

Регрессионный анализ, часть 1

Математические методы в зоологии - на R, осень 2013

Марина Варфоломеева
Каф. Зоологии беспозвоночных, СПбГУ

Знакомимся с линейными моделями

- Модель простой линейной регрессии
- Проверка валидности модели
- Оценка качества подгонки модели

Вы сможете

- подобрать модель линейной регрессии и записать ее в виде уравнения
- проверить валидность модели при помощи t - или F -теста
- оценить долю изменчивости, которую объясняет модель, при помощи R^2

Модель простой линейной регрессии

Линейная регрессия

- простая

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

- множественная

$$Y_i = \beta_0 + \beta x_{1i} + \beta x_{2i} + \dots + \epsilon_i$$

Запись моделей в R

`зависимая_переменная ~ модель`

$\hat{y}_i = b_0 + bx_i$ (простая линейная регрессия со свободным членом (intercept))

- `Y ~ X`
- `Y ~ 1 + X`
- `Y ~ X + 1`

$\hat{y}_i = bx_i$ (простая линейная регрессия без свободного члена)

- `Y ~ X - 1`
- `Y ~ -1 + X`

$\hat{y}_i = b_0$ (уменьшенная модель, линейная регрессия у от свободного члена)

- `Y ~ 1`
- `Y ~ 1 - X`

Запишите в нотации R

эти модели линейных регрессий

- $\hat{y}_i = b_0 + bx_{1i} + bx_{2i} + bx_{3i}$ (множественная линейная регрессия со свободным членом)
- $\hat{y}_i = b_0 + bx_{1i} + bx_{3i}$ (уменьшенная модель множественной линейной регрессии, без X_2)

Минимизируем остаточную изменчивость

$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ - модель регрессии

$\hat{y}_i = b_0 + b_1 x_i$ - оценка модели

нужно оценить β_0 , β_1 и σ^2

- Метод наименьших квадратов (Ordinary Least Squares, см. рис.)

Еще есть методы максимального правдоподобия (Maximum Likelihood, REstricted Maximum Likelihood)

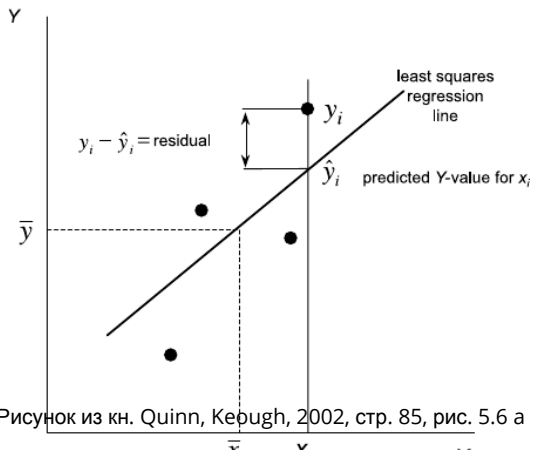


Рисунок из кн. Quinn, Kebugh, 2002, стр. 85, рис. 5.6 а

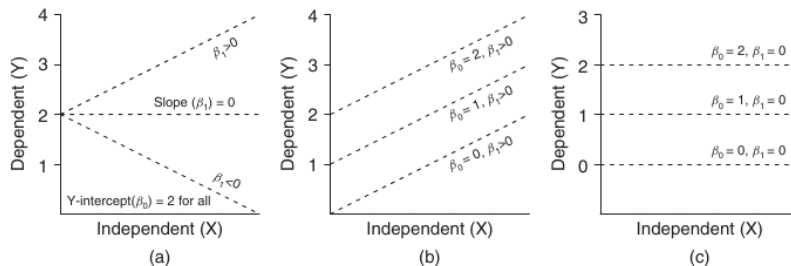
Оценки параметров линейной регрессии

минимизируют $\sum (y_i - \hat{y}_i)^2$, т.е. остатки.

