Регрессионный анализ, часть 2

Математические методы в зоологии - на R, осень 2013

Марина Варфоломеева
Каф. Зоологии беспозвоночных, СПбГУ

Когда и какую регрессию можно применять

- Условия применимости регрессионного анализа
- Мощность линейной регрессии

Вы сможете

- Проверить условия применимости простой линейной регрессии
- Рассчитать мощность линейной регрессии

Пример: усыхающие личинки мучных хрущаков

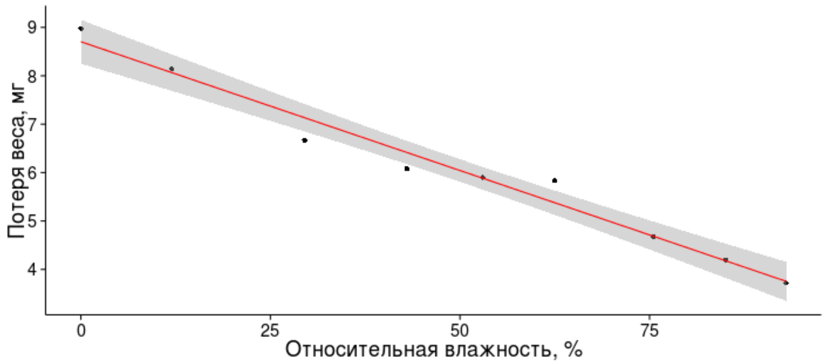
Как зависит потеря влаги личинками малого мучного хрущака *Tribolium confusum* от влажности воздуха? (Nelson, 1964)

```
# Внимание, установите рабочую директорию,
# или используйте полный путь к файлу
setwd("C:/mathmethr/week2")
## из .xlsx
library(XLConnect)
wb <- loadWorkbook("./data/nelson.xlsx")
nelson <- readWorksheet(wb, sheet = 1)
## или из .csv
# nelson <- read.table(file="./data/nelson.csv",
#                      header = TRUE, sep = "\t",
#                      dec = ".")
```

Данные из Sokal, Rohlf, 1997, табл. 14.1 по Logan, 2010. глава 8, пример 8с

Как зависит потеря веса от влажности? График рассеяния.

```
library(ggplot2)
theme_set(theme_classic()) # устанавливаем понравившуюся тему до конца сессии
p_nelson <- ggplot(data=nelson, aes(x = humidity, y = weightloss)) +
  geom_point() +
  geom_smooth(method = "lm", colour = "red") +
  labs(x = "Относительная влажность, %", y = "Потеря веса, мг")
p_nelson
```



Проверяем, есть ли зависимость потери веса от влажности с помощью линейной регрессии

```
# линейная регрессия из прошлой лекции
nelson_lm <- lm(weightloss ~ humidity, nelson)
summary(nelson_lm)
```

```
##
## Call:
## lm(formula = weightloss ~ humidity, data = nelson)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4640 -0.0344  0.0167  0.0746  0.4524
##
## Coefficients:
##             Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  8.70403    0.19156    45.4 0.00000000065 ***
## humidity    -0.05322    0.00326   -16.4 0.00000078161 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.297 on 7 degrees of freedom
## Multiple R-squared:  0.974,  Adjusted R-squared:  0.971
## F-statistic:  267 on 1 and 7 DF,  p-value: 0.000000782
```

Зависимость потери веса от влажности можно описать уравнением

Для этого подставим коэффициенты в уравнение линейной регрессии $y = b_0 + b_1 x$

```
coef(nelson_lm) # Коэффициенты регрессии
```

```
## (Intercept)    humidity
##      8.7040     -0.0532
```

$weightloss =$ 8.7 -0.05 $humidity$

Чаще более академические обозначения:

$y =$ 8.7 -0.05 x, $R^2 =$ 0.974

Потеря веса мучными хрущаками в результате высыхания достоверно зависит от относительной влажности ($\beta_1 =$ -0.05 $\pm$ 0.01, $p < 0.01$)

Насколько можно доверять оценкам коэффициентов, которые мы получили?

