Регрессионный анализ, часть 1

Математические методы в зоологии с использованием R

Марина Варфоломеева

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: стерильность пыльцы гибридов

Гибриды отдаленных видов часто бывают стерильны.
Но насколько они должны быть разными для этого?
Как зависит плодовитость межвидовых гибридов смолевок рода *Silene* от их генетической удаленности?

- `proportionSterile` — доля стерильных пыльцевых зерен
- `geneticDistance` — генетическая удаленность видов

Смолевка обыкновенная *Silene vulgaris*, by Rhododendrites [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], from Wikimedia Commons

Moyle et al. 2004; данные из Whitlock, Schluter, 2015, глава 17, упр.10; Данные в файлах HybridPollenSterility.xlsx и HybridPollenSterility.csv

Читаем данные из файла

Чтение из xlsx:

```
library(readxl)
hybrid <- read_excel(path = 'data/HybridPollenSterility.xlsx', sheet = 1)
```

Чтение из csv:

```
hybrid <- read.table(file = 'data/HybridPollenSterility.csv', header = TRUE,
```

Все ли правильно открылось?

```
str(hybrid)      # Структура данных
```

```
# 'data.frame': 23 obs. of  2 variables:
#  $ geneticDistance  : num  0 0 0 0 0 0.03 0.02 0.03 0.04 0.04 ...
# $ proportionSterile: num  0.02 0.06 0.14 0.24 0.3 0.62 0.28 0.23 0.15 0.45 ...
```

```
head(hybrid)     # Первые несколько строк файла
```

```
#   geneticDistance proportionSterile
# 1            0.00              0.02
# 2            0.00              0.06
# 3            0.00              0.14
# 4            0.00              0.24
# 5            0.00              0.30
# 6            0.03              0.62
```

Сделаем более короткие имена

Сейчас переменные называются так:

```
colnames(hybrid)
```

```
# [1] ”geneticDistance”   ”proportionSterile”
```

Сделаем более удобные короткие названия:

```
colnames(hybrid) <- c('Distance', 'Sterile')
```

Теперь переменные стали называться так:

```
colnames(hybrid)
```

```
# [1] ”Distance” ”Sterile”
```

Знакомимся с данными

Есть ли пропущенные значения?

```
colSums(is.na(hybrid))
```

```
# Distance  Sterile
#        0         0
```

Каков объем выборки?

Поскольку пропущенных значений нет, можем просто посчитать число строк:

```
nrow(hybrid)
```

```
# [1] 23
```

Теперь все готово, чтобы мы могли ответить на вопрос исследования.

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

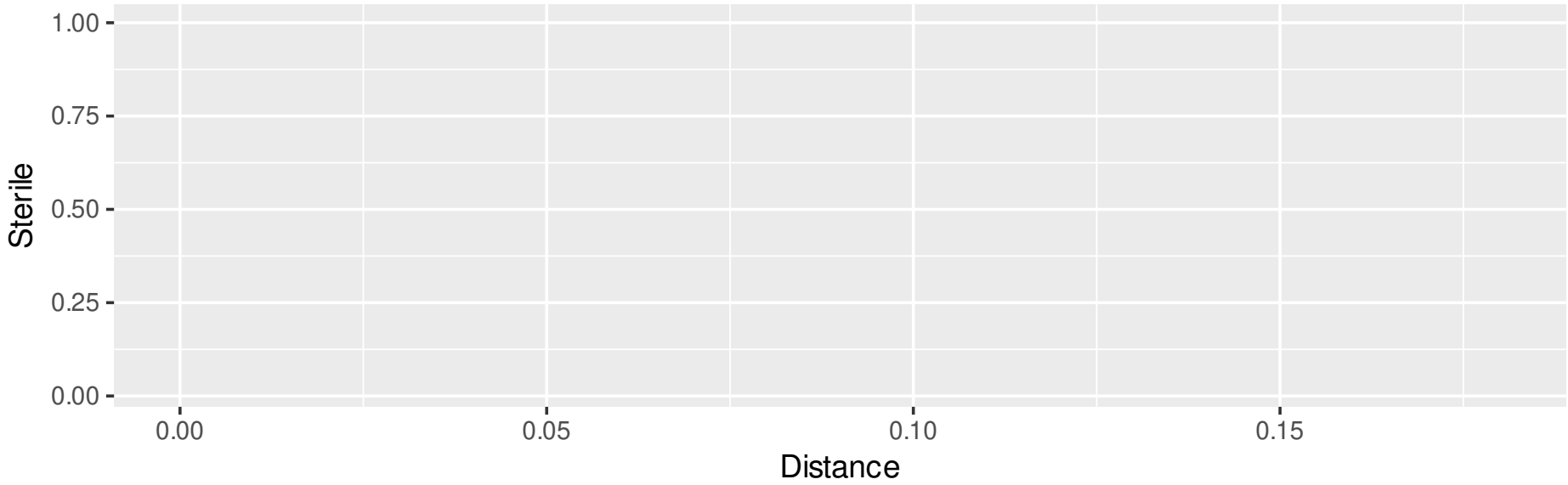
```
ggplot(data = hybrid)
```

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 = hybrid, aes(x = Distance, y = Sterile))
```

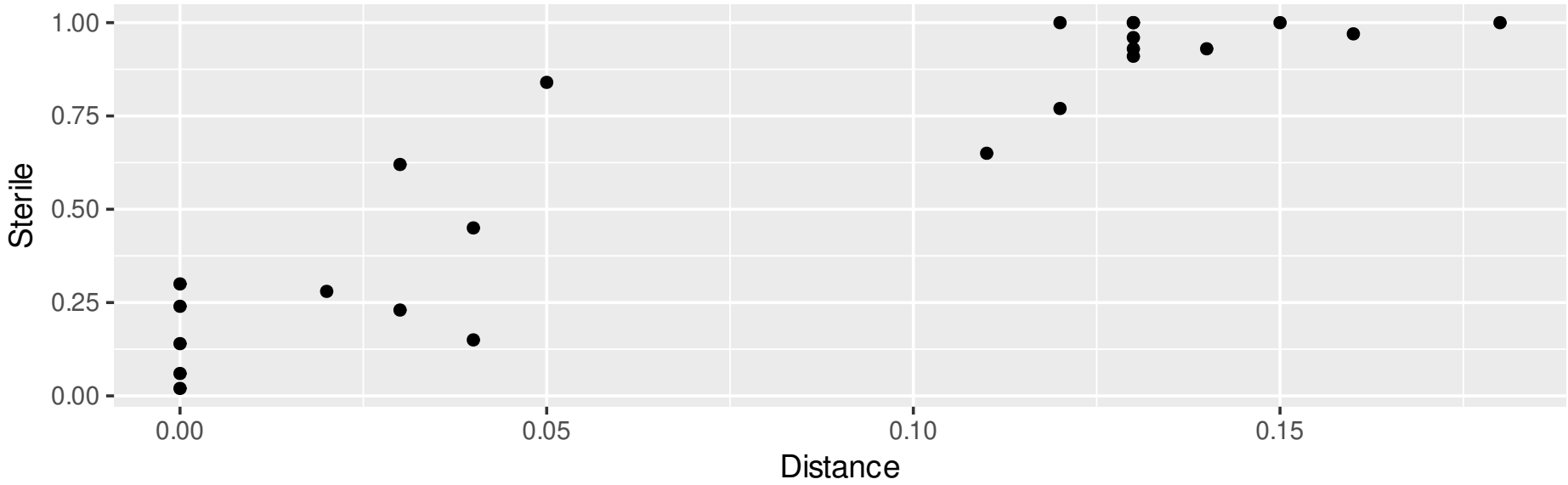


В виде чего изображать?

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

`geom_point()` — точки

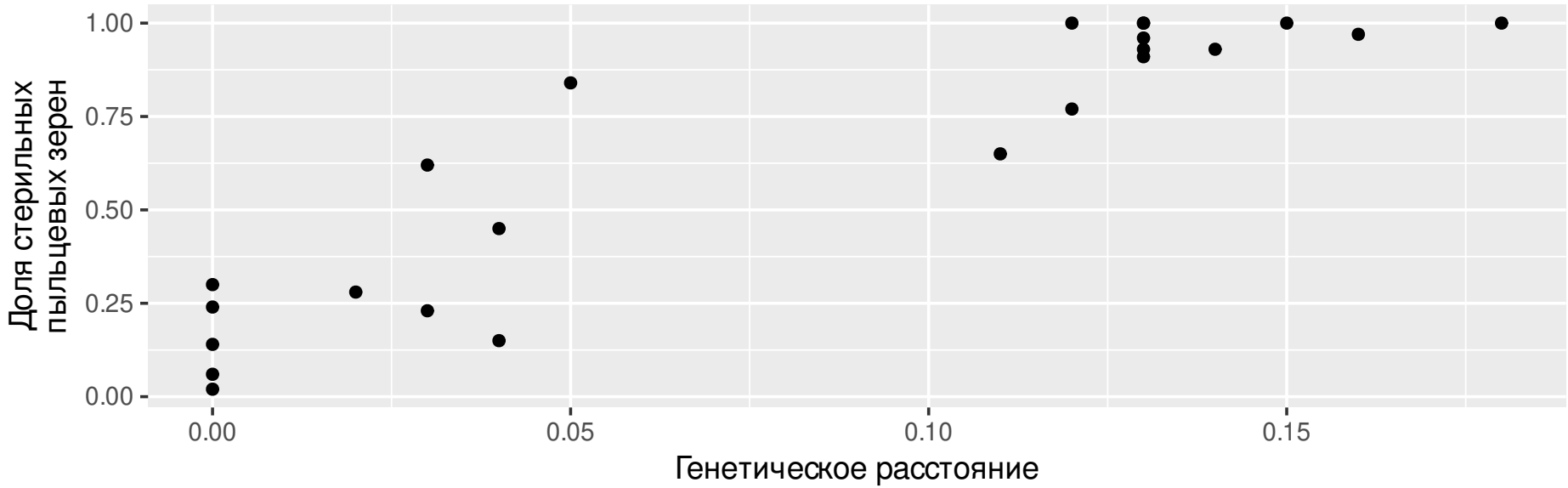
```
ggplot(data = hybrid, aes(x = Distance, y = Sterile)) +
  geom_point()
```



Подписи осей, заголовок и т.д.

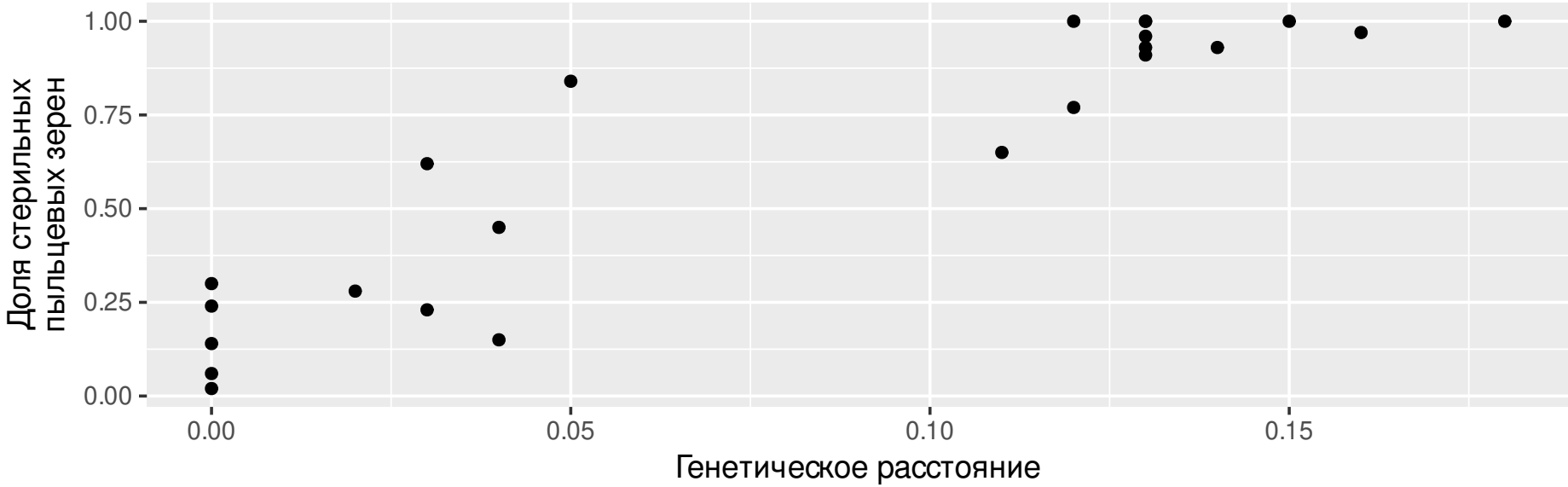
Элемент `labs()` — создает подписи. Аргументы — это имена эстетик, например, x, y и т.д. Заголовок графика называется `title`