ПАРАМЕТРЫ	ОЦЕНКИ ПАРАМЕТРОВ	СТАНДАРТНЫЕ ОШИБКИ ОЦЕНОК
β_1	$b_1 = \frac{\sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^n (x_i - \bar{x})^2}$	$SE_{b_1} = \sqrt{\frac{MS_e}{\sum_{i=1}^n (x_i - \bar{x})^2}}$
β_0	$b_0 = \bar{y} - b_1 \bar{x}$	$SE_{b_0} = \sqrt{MS_e \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}$
ϵ_i	$e_i = y_i - \hat{y}_i$	$\approx \sqrt{MS_e}$

- Стандартные ошибки коэффициентов нужны
 - для построения доверительных интервалов
 - для статистических тестов

Коэффициенты регрессии



- Если нужно сравнивать - лучше стандартизованные (= "бета коэффициенты") коэффициенты (на след. лекции про сравнение моделей)
 - $b_1^* = \frac{b_1 \sigma_x}{\sigma_y}$
 - не зависят от масштаба

Пример: усыхающие личинки мучных хрущаков

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

9 экспериментов, продолжительность 6 дней

- разная относительная влажность воздуха, % (humidity)
- измерена потеря влаги, мг (weightloss)

Данные в файлах `nelson.xlsx` и `nelson.csv`



Читаем данные из файла и знакомимся с ними

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.xlsx", header = TRUE, sep = "\\t", dec = ".")
```

```
str(nelson)
```

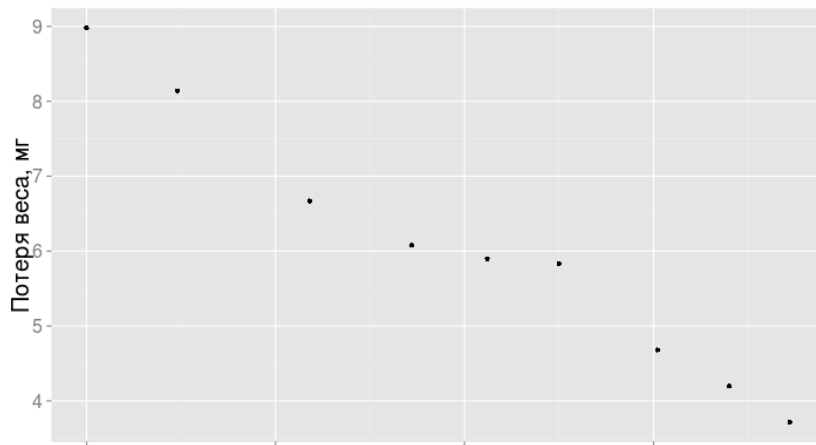
```
## 'data.frame':   9 obs. of  2 variables:
## $ humidity : num  0 12 29.5 43 53 62.5 75.5 85 93
## $ weightloss: num  8.98 8.14 6.67 6.08 5.9 5.83 4.68 4.2 3.72
```

```
head(nelson)
```

```
##   humidity weightloss
## 1      0.0         8.98
## 2     12.0         8.14
```

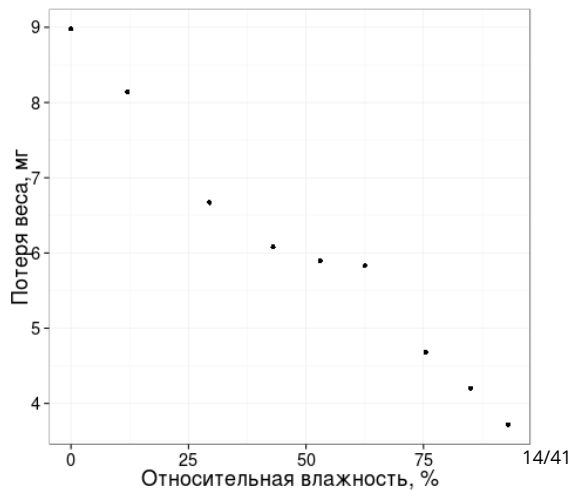
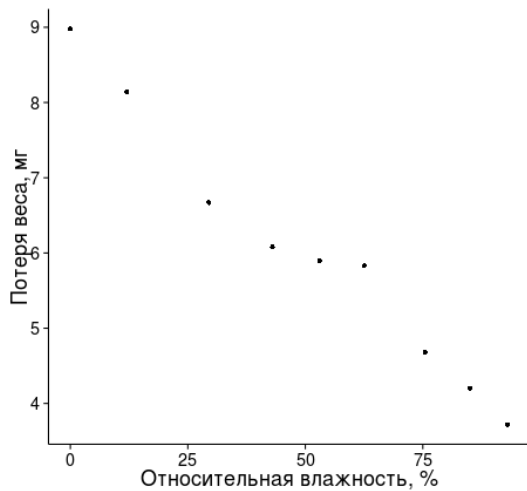
Как зависит потеря веса от влажности? График рассеяния.

