Регрессионный анализ, часть 1

Математические методы в зоологии с использованием R

Марина Варфоломеева

1. **Графики средствами пакета ggplot2**
2. **Корреляция**
3. **Линейная регрессия**
4. **Подбор коэффициентов линейной регрессии**
5. **Линейная регрессия в R**
6. **Тестирование значимости модели и ее коэффициентов**
7. **График линейной регрессии**
8. **Оценка качества подгонки модели**
9. **Использование линейной регрессии для предсказаний (для самостоятельного разбора)**

Вы сможете

- посчитать и протестировать различные коэффициенты корреляции между переменными
- подобрать модель линейной регрессии и записать ее в виде уравнения
- интерпретировать коэффициенты простой линейной регрессии
- протестировать значимость модели и ее коэффициентов при помощи t- или F-теста
- оценить долю изменчивости, которую объясняет модель, при помощи R^2

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 8c; Данные в файлах nelson.xlsx и nelson.csv

Читаем данные из файла

Чтение из xlsx

```
library(readxl)
nelson <- read_excel(path = "data/nelson.xlsx", sheet = 1)
```

Все ли правильно открылось?

```
str(nelson)      # Структура данных
```

```
# 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)     # Первые несколько строк файла
```

```
# # 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
# 5     53.0       5.90
# 6     62.5       5.83
```

Знакомимся с данными

Есть ли пропущенные значения?

```
colSums(is.na(nelson))
```

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

Каков объем выборки?

Поскольку пропущенных значений нет, можем просто посчитать число строк

```
nrow(nelson)
```

```
# [1] 9
```

Теперь все готово, чтобы мы могли ответить на вопрос, как зависит потеря веса от влажности?

Графики средствами пакета ggplot2

Грамматика графиков

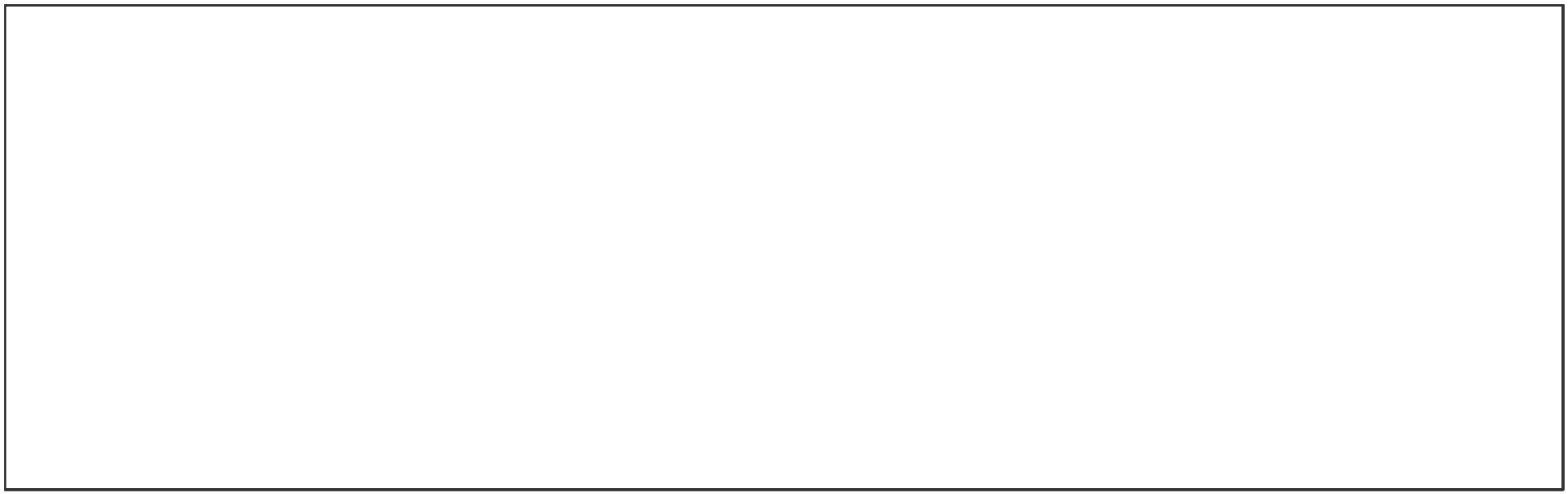
1. Откуда брать данные?
2. Какие переменные изображать на графике?
3. В виде чего изображать?
4. Какие подписи нужны?
5. Какую тему оформления нужно использовать?

Давайте поэтапно построим график

С чего начинаются графики?

- `library(ggplot2)` — активирует пакет ggplot2 со всеми его функциями
- `ggplot()` — создает пустой “базовый” слой — основу графика

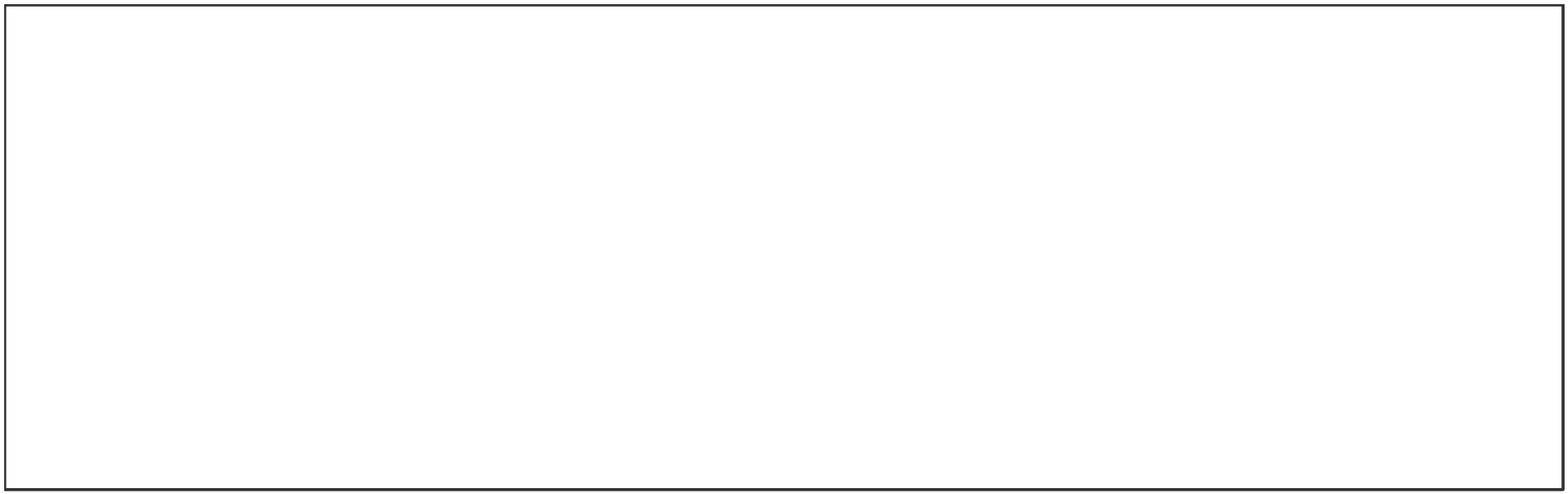
```
library(ggplot2)
ggplot()
```



Откуда брать данные?

Обычно в основе графика пишут, откуда брать данные

```
ggplot(data = nelson)
```

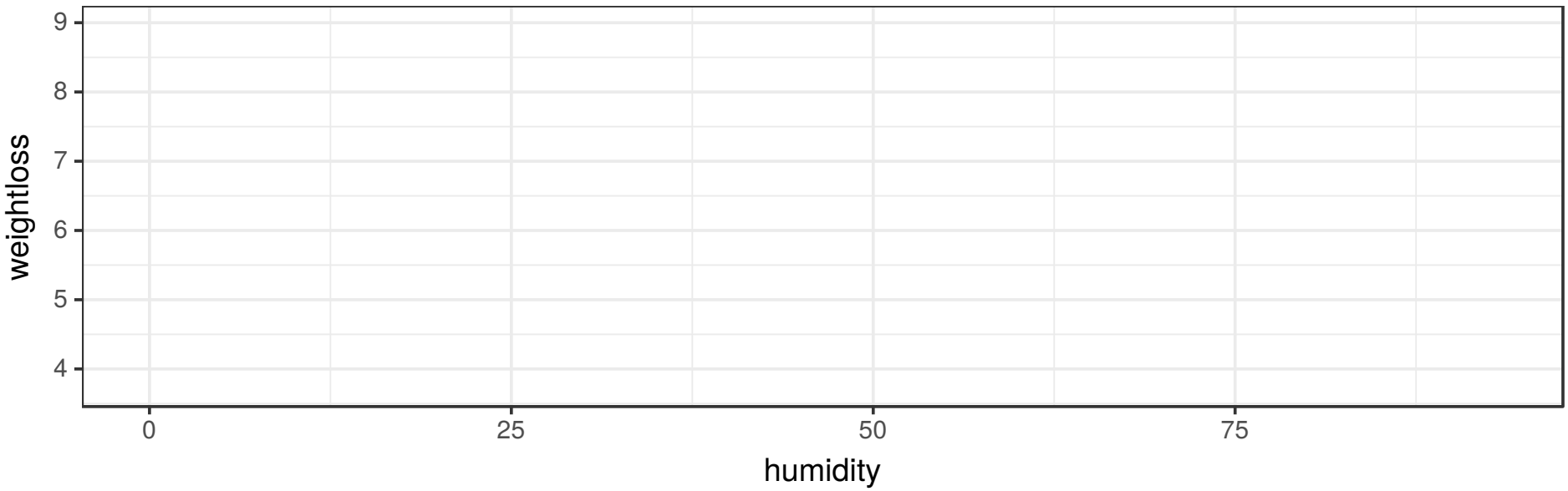


Какие переменные изображать на графике?

Эстетики — это свойства будущих элементов графика, которые будут изображать данные (`x`, `y`, `colour`, `fill`, `size`, `shape`, и т.д.)