```
ggplot(data = hybrid, aes(x = Distance, y = Sterile)) +
  geom_point() +
  labs(x = 'Генетическое расстояние',
       y = 'Доля стерильных\nпыльцевых зерен')
```



Графики ggplot можно сохранять в переменные

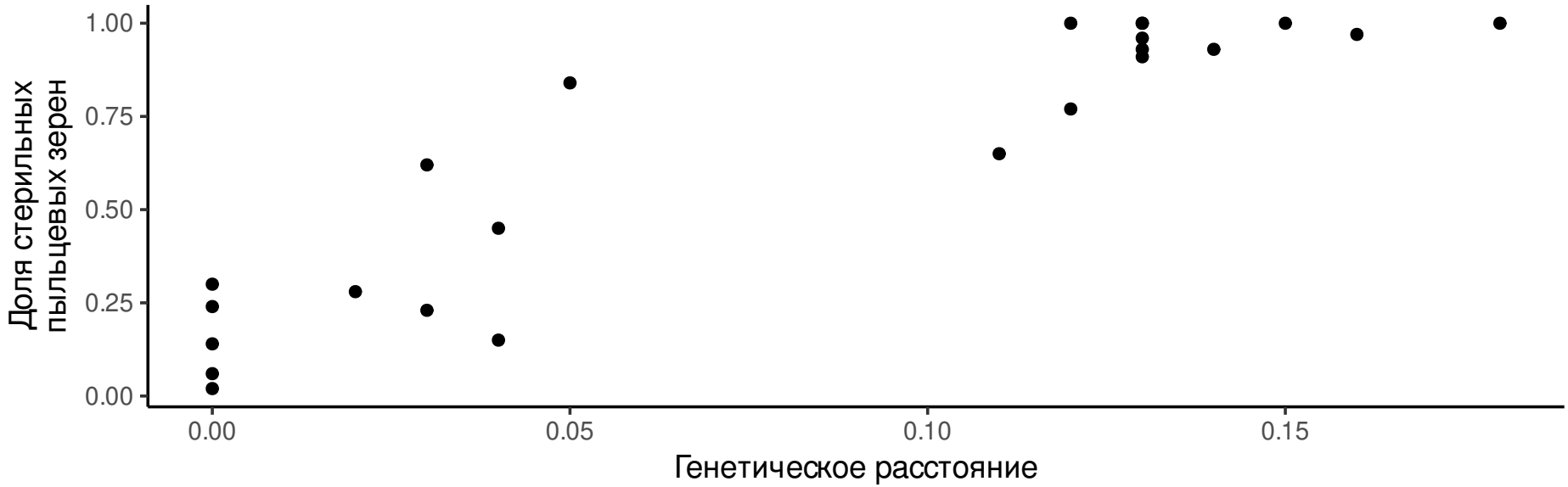
```
gg_hybrid <- ggplot(data = hybrid, aes(x = Distance, y = Sterile)) +
  geom_point() +
  labs(x = 'Генетическое расстояние',
       y = 'Доля стерильных\nпыльцевых зерен')
gg_hybrid
```



Темы оформления графиков можно менять и настраивать

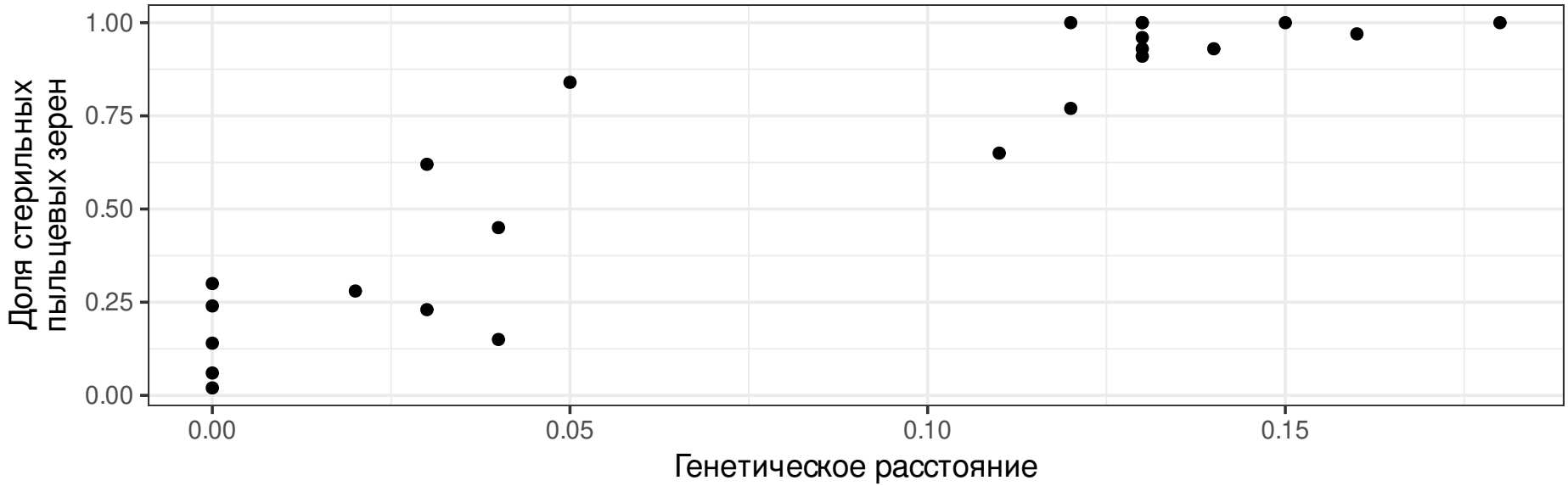
`theme()` — меняет отдельные элементы (см. справку) `theme_bw()`, `theme_classic()` и т.д. — стили оформления целиком