Условия применимости простой линейной регрессии и анализ остатков

Условия применимости простой линейной регрессии

должны выполняться, чтобы тестировать гипотезы

1. Независимость
2. Линейность
3. Нормальное распределение
4. Гомогенность дисперсий

1. Независимость

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

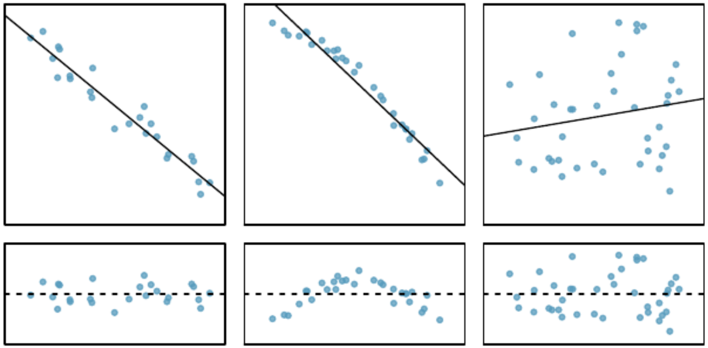


Рисунок из кн. Diez et al., 2010, стр. 332, рис. 7.8

2. Линейность связи

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

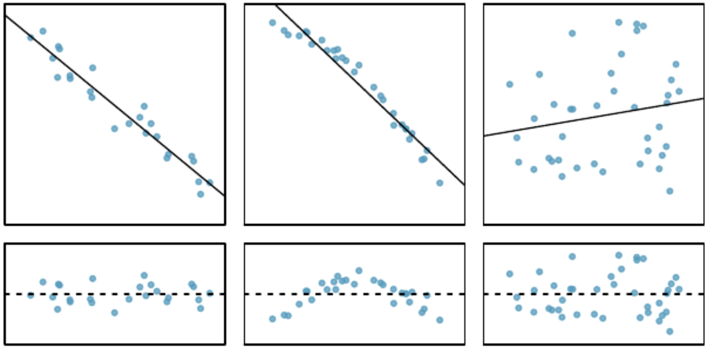


Рисунок из кн. Diez et al., 2010, стр. 332, рис. 7.8

Вот, что бывает, если неглядя применять линейную регрессию

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

$y_i = 3.0 + 0.5x_i,$

$r^2 = 0.68,$

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

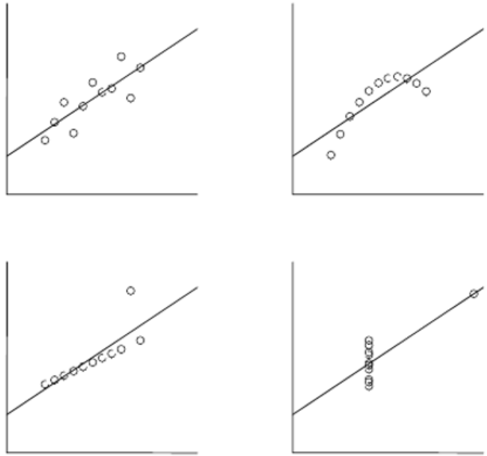


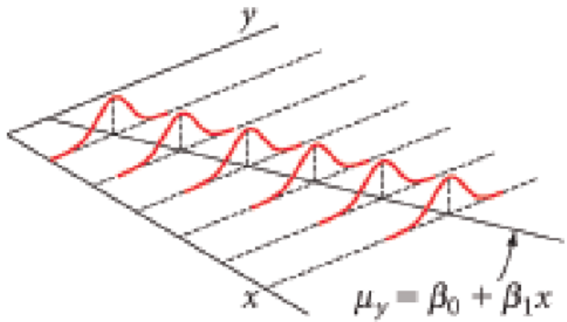
Рисунок из кн. Quinn, Keough, 2002, стр. 97, рис. 5.9

