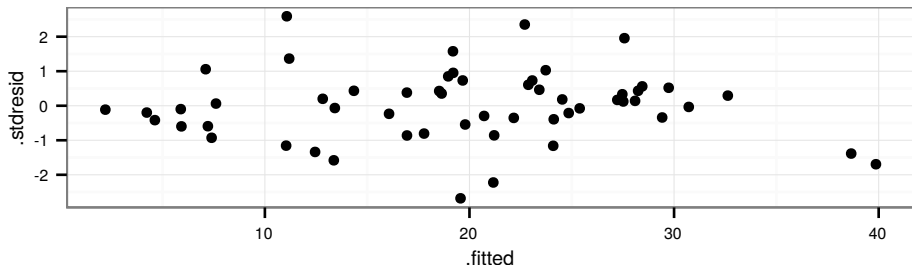
$$\frac{y_i - \hat{y}_i}{\sqrt{MS_e}}$$

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

## Решение:

График зависимости стандартизованных остатков от предсказанных значений

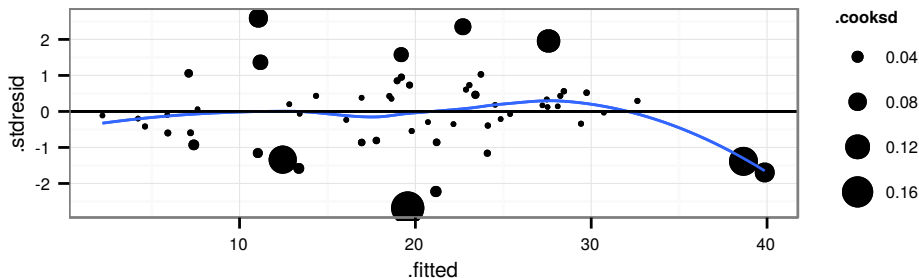
```
theme_set(theme_bw(base_size = 8) + theme(legend.key = element_blank()))  
gg_resid <- ggplot(data = bird_diag, aes(x = .fitted, y = .stdresid))  
gg_resid + geom_point()
```



# График стандартизованных остатков от предсказанных значений

График станет информативнее, если кое-что добавить

```
gg_resid <- gg_resid + geom_point(aes(size = .cooks_d)) + # расстояние Кука
  geom_smooth(method="loess", se = FALSE) + # линия тренда
  geom_hline(yintercept = 0) # горизонтальная линия y = 0
gg_resid
```

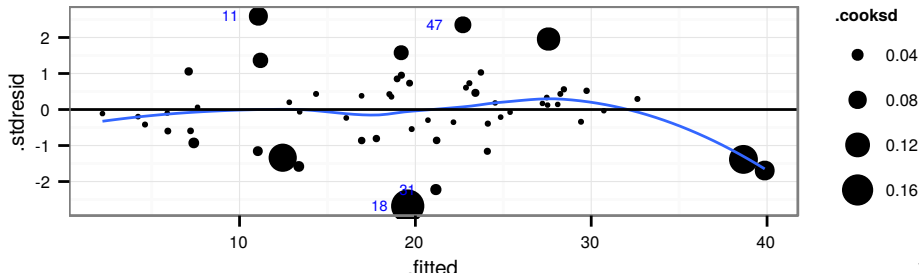


## Можно подписать остатки больше двух стандартных отклонений на предыдущем графике

Стандартизованные остатки  $\leq 2SD$  - обычные,  $> 3SD$  - редкие

```
# 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

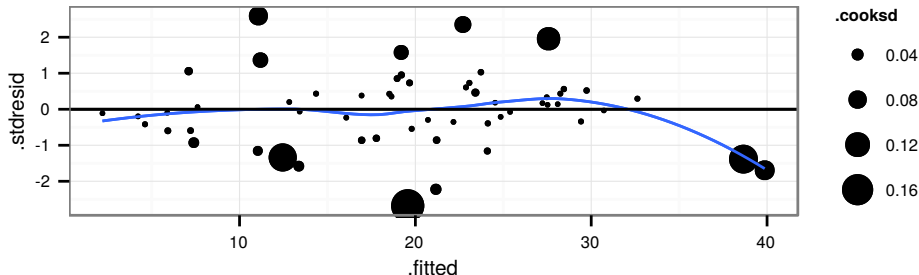
gg_resid_lab <- gg_resid +
  geom_text(aes(label = labs), hjust=2, colour = "blue", size = 2)
gg_resid_lab
```





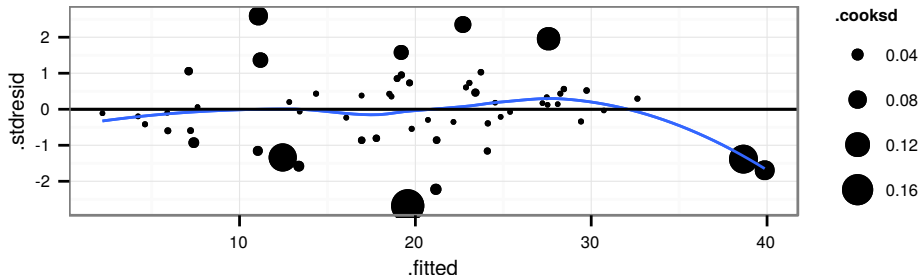
# Интерпретируем график стандартизованных остатков от предсказанных значений

Какие выводы можно сделать по графику остатков?