```
gg_hybrid + theme_classic()
```



Можно установить любимую тему для всех последующих графиков

```
theme_set(theme_bw())
gg_hybrid
```



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 = 'hybrids_Sterile.png', plot = gg_hybrid)
ggsave(filename = 'hybrids_Sterile.pdf', plot = gg_hybrid)
```

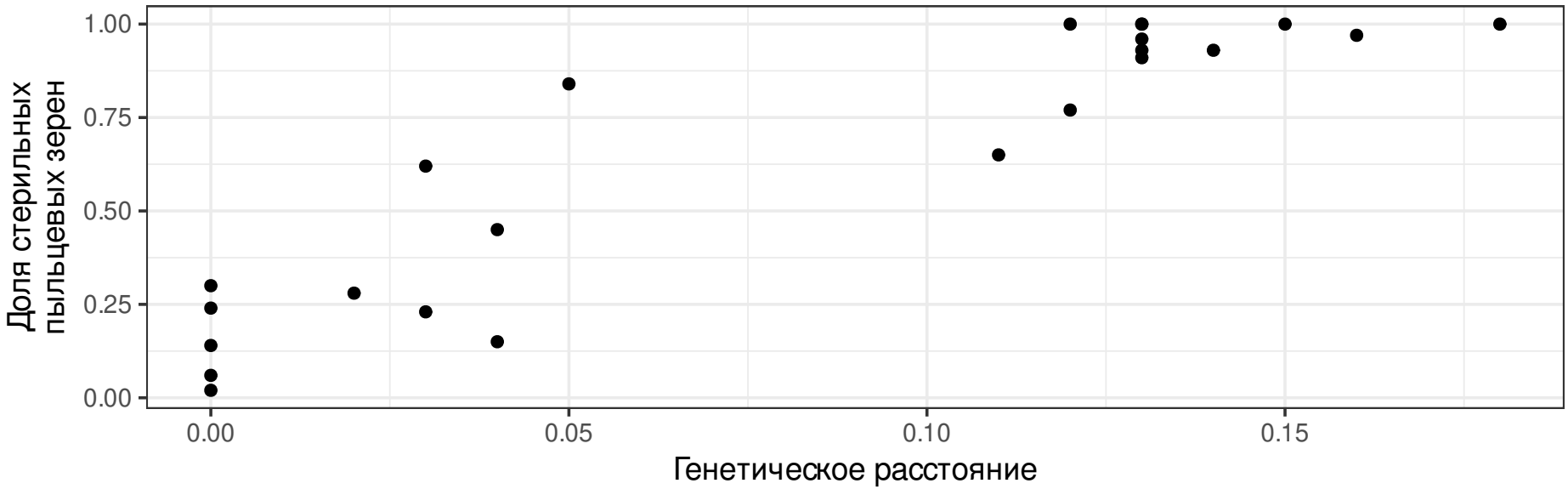
Корреляция

Есть ли связь между переменными?

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

Но насколько сильна эта связь?

```
gg_hybrid
```



Коэффициент корреляции — способ оценки силы связи между двумя переменными

Коэффициент корреляции Пирсона

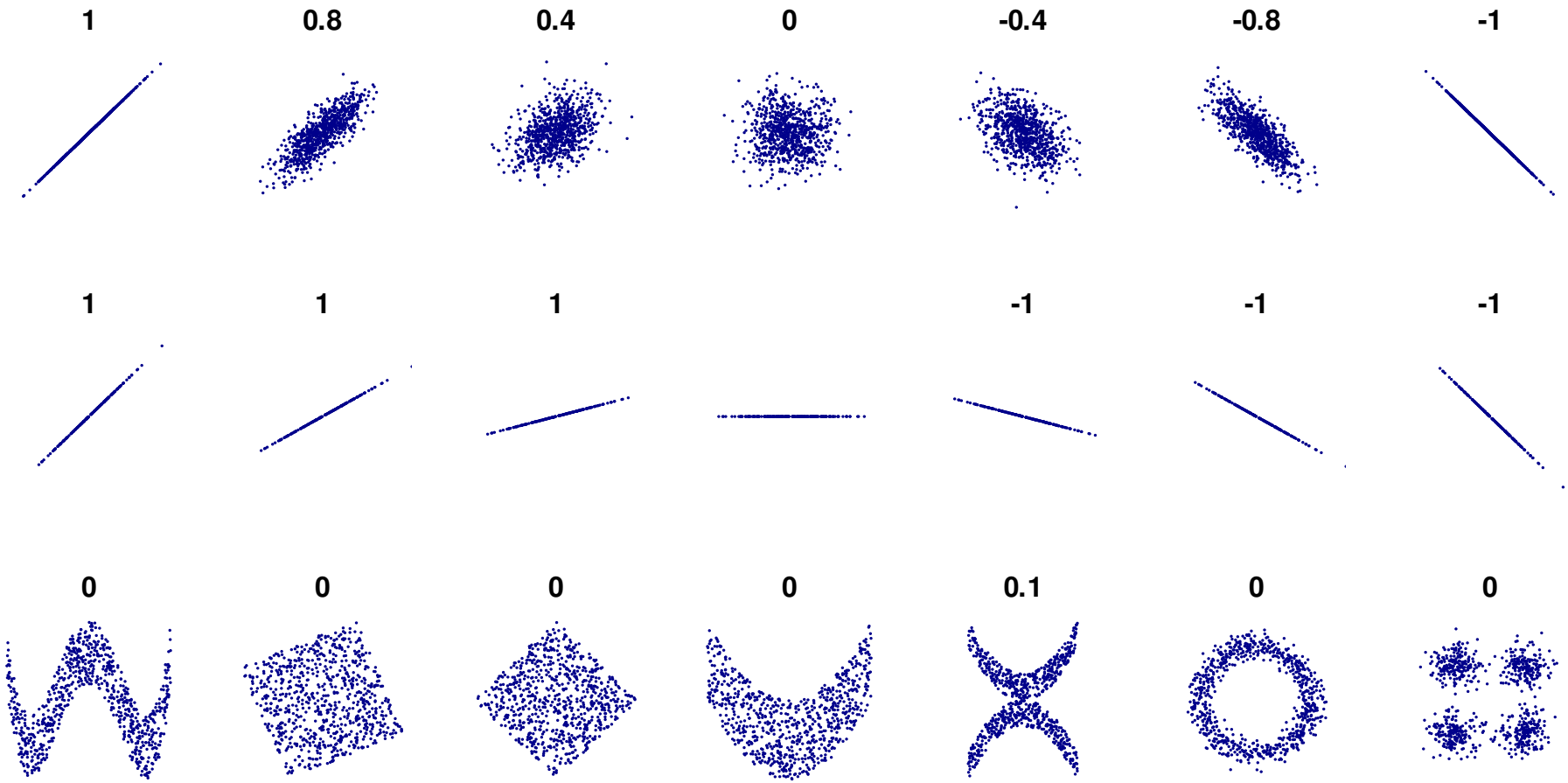
- Оценивает только линейную составляющую связи
- Параметрические тесты значимости (t-тест) применимы если переменные распределены нормально

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

Задание 1

Дополните код, чтобы вычислить корреляцию Пирсона между долей стерильной пыльцы и генетическим расстоянием.

Используйте нужные переменные из датасета `hybrid` и функцию `cor.test()`

```
p_cor <- cor.test(x = , y = ,
         alternative =  , method =  )
p_cor
```

Решение: корреляция между долей стерильной пыльцы и генетическим расстоянием

```
p_cor <- cor.test(x = hybrid$Distance, y = hybrid$Sterile,
         alternative = 'two.sided', method = 'pearson')
p_cor
```

```
#
#   Pearson's product-moment correlation
#
# data:  hybrid$Distance and hybrid$Sterile
# t = 10.54, df = 21, p-value = 0.0000000007659
# alternative hypothesis: true correlation is not equal to 0
# 95 percent confidence interval:
#  0.8116979 0.9646238
# sample estimates:
#       cor
# 0.9170651
```

Можно описать результаты несколькими способами:

- Доля стерильной пыльцы у межвидовых гибридов смолевок положительно коррелирует с генетическим расстоянием ($r = 0.92$, $p < 0.01$)
- Стерильной пыльцы у межвидовых гибридов смолевок становится больше с увеличением генетического расстояния между родителями ($r = 0.92$, $p < 0.01$)