`aes()` — функция, которая сопоставляет значения эстетик и переменные из источника данных (название происходит от англ. *aesthetics*)

```
ggplot(data = nelson, aes(x = humidity, y = weightloss))
```

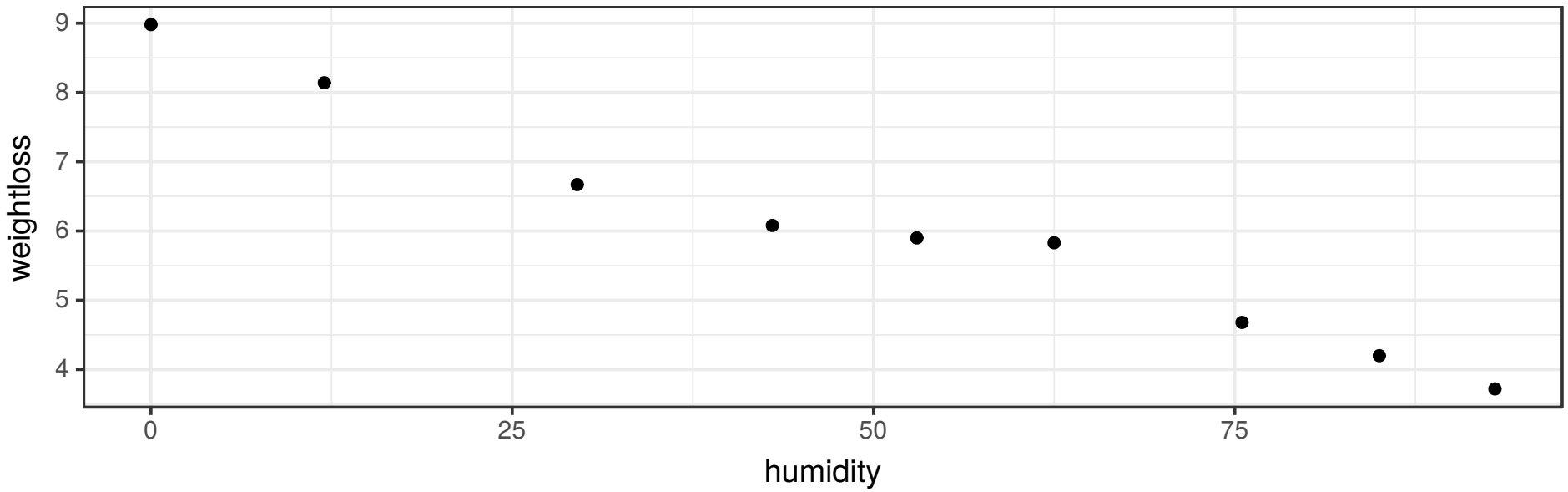


В виде чего изображать?

Геомы — графические элементы (`geom_point()`, `geom_line()`, `geom_bar()`, `geom_smooth()` и т.д., их очень много)

`geom_point()` — точки

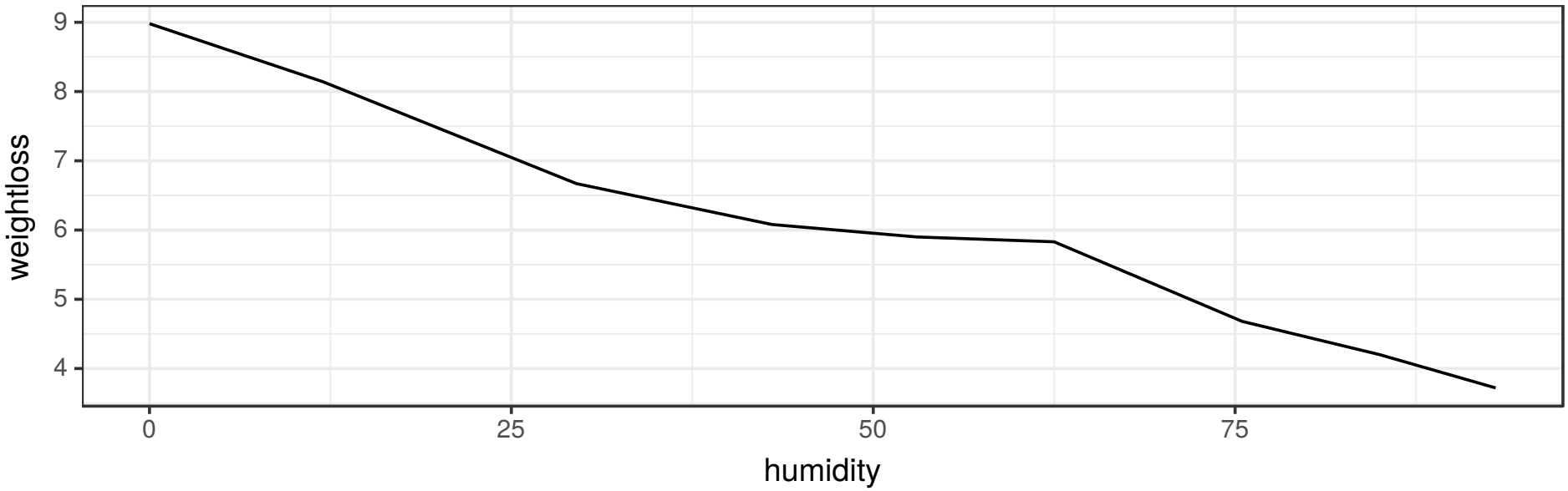
```
ggplot(data = nelson, aes(x = humidity, y = weightloss)) +
  geom_point()
```



В виде чего изображать?

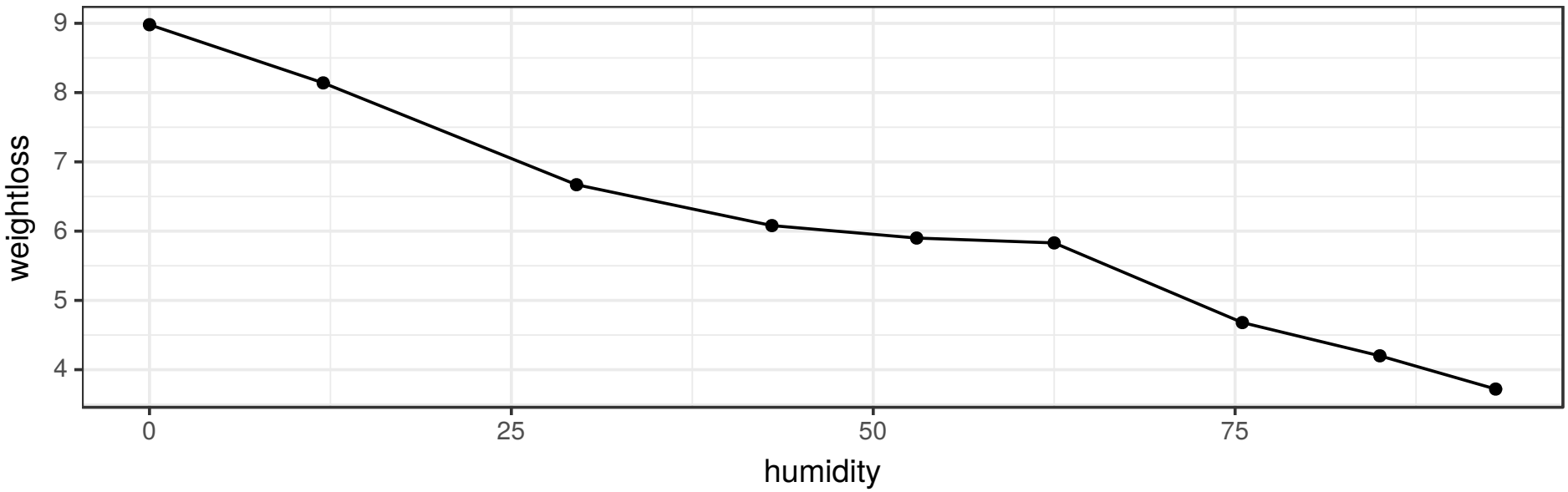
geom_line() — линии

```
ggplot(data = nelson, aes(x = humidity, y = weightloss)) +
  geom_line()
```



Можно использовать несколько геомов одновременно

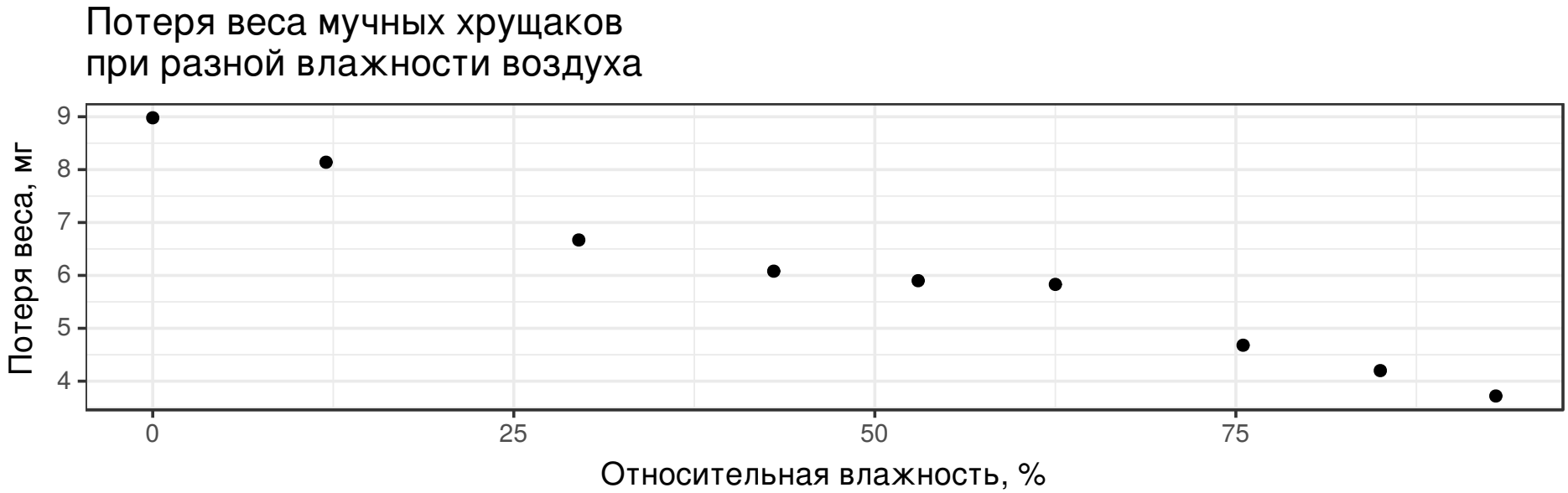
```
ggplot(data = nelson, aes(x = humidity, y = weightloss)) +
  geom_point() +
  geom_line()
```



Подписи осей, заголовок и т.д.