```
library(ggplot2)
p_nelson <- ggplot(data=nelson, aes(x = humidity, y = weightloss)) +
  geom_point() +
  labs(x = "Относительная влажность, %", y = "Потеря веса, мг")
p_nelson
```



Внешний вид графиков можно менять при помощи тем

```
p_nelson + theme_classic()  
p_nelson + theme_bw()  
theme_set(theme_classic()) # устанавливаем понравившуюся тему до конца сессии
```



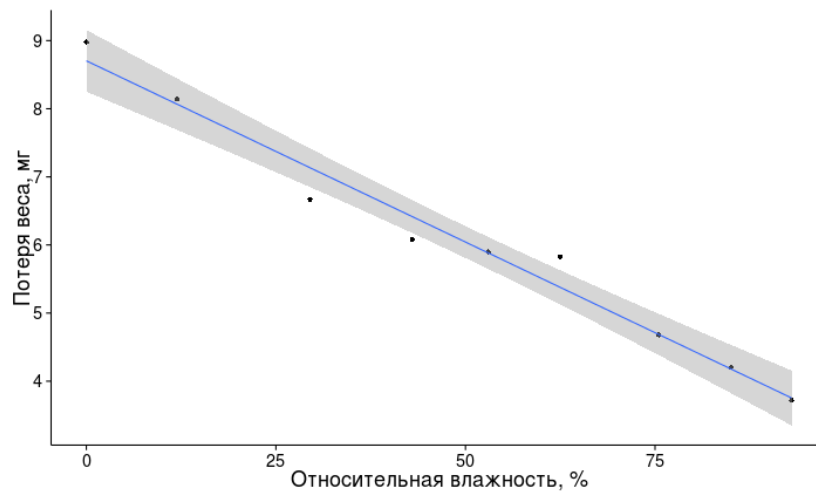
Подбираем параметры линейной модели

```
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
```

Добавим линию регрессии на график

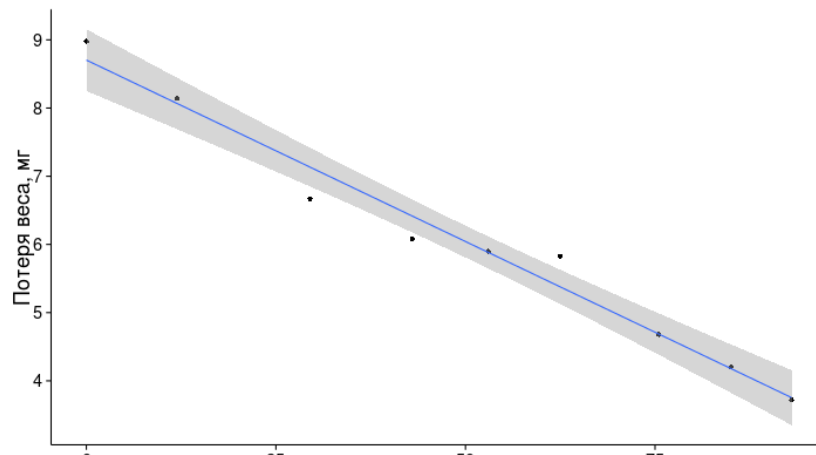
```
p_nelson + geom_smooth(method = "lm")
```



Как вы думаете,

что это за серая область вокруг линии регрессии?