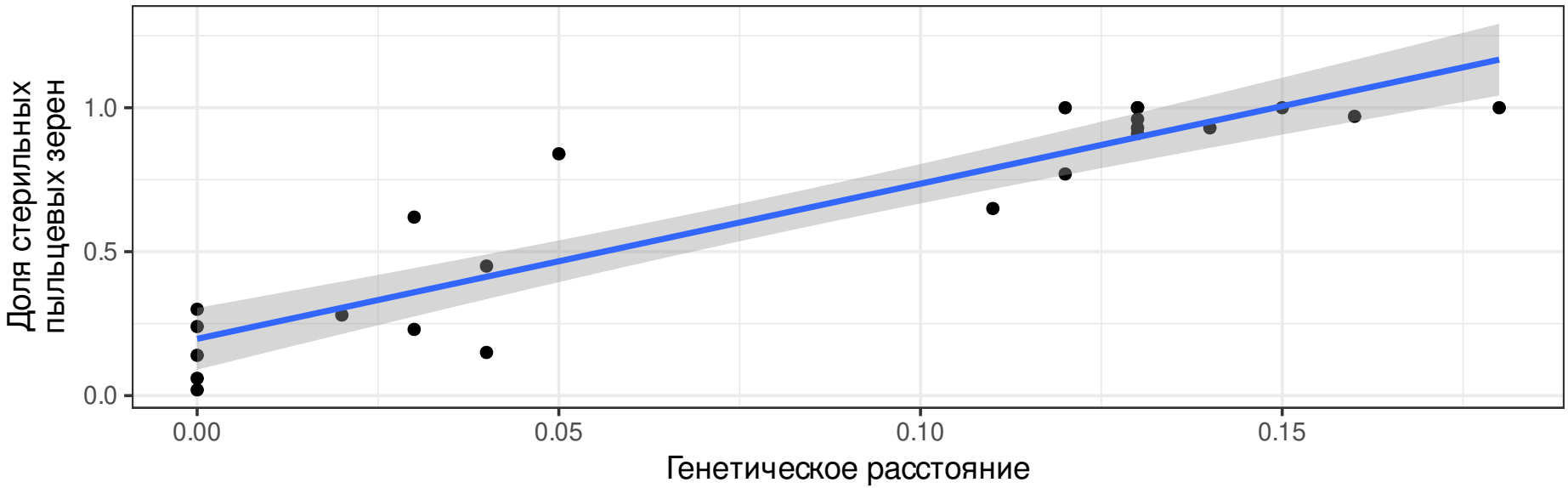
Т.е. в этой системе видов степень репродуктивной изоляции возрастает с увеличением генетического расстояния между видами.

Линейная регрессия

Линейная регрессия

- позволяет описать зависимость между количественными величинами
- позволяет предсказать значение одной величины, зная значения других

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$



- В этой картинке есть подвох, но о нем мы поговорим позже…

Линейная регрессия бывает простая и множественная

- простая

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

- множественная

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \varepsilon_i$$

Линейная регрессия в генеральной совокупности и в выборке

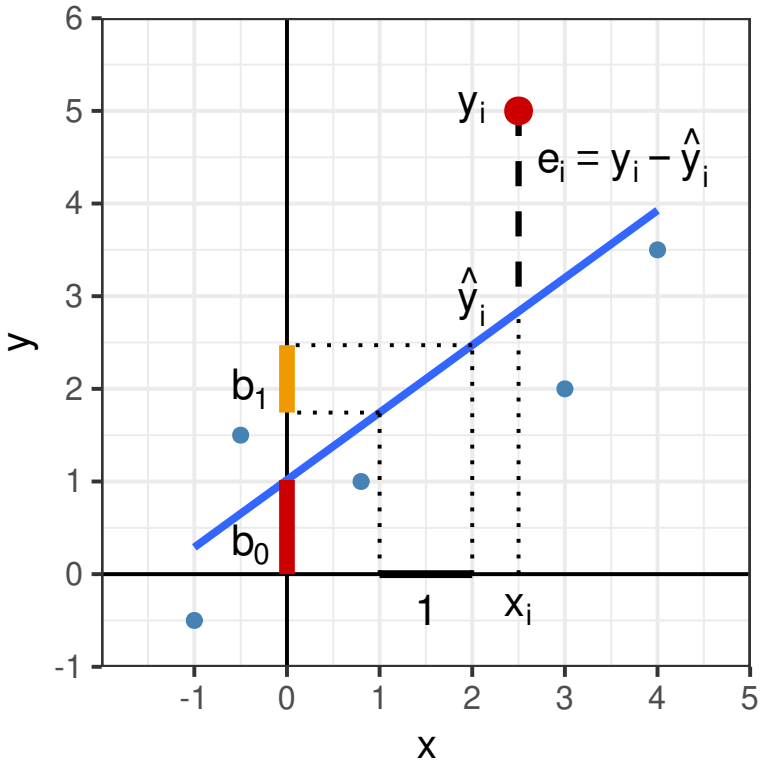
В уравнении линейной регрессии, описывающей зависимость в генеральной совокупности, обозначения записываются греческими буквами:

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

Обозначения в уравнении модели, построенной по выборке — латинскими:

$$y_i = b_0 + b_1 x_i + e_i$$

Что есть что в уравнении линейной регрессии



$$y_i = b_0 + b_1 x_i + e_i$$

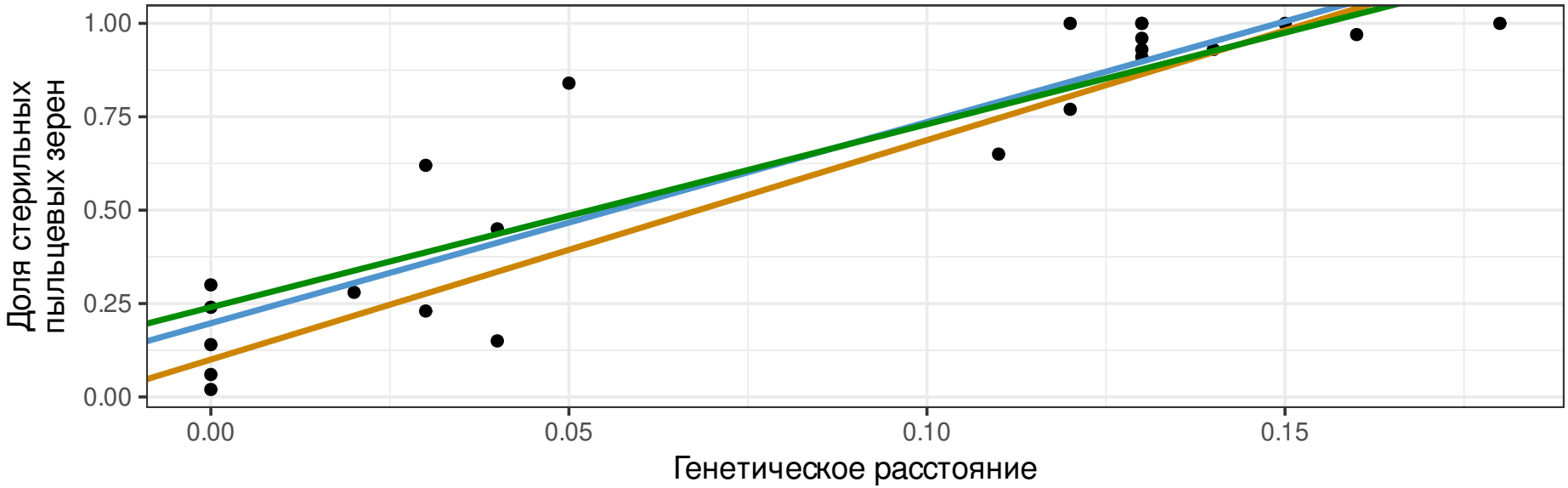
- y_i — наблюдаемое значение зависимой переменной
- $\hat{y}_i$ — предсказанное значение зависимой переменной
- e_i — остатки (отклонения наблюдаемых от предсказанных значений)

- b_0 — отрезок (Intercept), отсекаемый регрессионной прямой на оси y
- b_1 — коэффициент угла наклона регрессионной прямой

Подбор коэффициентов линейной регрессии

Как провести линию регрессии?

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



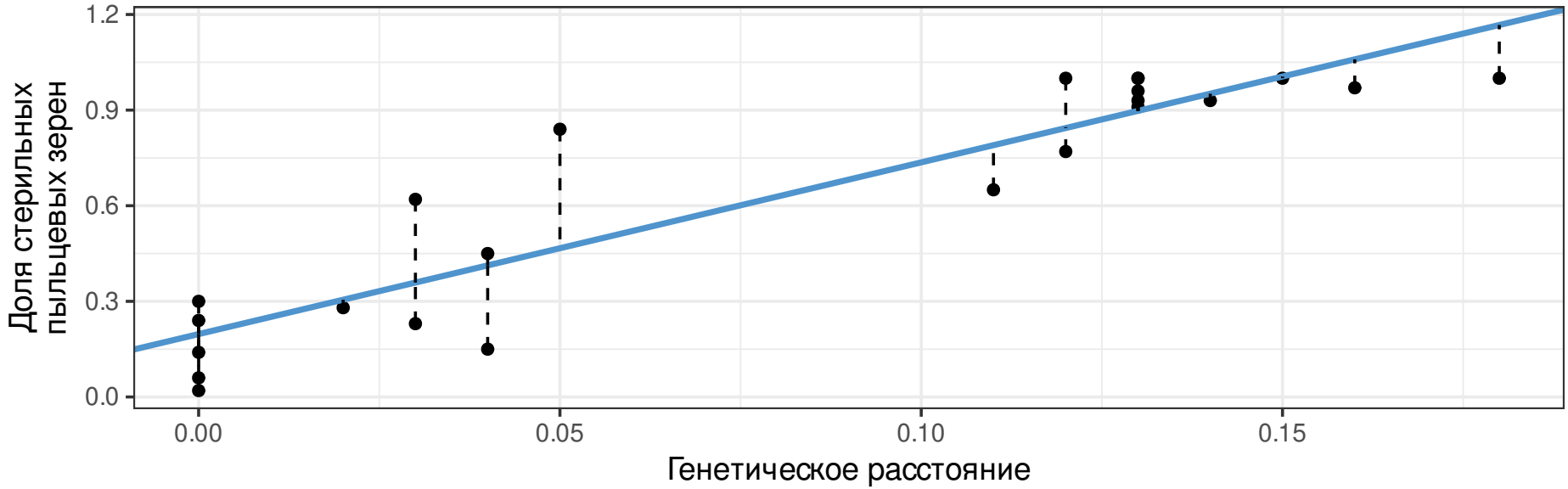
Нужно получить оценки b_0 и b_1 значений параметров линейной модели β_0 и β_1.