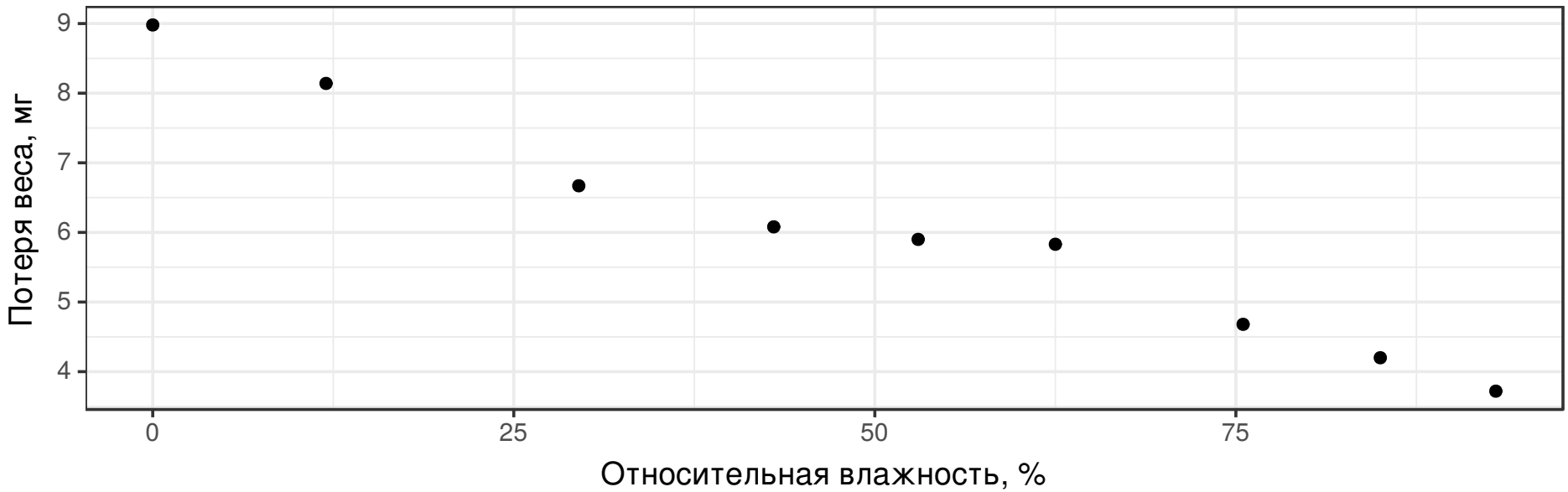
Элемент `labs()` — создает подписи. Аргументы — это имена эстетик, например, `x`, `y` и т.д. Заголовок графика называется `title`

```
ggplot(data = nelson, aes(x = humidity, y = weightloss)) +
  geom_point() +
  labs(x = "Относительная влажность, %", y = "Потеря веса, мг",
       title = "Потеря веса мучных хрущаков \nпри разной влажности воздуха")
```



Графики ggplot можно сохранять в переменные

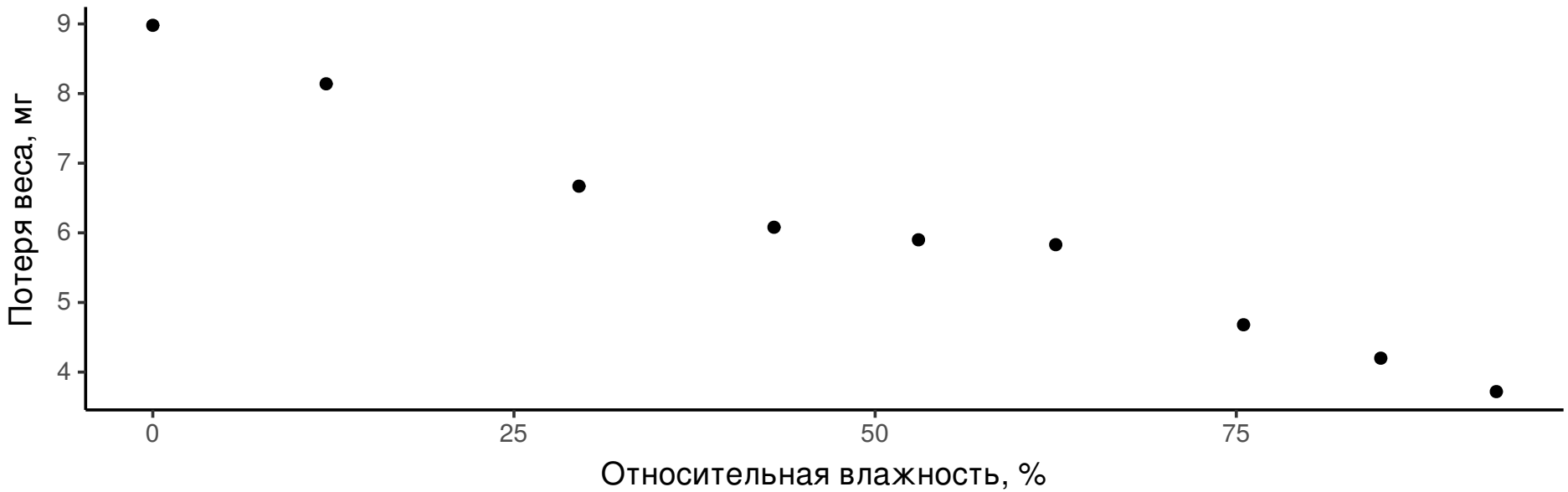
```
gg_nelson <- ggplot(data = nelson, aes(x = humidity, y = weightloss)) +
  geom_point() +
  labs(x = "Относительная влажность, %", y = "Потеря веса, мг")
gg_nelson
```



Темы оформления графиков можно менять и настраивать

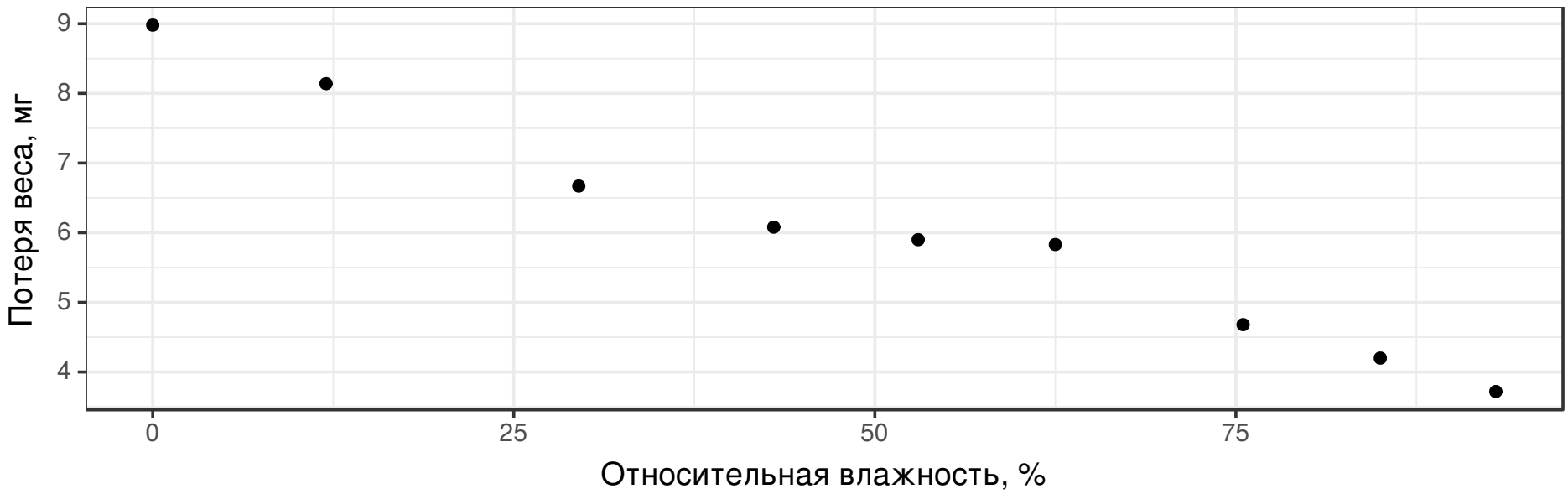
`theme()` — меняет отдельные элементы (см. справку) `theme_bw()`, `theme_classic()` и т.д. — стили оформления целиком

```
gg_nelson + theme_classic()
```



Можно установить любимую тему для всех последующих графиков

```
theme_set(theme_bw())
gg_nelson
```



Графики можно сохранять в файлы

Функция `ggsave()` позволяет сохранять графики в виде файлов во множестве разных форматов (“eps”, “ps”, “tex”, “pdf”, “jpeg”, “tiff”, “png”, “bmp”, “svg” или “wmf”). Параметры изображений настраиваются (см. справку)

```
ggsave(filename = ”bugs_weightloss.png”, plot = gg_nelson)
ggsave(filename = ”bugs_weightloss.pdf”, plot = gg_nelson)
```

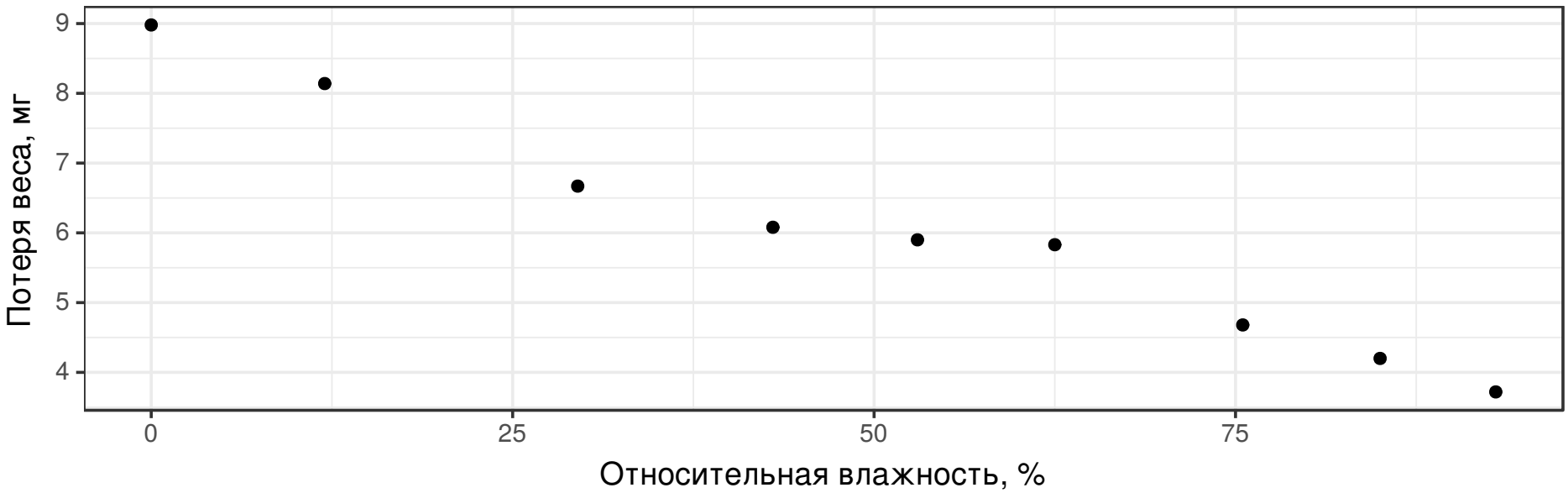
Корреляция

Есть ли связь между переменными?

