

Корреляция и регрессия

Основы биостатистики, осень 2022

Марина Варфоломеева

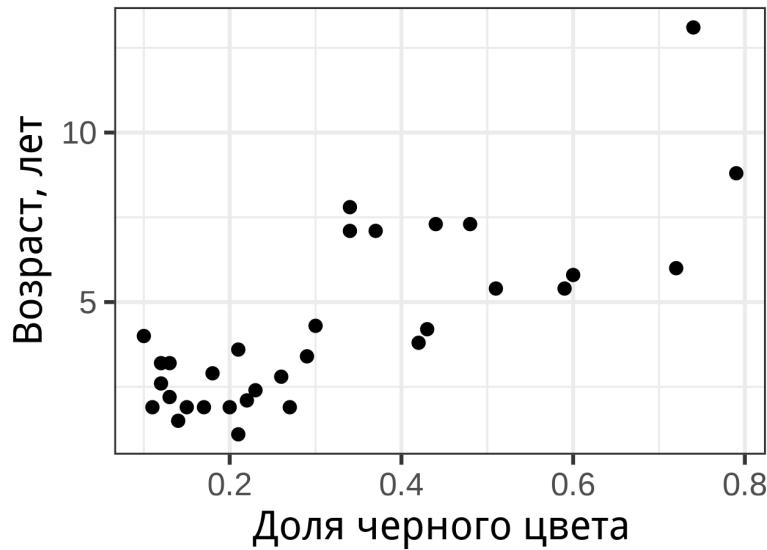
- Корреляция
- Простая линейная регрессия
- Условия применимости линейной регрессии

Корреляция

Пример: львийные носы

Определение возраста львов на расстоянии важно, чтобы решить, на кого можно охотиться.

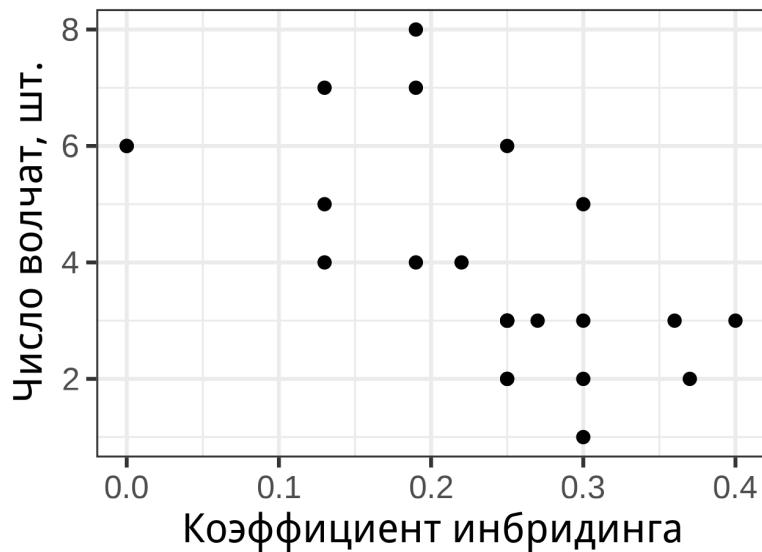
Есть ли связь между степенью пигментации львиного носа и возрастом льва? (данные Whitman et al., 2004)



Пример: инбридинг у волков

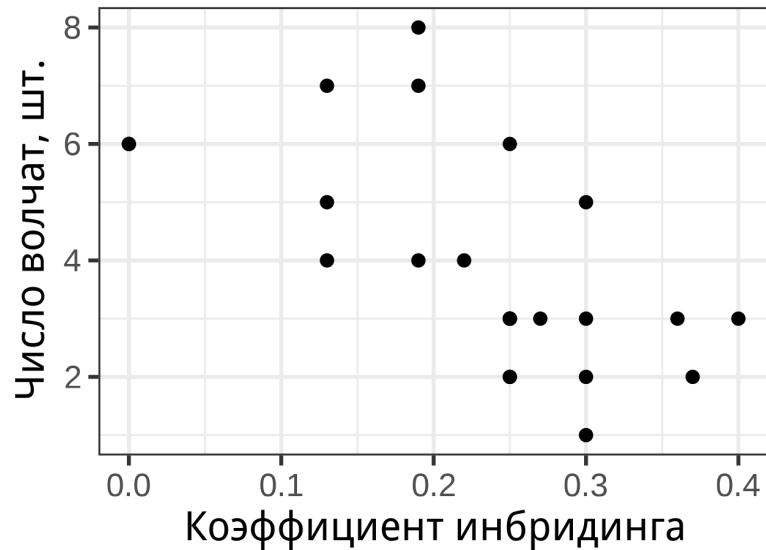
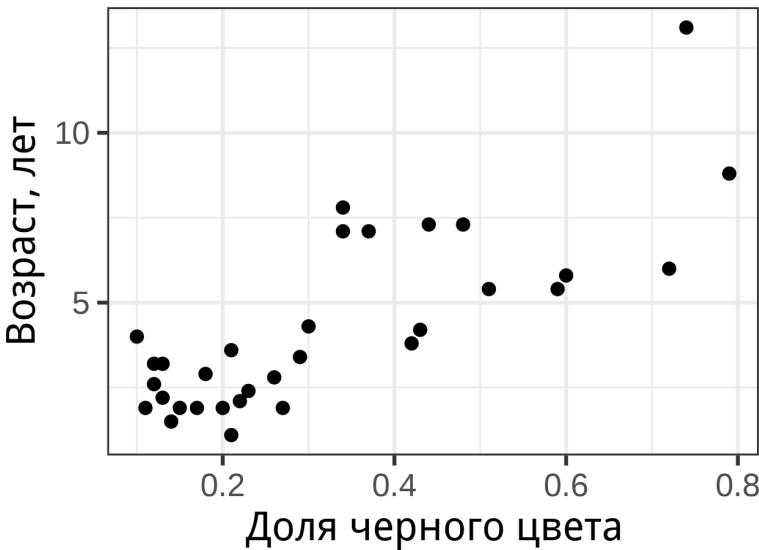
В 70-80 волки в Норвегии и Швеции прошли через бутылочное горлышко. Популяция восстановилась всего от пары особей, поэтому можно ожидаемо наблюдать последствия инбридинга.

Связан ли коэффициент инбридинга и число волчат в выводке, переживших свою первую зиму? (данные Liberg et al, 2005)



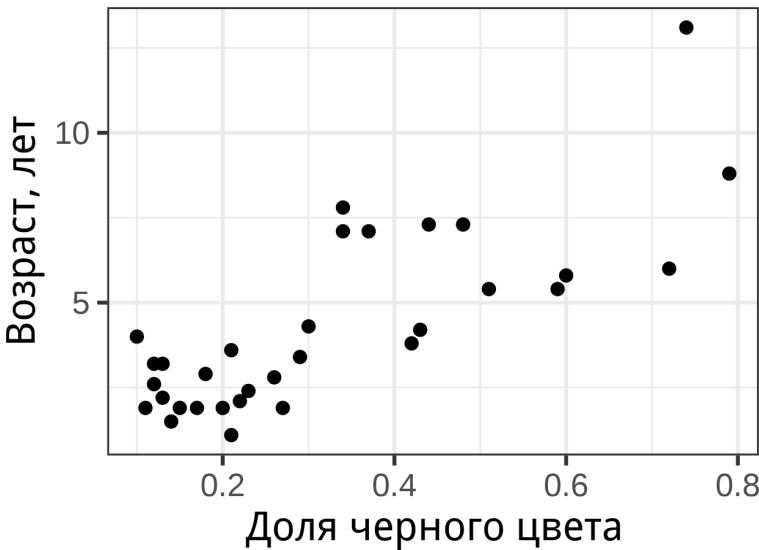
Корреляция

Когда переменные взаимосвязаны друг с другом, говорят, что между ними есть корреляция.

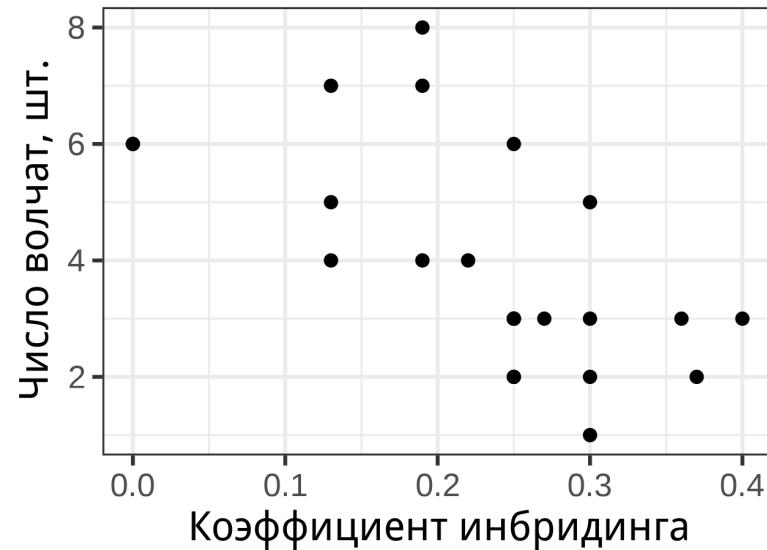


Корреляция

Когда переменные взаимосвязаны друг с другом, говорят, что между ними есть корреляция.



Положительная корреляция — чем больше одна величина, тем больше другая.



Отрицательная корреляция — чем больше одна величина, тем меньше другая.

Коэффициент корреляции Пирсона

— оценивает силу и направление связи между численными величинами.

