Регрессионный анализ, часть 1

Математические методы в зоологии с использованием R

Марина Варфоломеева

1. **Описание зависимости между переменными**

2. **Линейная регрессия**

3. **Неопределенность оценок коэффициентов**

4. **Тестирование значимости модели и ее коэффициентов**

5. **Оценка качества подгонки модели**

Вы сможете

- посчитать и протестировать различные коэффициенты корреляции между переменными
- 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

Описание зависимости между переменными

Пример: потеря влаги личинками мучных хрущаков

Как зависит потеря влаги личинками малого мучного хрущака *Tribolium confusum* от влажности воздуха?

- 9 экспериментов, продолжительность 6 дней
- разная относительная влажность воздуха, %
- измерена потеря влаги, мг

Малый мучной хрущак *Tribolium confusum*, photo by Sarefo, CC BY-SA

Nelson, 1964; данные из Sokal, Rohlf, 1997, табл. 14.1 по Logan, 2010. глава 8, пример 8с; Данные в файлах nelson.xlsx и nelson.csv

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 файлы
download(
  url = "https://varmara.github.io/mathmethr/data/nelson.xlsx",
  destfile = "data/nelson.xlsx")
## или .csv
# download(
#   url = "https://varmara.github.io/mathmethr/data/nelson.csv",
#   destfile = "data/nelson.csv")
```

Читаем данные из файла

```
library(readxl)
nelson <- read_excel("data/nelson.xlsx", sheet = 1)
## или из .csv
# nelson <- read.table(file="data/nelson.csv", header = TRUE, sep = "\t", dec

# Все ли правильно открылось
str(nelson)
```

```
# Classes 'tbl_df', 'tbl' and '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)
```

```
# # A tibble: 6 × 2
#   humidity weightloss
#      <dbl>      <dbl>
# 1      0.0       8.98
# 2     12.0       8.14
# 3     29.5       6.67
# 4     43.0       6.08
```

Знакомимся с данными

```
# Есть ли пропущенные значения
sapply(nelson, function(x)sum(is.na(x)))
```

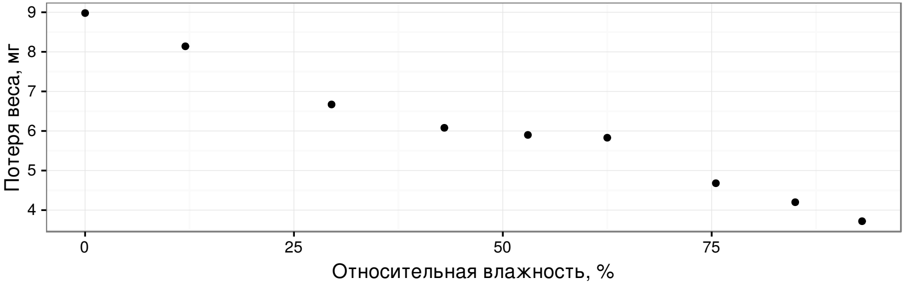
```
#   humidity weightloss
#          0          0
```

