

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
- Регрессия по I и II модели

Вы сможете

- Проверить условия применимости простой линейной регрессии
- Рассчитать мощность линейной регрессии
- Объяснить, какие данные подходят для расчета регрессии по I и II модели
- Отличать случаи, когда обычная регрессия методом наименьших квадратов применима к данным, собранным для II модели
- Рассчитывать коэффициенты регрессии по II модели методом RMA (Ranged Major Axis), их стандартные ошибки, и записывать их в виде уравнения.

Пример: усыхающие личинки мучных хрущаков

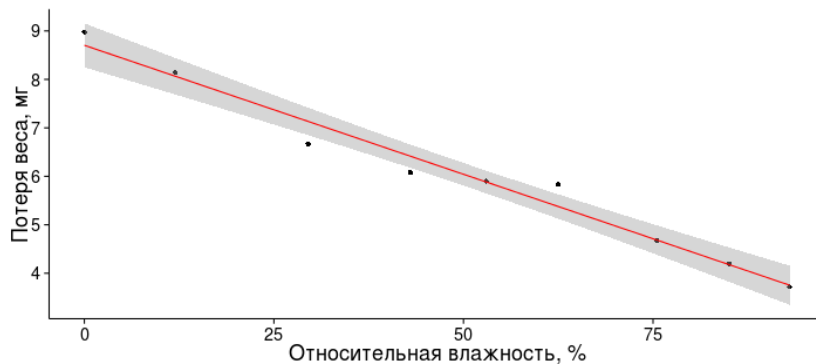
Как зависит потеря влаги личинками [малого мучного хрущака](#) *Tribolium confusum* от влажности воздуха? (Nelson, 1964)

```
# 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 = ".")
```