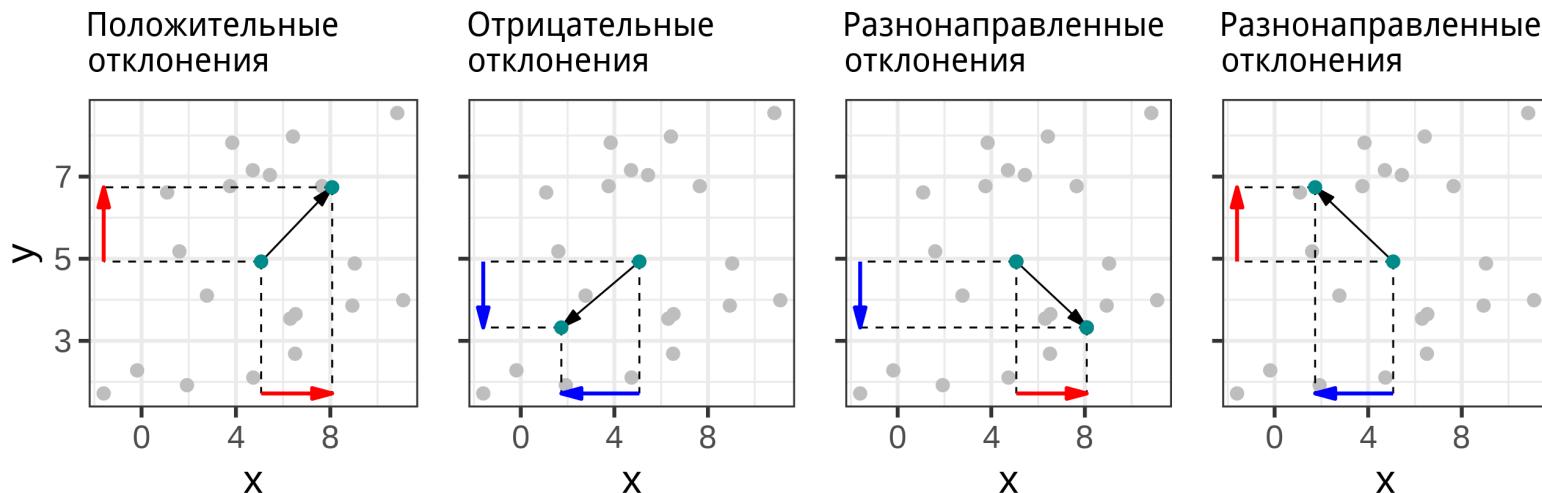
$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

Коэффициент корреляции Пирсона

— оценивает силу и направление связи между численными величинами.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

В числителе — сумма произведений **отклонений** переменных от их средних.

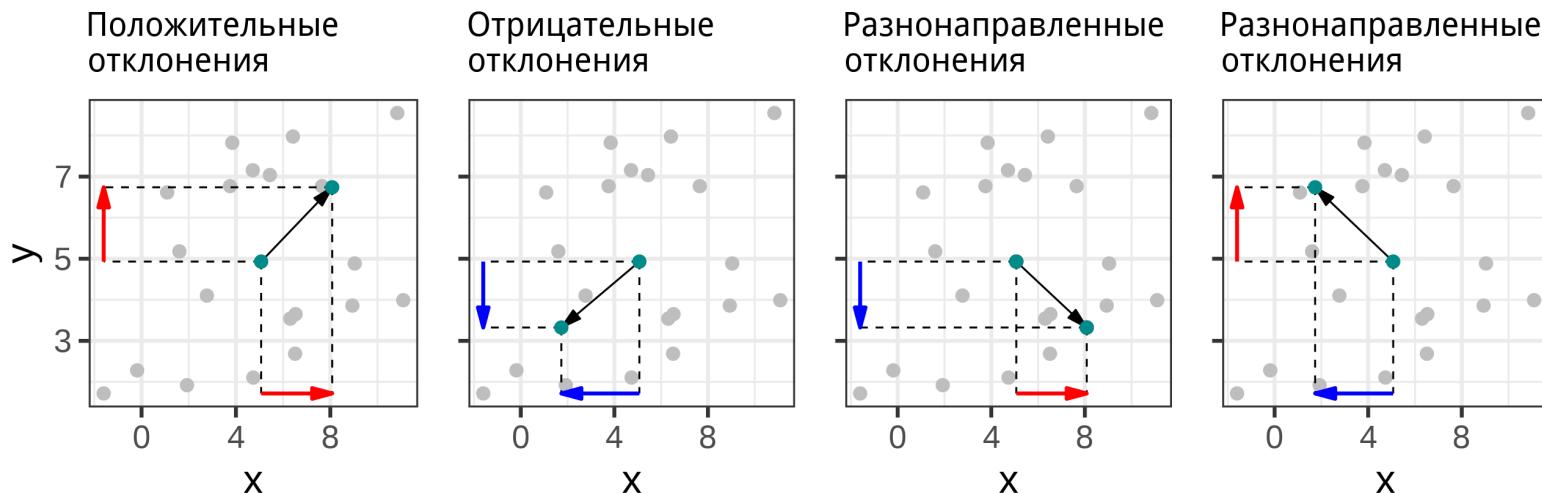


Коэффициент корреляции Пирсона

— оценивает силу и направление связи между численными величинами.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

В числителе — сумма произведений **отклонений** переменных от их средних.



$$-1 < r < 1$$

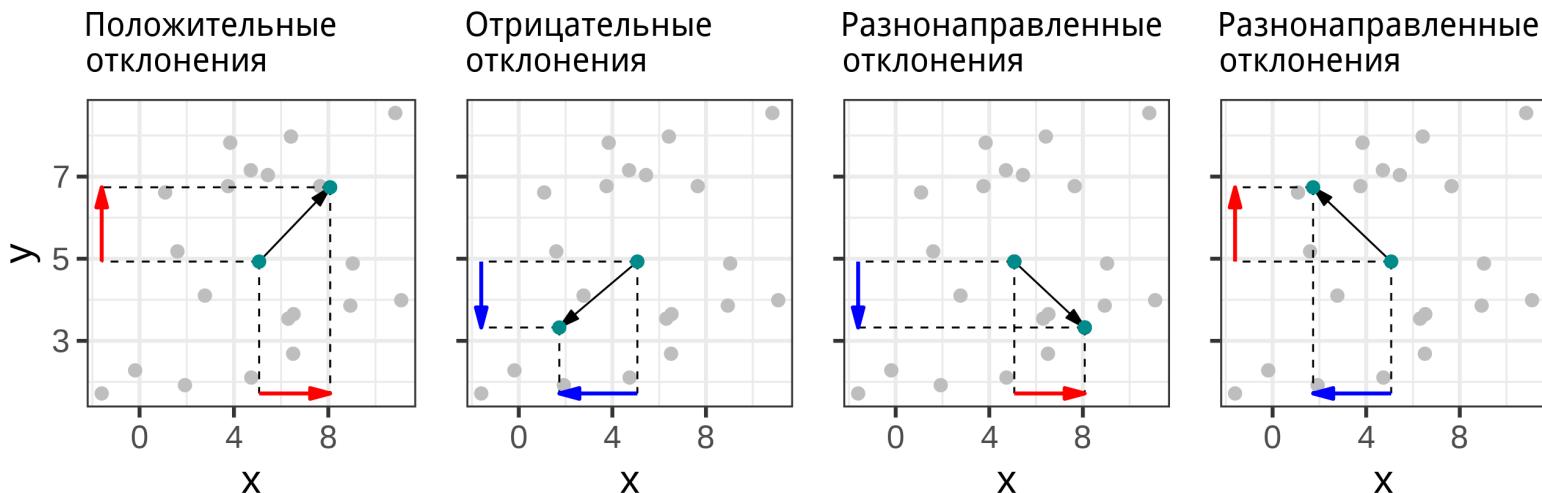
- $|r| = 1$ — сильная связь

Коэффициент корреляции Пирсона

— оценивает силу и направление связи между численными величинами.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

В числителе — сумма произведений **отклонений** переменных от их средних.



$$-1 < r < 1$$

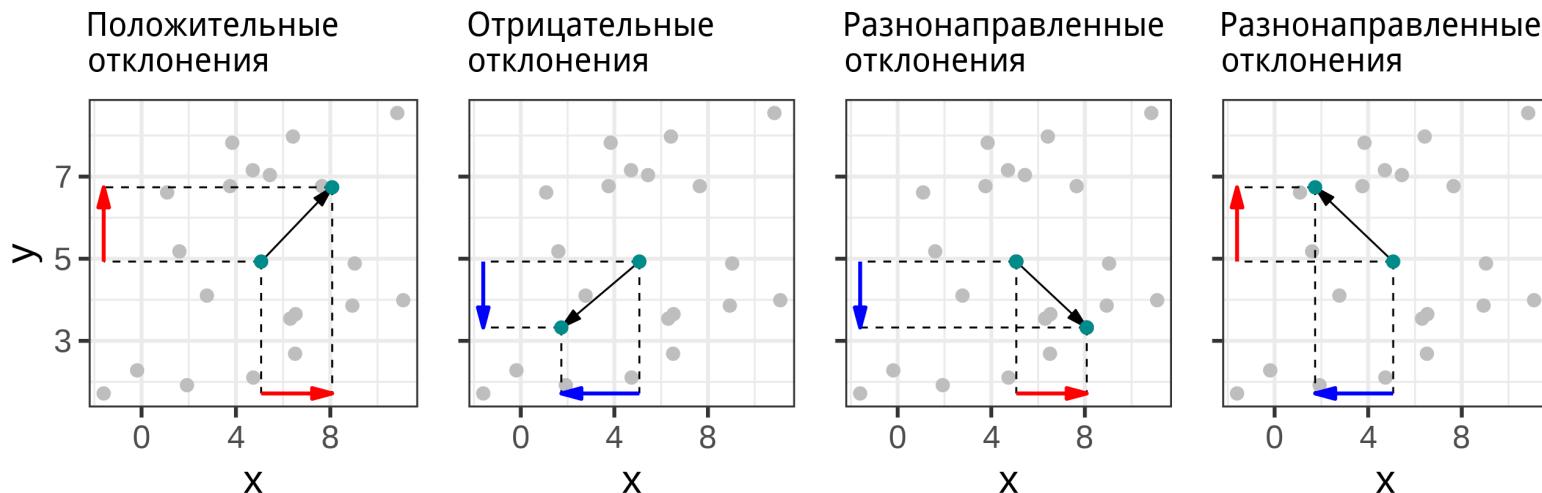
- $|r| = 1$ — сильная связь: $r > 0$ — положительная, $r < 0$ — отрицательная