```
# Какой объем выборки?
nrow(nelson)
```

```
# [1] 9
```

Как зависит потеря веса от влажности?

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



Коэффициент корреляции — способ оценки силы связи между двумя переменными

Коэффициент корреляции Пирсона

- Оценивает только линейную составляющую связи
- Параметрические тесты (t-критерий) значимости применимы если переменные распределены нормально

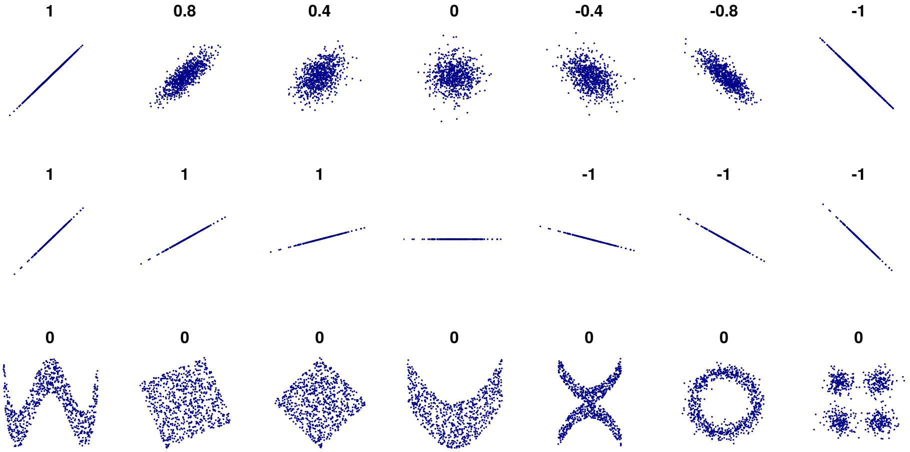
Ранговые коэффициенты корреляции (кор. Кендалла и кор. Спирмена)

- Не зависят от формы распределения переменных
- Тест на значимость непараметрический

Интерпретация коэффициента корреляции

$-1 < \rho < 1$ $|\rho| = 1$ — сильная связь $\rho = 0$ — нет связи

- В тестах для проверки значимости тестируется гипотеза $H_0 : \rho = 0$



Можно расчитать значение коэффициента корреляции между потерей веса и влажностью

```
p_cor <- cor.test(nelson$humidity, nelson$weightloss,
         alternative = "two.sided", method = "pearson")
p_cor
```

```
#
#   Pearson's product-moment correlation
#
# data:  nelson$humidity and nelson$weightloss
# t = -20, df = 7, p-value = 0.0000008
# alternative hypothesis: true correlation is not equal to 0
# 95 percent confidence interval:
#  -0.997 -0.938
# sample estimates:
#    cor
# -0.987
```

Можно расчитать значение коэффициента корреляции между потерей веса и влажностью

```
p_cor <- cor.test(nelson$humidity, nelson$weightloss,
         alternative = "two.sided", method = "pearson")
p_cor
```