Но как это сделать?

Метод наименьших квадратов — один из способов подбора параметров

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

Оценки параметров линейной регрессии b_0 и b_1 подбирают так, чтобы минимизировать сумму квадратов остатков $\sum \varepsilon_i^2$, т.е. $\sum (y_i - \hat{y}_i)^2$.



Оценки параметров линейной регрессии

Параметр	Оценка	Стандартная ошибка
β_0	$b_0 = \bar{y} - b_1 \bar{x}$	$SE_{b_0} = \sqrt{MS_e [\frac{1}{n} + \frac{\bar{x}}{\sum (x_i - \bar{x})^2}]}$
β_1	$b_1 = \frac{\sum [(x_i - \bar{x})(y_i - \bar{y})]}{\sum (x_i - \bar{x})^2}$	$SE_{b_1} = \sqrt{\frac{MS_e}{\sum (x_i - \bar{x})^2}}$
ε_i	$e_i = y_i - \hat{y}_i$	$\approx \sqrt{MS_e}$

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

Стандартные ошибки коэффициентов

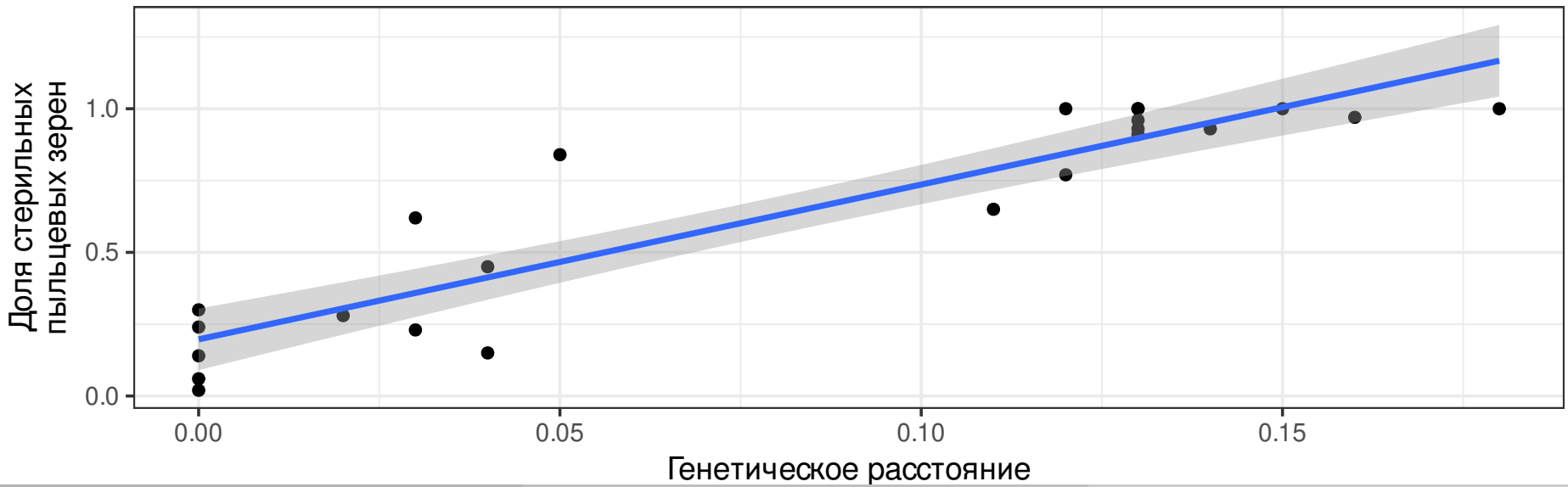
- используются для построения доверительных интервалов
- нужны для статистических тестов

Неопределенность оценки положения регрессии

Доверительный интервал коэффициента — это зона, в которой при повторных выборках из генеральной совокупности с заданной вероятностью будет лежать среднее значение оценки коэффициента. Если $\alpha = 0.05$, то получается 95% доверительный интервал.

$$b_1 \pm t_{\alpha, df=n-2} \cdot SE_{b_1}$$

Доверительная зона регрессии — это зона, в которой при повторных выборках из генеральной совокупности с заданной вероятностью лежит регрессионная прямая.

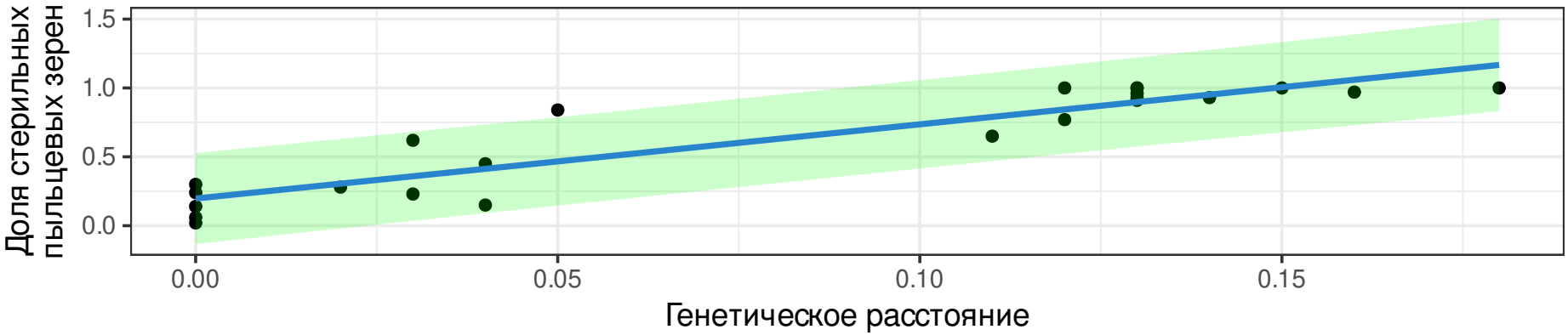


Неопределенность оценок предсказанных значений

Доверительный интервал к предсказанному значению — это зона, в которую попадает заданная доля значений $\hat{y}_i$ при данном x_i

$$\hat{y}_i \pm t_{\alpha,n-2} \cdot SE_{\hat{y}_i}, \; SE_{\hat{y}} = \sqrt{MS_e[1 + \frac{1}{n} + \frac{(x_{prediction} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]}$$

Доверительная область значений регрессии — это зона, в которую попадает $(1-\alpha) \cdot 100\%$ всех предсказанных значений



Линейная регрессия в R

Как в R задать формулу линейной регрессии

`lm(formula = формула_модели, data = данные)` - функция для подбора регрессионных моделей

Формат формулы модели: `зависимая_переменная ~ независимые_переменные`

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

Другие 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

Задание 2

Используя данные из датасета `hybrid` подберите модель линейной регрессии, описывающую зависимость доли стерильной пыльцы `Sterile` от генетического расстояния `Distance`.

Запишите коэффициенты модели и уравнение линейной регрессии.

Подсказки:

`lm(formula = формула_модели, data = данные)` — функция для подбора регрессионных моделей

Формат формулы модели: `зависимая_переменная ~ независимые_переменные`

`summary(модель)` — функция, показывающая краткую информацию о модели в виде таблицы

`coef(модель)` — функция, показывающая только коэффициенты модели

```
hybrid_lm <- lm(formula = , data = )
```

Решение: Подбираем параметры линейной модели

```
hybrid_lm <- lm(formula = Sterile ~ Distance, data = hybrid)
summary(hybrid_lm)
```

```
#
# Call:
# lm(formula = Sterile ~ Distance, data = hybrid)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -0.26276 -0.10907 -0.00538  0.08237  0.37336
#
# Coefficients:
#             Estimate Std. Error t value       Pr(>|t|)
# (Intercept)  0.19726    0.05149   3.831       0.000973 ***
# Distance     5.38747    0.51117  10.540 0.000000000766 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.1501 on 21 degrees of freedom
# Multiple R-squared:  0.841,   Adjusted R-squared:  0.8334
# F-statistic: 111.1 on 1 and 21 DF,  p-value: 0.0000000007659
```

Коэффициенты линейной регрессии:

- $b_0 = 0.2 \pm 0.05$
- $b_1 = 5.4 \pm 0.5$

Решение: Записываем уравнение линейной регрессии

Модель:

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

Коэффициенты:

```
coef(hybrid_lm)
```

```
# (Intercept)    Distance
#   0.1972632   5.3874710
```

Уравнение регрессии:

$$\widehat{Sterile}_i = 0.2 + 5.4 Distance_i$$

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 случайная (=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 (=остаточная)?