3. Нормальное распределение

Нужно, т.к. в модели $Y_i = \beta_0 + \beta x_i + \epsilon_i$

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

- К счастью, это значит, что $\epsilon_i \sim N(0, \sigma^2)$



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

Рисунок из кн. Watkins et al., 2008, стр. 743, рис. 11.4

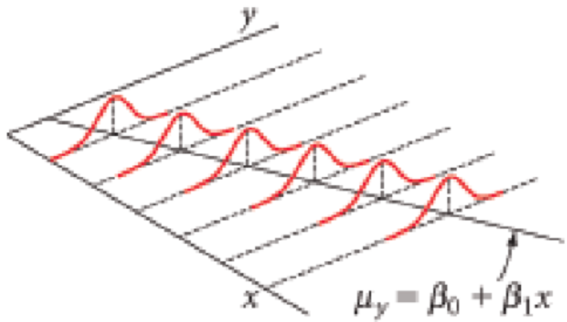
4. Гомогенность дисперсий

Нужно, т.к. в модели $Y_i = \beta_0 + \beta x_i + \epsilon_i$

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

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

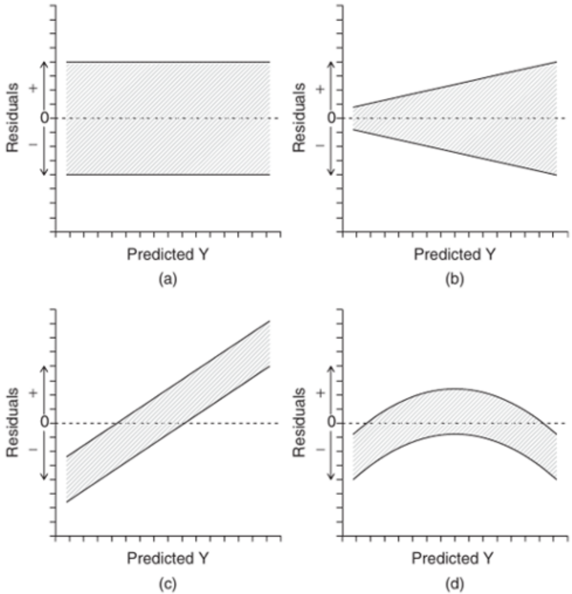
- К счастью, поскольку $\epsilon_i \sim N(0, \sigma^2)$, можно проверить равенство дисперсий остатков ϵ_i



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

Рисунок из кн. Watkins et al., 2008, стр. 743, рис. 11.4

Диагностика регрессии по графикам остатков

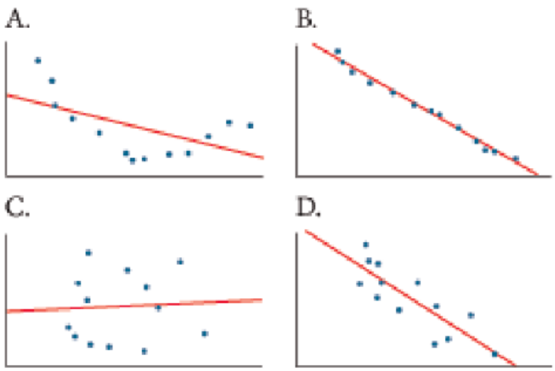


- условия:
  - a - все выполнены
  - b - разброс остатков разный (wedge-shaped pattern)
  - c - разброс остатков одинаковый, но нужны дополнительные предикторы
  - d - к нелинейной зависимости применили линейную регрессию

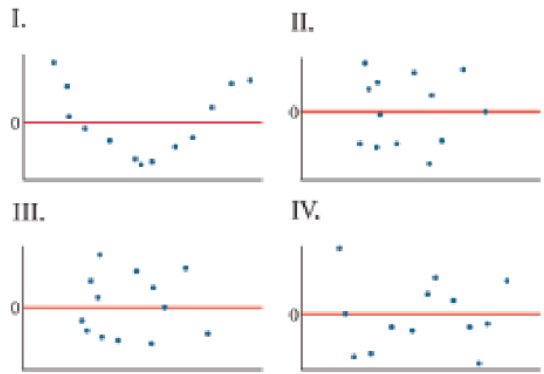
Рисунок из кн. Logan, 2010, стр. 174, рис. 8.5 d

Скажите,

- 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.



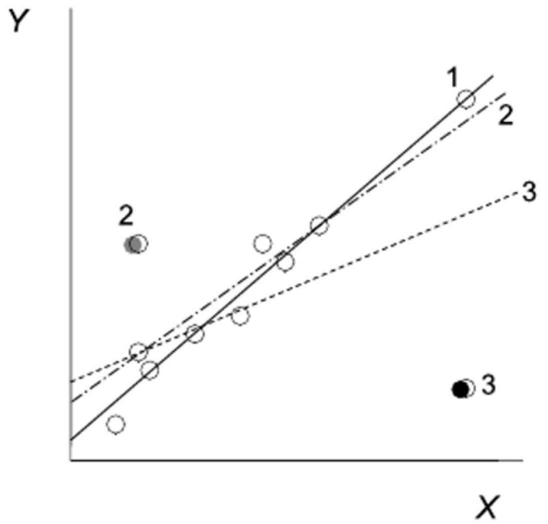
Display 3.85 Four residual plots.