```
#
#   Pearson's product-moment correlation
#
# data:  nelson$humidity and nelson$weightloss
# t = -20, df = 7, p-value = 0.0000008
# alternative hypothesis: true correlation is not equal to 0
# 95 percent confidence interval:
#  -0.997 -0.938
# sample estimates:
#    cor
# -0.987
```

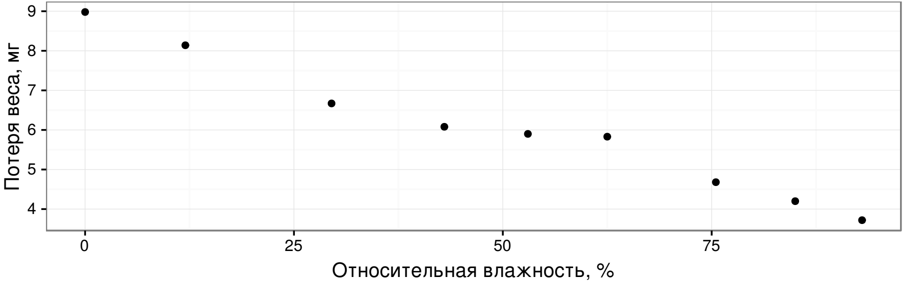
Можно описать результаты несколькими способами:

- Величина потери веса мучных хрущаков коррелирует с относительной влажностью воздуха ($r = -0.99$, $p < 0.01$)
- Мучные хрущаки теряют вес при уменьшении относительной влажности воздуха ($r = -0.99$, $p < 0.01$)

Коэффициент корреляции не позволяет предсказать значение одной переменной, зная знаячение другой

Нам бы хотелось описать функциональную зависимость

$$weightloss_i = b_0 + b_1 humidity_i$$



Линейная регрессия

Линейная регрессия

- простая

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- множественная

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \varepsilon_i$$

Как провести линию регрессии?

Линейная модель:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Оценка модели:

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

Нужно оценить параметры линейной модели:

- β_0
- β_1

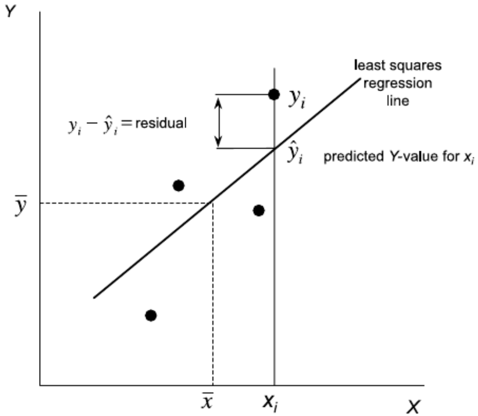
Методы оценки параметров:

- Метод наименьших квадратов (Ordinary Least Squares)
- Методы максимального правдоподобия (Maximum Likelihood, REstricted Maximum Likelihood)

Метод наименьших квадратов

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

Оценки параметров линейной регрессии подбирают так, чтобы минимизировать остатки $\sum (y_i - \hat{y}_i)^2$



Линия регрессии по методу наименьших квадратов
из кн. Quinn, Keough, 2002, стр. 85, рис. 5.6 а

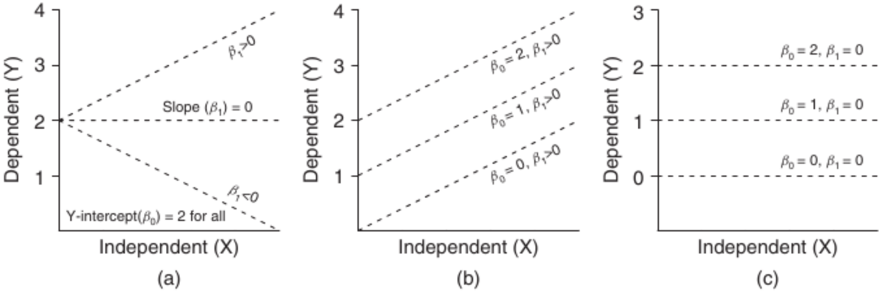
Оценки параметров линейной регрессии

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	$b_0 = \bar{y} - b_1\bar{x}$	$SE_{b_0} = \sqrt{MS_e[\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]}$

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

Таблица из кн. Quinn, Keough, 2002, стр. 86, табл. 5.2

Интерпретация коэффициентов регрессии



Интерпретация коэффициентов регрессии

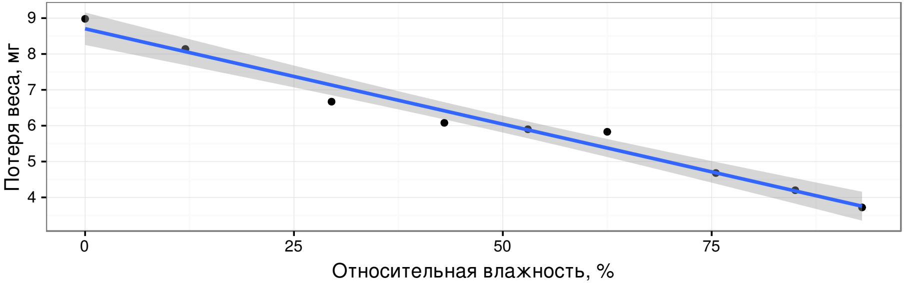
Рисунок из кн. Logan, 2010, стр. 170, рис. 8.2

Для сравнения разных моделей - стандартизованные коэффициенты

- Не зависят от масштаба измерений x и y
- Можно вычислить, зная обычные коэффициенты и их стандартные отклонения $b_1^* = b_1 \frac{\sigma_x}{\sigma_y}$
- Можно вычислить, посчитав регрессию по стандартизованным данным

Добавим линию регрессии на график

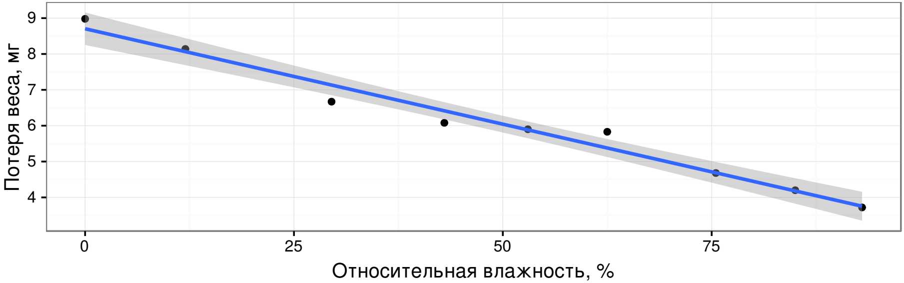
```
gg_nelson + geom_smooth(method = "lm")
```



Что это за серая область вокруг линии регрессии?

Добавим линию регрессии на график

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



Что это за серая область вокруг линии регрессии?

Доверительная зона регрессии

- 95% доверительная зона регрессии
- В ней с 95% вероятностью лежит регрессионная прямая
- Возникает из-за неопределенности оценок коэффициентов регрессии

Как в R задать формулу линейной регрессии

`lm(формула, данные)` - функция для подбора регрессионных моделей

Формат формулы: `зависимая_переменная ~ модель`

- $\hat{y}_i = b_0 + b_1 x_i$ (простая линейная регрессия с b_0 (intercept))
  - Y ~ X
  - Y ~ 1 + X
  - Y ~ X + 1
- $\hat{y}_i = b_1 x_i$ (простая линейная регрессия без b_0)
  - Y ~ X - 1
  - Y ~ -1 + X
- $\hat{y}_i = b_0$ (уменьшенная модель, линейная регрессия Y от b_0)
  - Y ~ 1
  - Y ~ 1 - X

Задача

Запишите в нотации R эти модели линейных регрессий

- $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i}$

(множественная линейная регрессия с b_0)

- $\hat{y}_i = b_0 + b_1 x_{1i} + b_3 x_{3i}$

(уменьшенная модель множественной линейной регрессии, без x_2)

Решение

- $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i}$

(множественная линейная регрессия с b_0)