Коэффициент корреляции Пирсона

— оценивает силу и направление связи между численными величинами.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

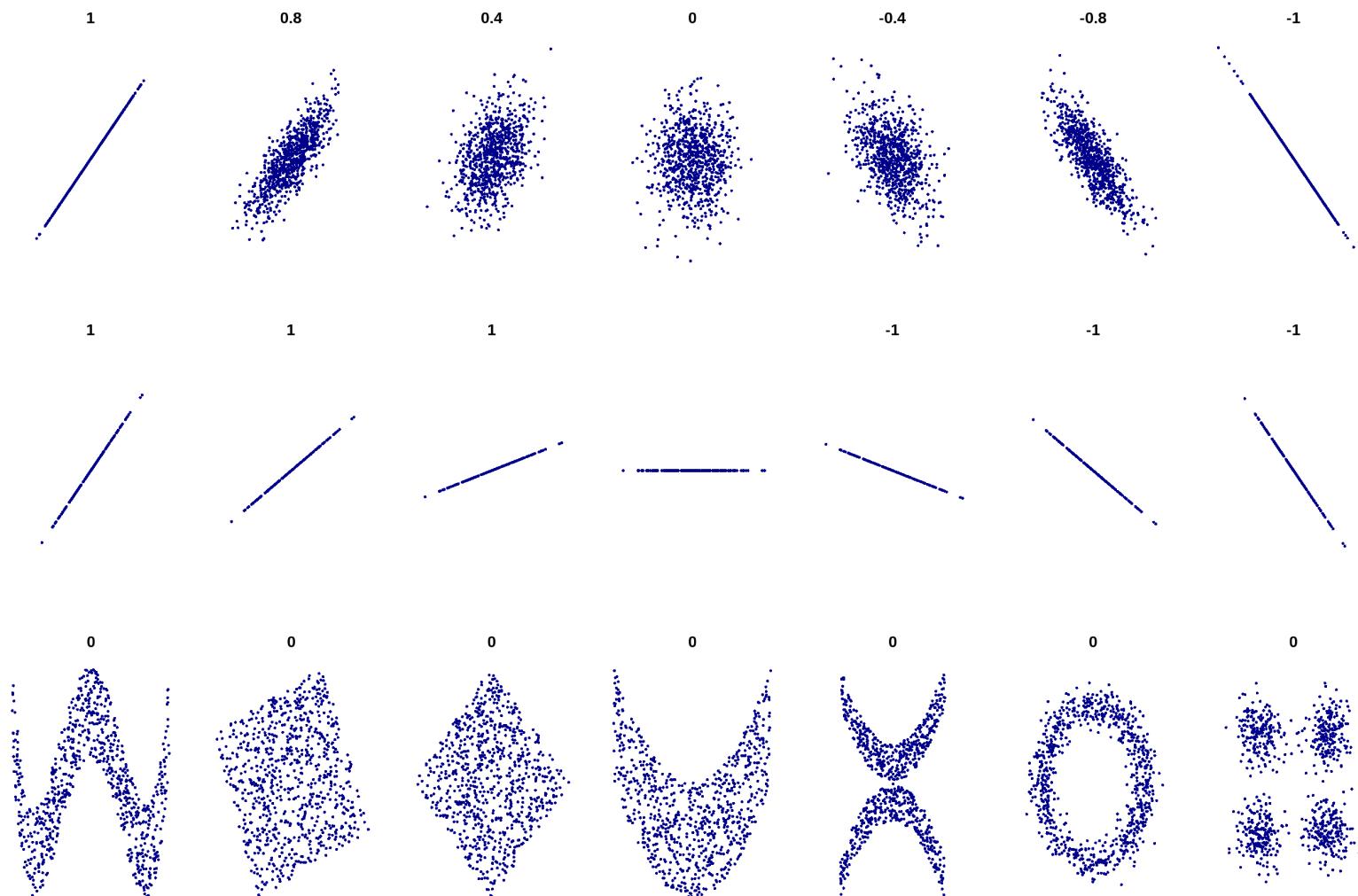
В числителе — сумма произведений **отклонений** переменных от их средних.



$$-1 < r < 1$$

- $|r| = 1$ — сильная связь: $r > 0$ — положительная, $r < 0$ — отрицательная
- $r = 0$ — нет связи

Корреляция на графике



Стандартная ошибка коэффициента корреляции

r — корреляция, рассчитаная по данным, это оценка истинного значения корреляции ρ в генеральной совокупности.

Стандартная ошибка этой оценки:

$$SE_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

Ее не получится использовать для доверительного интервала, т.к. ее выборочное распределение не нормально.

Но ее можно использовать для тестов.

Приблизительный доверительный интервал

Z-преобразование Фишера:

$$z = 0.5 \ln \left(\frac{1+r}{1-r} \right)$$

Приблизительный доверительный интервал

Z-преобразование Фишера:

$$z = 0.5 \ln\left(\frac{1+r}{1-r}\right)$$

Стандартная ошибка выборочного распределения z

$$SE_z = \sqrt{\frac{1}{n-3}}$$

Приблизительный доверительный интервал

Z-преобразование Фишера:

$$z = 0.5 \ln \left(\frac{1+r}{1-r} \right)$$

Стандартная ошибка выборочного распределения z

$$SE_z = \sqrt{\frac{1}{n-3}}$$

Доверительный интервал для Z-преобразованного значения корреляции

$$z - 1.96 \cdot SE_z < z < z + 1.96 \cdot SE_z$$

Приблизительный доверительный интервал

Z-преобразование Фишера:

$$z = 0.5 \ln\left(\frac{1+r}{1-r}\right)$$

Стандартная ошибка выборочного распределения z

$$SE_z = \sqrt{\frac{1}{n-3}}$$

Доверительный интервал для Z-преобразованного значения корреляции

$$z - 1.96 \cdot SE_z < z < z + 1.96 \cdot SE_z$$

Границы нужно трансформировать обратно из z шкалы в r , чтобы получить доверительный интервал для r

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Тестирование значимости коэффициента корреляции

$H_0 : \rho = 0$ — нет связи между переменными (в генеральной совокупности корреляция ρ между ними равна нулю)

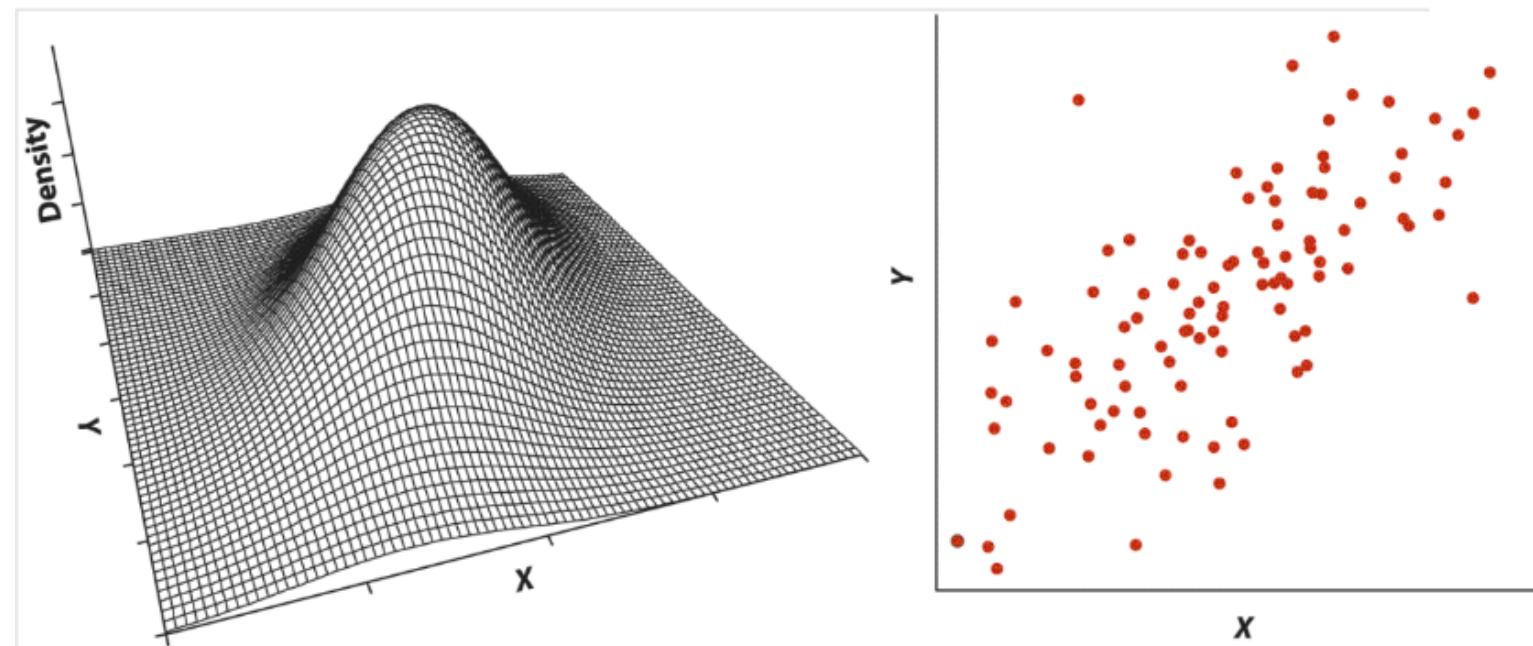
$H_A : \rho \neq 0$ — между переменными есть связь