```
p_nelson + geom_smooth(method = "lm")
```



Неопределенность оценок коэффициентов и предсказанных значений

Неопределенность оценок коэффициентов

- **Доверительный интервал коэффициента**
 - зона, в которой с $(1 - \alpha) \cdot 100\%$ вероятностью содержится среднее значение коэффициента
 - $b_1 \pm t_{\alpha, df=n-2} SE_{b_1}$
 - $\alpha = 0.05 \Rightarrow (1 - 0.05) \cdot 100\% = 95\%$ интервал
- **Доверительная зона регрессии**
 - зона, в которой с $(1 - \alpha) \cdot 100\%$ вероятностью лежит регрессионная прямая

Находим доверительные интервалы коэффициентов

```
# Вспомните, в выдаче summary(nelson_lm) были только оценки коэффициентов  
# и стандартные ошибки
```

```
# оценки коэффициентов отдельно  
coef(nelson_lm)
```

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

```
# доверительные интервалы коэффициентов  
confint(nelson_lm)
```

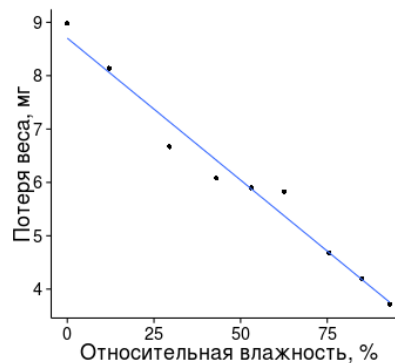
```
##           2.5 % 97.5 %  
## (Intercept) 8.2510 9.1570  
## humidity    -0.0609 -0.0455
```

Оценим, какова средняя потеря веса при заданной влажности

Нельзя давать оценки вне интервала значений X !

```
# новые данные для предсказания значений
newdata <- data.frame(humidity = c(50, 100))
predict(nelson_lm, newdata,
        interval = "confidence", se = TRUE)
```

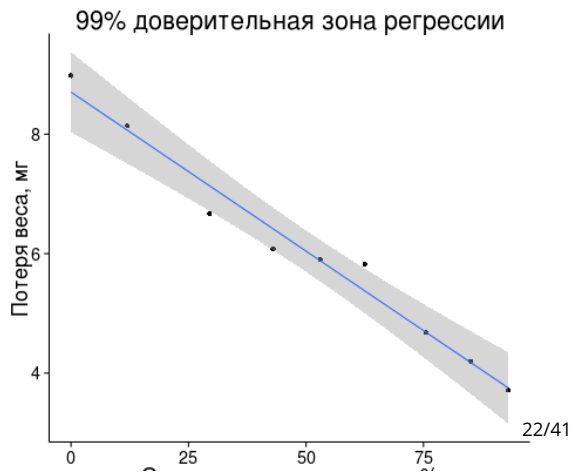
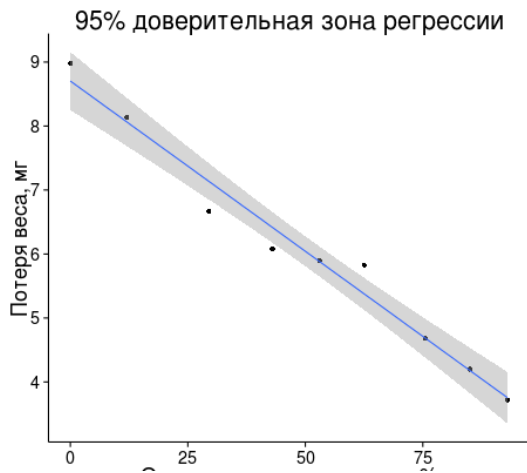
```
## $fit
##   fit lwr upr
## 1 6.04 5.81 6.28
## 2 3.38 2.93 3.83
##
## $se.fit
##      1      2
## 0.0989 0.1894
##
## $df
## [1] 7
##
## $residual.scale
## [1] 0.297
```



- При 50 и 100% относительной влажности ожидаемая средняя потеря веса жуков будет 6 ± 0.2 и 3.4 ± 0.4 , соответственно.

Строим доверительную зону регрессии