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
```

Зависимость потери веса от влажности можно описать уравнением

Для этого подставим коэффициенты в уравнение линейной регрессии $y = b_0 + b_1x$

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

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

$weightloss = 8.7 - 0.05 \text{ humidity}$

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

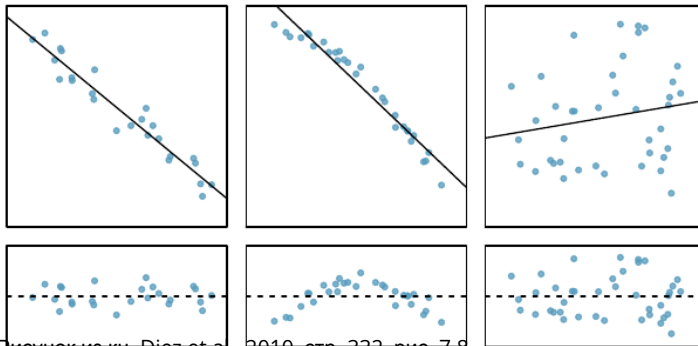


Рисунок из кн. Diez et al., 2010, стр. 332, рис. 7.8

2. Линейность связи

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

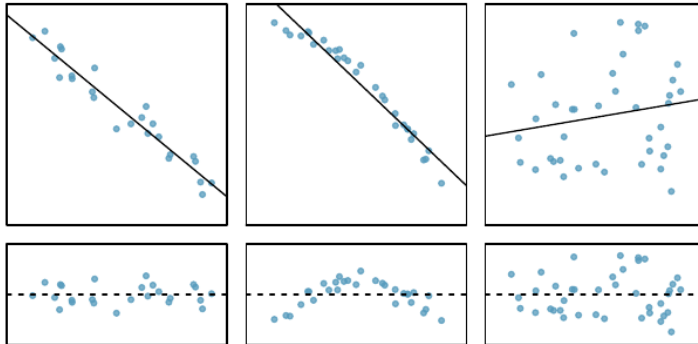


Рисунок из кн. Diez et al., 2010, стр. 332, рис. 7.8

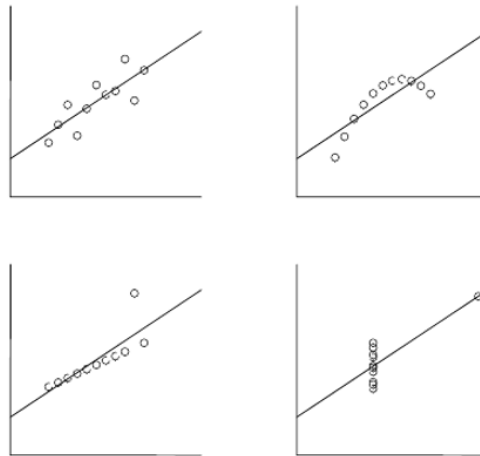
Вот, что бывает, если неглядя применять линейную регрессию

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

$$y_i = 3.0 + 0.5x_i,$$

$$r^2 = 0.68,$$

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



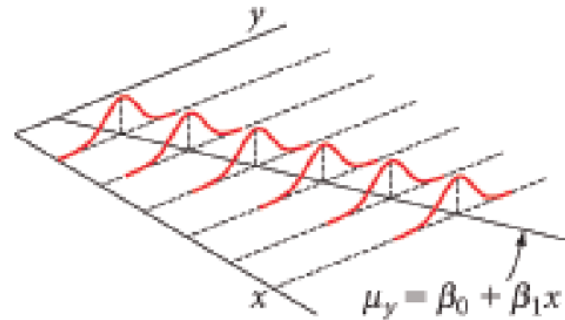
3. Нормальное распределение

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



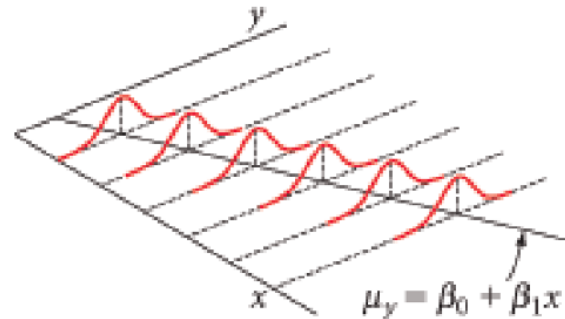
4. Гомогенность дисперсий

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

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

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

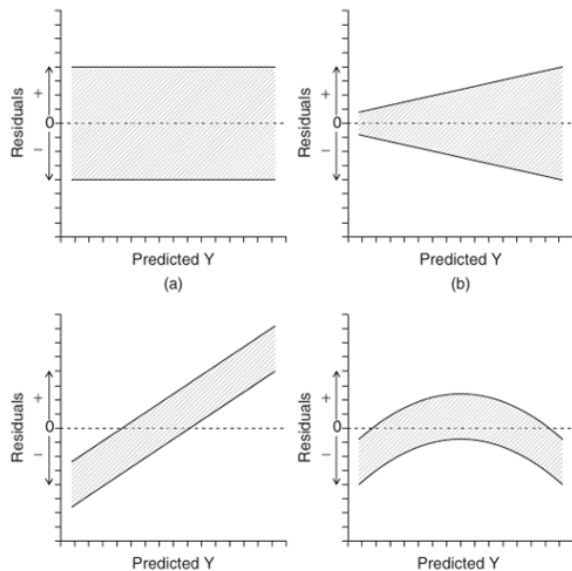
- К счастью, поскольку $\epsilon_i \sim N(0, \sigma^2)$, можно проверить равенство дисперсий остатков ϵ_i



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