$$t = \frac{r - 0}{\text{SE}_r}$$

$$df = n - 2$$

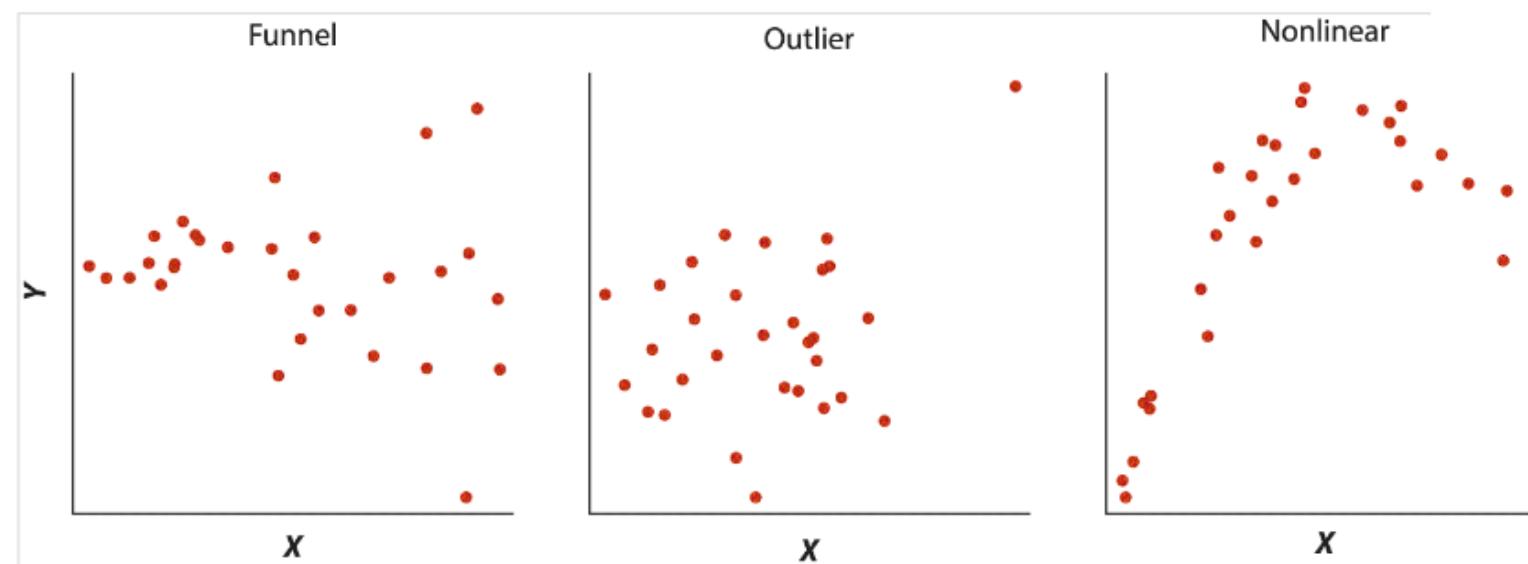
Условия применимости коэффициента корреляции Пирсона

(1) Двумерное нормальное распределение переменных (требуется для работы тестов значимости)



Условия применимости коэффициента корреляции Пирсона

- (2) Не должно быть гетерогенности дисперсий
- (3) В данных не должно быть выбросов (= outliers)
- (4) Связь должна быть линейной. Если связь нелинейна, то коэффициент корреляции Пирсона оценит только ее линейную составляющую.



Если нарушены условия применимости

Что можно сделать:

- выбросы можно удалить, если есть аргументы в пользу этого
- трансформация данных может помочь:
 - для линеаризации зависимости
 - для нормализации формы распределения

Если нарушены условия применимости

Что можно сделать:

- выбросы можно удалить, если есть аргументы в пользу этого
- трансформация данных может помочь:
 - для линеаризации зависимости
 - для нормализации формы распределения

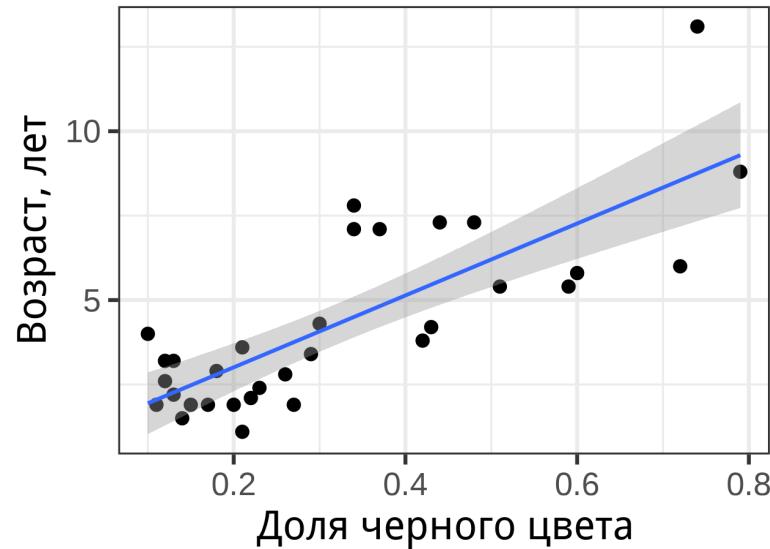
В других случаях ранговые коэффициенты корреляции:

- кор. Кендалла
- кор. Спирмена,
и т.д.

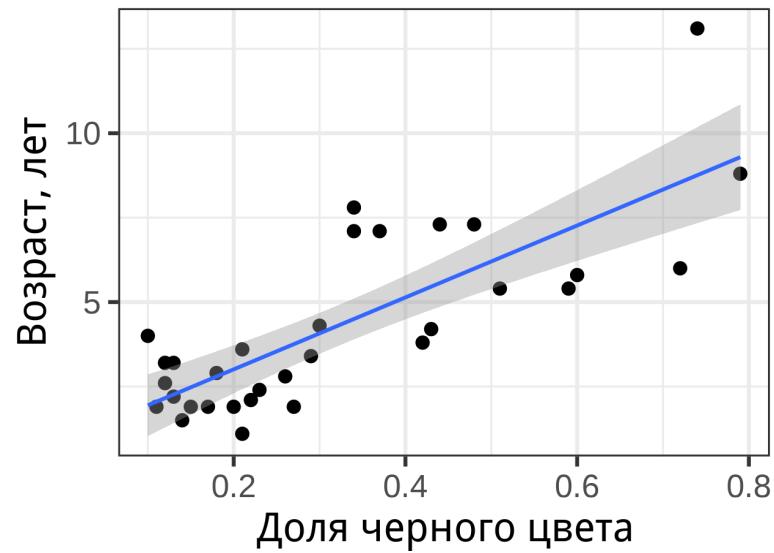
Линейная регрессия

Линейная регрессия

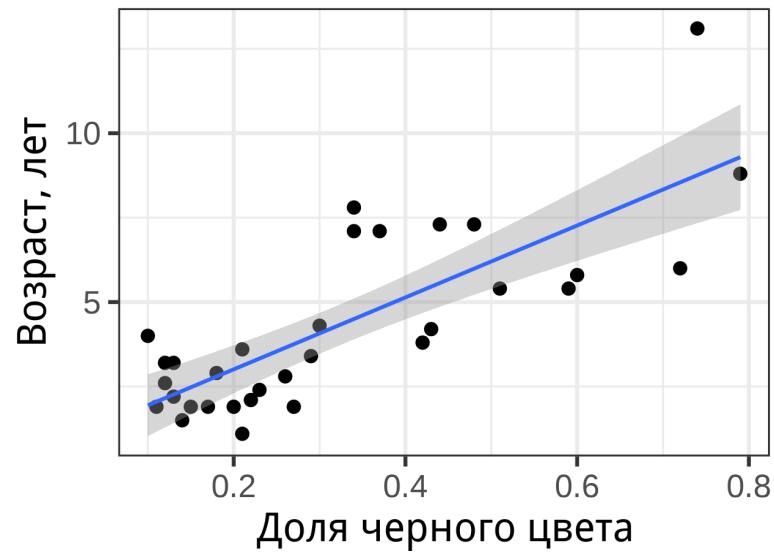
- позволяет описать зависимость между количественными величинами
- позволяет предсказать значение одной величины, зная значения других