Судя по всему, да, скажем мы, глядя на график.

Но насколько сильна эта связь?

```
gg_nelson
```



Коэффициент корреляции — способ оценки силы связи между двумя переменными

Коэффициент корреляции Пирсона

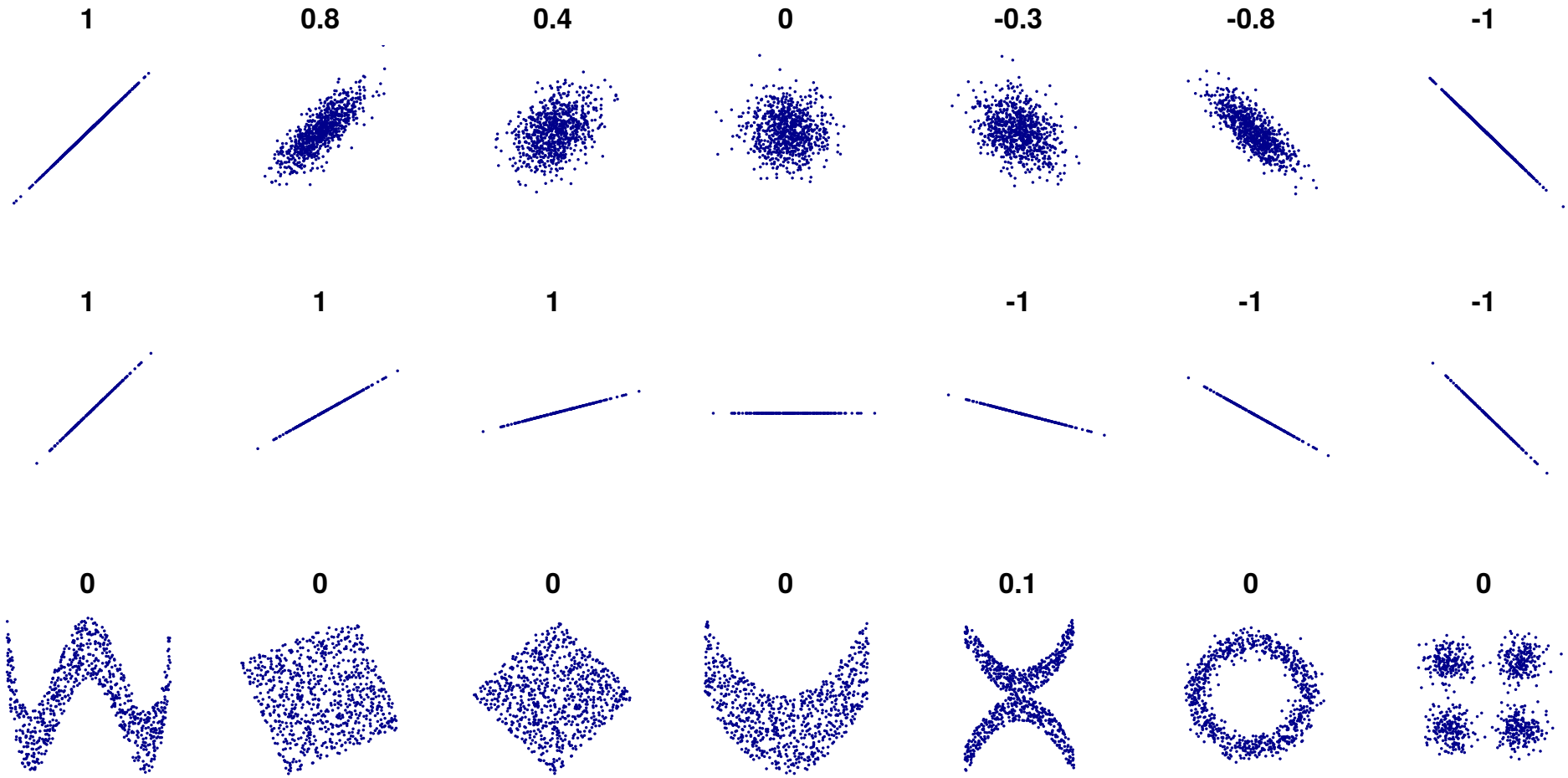
- Оценивает только линейную составляющую связи
- Параметрические тесты значимости (t-критерий) применимы если переменные распределены нормально

В других случаях используются ранговые коэффициенты корреляции (например, кор. Кендалла и кор. Спирмена).

Интерпретация коэффициента корреляции

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

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



By DenisBoigelot, original uploader was Imagecreator [CC0], via Wikimedia Commons