Тестируем значимость коэффициентов t-критерием

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

$H_0 : b_1 = \theta$, для $\theta = 0$
$H_A : b_1 \neq \theta$

t-статистика подчиняется t-распределению с числом степеней свободы $df = n - p$, где p — число параметров.

Для простой линейной регрессии $df = n - 2$.

Тестируем значимость коэффициентов t-критерием

```
summary(hybrid_lm)
```

```
#
# Call:
# lm(formula = Sterile ~ Distance, data = hybrid)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -0.26276 -0.10907 -0.00538  0.08237  0.37336
#
# Coefficients:
#             Estimate Std. Error t value       Pr(>|t|)
# (Intercept)  0.19726    0.05149   3.831       0.000973 ***
# Distance     5.38747    0.51117  10.540 0.000000000766 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.1501 on 21 degrees of freedom
# Multiple R-squared:  0.841,   Adjusted R-squared:  0.8334
# F-statistic: 111.1 on 1 and 21 DF,  p-value: 0.0000000007659
```

Результаты можно описать в тексте так:

- Доля стерильной пыльцы у межвидовых гибридов значимо возрастает с увеличением генетического расстояния ($b_1 = 5.39$, $t = 10.54$, $p < 0.01$)

Тестируем значимость модели целиком при помощи F-критерия

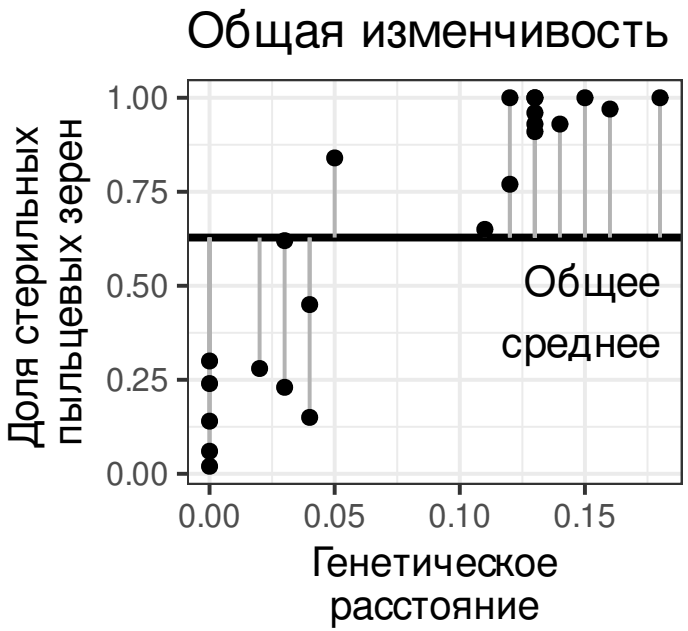
$$F = \frac{MS_{regression}}{MS_{error}}$$

$H_0 : \beta_1 = 0$

Число степеней свободы $df_{regression}$, df_{error}

Общая изменчивость

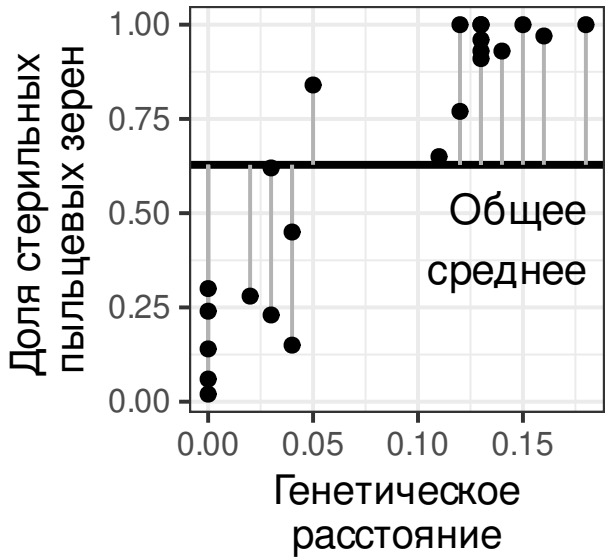
Общая изменчивость — SS_{total}, сумма квадратов отклонений от общего среднего значения



Общая изменчивость делится на объясненную и остаточную

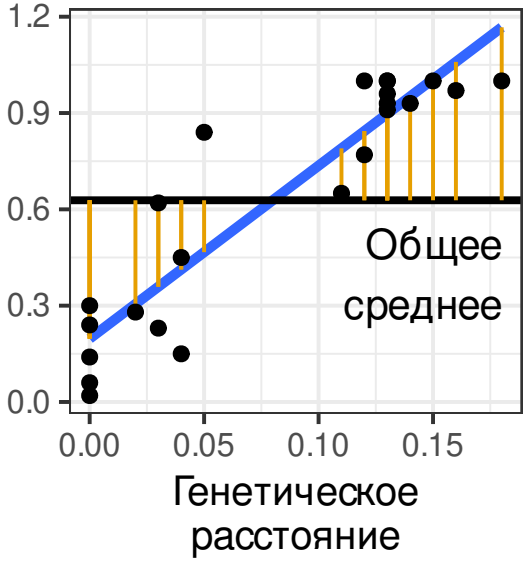
$$SS_t = SS_r + SS_e \qquad MS_t \neq MS_r + MS_e$$

Общая



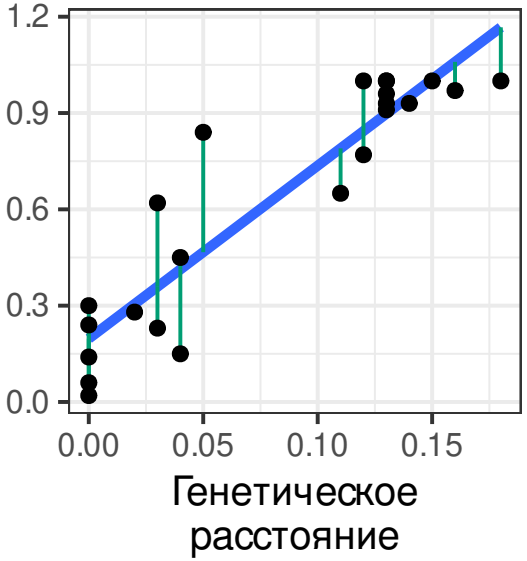
$$SS_t = \sum (y_i - \bar{y})^2$$
$$df_t = n - 1$$
$$MS_t = \frac{SS_t}{df_t}$$

Объясненная



$$SS_r = \sum (\hat{y} - \bar{y})^2$$
$$df_r = p - 1$$
$$MS_r = \frac{SS_r}{df_r}$$

Случайная



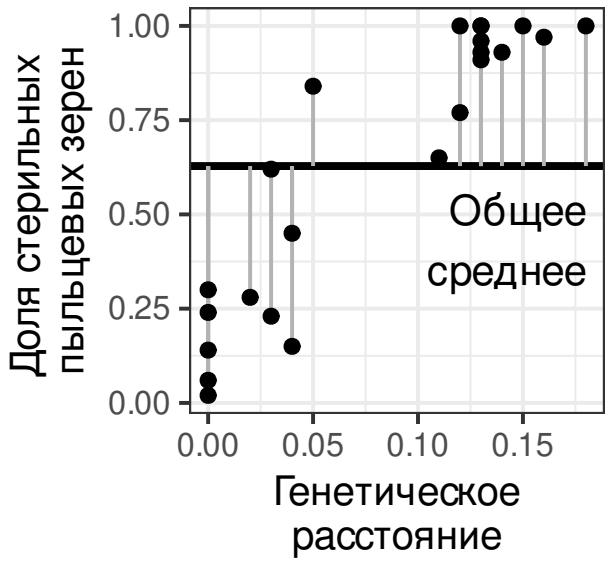
$$SS_e = \sum (y_i - \hat{y})^2$$
$$df_e = n - p$$
$$MS_e = \frac{SS_e}{df_e}$$

Если зависимости нет, то коэффициент $b_1 = 0$

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

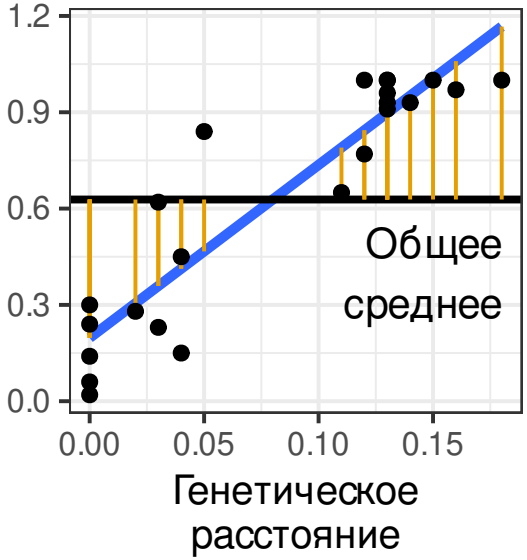
Это можно использовать при тестировании гипотезы $\beta_1 = 0$.