```
Y ~ X1 + X2 + X3

Y ~ 1 + X1 + X2 + X3
```

- $\hat{y}_i = b_0 + b_1 x_{1i} + b_3 x_{3i}$

(уменьшенная модель множественной линейной регрессии, без x_2)

```
Y ~ X1 + X3

Y ~ 1 + X1 + X3
```

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
```

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
```

Коэффициенты линейной регрессии:

- $b_0 = 8.7$
- $b_1 = -0.05$

Неопределенность оценок коэффициентов

Неопределенность оценок коэффициентов

Доверительный интервал коэффициента

- зона, в которой с $(1-\alpha)\cdot 100\%$ вероятностью содержится среднее значение коэффициента
- $b_1 \pm t_{\alpha,df=n-2}SE_{b_1}$
- $\alpha = 0.05 => (1-0.05)\cdot 100\% = 95\%$ интервал

Доверительная зона регрессии

- зона, в которой с $(1-\alpha)\cdot 100\%$ вероятностью лежит регрессионная прямая

Находим доверительные интервалы коэффициентов

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

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

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

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

Предсказываем Y при заданном X

Какова средняя потеря веса при заданной влажности?

```
newdata <- data.frame(humidity = c(50, 100)) # значения, для которых предсказываем
(pr1 <- 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
```

Предсказываем Y при заданном X

Какова средняя потеря веса при заданной влажности?