Рисунок из кн. Watkins et al., 2008, стр. 743, рис. 17.4

Диагностика регрессии по графикам остатков



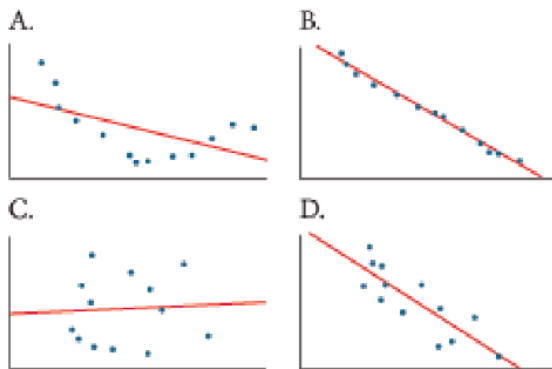
• условия:

- а - 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

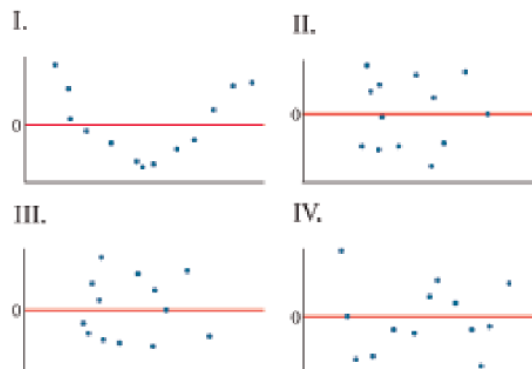
Скажите,

- 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 scatter plots.

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

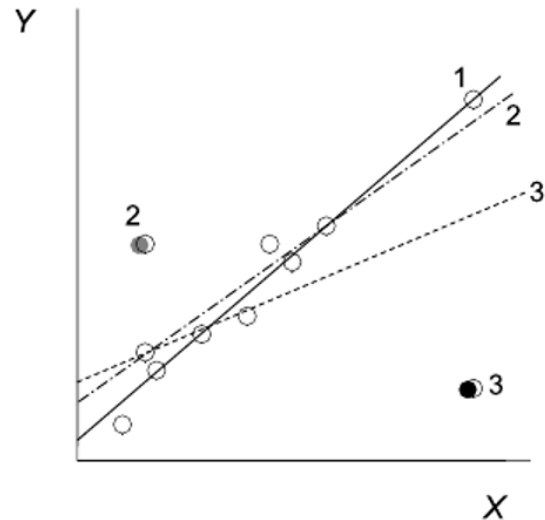


Display 3.85 Four residual plots.

Какие наблюдения влияют на ход регрессии больше других?

Влиятельные наблюдения, выбросы, outliers

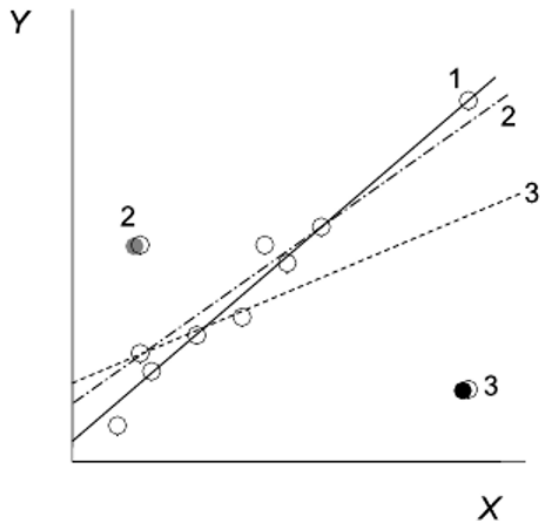
- большая абсолютная величина остатка
- близость к краям области определения ([leverage](#) - рычаг, "сила"; иногда называют \hat{h})
- 1 - не влияет
- 2 - умеренно влияет (большой остаток, малая сила влияния)
- 3 - очень сильно влияет (большой остаток, большая сила влияния)



Как оценить влияние наблюдений

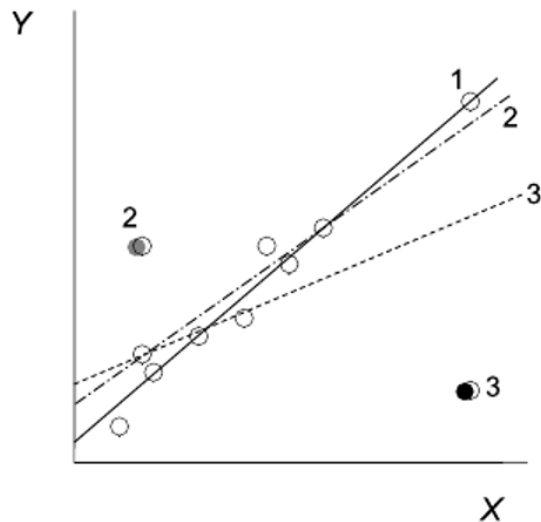
Расстояние Кука (Cook's d) (Cook, 1977)

- 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).
 - Что делать с влиятельными точками?
 - Проверить, не ошибка ли это. Если это не ошибка, не удалять - обсуждать!
- 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



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

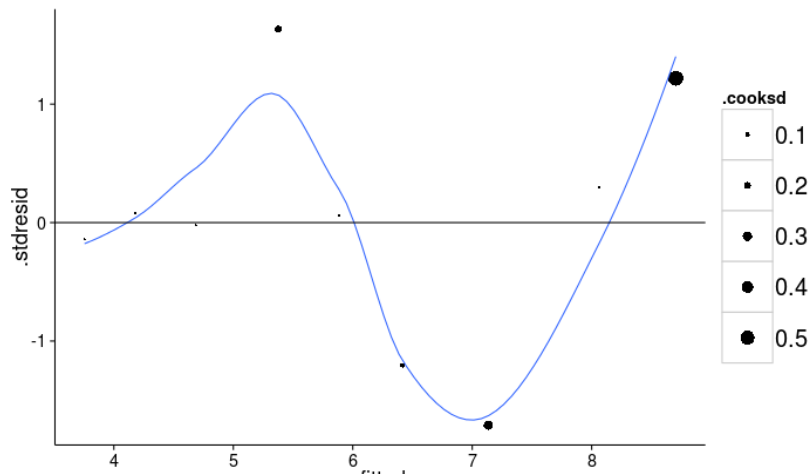
Используйте данные из `nelson_diag`

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

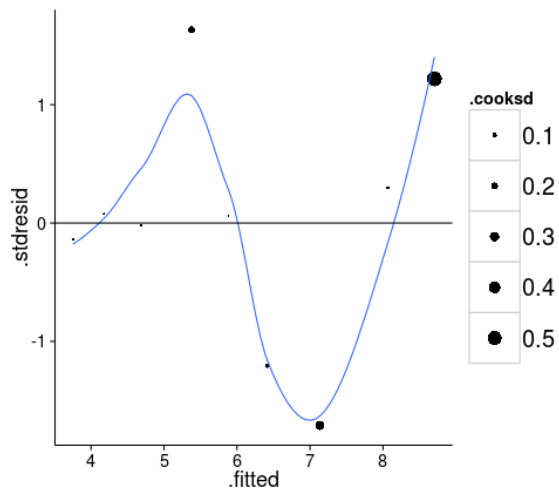
```
ggplot()
```


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  $y = 0$ 
```