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 = -16.346, df = 7, p-value = 0.0000007816
# alternative hypothesis: true correlation is not equal to 0
# 95 percent confidence interval:
#  -0.9973935 -0.9379224
# sample estimates:
#        cor
# -0.9871523
```

Можно описать результаты несколькими способами:

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

Линейная регрессия

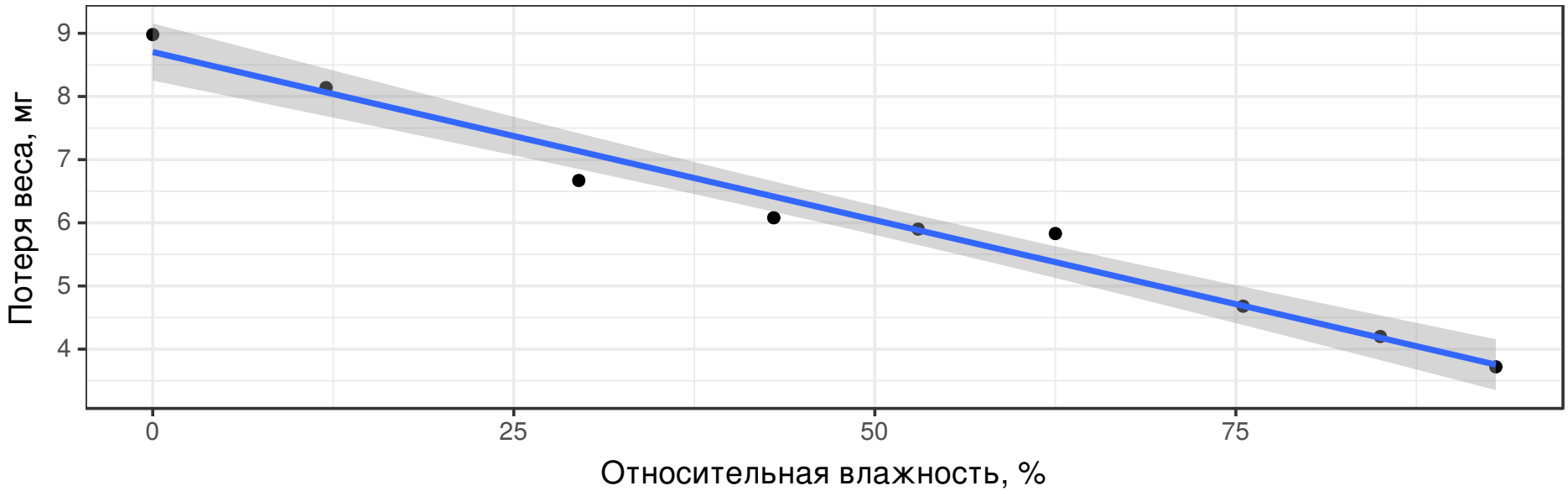
Линейная регрессия

- позволяет описать функциональную зависимость между переменны
- позволяет предсказать значение одной переменной, зная значение другой

Зависимая переменная называется отклик

Те переменные, от которых она зависит — предикторы

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



Линейная регрессия

- простая

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

- множественная

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

Интерпретация коэффициентов регрессии

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

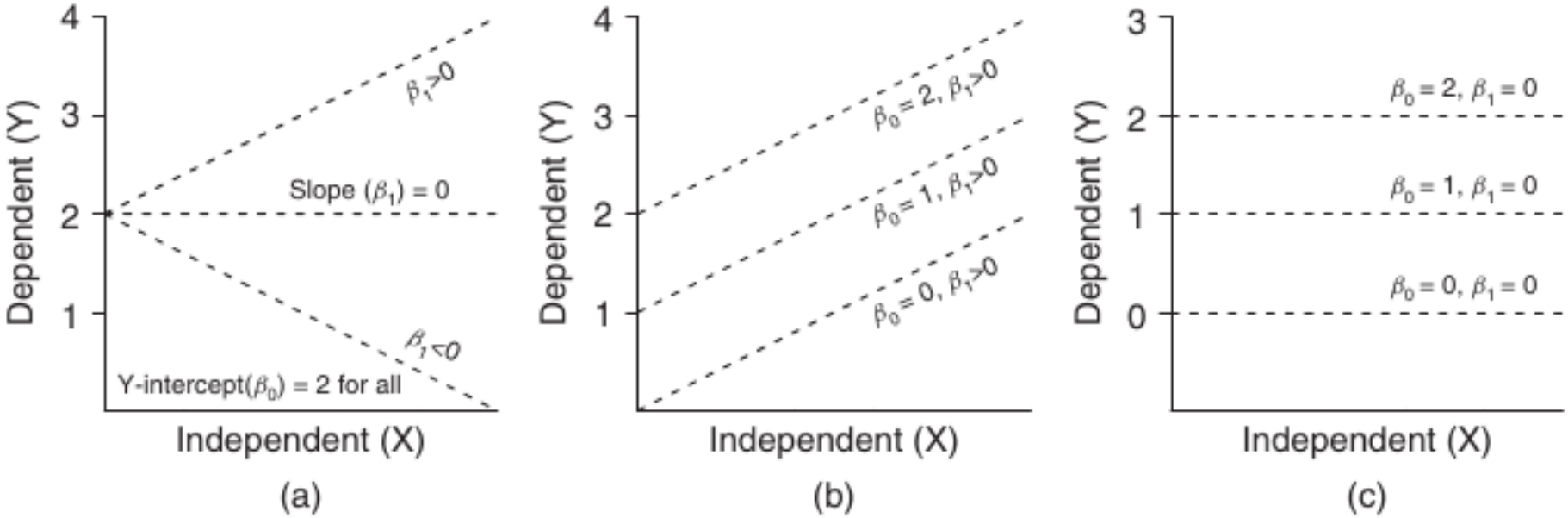


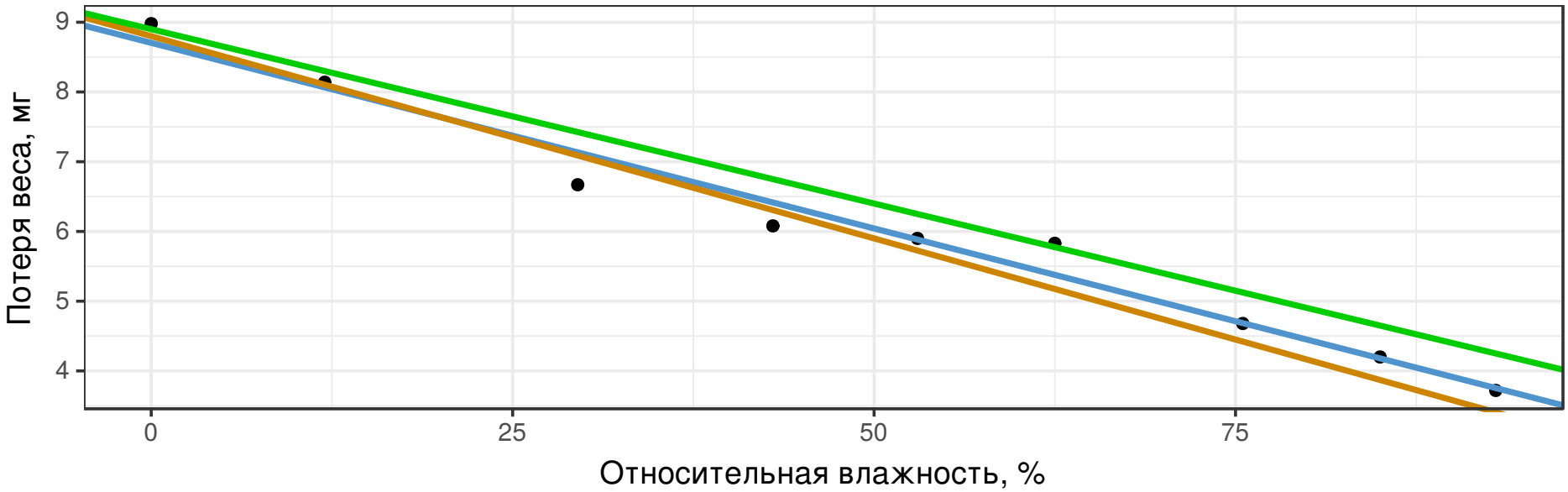
Рисунок из кн. Logan, 2010, стр. 170, рис. 8.2

- b_0 — Отрезок (Intercept), отсекаемый регрессионной прямой на оси y. Значение зависимой переменной y, если предиктор $x = 0$.
- b_1 — Коэффициент угла наклона регрессионной прямой. Показывает на сколько единиц изменяется отклик (y), при увеличении значения предиктора (x) на единицу.

Подбор коэффициентов линейной регрессии

Как провести линию регрессии?

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



Нужно оценить параметры линейной модели:

- β_0
- β_1

Но как это сделать?

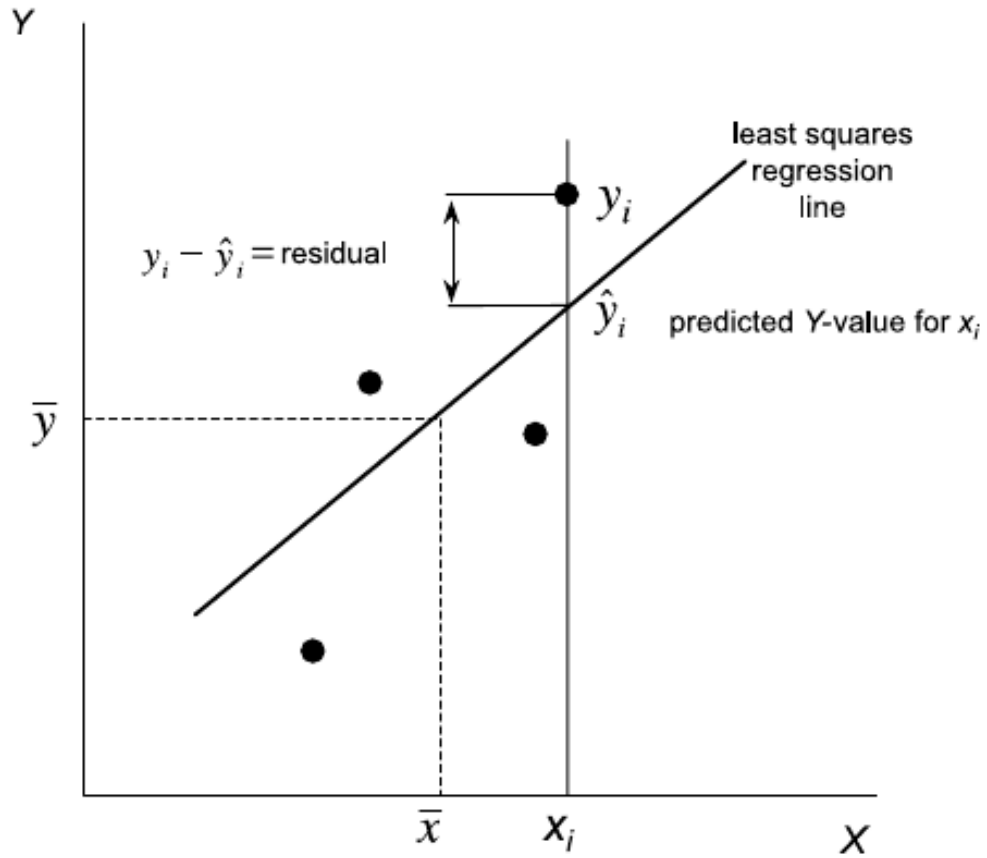
Метод наименьших квадратов — один из способов подбора параметров

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

Нужно получить оценки параметров линейной модели:

- b_0
- b_1

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



Линия регрессии по методу наименьших квадратов

из кн. Quinn, Keough, 2002, стр. 85, рис. 5.6 a

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

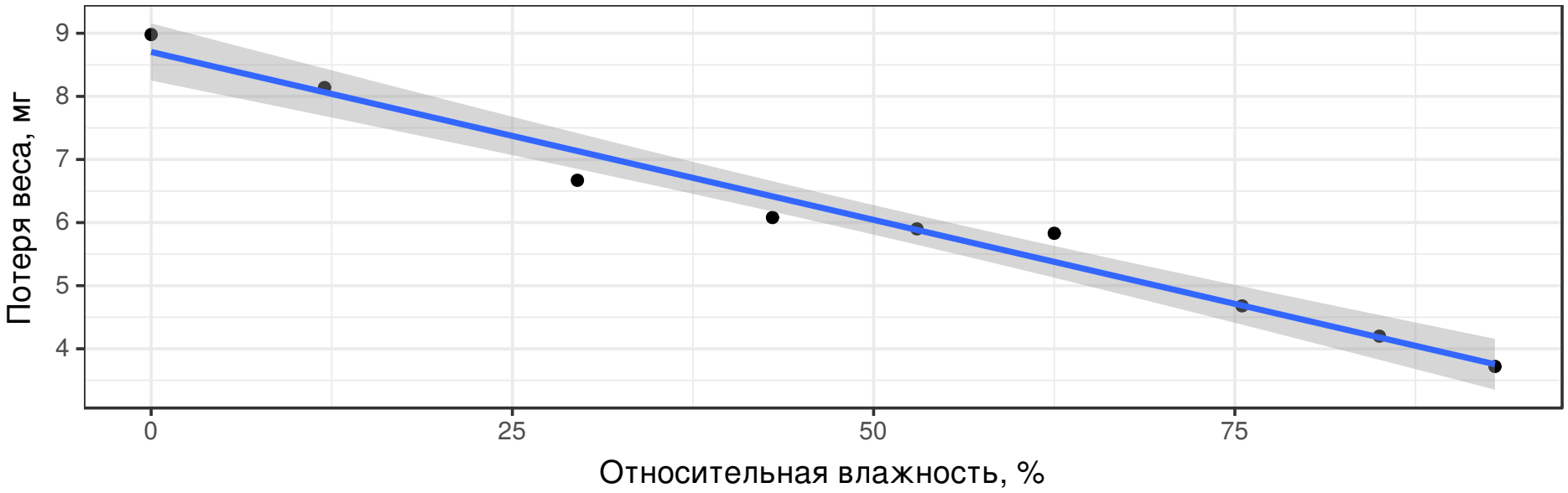
Таблица из кн. Quinn, Keough, 2002, стр. 86, табл. 5.2

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

Доверительный интервал коэффициента

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

Доверительная зона регрессии



Доверительная зона регрессии

- $(1-\alpha)\cdot 100\%$ доверительная зона регрессии
- зона, в которой с $(1-\alpha)\cdot 100\%$ вероятностью лежит регрессионная прямая
- Возникает из-за неопределенности оценок коэффициентов регрессии

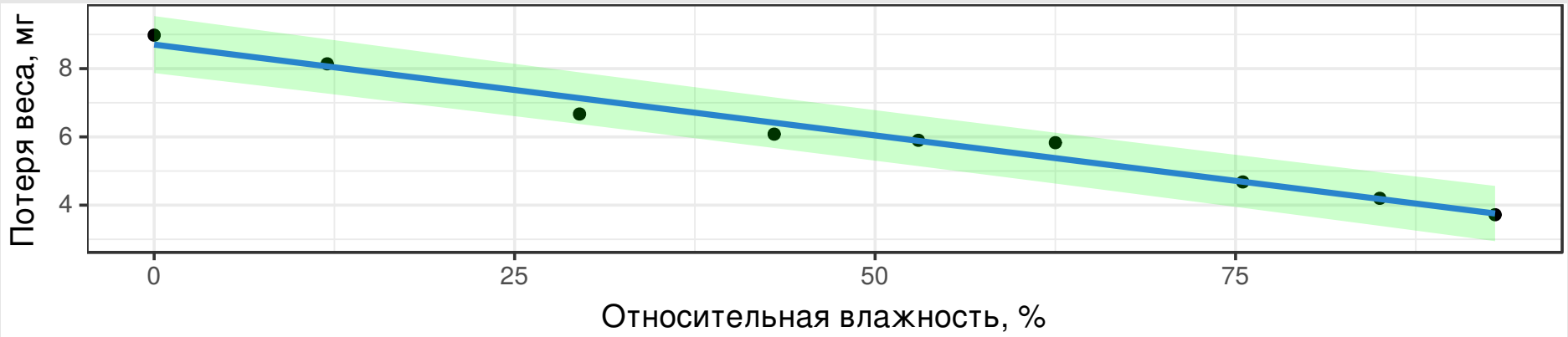
Неопределенность оценок предсказанных значений

Доверительный интервал к предсказанному значению

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

Доверительная область значений регрессии

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



Линейная регрессия в R

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

Примеры формул линейной регрессии

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

(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
# -0.46397 -0.03437  0.01675  0.07464  0.45236
#
# Coefficients:
#              Estimate Std. Error t value       Pr(>|t|)
# (Intercept)  8.704027   0.191565   45.44 0.000000000654 ***
# humidity    -0.053222   0.003256  -16.35 0.000000781615 ***
# ---
# Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
#
# Residual standard error: 0.2967 on 7 degrees of freedom
# Multiple R-squared:  0.9745,  Adjusted R-squared:  0.9708
# F-statistic: 267.2 on 1 and 7 DF,  p-value: 0.0000007816
```