Общая



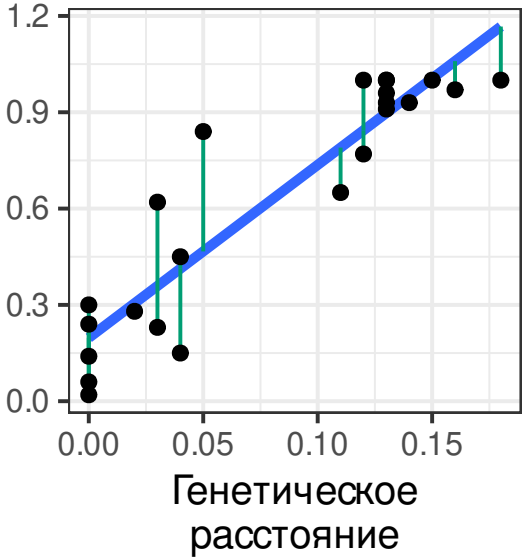
$$SS_t = \sum (y_i - \bar{y})^2$$
$$df_t = n - 1$$
$$MS_t = \frac{SS_t}{df_t}$$

Объясненная



$$SS_r = \sum (\hat{y} - \bar{y})^2$$
$$df_r = p - 1$$
$$MS_r = \frac{SS_r}{df_r}$$

Случайная



$$SS_e = \sum (y_i - \hat{y})^2$$
$$df_e = n - p$$
$$MS_e = \frac{SS_e}{df_e}$$

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:

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

Подчиняется F-распределению с параметрами df_r и df_e.

Для простой линейной регрессии $df_r = 1$ и $df_e = n - 2$.

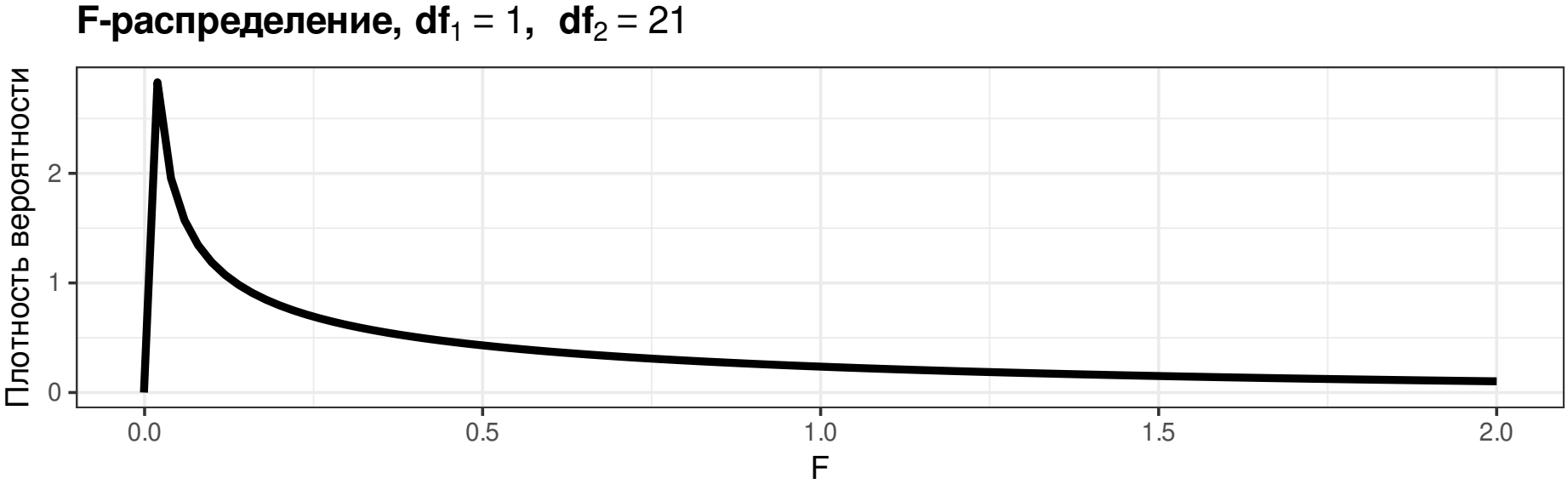


Таблица результатов дисперсионного анализа

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

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

Проверяем значимость модели при помощи F-критерия

```
library(car)
Anova(hybrid_lm)
```

```
# Anova Table (Type II tests)
#
# Response: Sterile
#           Sum Sq Df F value          Pr(>F)
# Distance  2.50194  1  111.08 0.0000000007659 ***
# Residuals 0.47299 21
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

- Доля стерильной пыльцы межвидовых гибридов смолевок значимо зависит от генетического расстояния ($F_{1,21} = 111.08$, $p < 0.001$).

График линейной регрессии

Задание 3

Дополните график `gg_hybrid`, чтобы построить - 95% доверительную зону регрессии - 99% доверительную зону регрессии

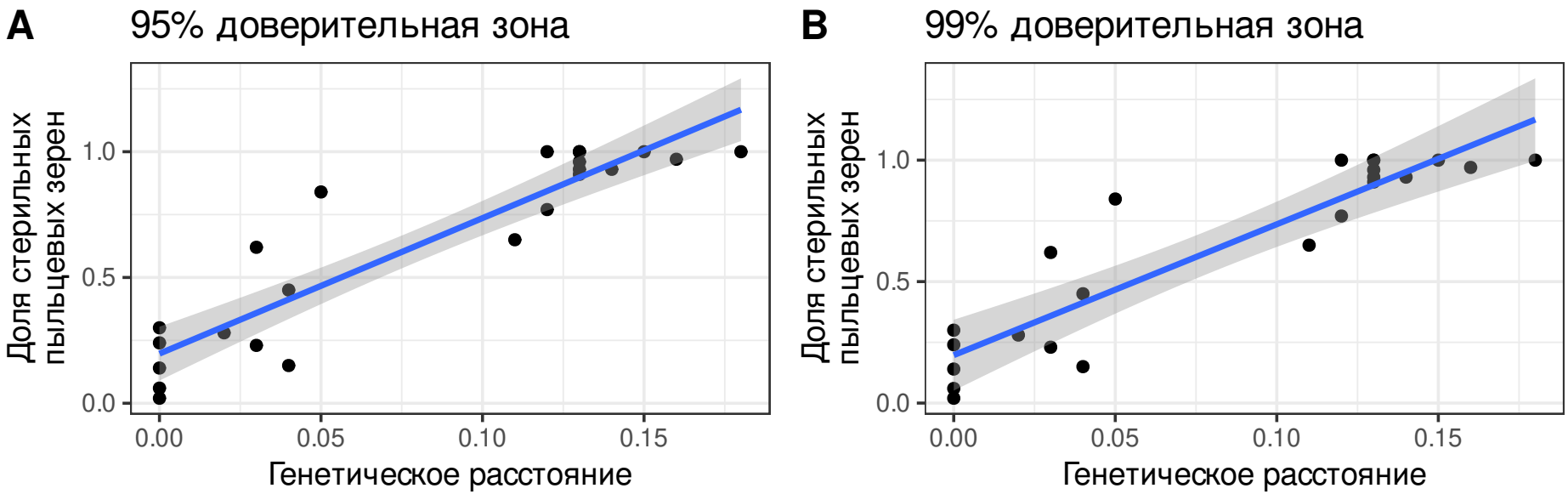
Используйте `geom_smooth()` и его аргументы `method` и `level`

```
gg1 <- gg_hybrid +
  labs(title = '95% доверительная зона')
gg1
gg2 <- gg_hybrid +
  labs(title = '99% доверительная зона')
gg2

library(cowplot)
plot_grid(gg1, gg2, nrow = 1, labels = 'AUTO')
```

Решение: Строим доверительную зону регрессии

```
gg1 <- gg_hybrid + geom_smooth(method = 'lm') +
  labs(title = '95% доверительная зона')
gg2 <- gg_hybrid + geom_smooth(method = 'lm', level = 0.99) +
  labs(title = '99% доверительная зона')
library(cowplot)
plot_grid(gg1, gg2, nrow = 1, labels = 'AUTO')
```



Оценка качества подгонки модели

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(hybrid_lm)
```

```
#
# Call:
# lm(formula = Sterile ~ Distance, data = hybrid)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -0.26276 -0.10907 -0.00538  0.08237  0.37336
#
# Coefficients:
#             Estimate Std. Error t value       Pr(>|t|)
# (Intercept)  0.19726    0.05149   3.831       0.000973 ***
# Distance     5.38747    0.51117  10.540 0.000000000766 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.1501 on 21 degrees of freedom
# Multiple R-squared:  0.841,   Adjusted R-squared:  0.8334
# F-statistic: 111.1 on 1 and 21 DF,  p-value: 0.0000000007659
```

Сравнение качества подгонки моделей при помощи R^2_{adj}

R^2_{adj} — скорректированный R^2

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

где $df_e = n - p$, $df_t = n - 1$

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

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 доля стерильной пыльцы межвидового гибрида, если генетическое расстояние между родителями 0.07 или 0.055?