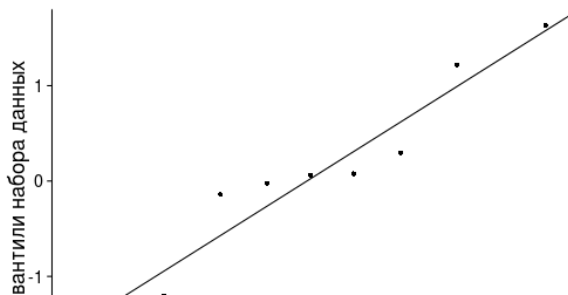
Какие выводы можно сделать по графику остатков?



- 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

Мощность линейной регрессии

Величина эффекта из общих соображений

```
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()
```

Регрессия по I и II модели

Регрессия по I и II модели

- I модель
 - x_i - фиксированные факторы, заранее заданные значения
 - предсказывать можно только для существующих в модели значений x_i
 - используется
 - метод наименьших квадратов (Ordinary Least Squares, **OLS**)
 - Предсказания и тесты гипотез по I модели иногда применимы и к случайным факторам (Quinn Keough 2002).
- II модель
 - x_i - случайные факторы, значения неизвестны заранее
 - предсказывать можно для любых значений x_i
 - используется
 - метод главных осей (Major Axis, **MA**)
 - метод сжатых главных осей (Ranged Major Axis, **RMA**)
 - Если главная цель точные **оценки коэффициентов и их сравнение, обязательно II модель.**