Коэффициенты линейной регрессии:

- $b_0 = 8.7 \pm 0.2$
- $b_1 = -0.053 \pm 0.003$

Записываем уравнение линейной регрессии

Коэффициенты модели:

```
coef(nelson_lm)
```

```
# (Intercept)    humidity
#  8.70402730 -0.05322215
```

Уравнение регрессии:

weightloss = 8.70 - 0.05 humidity

Более формальная запись:

Y = 8.70 - 0.05 X1

Тестирование значимости модели и ее коэффициентов

Способы проверки значимости модели и ее коэффициентов

Существует несколько способов проверки значимости модели

- Значима ли модель целиком?
  - F критерий: действительно ли объясненная моделью изменчивость больше, чем остаточная изменчивость
- Значима ли связь между предиктором и откликом?
  - t-критерий: отличается ли от нуля коэффициент при этом предикторе
  - F-критерий: действительно ли объясненная предиктором изменчивость больше, чем случайная?

Тестируем значимость коэффициентов 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
# -0.46397 -0.03437  0.01675  0.07464  0.45236
#
# Coefficients:
#              Estimate Std. Error t value       Pr(>|t|)
# (Intercept)  8.704027   0.191565   45.44 0.000000000654 ***
# humidity    -0.053222   0.003256  -16.35 0.000000781615 ***
# ---
# Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
#
# Residual standard error: 0.2967 on 7 degrees of freedom
# Multiple R-squared:  0.9745,  Adjusted R-squared:  0.9708
# F-statistic: 267.2 on 1 and 7 DF,  p-value: 0.0000007816
```

Результаты можно описать в тексте так:

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

Тестируем значимость модели целиком 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}, сумма квадратов отклонений от общего среднего значения

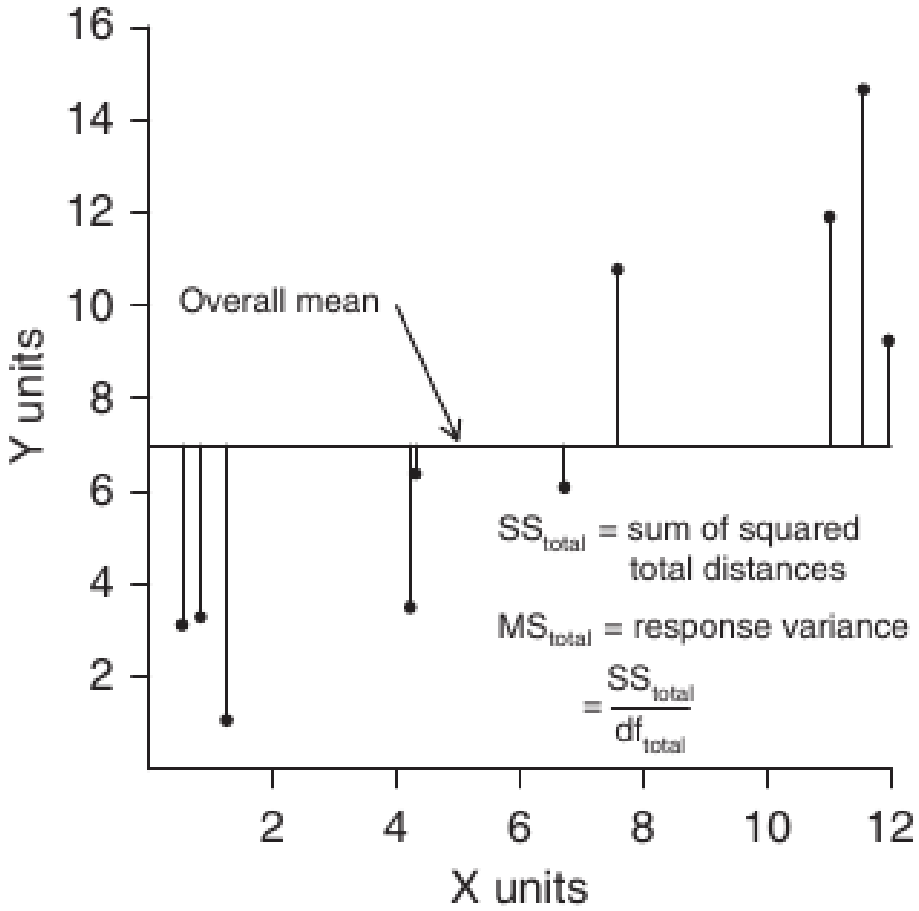
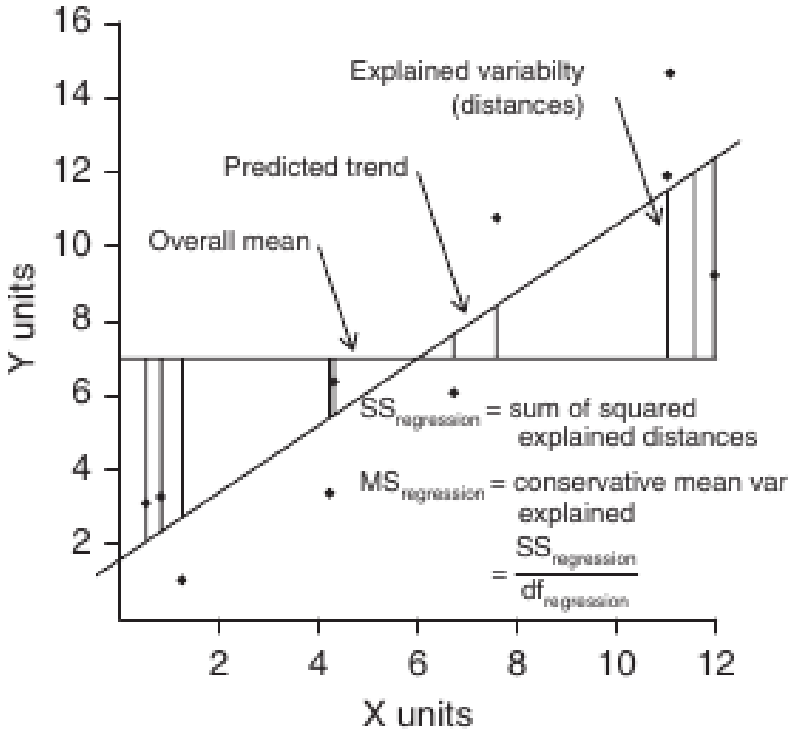


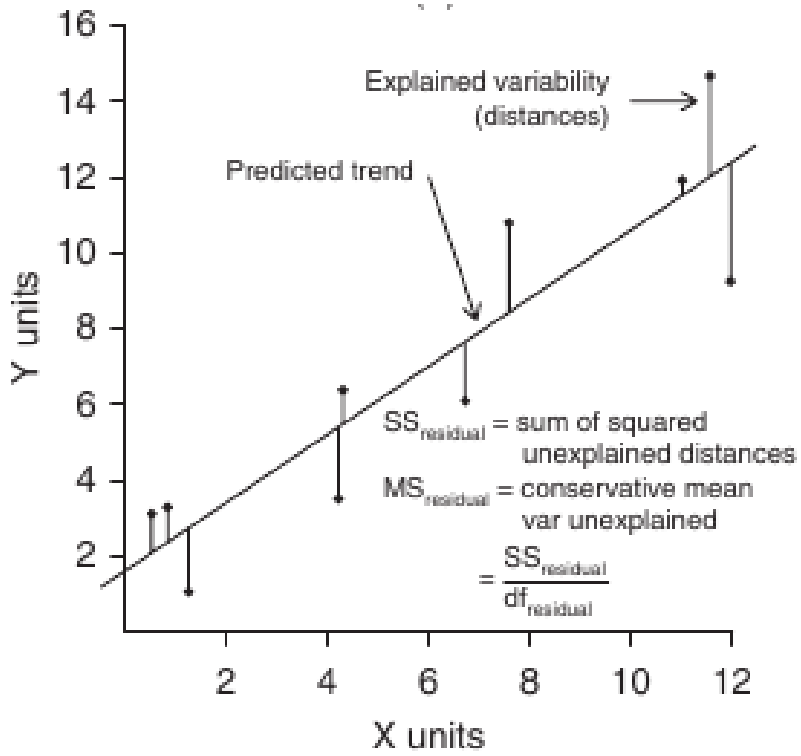
Рис. из кн. Logan, 2010, стр. 172, рис. 8.3

Общая изменчивость делится на объясненную и остаточную

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



Объясненная изменчивость

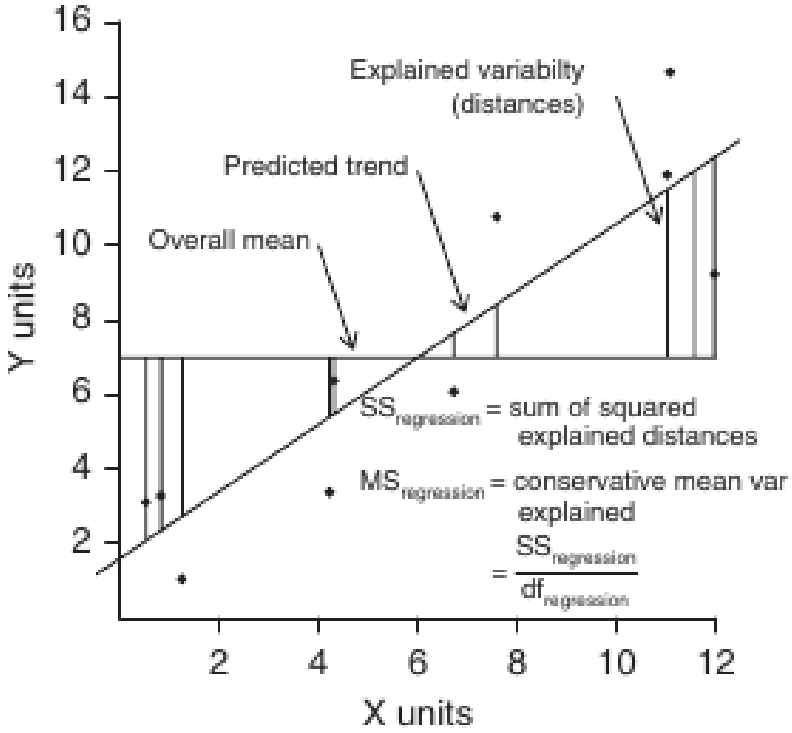


Остаточная изменчивость

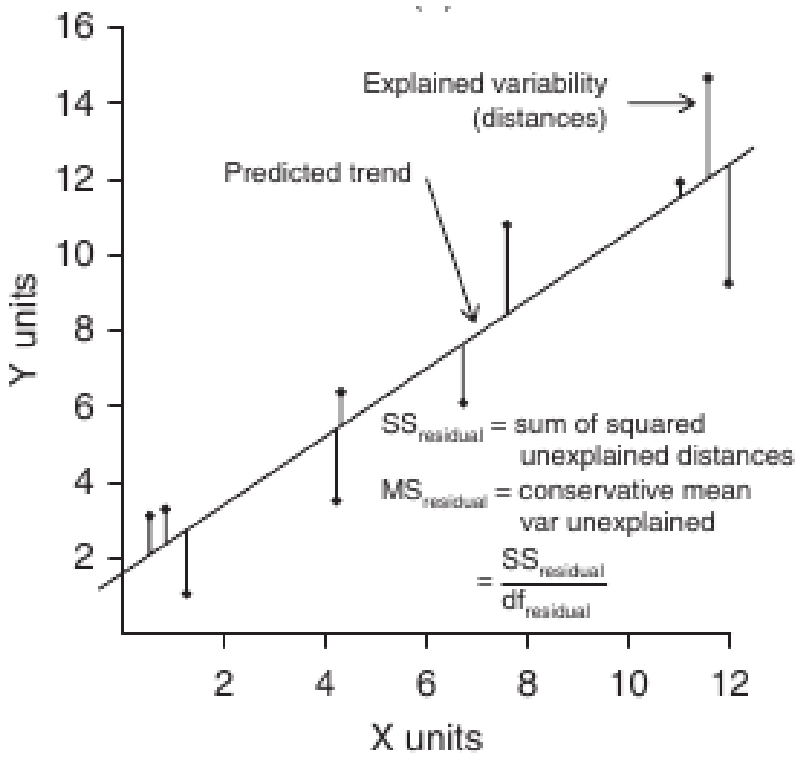
Рис. из кн. Logan, 2010, стр. 172, рис. 8.3

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

F-критерий и распределение F-статистики

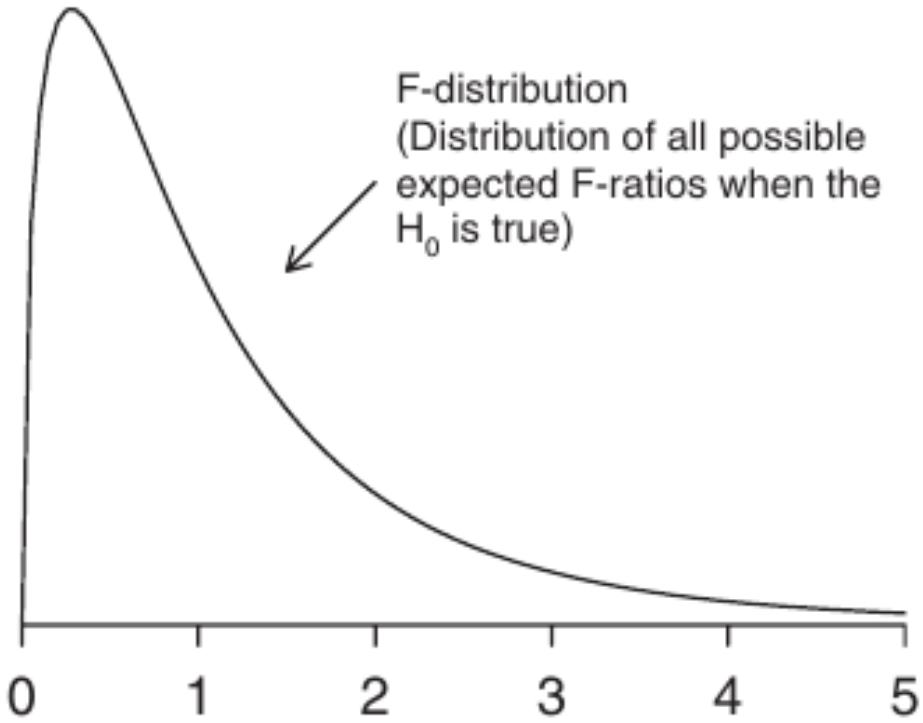
Если $b_1 = 0$, тогда $\hat{y}_i = \bar{y}_i$ и $MS_r \approx MS_e$

F - соотношение объясненной и не объясненной изменчивости

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

Зависит от

- α
- df_r
- df_e



Распределение F-статистики при справедливой H_0

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

Таблица результатов дисперсионного анализа

Источник изменчивости	df	SS	MS	F
Регрессия	$df_r = 1$	$SS_r = \sum (\bar{y} - \hat{y}_i)^2$	$MS_r = \frac{SS_r}{df_r}$	$F_{df_r, df_e} = \frac{MS_r}{MS_e}$
Остаточная	$df_e = n - 2$	$SS_e = \sum (y_i - \hat{y}_i)^2$	$MS_e = \frac{SS_e}{df_e}$	
Общая	$df_t = n - 1$	$SS_t = \sum (\bar{y} - y_i)^2$		

Минимальное упоминание результатов в тексте должно содержать F_{df_r, df_e} и p.

Проверяем значимость модели при помощи F-критерия