```
p_nelson + geom_smooth(method = "lm") +  
  ggtitle ("95% доверительная зона регрессии")  
p_nelson + geom_smooth(method = "lm", level = 0.99) +  
  ggtitle ("99% доверительная зона регрессии")
```



Неопределенность оценок предсказанных значений

- **Доверительный интервал к предсказанному значению**

- зона в которую попадают $(1 - \alpha) \cdot 100\%$ значений \hat{y}_i при данном x_i
- $\hat{y}_i \pm t_{0.05, n-2} SE_{\hat{y}_i}$
- $SE_{\hat{y}} = \sqrt{MS_e \left[1 + \frac{1}{n} + \frac{(x_{prediction} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}$

- **Доверительная область значений регрессии**

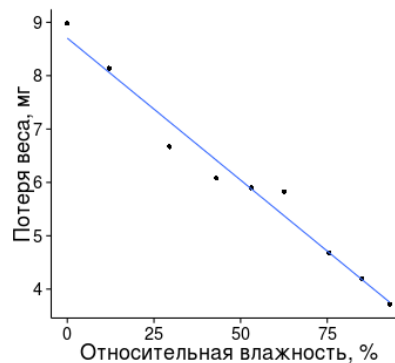
- зона, в которую попадает $(1 - \alpha) \cdot 100\%$ всех предсказанных значений

Предсказываем для новых значений

Нельзя использовать для предсказаний вне интервала значений X !

```
# новые данные для предсказания значений
newdata <- data.frame(humidity = c(50, 100))
predict(nelson_lm, newdata,
        interval = "prediction", se = TRUE)
```

```
## $fit
##   fit lwr upr
## 1 6.04 5.30 6.78
## 2 3.38 2.55 4.21
##
## $se.fit
##      1      2
## 0.0989 0.1894
##
## $df
## [1] 7
##
## $residual.scale
## [1] 0.297
```



- У 95% жуков при 50 и 100% относительной влажности будет потеря веса будет в пределах 6 ± 0.7 и 3.4 ± 0.8 , соответственно.

Данные для доверительной области значений

```
# предсказанные значения для исходных данных  
predict(nelson_lm, interval = "prediction")
```

```
## Warning: predictions on current data refer to _future_ responses
```

```
##   fit   lwr   upr  
## 1 8.70 7.87 9.54  
## 2 8.07 7.27 8.86  
## 3 7.13 6.38 7.89  
## 4 6.42 5.67 7.16  
## 5 5.88 5.14 6.62  
## 6 5.38 4.63 6.12  
## 7 4.69 3.92 5.45  
## 8 4.18 3.39 4.97  
## 9 3.75 2.95 4.56
```

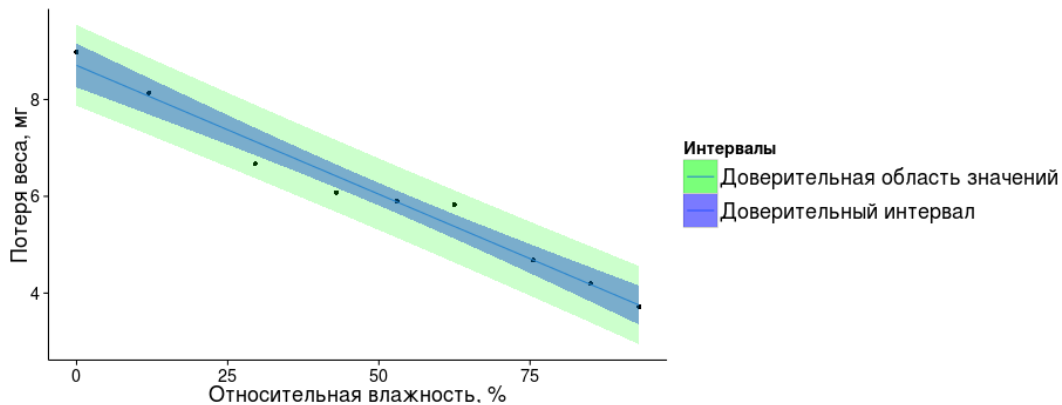
```
# объединим с исходными данными в новом датафрейме - для графиков  
nelson_with_pred <- data.frame(nelson, predict(nelson_lm, interval = "prediction"))
```