Сравнение OLS, MA и RMA регрессии

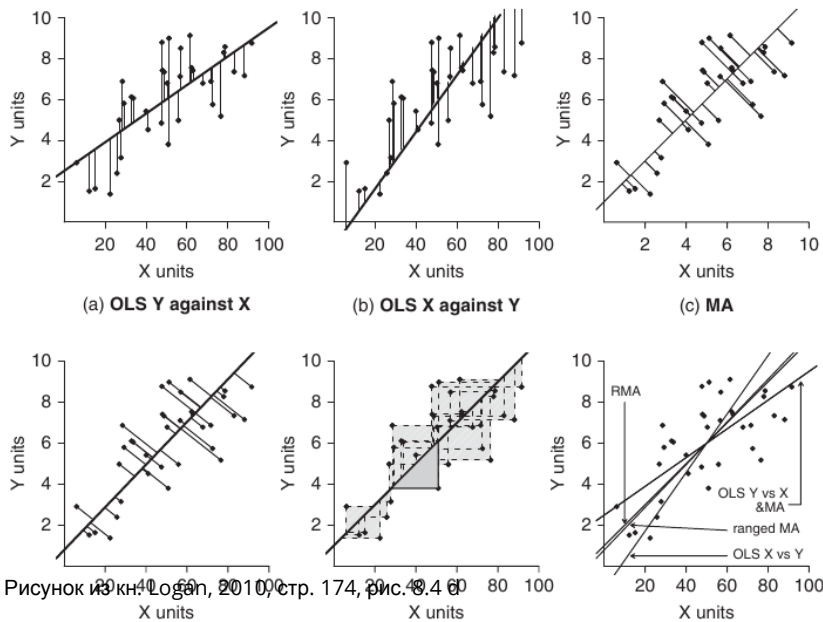


Рисунок из кн.: Logan, 2010, стр. 174, рис. 8.4 б

Пример: морфометрия поссумов

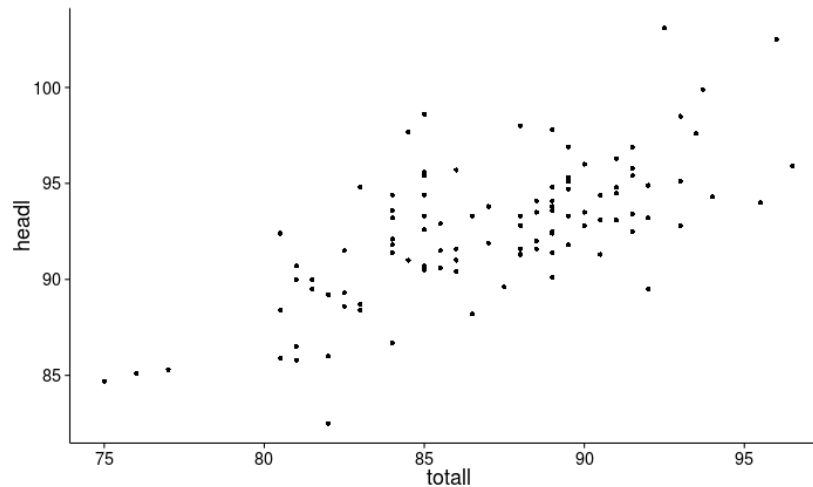
```
wb <- loadWorkbook("./data/possum-small.xls")
possum <- readWorksheet(wb, sheet = 1)
## или из .csv
# possum <- read.table(file="./data/possum-small.csv", header = TRUE,
#   sep = "\t", dec = ".")
```

```
str(possum)
```

```
## 'data.frame':   104 obs. of  8 variables:
## $ site   : num  1 1 1 1 1 1 1 1 1 1 ...
## $ pop    : chr  "Vic" "Vic" "Vic" "Vic" ...
## $ sex    : chr  "m" "f" "f" "f" ...
## $ age    : num  8 6 6 6 2 1 2 6 9 6 ...
## $ headl  : num  94.1 92.5 94 93.2 91.5 93.1 95.3 94.8 93.4 91.8 ...
## $ skullw: num  60.4 57.6 60 57.1 56.3 54.8 58.2 57.6 56.3 58 ...
## $ totall: num  89 91.5 95.5 92 85.5 90.5 89.5 91 91.5 89.5 ...
## $ taill  : num  36 36.5 39 38 36 35.5 36 37 37 37.5 ...
```

Зависит ли длина головы поссумов от общей длины тела?

```
ggplot(data = possum, aes(x = totall, y = headl)) + geom_point()
```