```
library(car)
nelson_aov <- Anova(nelson_lm, type = 3)
summary(nelson_aov)
```

```
#      Sum Sq                Df        F value           Pr(>F)
#  Min.   :  0.6161   Min.   :1   Min.   : 267.2   Min.   :0.0000000
#  1st Qu.: 12.0653   1st Qu.:1   1st Qu.: 716.5   1st Qu.:0.0000002
#  Median : 23.5145   Median :1   Median :1165.8   Median :0.0000004
#  Mean   : 68.6077   Mean   :3   Mean   :1165.8   Mean   :0.0000004
#  3rd Qu.:102.6036   3rd Qu.:4   3rd Qu.:1615.2   3rd Qu.:0.0000006
#  Max.   :181.6926   Max.   :7   Max.   :2064.5   Max.   :0.0000008
#                                 NA's   :1        NA's   :1
```

Результаты дисперсионного анализа можно описать в тексте (или представить в виде таблицы):

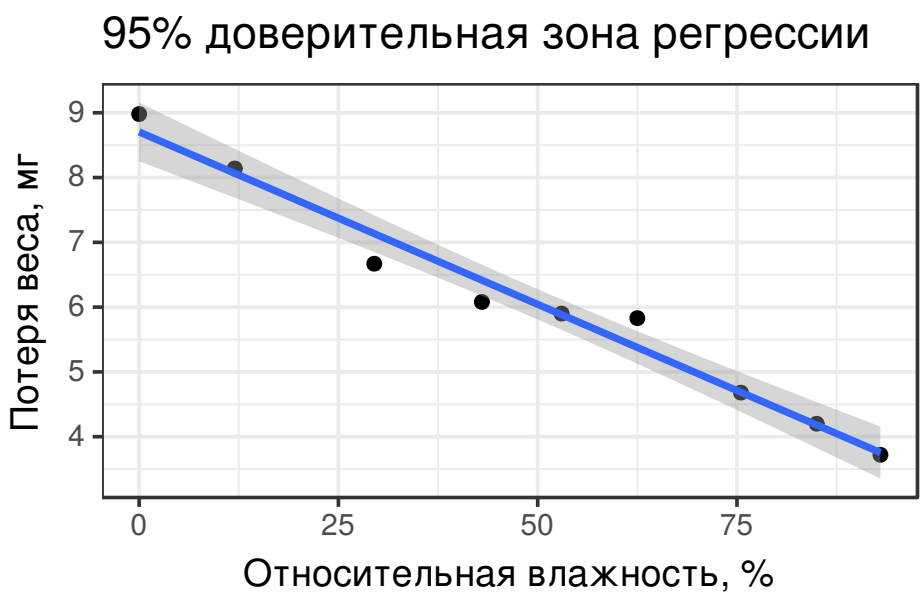
- Количество влаги, потерянной жуками в период эксперимента, достоверно зависело от уровня относительной влажности ($F_{1,7} = 267$, $p < 0.01$).

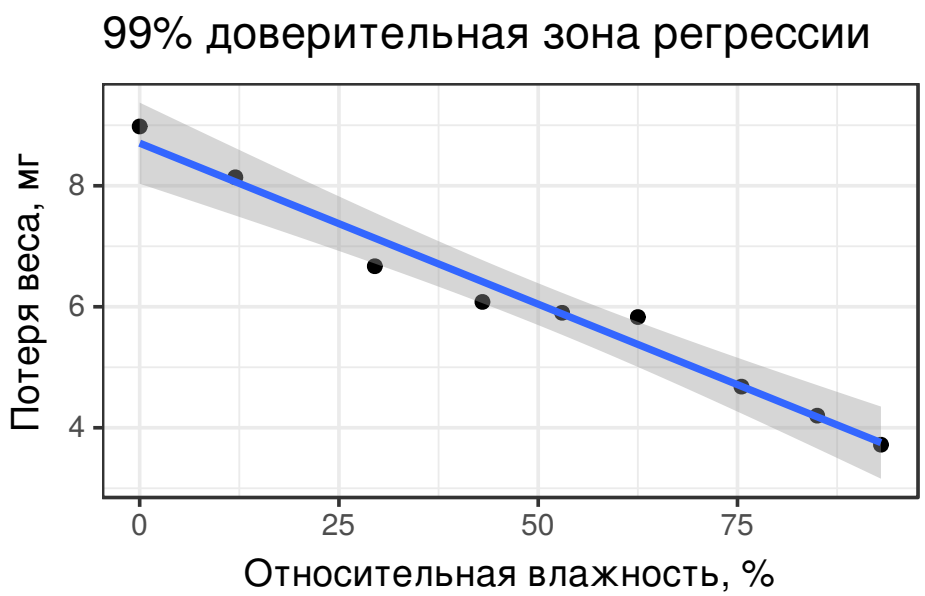
График линейной регрессии

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")
```