Значения, для которых предсказываем:

```
new_data1 <- data.frame(Distance = c(0.07, 0.055))
new_data1
```

```
#   Distance
# 1    0.070
# 2    0.055
```

Предсказания

```
(pr1 <- predict(hybrid_lm, newdata = new_data1,
                interval = 'confidence', se = TRUE))
```

```
# $fit
#         fit       lwr       upr
# 1 0.5743862 0.5084456 0.6403267
# 2 0.4935741 0.4232788 0.5638693
#
# $se.fit
#          1          2
# 0.03170808 0.03380207
#
# $df
# [1] 21
```

Предсказываем изменение Y для 95% наблюдений при заданном X

В каких пределах находится доля стерильной пыльцы, если генетическое расстояние между родителями 0.07 или 0.055?

```
# значения, для которых предсказываем
new_data1 <- data.frame(Distance = c(0.07, 0.055))
(pr2 <- predict(hybrid_lm, newdata = new_data1,
                interval = 'prediction', se = TRUE))
```

```
# $fit
#         fit       lwr       upr
# 1 0.5743862 0.2553929 0.8933794
# 2 0.4935741 0.1736523 0.8134959
#
# $se.fit
#          1          2
# 0.03170808 0.03380207
#
# $df
# [1] 21
#
# $residual.scale
# [1] 0.1500776
```

- У 95% межвидовых гибридов, у которых генетическое расстояние между родителями 0.07 или 0.055, доля стерильной пыльцы будет в пределах 0.6 ± 0.3 и 0.5 ± 0.3, соответственно.

Построим график доверительной области значений

Создадим данные для графика.

Для этого объединим в новом датафрейме:

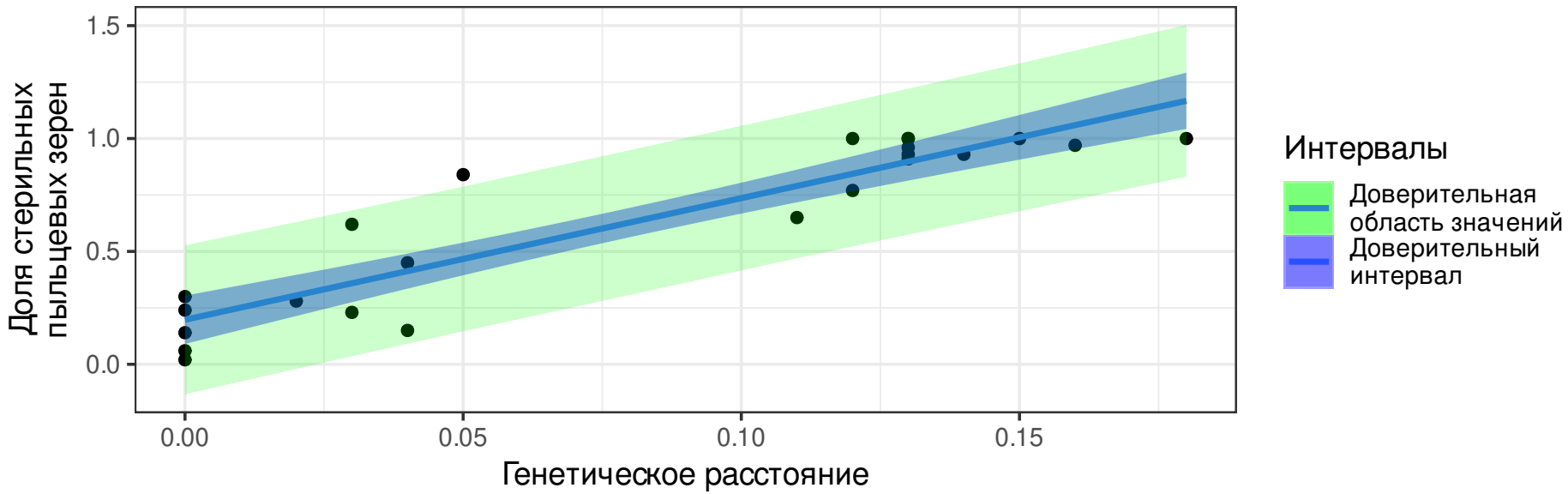
- исходные данные
- предсказанные значения для исходных данных

```
pr_all <- predict(hybrid_lm, interval = 'prediction')
hybrid_with_pred <- data.frame(hybrid, pr_all)
head(hybrid_with_pred)
```

```
#   Distance Sterile       fit          lwr       upr
# 1     0.00    0.02 0.1972632 -0.13270028 0.5272267
# 2     0.00    0.06 0.1972632 -0.13270028 0.5272267
# 3     0.00    0.14 0.1972632 -0.13270028 0.5272267
# 4     0.00    0.24 0.1972632 -0.13270028 0.5272267
# 5     0.00    0.30 0.1972632 -0.13270028 0.5272267
# 6     0.03    0.62 0.3588873  0.03567102 0.6821036
```

Строим доверительную область значений и доверительный интервал одновременно

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

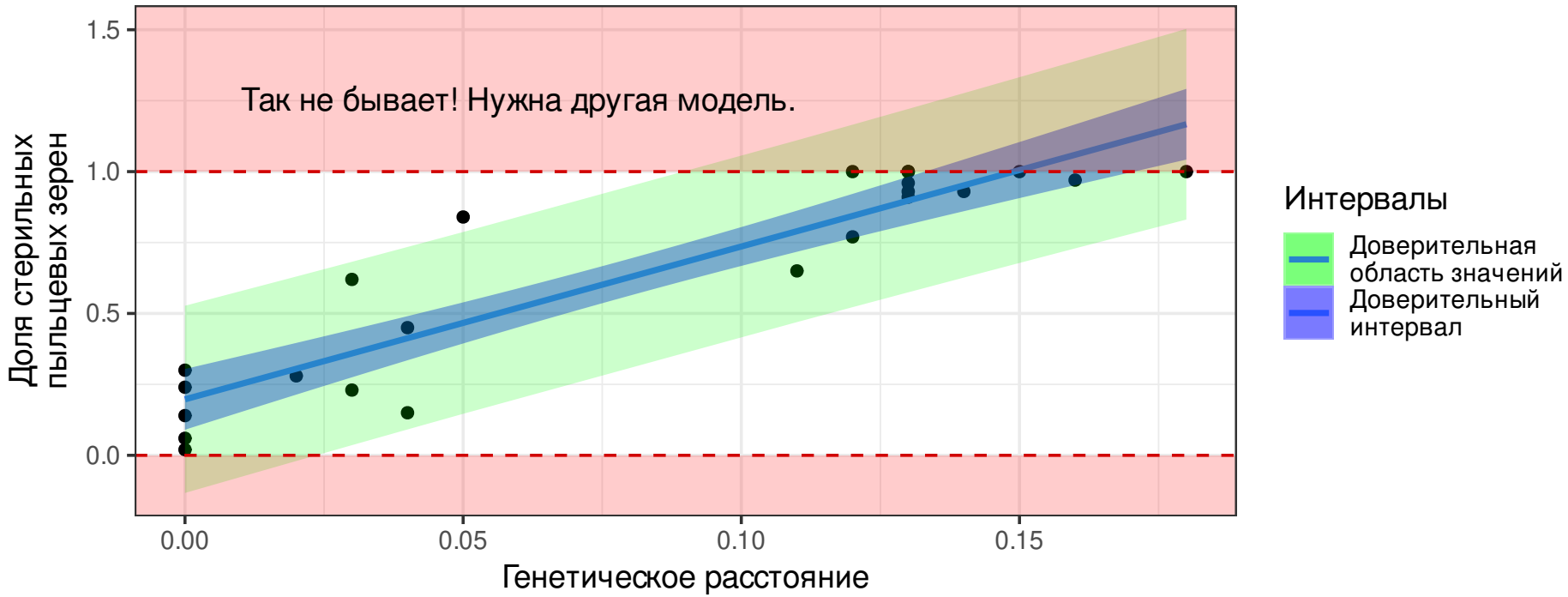


Осторожно! У такой линейной регрессии есть проблемы!

Для некоторых значений генетического расстояния построенная нами модель предсказывает больше 100% стерильной пыльцы.

Так не бывает!

Вместо простой линейной регрессии нужно использовать более сложную линейную модель (это за рамками курса)



Take home messages

- Модель простой линейной регрессии $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- В оценке коэффициентов регрессии (b_0 и b_1) и предсказанных значений ($\hat{y}_i$) существует неопределенность. Доверительные интервалы можно рассчитать, зная стандартные ошибки.
- Значимость всей регрессии и ее параметров можно проверить при помощи t- или F-теста. Для простой линейной регрессии $H_0 : \beta_1 = 0$.
- Качество подгонки модели можно оценить при помощи коэффициента детерминации R^2
- Не всякие данные можно описать при помощи простой линейной регрессии.

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