```r
newdata <- data.frame(humidity = c(50, 100)) # значения, для которых предсказываем
(pr1 <- predict(nelson_lm, newdata, interval = "confidence", se = TRUE))
```

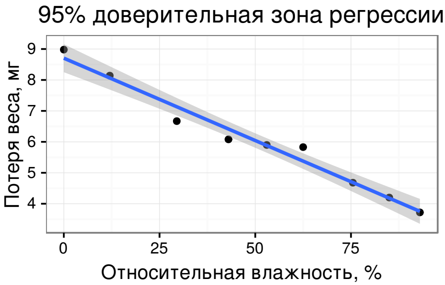
```
# $fit
#    fit  lwr  upr
# 1 6.04 5.81 6.28
# 2 3.38 2.93 3.83
#
# $se.fit
#      1      2
# 0.0989 0.1894
#
# $df
# [1] 7
#
# $residual.scale
# [1] 0.297
```

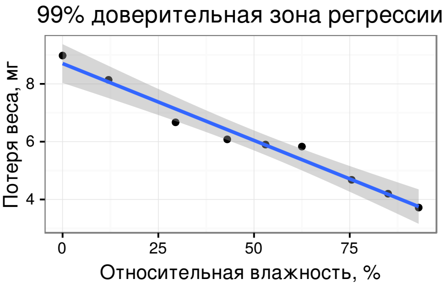
- При 50 и 100% относительной влажности ожидаемая средняя потеря веса жуков будет 6 $\pm$ 0.2 и 3.4 $\pm$ 0.4, соответственно.

Строим доверительную зону регрессии

```
gg_nelson + geom_smooth(method = "lm") +
  labs (title = "95% доверительная зона регрессии")

gg_nelson + geom_smooth(method = "lm", level = 0.99) +
  labs (title = "99% доверительная зона регрессии")
```





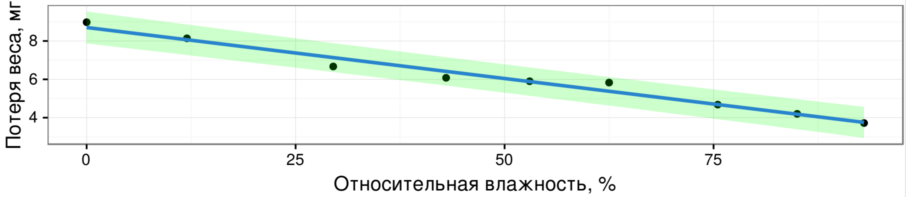
Неопределенность оценок предсказанных значений

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 изменение Y для 95% наблюдений при заданном X

В каких пределах находится потеря веса у 95% жуков при заданной влажности?

```
newdata <- data.frame(humidity = c(50, 100)) # новые данные для предсказания значений
(pr2 <- 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
```

Предсказываем изменение Y для 95% наблюдений при заданном X

В каких пределах находится потеря веса у 95% жуков при заданной влажности?

```
newdata <- data.frame(humidity = c(50, 100)) # новые данные для предсказания значений
(pr2 <- 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 объединим с исходными данными в новом датафрейме - для графиков

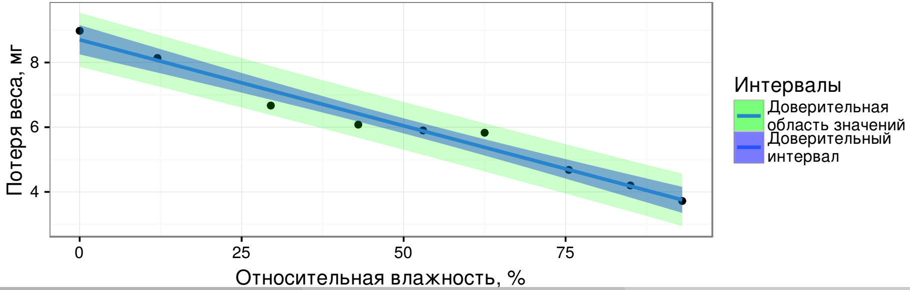
```
(pr_all <- predict(nelson_lm, interval = "prediction"))
```

```
#    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
```

```
nelson_with_pred <- data.frame(nelson, pr_all)
```