Уравнение линейной регрессии



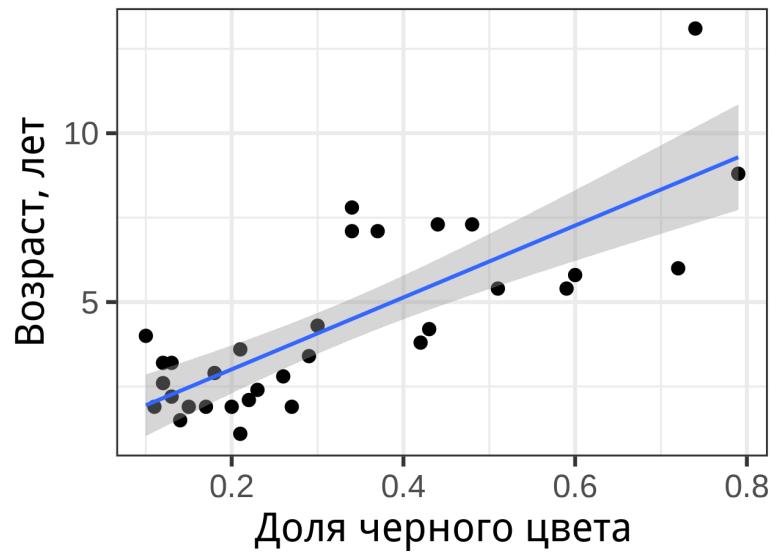
Уравнение линейной регрессии



В выборке

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

Уравнение линейной регрессии



В выборке

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

В генеральной совокупности

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

Линейная регрессия бывает простая и множественная

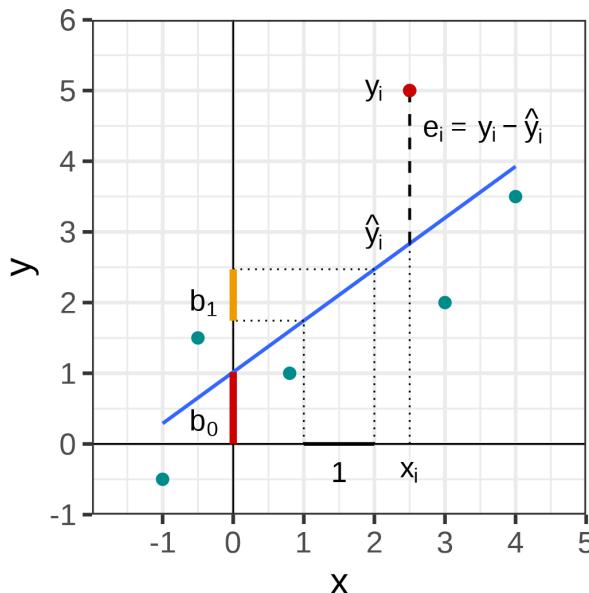
- простая

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

- множественная

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

Коэффициенты линейной регрессии

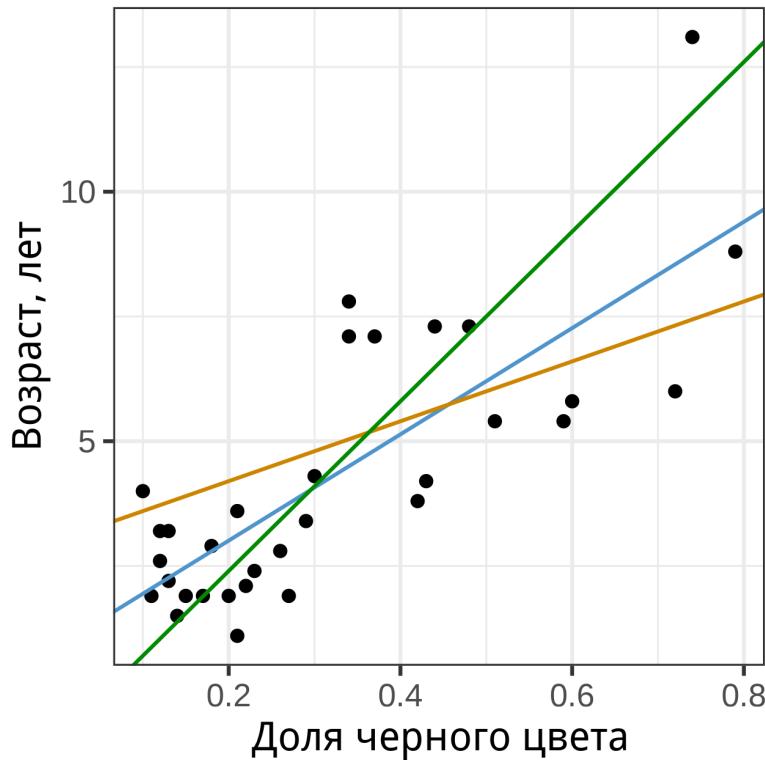


$$y_i = b_0 + b_1 x_i + e_i$$

- y_i — наблюдаемое значение (= observed) зависимой переменной, отклик
- x_i — значение независимой переменной, предиктор (=predictor)
- \hat{y}_i — предсказанное значение (= fitted, predicted) зависимой переменной
- e_i — остатки (= residuals), отклонения наблюдаемых от предсказанных значений

- b_0 — свободный член линейной модели, отрезок (intercept), отсекаемый регрессионной прямой на оси y
- b_1 — коэффициент угла наклона (slope) регрессионной прямой

Как провести линию регрессии?

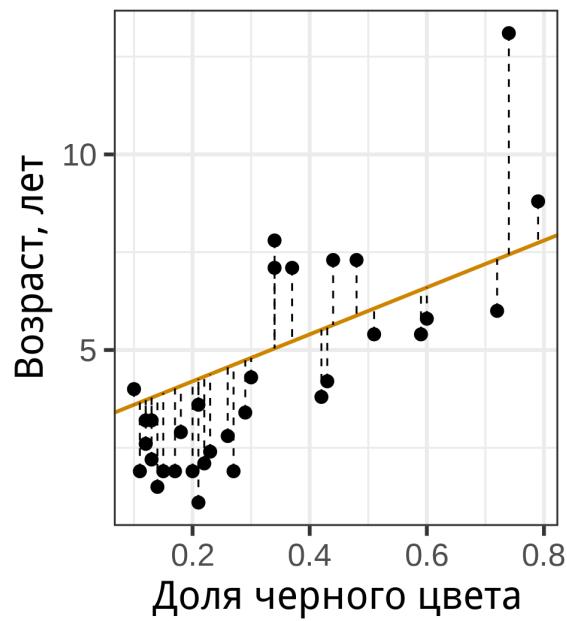


$$\hat{y}_i = b_0 + b_1 x_i$$

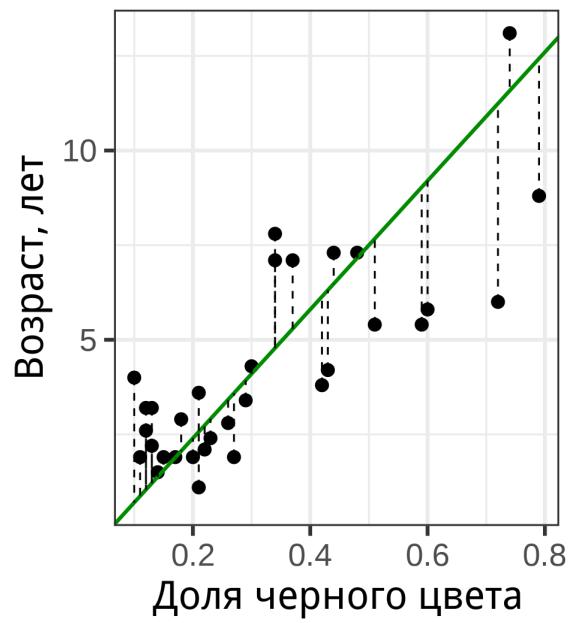
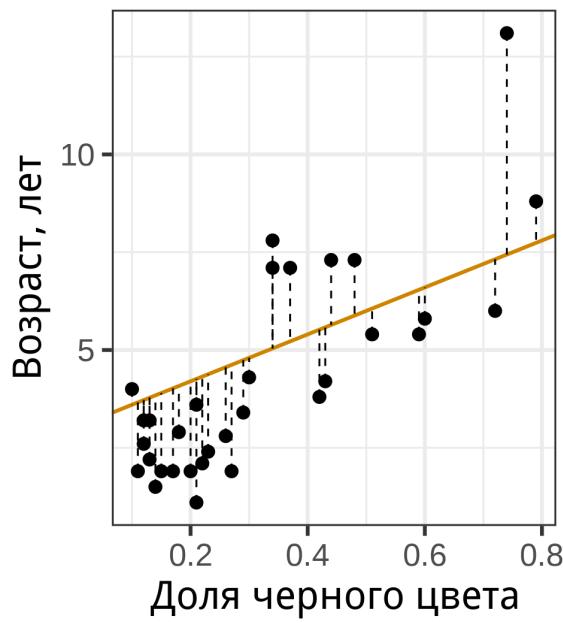
Нужно получить b_0 и b_1 — оценки истинных параметров линейной модели β_0 и β_1 .

Линия должна проходить как можно ближе к точкам

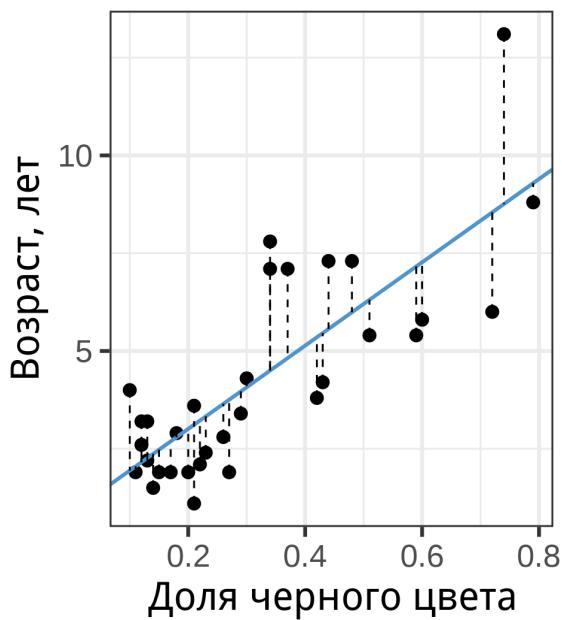
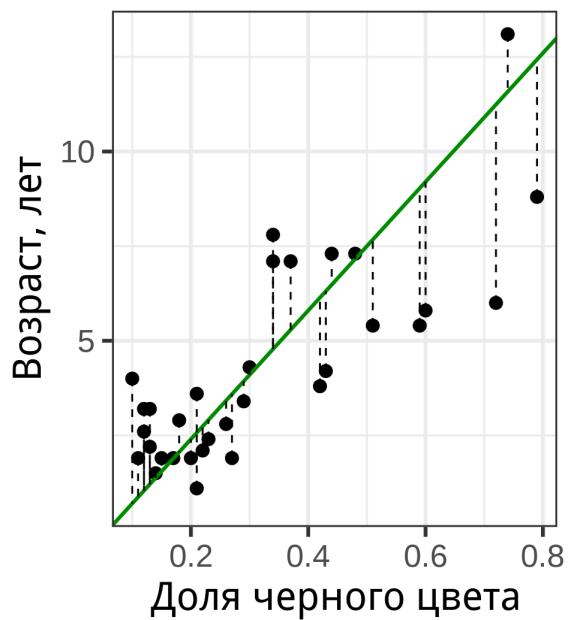
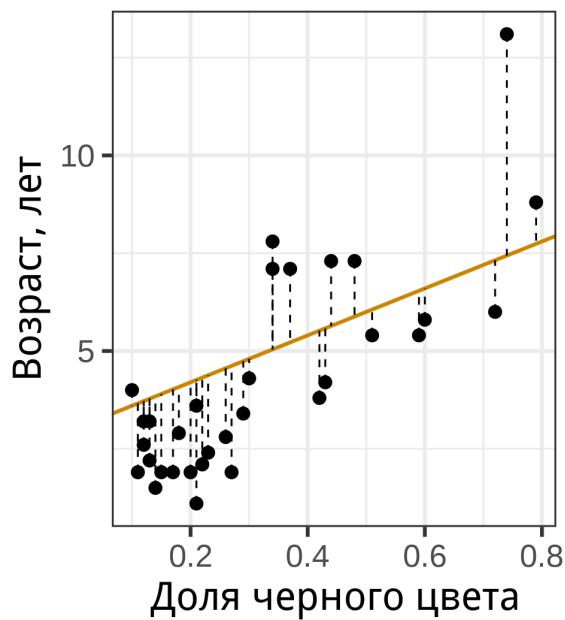
Линия должна проходить как можно ближе к точкам