Оценка качества подгонки модели

Коэффициент детерминации

Коэффициент детерминации R^2

доля общей изменчивости, объясненная линейной связью x и y

$$R^2 = \frac{SS_r}{SS_t} = 1 - \frac{SS_e}{SS_t}$$

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

Иначе рассчитывается как квадрат коэффициента корреляции $R^2 = r^2$
Не используйте обычный R^2 для множественной регрессии!

Коэффициент детерминации можно найти в сводке модели

```
summary(nelson_lm)
```

```
#
# Call:
# lm(formula = weightloss ~ humidity, data = nelson)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -0.46397 -0.03437  0.01675  0.07464  0.45236
#
# Coefficients:
#             Estimate Std. Error t value       Pr(>|t|)
# (Intercept)  8.704027   0.191565   45.44 0.000000000654 ***
# humidity    -0.053222   0.003256  -16.35 0.000000781615 ***
# ---
# Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
#
# Residual standard error: 0.2967 on 7 degrees of freedom
# Multiple R-squared:  0.9745,  Adjusted R-squared:  0.9708
# F-statistic: 267.2 on 1 and 7 DF,  p-value: 0.0000007816
```

Сравнение качества подгонки моделей

R^2_{adj} — скорректированный R^2

$$R^2_{adj} = 1 - \frac{SS_e / df_e}{SS_t / df_t}$$

где $df_e = n - p - 1$, $df_t = n - 1$
R^2_{adj} учитывает число переменных в модели, вводится штраф за каждый новый параметр.
Используйте R^2_{adj} для сравнения моделей с разным числом параметров.

Использование линейной регрессии для предсказаний (для самостоятельного разбора)

Использование линейной регрессии для предсказаний

Для конкретного значения предиктора мы можем сделать два типа предсказаний:

- предсказываем среднее значение отклика — это оценка точности положения линии регрессии
- предсказываем значение отклика у 95% наблюдений — это оценка точности предсказаний

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 6.042920 5.809068 6.276771
# 2 3.381812 2.933952 3.829672
#
# $se.fit
#         1         2
# 0.09889579 0.18940006
#
# $df
# [1] 7
#
# $residual.scale
# [1] 0.2966631
```

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

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 6.042920 5.303471 6.782368
# 2 3.381812 2.549540 4.214084
#
# $se.fit
#         1         2
# 0.09889579 0.18940006
#
# $df
# [1] 7
#
# $residual.scale
# [1] 0.2966631
```

- 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

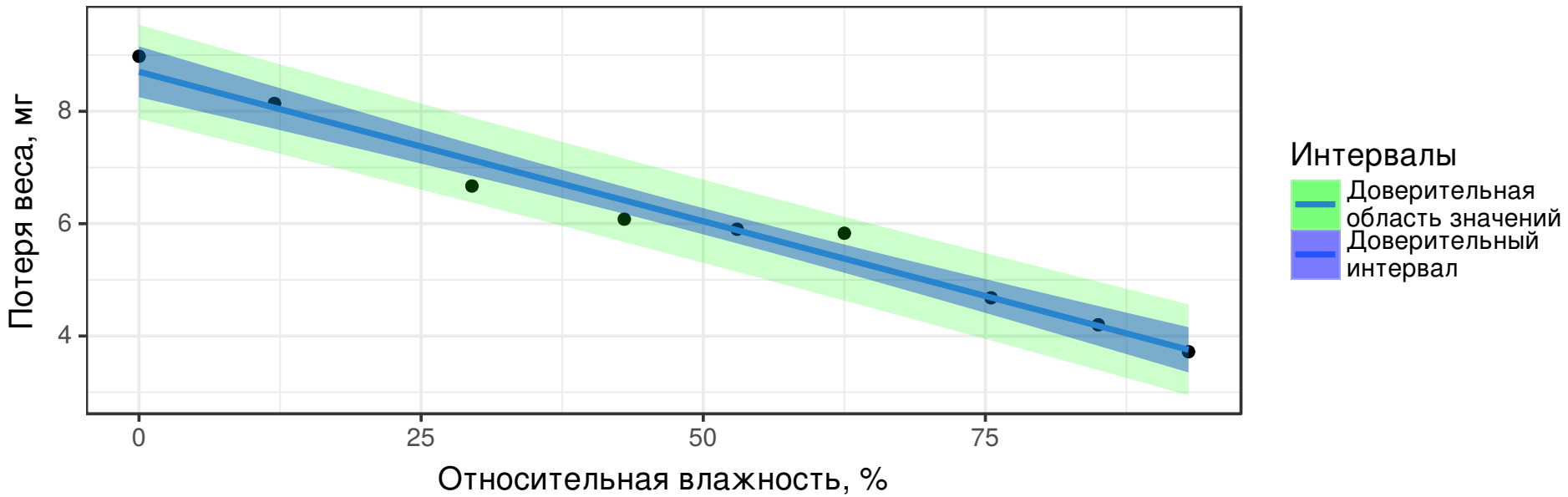
```
(pr_all <- predict(nelson_lm, interval = "prediction"))
```

```
#        fit      lwr      upr
# 1 8.704027 7.868990 9.539064
# 2 8.065361 7.269036 8.861687
# 3 7.133974 6.377243 7.890704
# 4 6.415475 5.673847 7.157102
# 5 5.883253 5.143538 6.622969
# 6 5.377643 4.632344 6.122941
# 7 4.685755 3.921455 5.450055
# 8 4.180144 3.394150 4.966139
# 9 3.754367 2.945412 4.563322
```

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

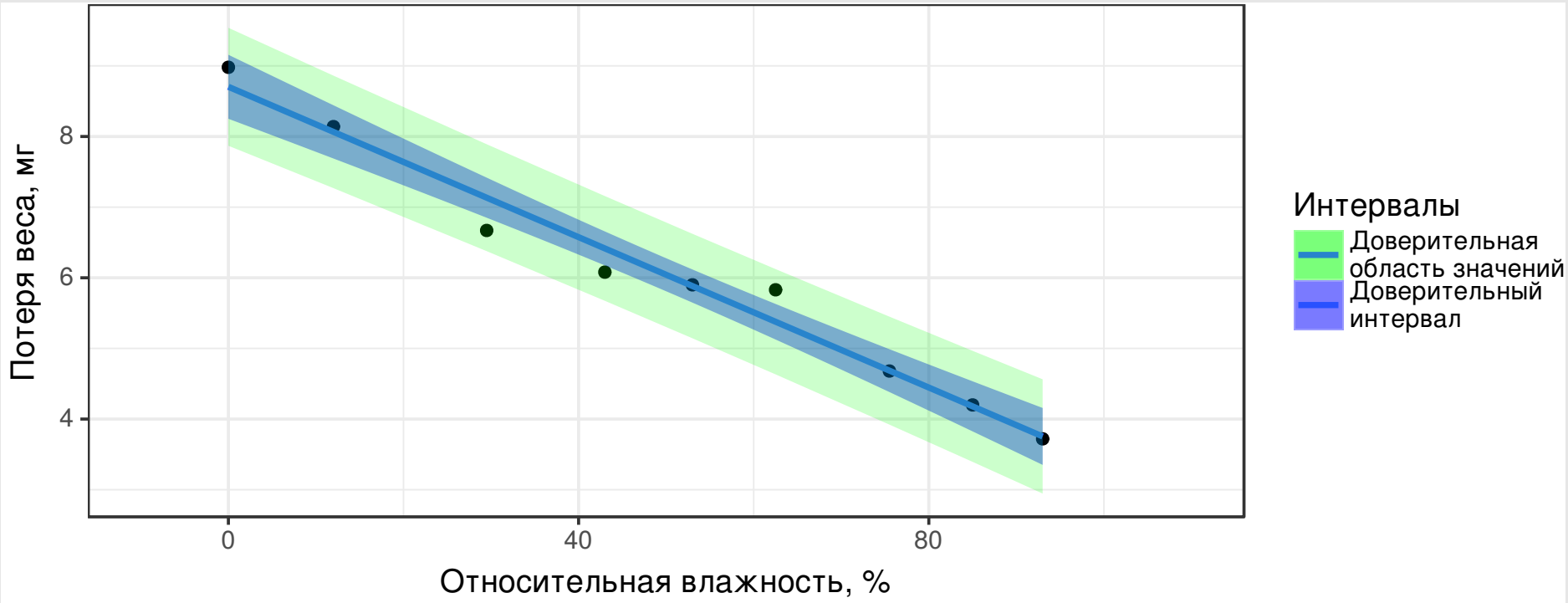
Строим доверительную область значений и доверительный интервал одновременно

```
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!



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

- Гланц, 1999, стр. 221-244
- OpenIntro: Statistics
- Quinn, Keough, 2002, pp. 78-110
- Logan, 2010, pp. 170-207
- Sokal, Rohlf, 1995, pp. 451-491
- Zar, 1999, pp. 328-355