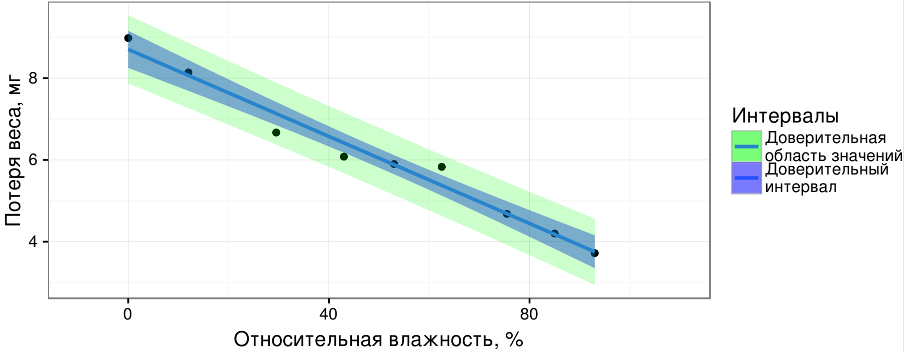
Строим доверительную область значений и доверительный интервал одновременно

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



Осторожно!

Вне интервала значений X ничего предсказать нельзя!



Тестирование значимости модели и ее коэффициентов

Тестируем коэффициенты t-критерием

t-критерий

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

$H_0 : b_1 = \theta$, для $\theta = 0$
Число степеней свободы $df = n - 2$

Тестируем значимость коэффициентов с помощью t-критерия

```
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
```

Тестируем значимость коэффициентов с помощью t-критерия

```r
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
```

- Увеличение относительной влажности привело к достоверному замедлению потери веса жуками ($b_1 = -0.053$, $t = -16.35$, $p < 0.01$)

Проверка при помощи F-критерия

F-критерий

$$F = \frac{MS_{regression}}{MS_{error}}$$

$H_0 : \beta_1 = 0$
Число степеней свободы $df_{regression}$, df_{error}

Общая изменчивость

Общая изменчивость - SS_{total}, отклонения от общего среднего значения

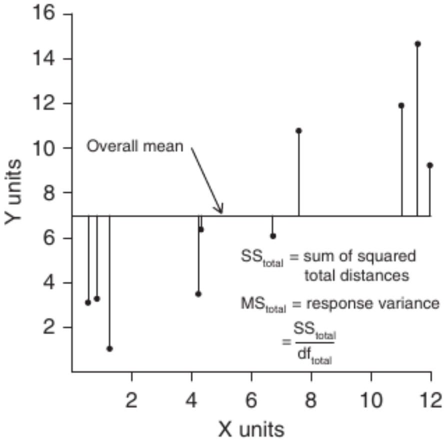
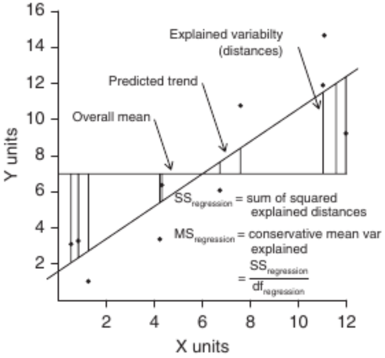


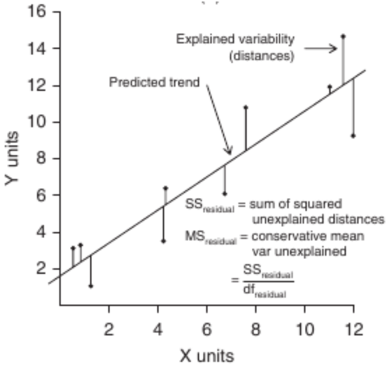
Рис. из кн. Logan, 2010, стр. 172, рис. 8.3