Линия должна проходить как можно ближе к точкам



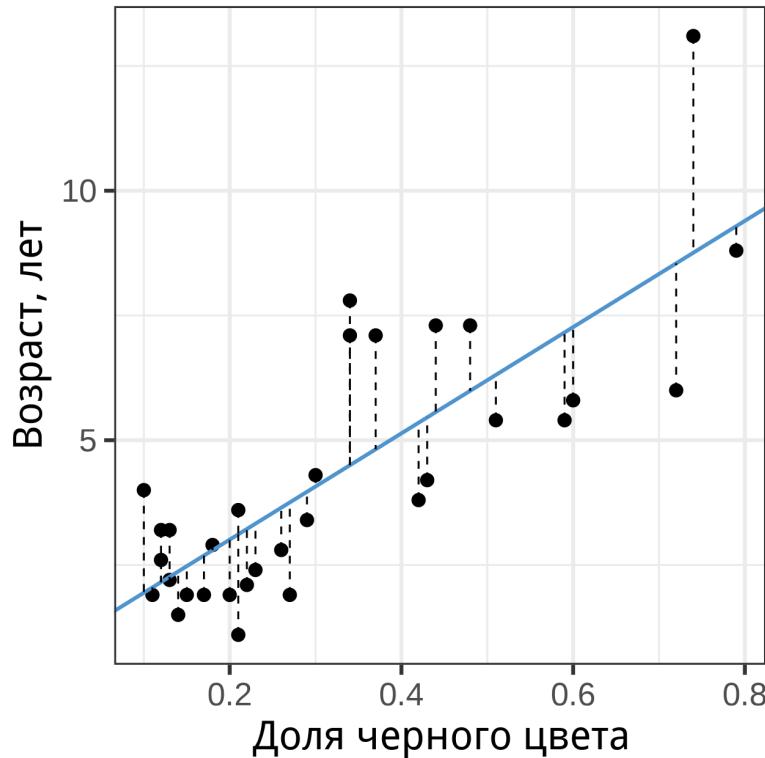
Линия должна проходить как можно ближе к точкам



Метод наименьших квадратов

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

Значения коэффициентов b_0 и b_1 подбирают так, чтобы минимизировать **сумму квадратов остатков** $\sum \varepsilon_i^2$, т.е. $MS_e = \sum (y_i - \hat{y}_i)^2$.



Оценки параметров линейной регрессии

Параметр	Оценка	Стандартная ошибка
β_0	$b_0 = \bar{y} - b_1 \bar{x}$	$SE_{b_0} = \sqrt{MS_e \left[\frac{1}{n} + \frac{\bar{x}}{\sum (x_i - \bar{x})^2} \right]}$
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

Оценки коэффициентов

- позволяют получить предсказанные значения

Стандартные ошибки коэффициентов

- используются для построения доверительных интервалов
- нужны для статистических тестов

Линейная регрессия в примере про львов

В общем виде линейная регрессия:

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

Линейная регрессия в примере про львов

В общем виде линейная регрессия:

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

В примере про львов:

$$\widehat{\text{Возраст}}_i = b_0 + b_1 \cdot \text{Доля черного}_i$$

Линейная регрессия в примере про львов

В общем виде линейная регрессия:

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

В примере про львов:

$$\widehat{\text{Возраст}}_i = b_0 + b_1 \cdot \text{Доля черного}_i$$

Мы подобрали коэффициенты модели методом наименьших квадратов:

	Значение коэффициента
` b_0 `	0.879
` b_1 `	10.647

Линейная регрессия в примере про львов

В общем виде линейная регрессия:

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

В примере про львов:

$$\widehat{\text{Возраст}}_i = b_0 + b_1 \cdot \text{Доля черного}_i$$

Мы подобрали коэффициенты модели методом наименьших квадратов:

Значение коэффициента	
` b_0 `	0.879
` b_1 `	10.647

Получилось уравнение:

$$\widehat{\text{Возраст}}_i = 0.88 + 10.65 \cdot \text{Доля черного}_i$$

Линейная регрессия в примере про львов

В общем виде линейная регрессия:

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

В примере про львов:

$$\widehat{\text{Возраст}}_i = b_0 + b_1 \cdot \text{Доля черного}_i$$

Мы подобрали коэффициенты модели методом наименьших квадратов:

Значение коэффициента	
` b_0 `	0.879
` b_1 `	10.647

Получилось уравнение:

$$\widehat{\text{Возраст}}_i = 0.88 + 10.65 \cdot \text{Доля черного}_i$$

Смысл коэффициентов:

- $b_0 = 0.88$ — Ожидаемый возраст льва с непигментированным носом 0.88 лет.
- $b_1 = 10.65$ — Если доля черного увеличится на единицу, ожидаемый возраст льва увеличится на 10.65 лет.

Доверительный интервал коэффициента 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.

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

Доверительный интервал коэффициента 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.

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

В примере про львов стандартные ошибки для каждого из коэффициентов:

	Значение коэффициента	SE
` b_0 `	0.879	0.569
` b_1 `	10.647	1.510

Доверительный интервал коэффициента 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.

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

В примере про львов стандартные ошибки для каждого из коэффициентов:

	Значение коэффициента	SE
` b_0 `	0.879	0.569
` b_1 `	10.647	1.510

$$df = 32 - 2 = 30$$

$$t_{30} = 2.04$$

Доверительные интервалы

Для b_0 :

$$0.879 \pm 2.04 \cdot 0.569$$
$$0.879 \pm 1.16$$

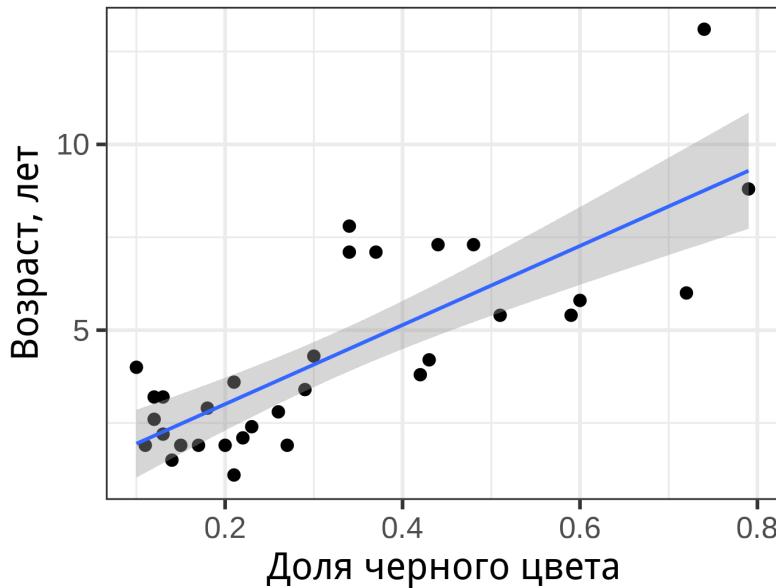
Для b_1 :