Рисунок 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)



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

Рисунок 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 $d \geq 4/(N - k - 1)$, где N - объем выборки, k - число предикторов.
- Дж. Фокс советует не обращать внимания на пороговые значения (Fox, 1991).
- Что делать с влиятельными точками?
  - Проверить, не ошибка ли это. Если это не ошибка, не удалять - обсуждать!
  - Проверить, что будет, если их исключить из модели

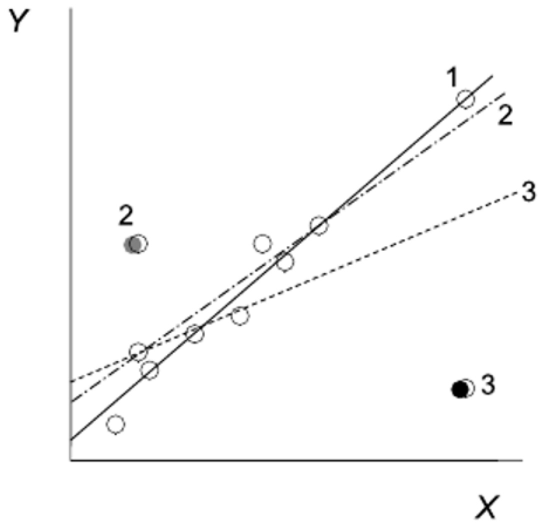


Рисунок из кн. Quinn, Keough, 2002, стр. 96, рис. 5.8

Что делать с выбросами?

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

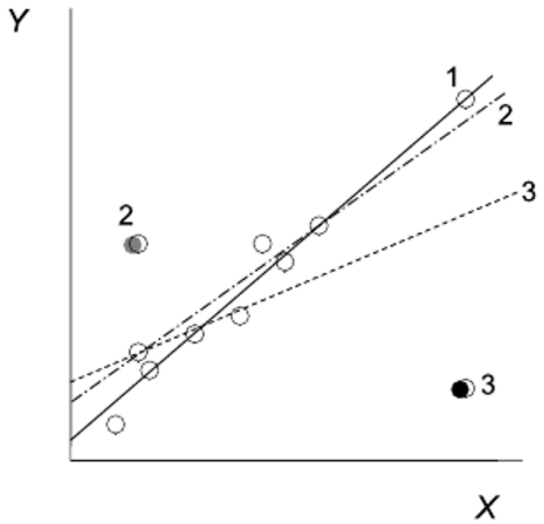


Рисунок из кн. Quinn, Keough, 2002, стр. 96, рис. 5.8

Проверим условия применимости

Проверьте линейность связи,

постройте для этого график рассеяния

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

Для анализа остатков выделим нужные данные в новый датафрейм

```
# нам нужна линейная регрессия из прошлой лекции
nelson_lm <- lm(weightloss ~ humidity, nelson) # линейная регрессия
# library(ggplot2) # функция fortify() находится в пакете ggplot2
nelson_diag <- fortify(nelson_lm)
names(nelson_diag) # названия переменных
```

```
## [1] "weightloss" "humidity"   ".hat"       ".sigma"     ".cooksd"
## [6] ".fitted"    ".resid"     ".stdresid"
```

- Кроме `weightloss` и `humidity` нам понадобятся
    - `.cooksd` - расстояние Кука
    - `.fitted` - предсказанные значения
    - `.resid` - остатки
    - `.stdresid` - стандартизованные остатки

Постройте график зависимости остатков от предиктора,

используя данные из `nelson_diag`

- `humidity` - относительная влажность (наш предиктор)
- `.resid` - остатки

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

- По абсолютным остаткам сложно сказать, большие они или маленькие. Нужна стандартизация

Постройте график зависимости стандартизованных остатков от предсказанных значений