- Общая длина тела (headl) - случайная переменная,

RMA-регрессия (Ranged Major Axis Regression, RMA)

```
# install.packages("lmodel2")  
library(lmodel2)  
possum_rma <- lmodel2(headl ~ totall, data = possum, range.y="relative",  
                      range.x = "relative", nperm = 100)  
possum_rma
```

```

##
## Model II regression
##
## Call: lmodel2(formula = headl ~ totall, data = possum, range.y =
## "relative", range.x = "relative", nperm = 100)
##
## n = 104    r = 0.691    r-square = 0.478
## Parametric P-values:  2-tailed = 4.68e-16    1-tailed = 2.34e-16
## Angle between the two OLS regression lines = 20.4 degrees
##
## Permutation tests of OLS, MA, RMA slopes: 1-tailed, tail corresponding to sign
## A permutation test of r is equivalent to a permutation test of the OLS slope
## P-perm for SMA = NA because the SMA slope cannot be tested
##
## Regression results
##   Method Intercept      Slope  Angle (degrees)  P-perm (1-tailed)
## 1    OLS      42.7      0.573         29.8         0.0099
## 2     MA      26.1      0.764         37.4         0.0099
## 3    SMA      20.4      0.829         39.7          NA
## 4    RMA      27.9      0.743         36.6         0.0099
##
## Confidence intervals
##   Method 2.5%-Intercept 97.5%-Intercept 2.5%-Slope 97.5%-Slope
## 1    OLS          32.45          53.0         0.455         0.691
## 2     MA          11.25          38.9         0.617         0.934
## 3    SMA           NA           NA          NA           NA
## 4    RMA           NA           NA          NA           NA

```

Подставим коэффициенты в уравнение линейной регрессии

$$y = b_0 + b_1x$$

```
possum_rma$regression.results # Коэффициенты регрессии, нас интересует RMA
```

##	Method	Intercept	Slope	Angle (degrees)	P-perm (1-tailed)
## 1	OLS	42.7	0.573	29.8	0.0099
## 2	MA	26.1	0.764	37.4	0.0099
## 3	SMA	20.4	0.829	39.7	NA
## 4	RMA	27.9	0.743	36.6	0.0099

$$headl = 27.89 + 0.74 \text{ totall}$$

или в более академических обозначениях:

$$y = 27.89 + 0.74 x, R^2 = 0.478$$

Длина головы достоверно зависит от общей длины туловища (RMA-регрессия, $\beta_1 = 0.74 \pm 0.15$, $p < 0.01$)

График RMA-регрессии

```
plot(possum_rma, "RMA", main = "",  
      xlab = "Общая длина, см", ylab = "Длина головы, мм")
```

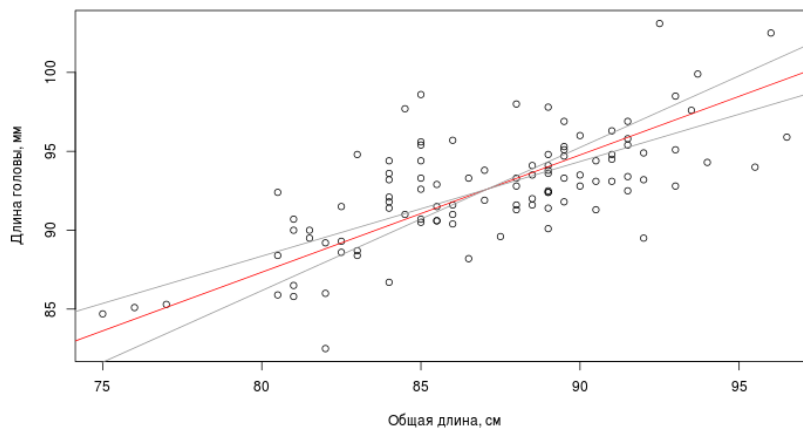
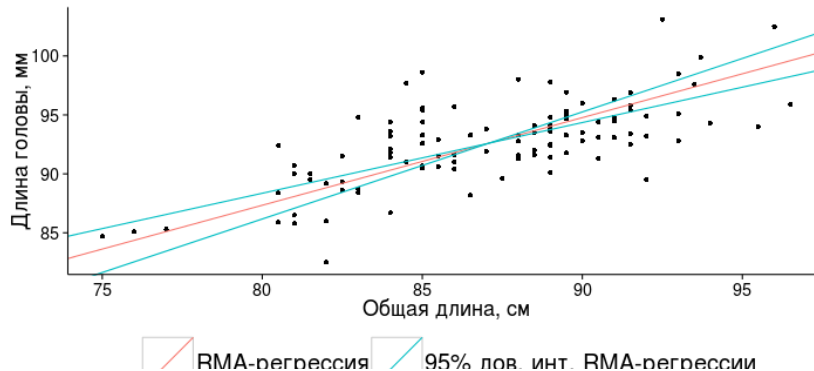