# Интерпретируем график стандартизованных остатков от предсказанных значений

Какие выводы можно сделать по графику остатков?

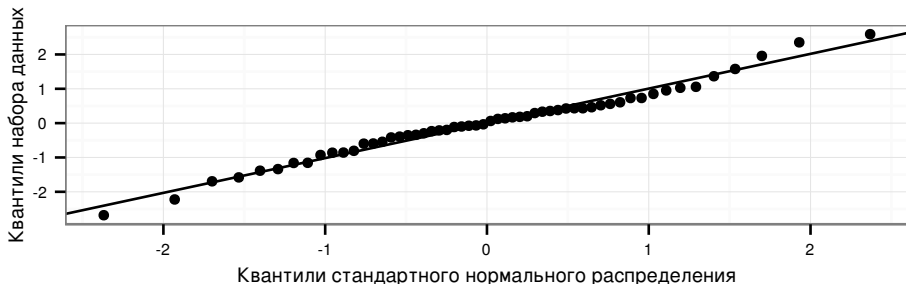


- Большая часть стандартизованных остатков в пределах двух стандартных отклонений. Есть отдельные влиятельные наблюдения, которые нужно проверить
- Разброс остатков не совсем одинаков. Похоже на гетерогенность дисперсий
- Тренда среди остатков нет

# Нормальноевероятностный график стандартизованных остатков

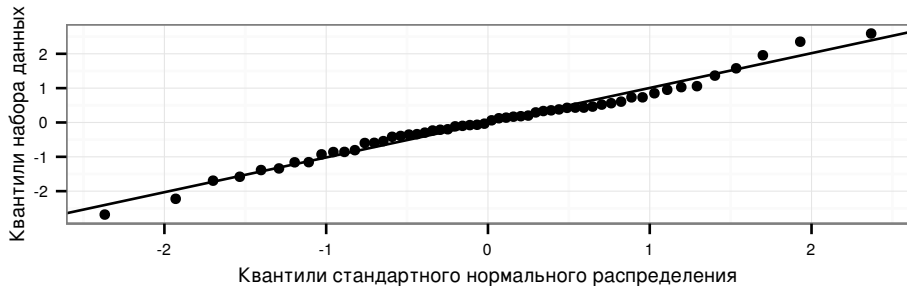
Используется, чтобы оценить форму распределения. Если точки лежат на одной прямой - нормальное распределение.

```
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 набора данных")
```



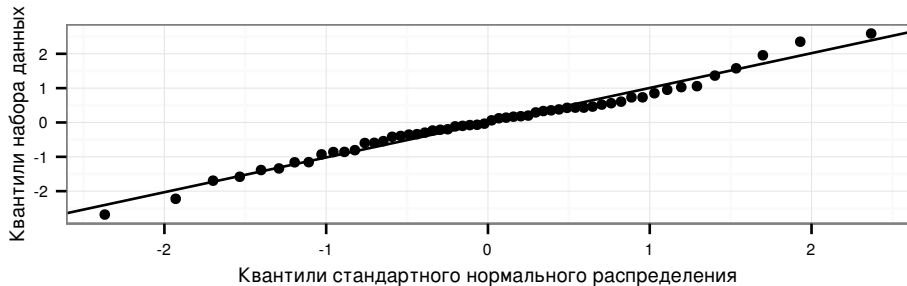
# Интерпретируем нормальновероятностный график

Какие выводы можно сделать по нормальновероятностному графику?



# Интерпретируем нормальновероятностный график

Какие выводы можно сделать по нормальновероятностному графику?



- Отклонений от нормального распределения нет

## Проверим, есть ли в этих данных коллинеарность предикторов

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

```
#  l10area  l10dist l10ldist  yr.isol
#    1.37    1.60    1.84    1.20
```

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

```
#  l10area  l10dist l10ldist  yr.isol
#   FALSE    FALSE    FALSE    FALSE
```

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

```
#  l10area  l10dist l10ldist  yr.isol
#    0.732    0.627    0.542    0.835
```

## Проверим, есть ли в этих данных коллинеарность предикторов

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

```
#  l10area  l10dist l10ldist  yr.isol
#    1.37    1.60    1.84    1.20
```

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

```
#  l10area  l10dist l10ldist  yr.isol
#   FALSE    FALSE    FALSE    FALSE
```

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

```
#  l10area  l10dist l10ldist  yr.isol
#    0.732    0.627    0.542    0.835
```

Все в порядке, предикторы независимы

# 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