```
## Warning: predictions on current data refer to _future_ responses
```

25/41

Строим доверительную область значений и доверительный интервал

```
p_nelson + geom_smooth(method = "lm", aes(fill = "Доверительный интервал"), alpha = 0.4) +  
  geom_ribbon(data = nelson_with_pred,  
            aes(y = fit, ymin = lwr, ymax = upr, fill = "Доверительная область значений"),  
            alpha = 0.2) +  
  scale_fill_manual("Интервалы", values = c('green', 'blue'))
```



$H_0 : \beta_1 = 0$
или t-, или F-тест

Проверка валидности модели

Проверка при помощи t-критерия

$$H_0 : b_1 = \theta, \theta = 0$$

$$t = \frac{b_1 - \theta}{SE_{b_1}}$$

$$df = n - 2$$

Проверка коэффициентов с помощью t-критерия есть в сводке модели

```
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
```

Проверка при помощи F-критерия

$$H_0 : \beta_1 = 0$$

- Та же самая нулевая гипотеза. Как так получается?

Общая изменчивость - отклонения от общего среднего значения

SS_{total}

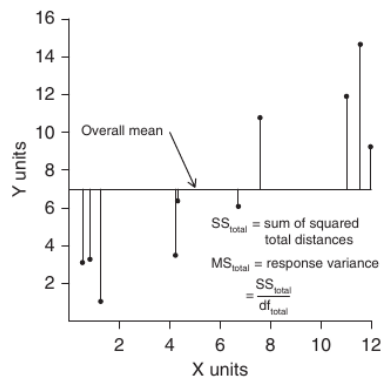


Рисунок из кн. Logan, 2010, стр. 172, рис. 8.3 а

$$SS_{total} = SS_{regression} + SS_{error}$$

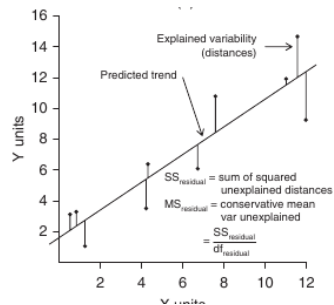
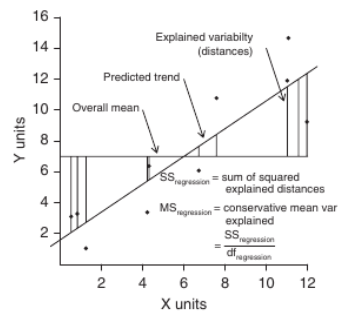
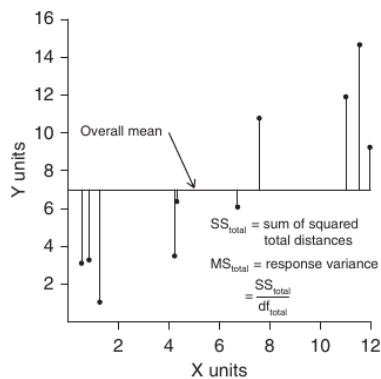


Рисунок из кн. Logan, 2010, стр. 172, рис. 8.3 а-с

Если зависимости нет, $b_1 = 0$

Тогда $\hat{y}_i = \bar{y}_i$
и $MS_{regression} \approx MS_{error}$

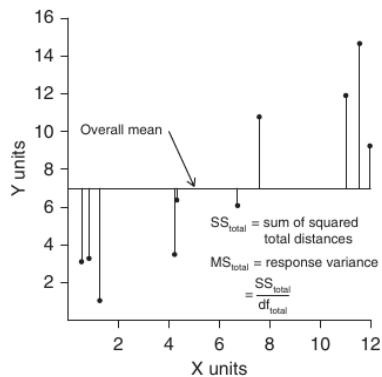
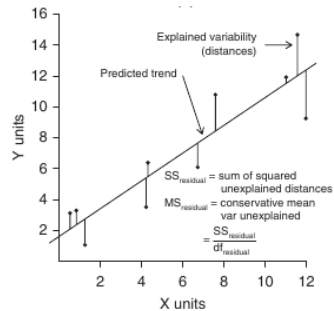
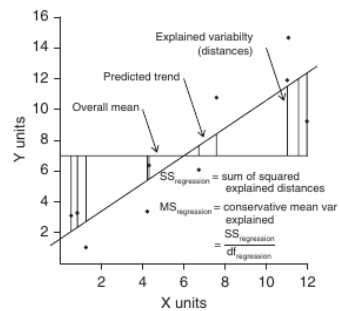
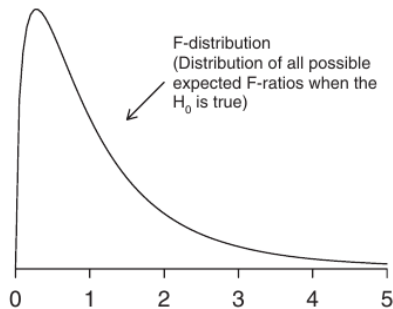


Рисунок из кн. Logan, 2010, стр. 172, рис. 8.3 а-с



F-критерий и распределение F-статистики

$$F = \frac{\text{Объясненная изменчивость}}{\text{Необъясненная изменчивость}} = \frac{MS_{\text{regression}}}{MS_{\text{error}}}$$



Зависит от

- α
- $df_{\text{regression}}$
- df_{error}

F-распределение при $H_0 : b_1 = 0$

Таблица результатов дисперсионного анализа

ИСТОЧНИК ИЗМЕНЧИВОСТИ	СУММЫ КВАДРАТОВ ОТКЛОНЕНИЙ, SS	ЧИСЛО СТЕПЕНЕЙ СВОБОДЫ, DF	СРЕДНИЙ КВАДРАТ ОТКЛОНЕНИЙ, MS	F
Регрессия	$SS_r = \sum (\bar{y} - \hat{y}_i)^2$	$df_r = 1$	$MS_r = \frac{SS_r}{df_r}$	$F_{df_r, df_e} = \frac{MS_r}{MS_e}$
Остаточная	$SS_e = \sum (y_i - \hat{y}_i)^2$	$df_e = n - 2$	$MS_e = \frac{SS_e}{df_e}$	
Общая	$SS_t = \sum (\bar{y} - y_i)^2$	$df_t = n - 1$		

- Минимальное упоминание в тексте - F_{df_r, df_e}, p

Проверяем валидность модели при помощи F-критерия