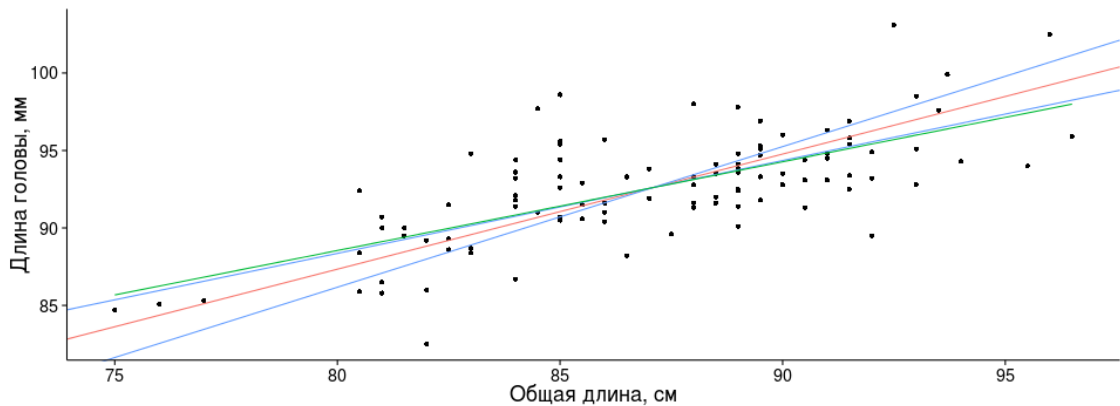
График RMA-регрессии




```
source(url("http://varmara.github.io/mathmethr-2013/w3-regression2/int_slope_lmodel2.R"))
reg_lines <- int_slope_lmodel2(possum_rma)
rma_plot <- ggplot(possum, aes(x = totalL, y = headL)) + geom_point() +
  geom_abline(data = reg_lines, aes(intercept = intercept, slope = slope,
    colour = c("blue", "red", "red")), show_guide = TRUE) +
  scale_color_discrete(name = "", labels = c("RMA-регрессия", "95% дов. инт. RMA-регрессии")) +
  labs(x = "Общая длина, см", y = "Длина головы, мм") + theme(legend.position = 'bottom')
rma_plot
```



Для сравнения - RMA- и обычная регрессия

```
rma_plot + geom_smooth(method = 'lm', se = FALSE, aes(colour = 'green'), show_guide = FALSE) +  
  scale_colour_discrete(name = "Линии:",  
    labels = c("RMA-регрессия", "OLS-регрессия", "95% дов. инт. RMA-регрессии"))
```



Линии:  RMA-регрессия  OLS-регрессия  95% дов. инт. RMA-регрессии

А можно ли использовать метод наименьших квадратов (OLS), если данные собраны по II модели, ?

- Можно, если :
 - Ошибка в оценке y_i >> ошибки в оценке x_i
 - Распределение y и x **не** многомерное нормальное
 - Зависимость y от x линейная
- Если цель предсказание y для x , то :
 - можно использовать OLS-оценки коэффициентов
 - нельзя - стандартные ошибки, доверительные интервалы, тесты параметров

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
- I модель. Фиксированные факторы, заранее заданные значения x_i , метод наименьших квадратов (OLS)
- II модель. Случайные факторы, значения x_i неизвестны заранее, метод главных осей (MA), метод сжатых главных осей (RMA)
- Предсказания и тесты гипотез по I модели иногда применимы и к случайным факторам (Quinn Keough 2002). Но если главная цель точные **оценки коэффициентов и их**

Дополнительные ресурсы

- Logan, 2010, pp. 170-207
- Quinn, Keough, 2002, pp. 92-104
- [Open Intro to Statistics](#), pp. 315-353.