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

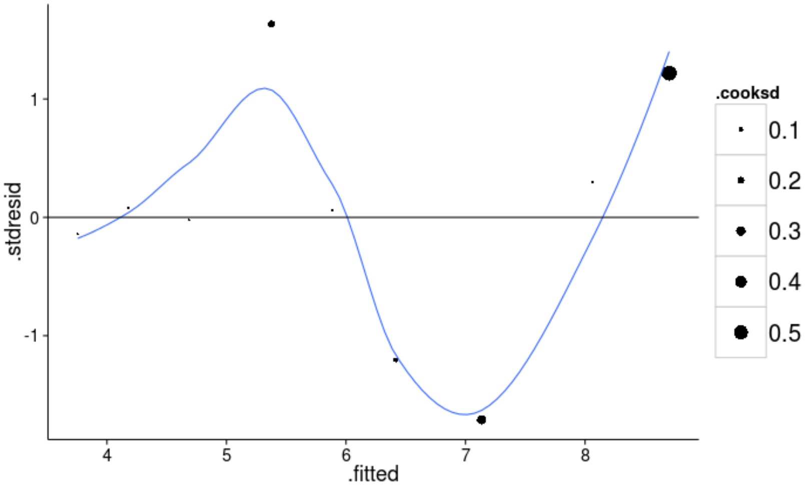
Используйте данные из `nelson_diag`

- `.fitted` - предсказанные значения
- `.resid` - остатки

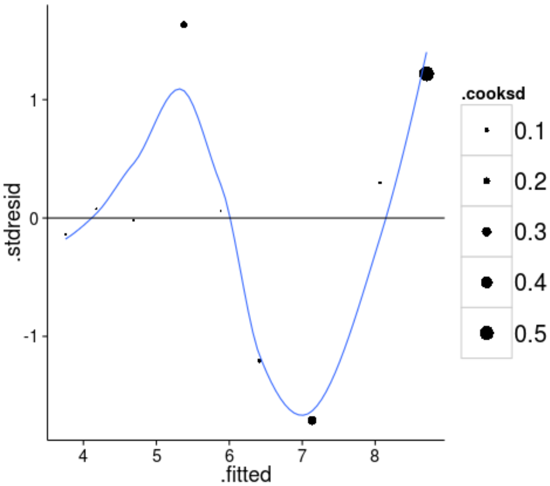
```
ggplot()
aes()
geom_point()
```

График станет информативнее, если кое-что добавить

```
ggplot(data = nelson_diag, aes(x = .fitted, y = .stdresid)) +
  geom_point(aes(size = .cooksd)) +          # расстояние Кука
  geom_smooth(method="loess", se = FALSE) +  # линия тренда, сглаживание локальной регрессией
  geom_hline(yintercept = 0)                 # горизонтальная линия на уровне y = 0
```



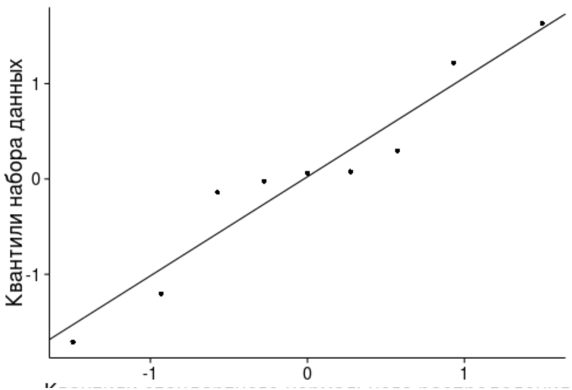
Какие выводы можно сделать по графику остатков?



- Стандартизованные остатки умеренной величины (в пределах двух стандартных отклонений), их разброс почти одинаков
- Мало точек, чтобы надежно оценить наличие трендов среди остатков

Нормально-вероятностный график стандартизованных остатков

```
mean_val <- mean(nelson_diag$.stdresid)
sd_val <- sd(nelson_diag$.stdresid)
quantile_plot <- ggplot(nelson_diag, aes(sample = .stdresid)) +
  geom_point(stat = "qq") +
geom_abline(intercept = mean_val, slope = sd_val) + # на эту линию должны ложиться значения
  labs(x = "Квантили стандартного нормального распределения", y = "Квантили набора данных")
quantile_plot
```



Используется, чтобы оценить форму распределения.

Если точки лежат на одной прямой - нормальное распределение.

- Небольшие отклонения от нормального распределения, но мало точек, чтобы оценить с уверенностью

Мощность линейной регрессии

Величина эффекта из общих соображений

```r
library(pwr)
cohen.ES(test="f2",size="large")
```

```
##
##      Conventional effect size from Cohen (1982)
##
##             test = f2
##             size = large
##      effect.size = 0.35
```

Величину эффекта можно оценить по R^2

$$f^2 = \frac{R^2}{1 - R^2}$$

R^2 - коэффициент детерминации

Посчитайте

какой нужен объем выборки, чтобы с вероятностью 0.8 обнаружить зависимость при помощи простой линейной регрессии, если ожидается $R^2 = 0.6$?

$$f^2 = \frac{R^2}{1 - R^2}$$

```
pwr.f2.test()
```

Take home messages

- Условия применимости простой линейной регрессии должны выполняться, чтобы тестировать гипотезы

    1. Независимость
    2. Линейность
    3. Нормальное распределение
    4. Гомогенность дисперсий

- Мощность линейной регрессии можно рассчитать как мощность F-критерия. Величину эффекта можно оценить по R^2

Дополнительные ресурсы

- Logan, 2010, pp. 170-207
- Quinn, Keough, 2002, pp. 92-104
- Open Intro to Statistics, pp. 315-353.