$$10.647 \pm 2.04 \cdot 1.51$$
$$10.647 \pm 3.08$$

Доверительная зона регрессии

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

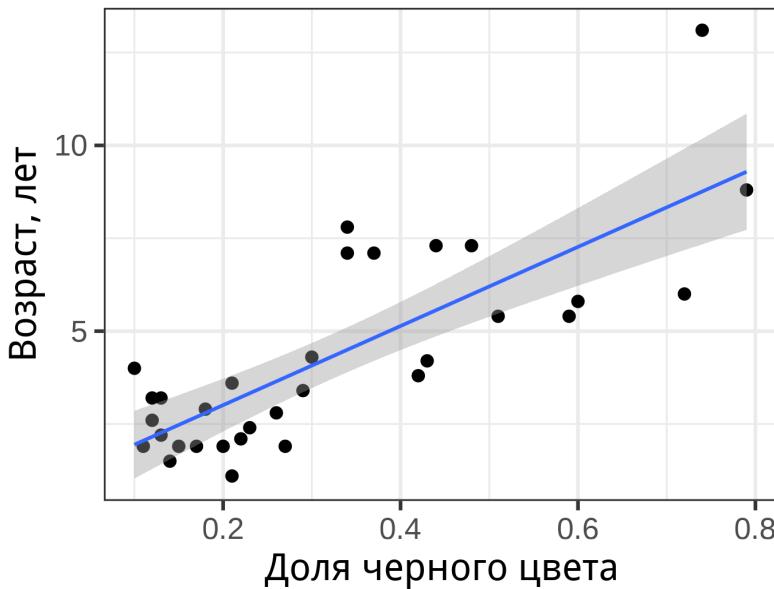
95% доверительная зона



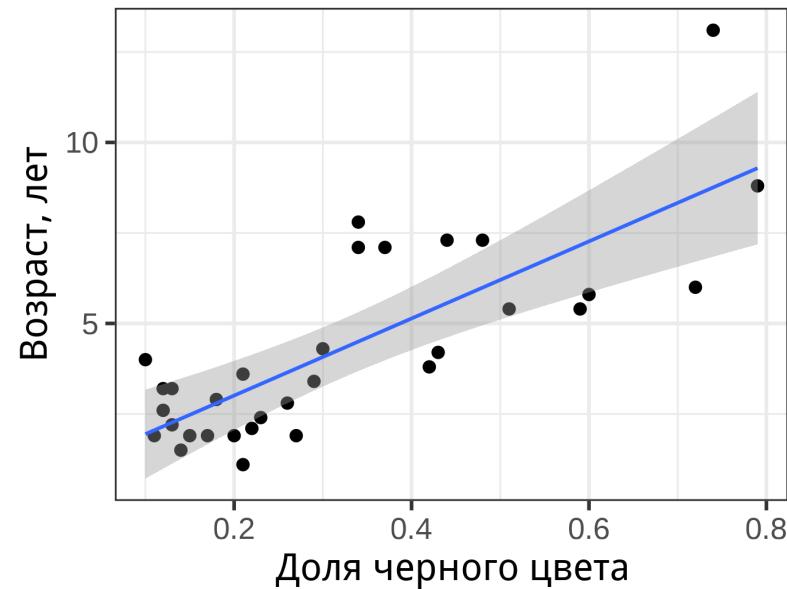
Доверительная зона регрессии

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

95% доверительная зона



99% доверительная зона



Предсказываем с помощью регрессии

Предсказания при помощи регрессии

В примере про львов получилось уравнение:

$$\widehat{\text{Возраст}}_i = 0.88 + 10.65 \cdot \text{Доля черного}_i$$

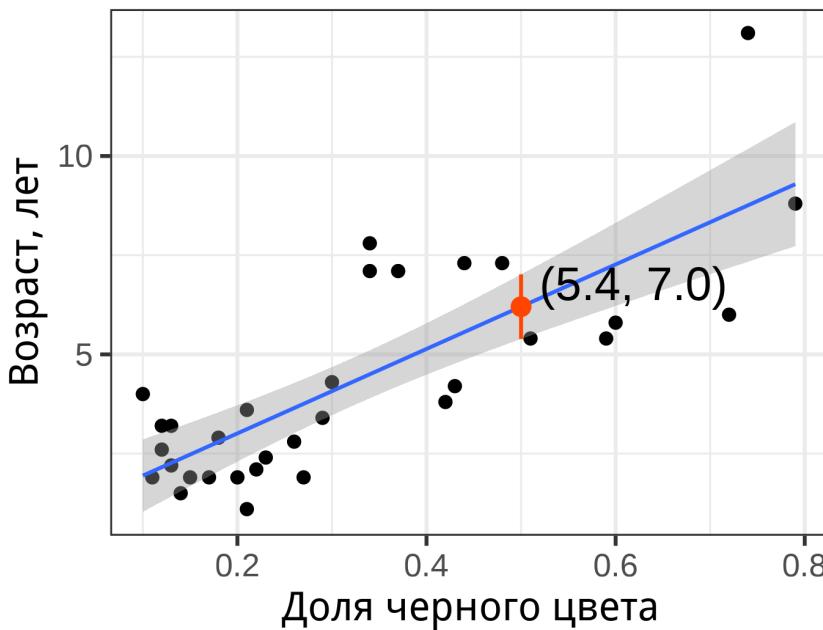
Предсказания при помощи регрессии

В примере про львов получилось уравнение:

$$\widehat{\text{Возраст}}_i = 0.88 + 10.65 \cdot \text{Доля черного}_i$$

Исследователь мог заметить в бинокль нового льва с долей черного на носу 0.5.

Ожидаемый возраст: $0.88 + 10.65 \cdot 0.5 = 6.2$



На линии регрессии лежат ожидаемые средние значения y при разных значениях x . В доверительный интервал для этих средних попадет **среднее значение возраста львов в 95% повторных выборок**.

Неопределенность оценки предсказаний

Доверительный интервал для предсказаний — это зона, в которую попадут индивидуальные наблюдения \hat{y}_i при данном x_i в заданной доле повторных выборок.

$$\hat{y}_i \pm t_{\alpha, n-2} \cdot SE_{\hat{y}_i},$$

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

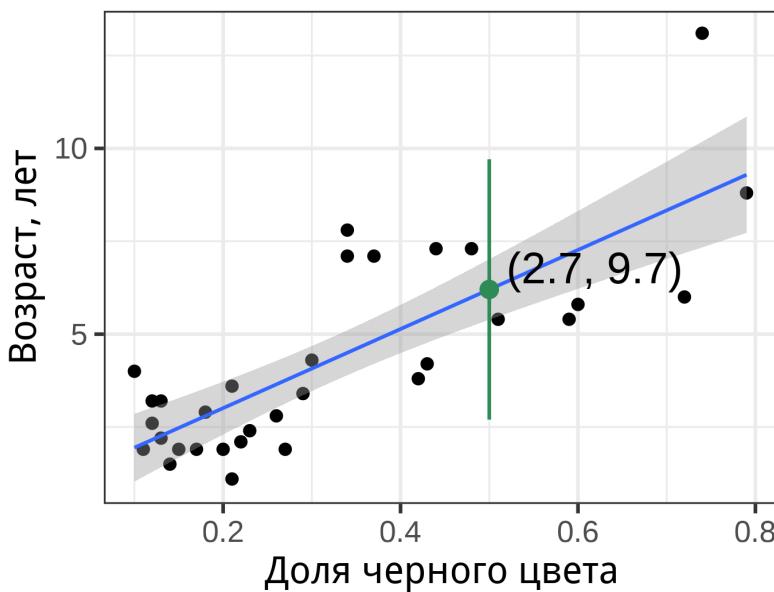
Неопределенность оценки предсказаний

Доверительный интервал для предсказаний — это зона, в которую попадут индивидуальные наблюдения \hat{y}_i при данном x_i в заданной доле повторных выборок.