Общая изменчивость

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



Объясненная изменчивость

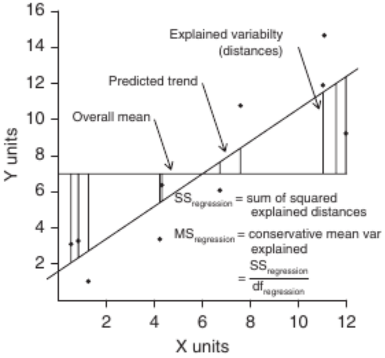


Остаточная изменчивость

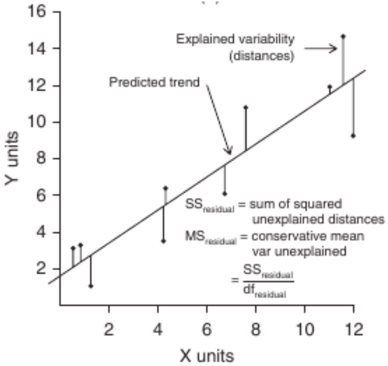
Рис. из кн. Logan, 2010, стр. 172, рис. 8.3

Если зависимости нет, $b_1 = 0$

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



Объясненная изменчивость



Остаточная изменчивость

Рис. из кн. Logan, 2010, стр. 172, рис. 8.3

Что оценивают средние квадраты отклонений?

Источник изменчивости	Суммы квадратов отклонений SS	Число степеней свободы df	Средний квадрат отклонений MS	Ожидаемый средний квадрат
Регрессия	$\sum(\bar{y} - \hat{y}_i)^2$	1	$\frac{\sum_{i=1}^{n}(\bar{y}-\hat{y}_i)^2}{1}$	$\sigma_\varepsilon^2 + {\beta_1}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$
Остаточная	$\sum(y_i - \hat{y}_i)^2$	$n - 2$	$\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$	σ_ε^2
Общая	$\sum(\bar{y} - y_i)^2$	$n - 1$		

Если $b_1 = 0$, тогда $\hat{y}_i = \bar{y}_i$ и $MS_{regression} \approx MS_{error}$

Тестируем:

$$F = \frac{MS_{regression}}{MS_{error}}$$

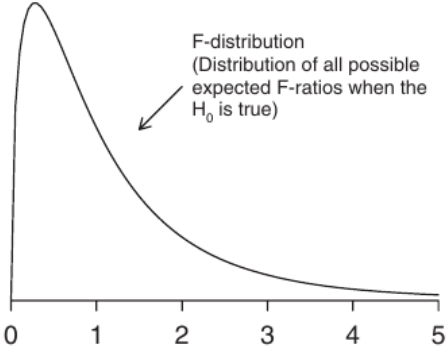
F-критерий и распределение F-статистики

F - соотношение объясненной и не объясненной изменчивости

$$F = \frac{MS_{regression}}{MS_{error}}$$

Зависит от

- α
- $df_{regression}$
- df_{error}



Распределение F-статистики при справедливой H_0
Рис. из кн. Logan, 2010, стр. 172, рис. 8.3

Таблица результатов дисперсионного анализа

Источник изменчивости	SS	df	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 результатов в тексте должно содержать F_{df_r,df_e} и 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

Коэффициент детерминации R^2

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 квадрат коэффициента корреляции $R^2 = r^2$

Коэффициент детерминации можно найти в сводке модели

```r
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
```

Сравнение качества подгонки моделей

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

Take home messages

- Модель простой линейной регрессии $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- В оценке коэффициентов регрессии и предсказанных значений существует неопределенность. Доверительные интервалы можно расчитать, зная стандартные ошибки.
- Значимость всей регрессии и ее параметров 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

- Учебники
  - Гланц, 1999, стр. 221-244
  - Open Intro to Statistics: Chapter 7. Introduction to linear regression, pp. 315-353.
  - Quinn, Keough, 2002, pp. 78-110
  - Logan, 2010, pp. 170-207
  - Sokal, Rohlf, 1995, pp. 451-491
  - Zar, 1999, pp. 328-355
- Упражнения для тренировки
  - OpenIntro Labs, Lab 7: Introduction to linear regression (Осторожно, они используют базовую графику а не `ggplot`)
    - Обычный вариант, упражнения 1—4
    - Интерактивный вариант на Data Camp, до вопроса 4