```
nelson_aov <- aov(nelson_lm)
summary(nelson_aov)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## humidity     1  23.51   23.51     267 0.00000078 ***
## Residuals     7   0.62    0.09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Количество влаги, потерянной жуками в период эксперимента, достоверно зависело от уровня относительной влажности ($F_{1,7} = 267, p < 0.01$).

Оценка качества подгонки модели

Коэффициент детерминации

доля общей изменчивости, объясненная линейной связью x и y

$$R^2 = \frac{SS_r}{SS_t}$$

$$0 \leq R^2 \leq 1$$

Иначе рассчитывается как $R^2 = r^2$

Коэффициент детерминации

можно найти в сводке модели

- Осторожно, не сравнивайте R^2 моделей с разным числом параметров, для этого есть $R^2_{adjusted}$

```
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
##
```

Take home messages

- Модель простой линейной регрессии $y_i = \beta_0 + \beta_1 \dot{x}_i + \epsilon_i$
- В оценке коэффициентов регрессии и предсказанных значений существует неопределенность. Доверительные интервалы можно рассчитать, зная стандартные ошибки.
- Валидность модели линейной регрессии можно проверить при помощи t- или F-теста.
 $H_0 : \beta_1 = 0$
- Качество подгонки модели можно оценить при помощи коэффициента детерминации R^2

Дополнительные ресурсы

- Гланц, 1999, стр. 221-244
- Logan, 2010, pp. 170-207
- Quinn, Keough, 2002, pp. 78-110
- [Open Intro to Statistics: Chapter 7. Introduction to linear regression](#), pp. 315-353.
- Sokal, Rohlf, 1995, pp. 451-491
- Zar, 1999, pp. 328-355