$$\hat{y}_i \pm t_{\alpha, n-2} \cdot SE_{\hat{y}_i},$$

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

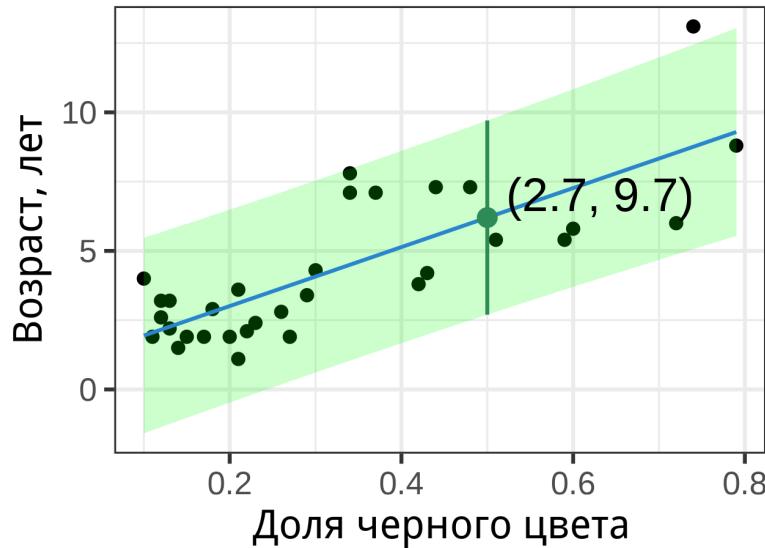
В примере про львов



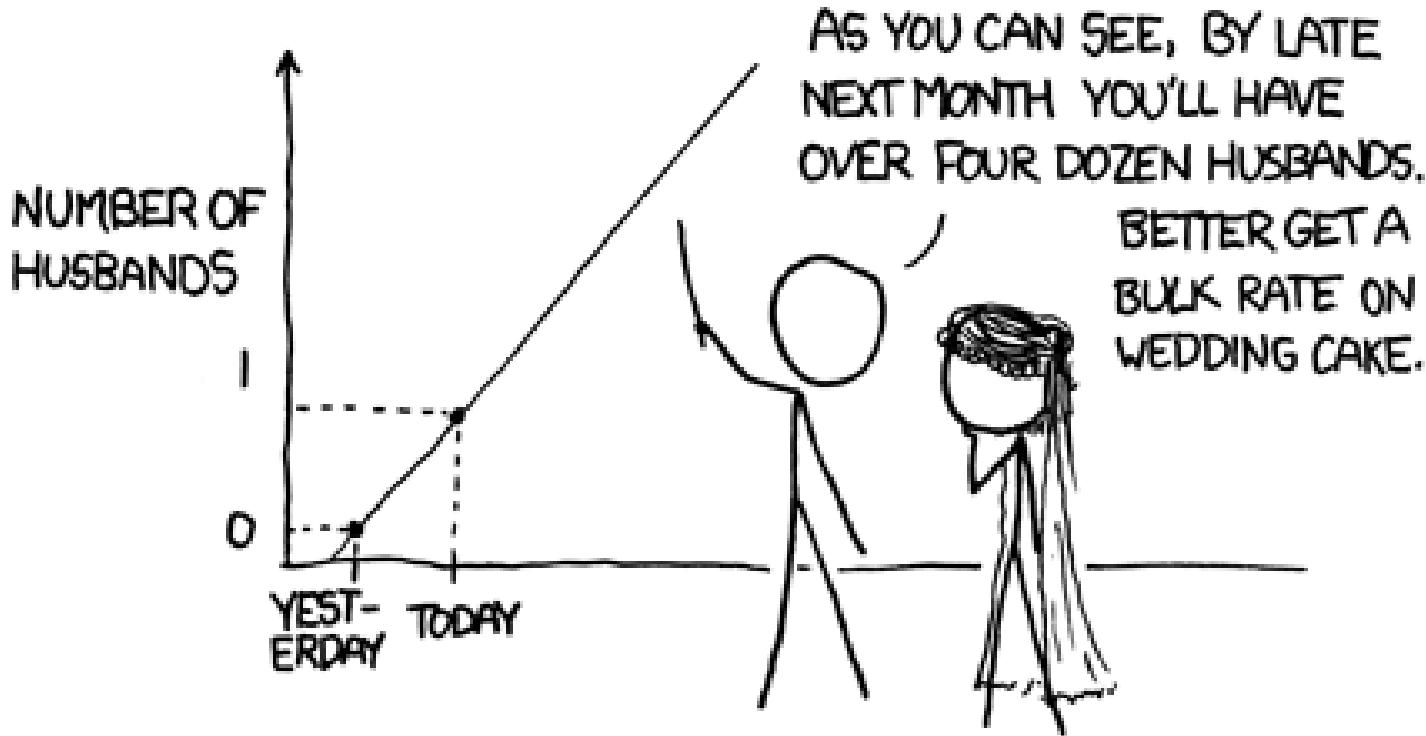
На линии регрессии лежат ожидаемые средние значения y при разных значениях x . В доверительный интервал предсказаний попадет **наблюденный возраст львов в 95% повторных выборок**.

Доверительная область значений

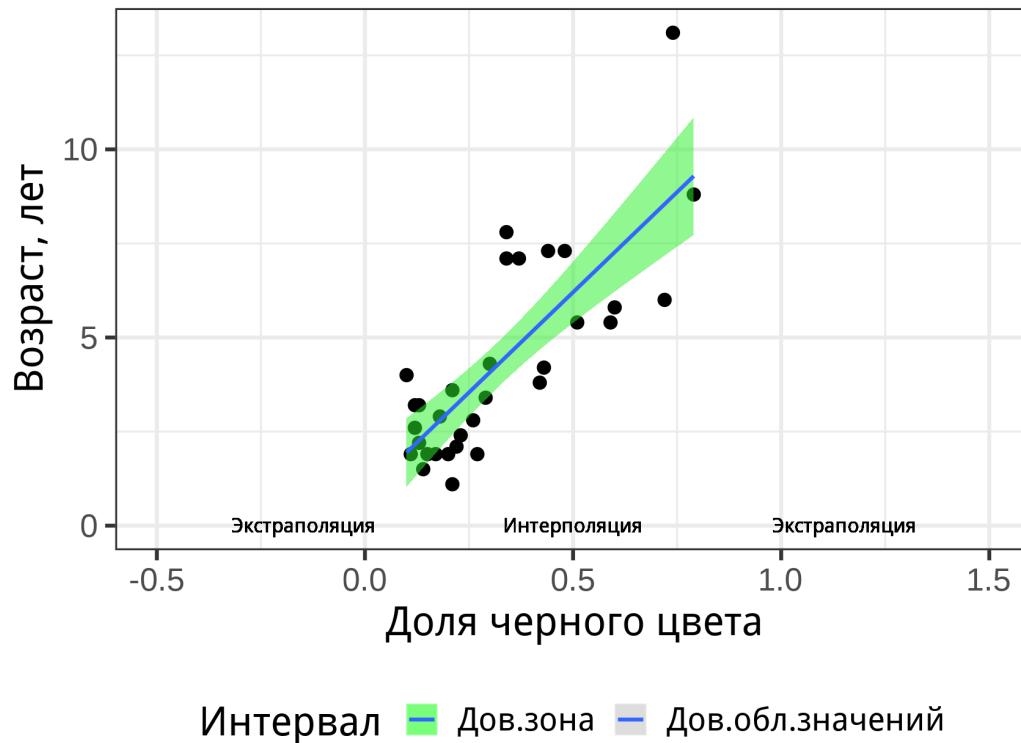
Доверительная область значений — это зона, в которую попадают наблюдения в $(1 - \alpha) \cdot 100\%$ повторных выборок



MY HOBBY: EXTRAPOLATING



Интерполяция и экстраполяция



Модель “работает” только в том диапазоне значений независимой переменной x , для которой она построена (интерполяция).

Экстраполяцию надо применять с большой осторожностью.

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: значимо ли отличаются модели с данным предиктором и без него?

Тестируем значимость коэффициентов t-критерием

$$t = \frac{b_1 - 0}{SE_{b_1}}$$

$$H_0 : b_1 = 0$$

$$H_A : b_1 \neq 0$$

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

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.879	0.569	1.54	0.133
proportionBlack	10.647	1.510	7.05	0.000

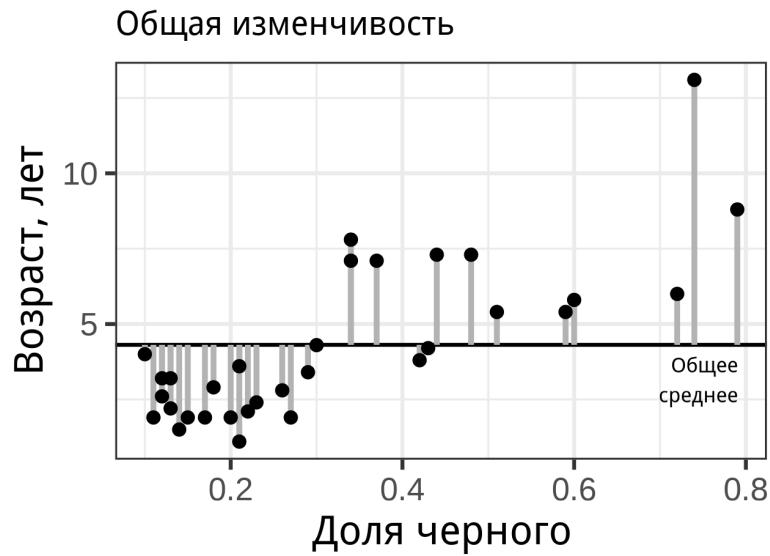
Результаты можно описать в тексте так:

Возраст льва статистически значимо зависит от количества черного пигмента на носу ($b_1 = 10.65, t_{df=30} = 7.05, p < 0.01$)

Тестирование гипотез при помощи F-критерия

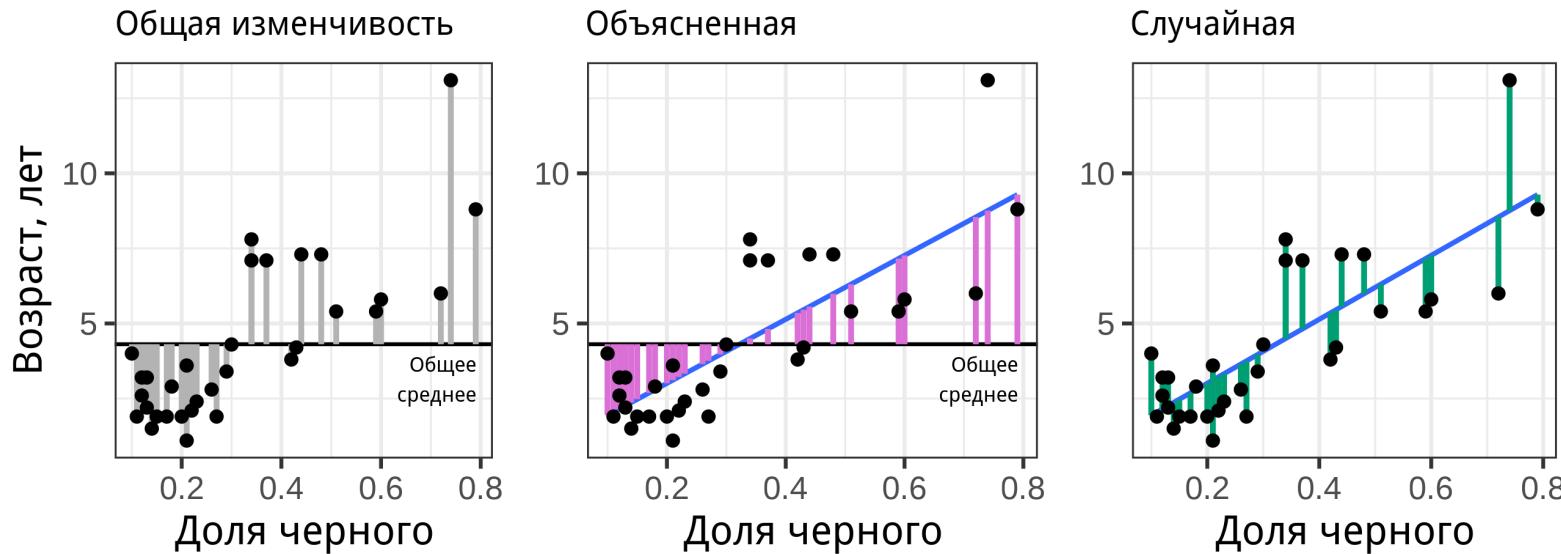
Общая изменчивость

Общая изменчивость SS_t — это сумма квадратов отклонений наблюдаемых значений y_i от общего среднего \bar{y}

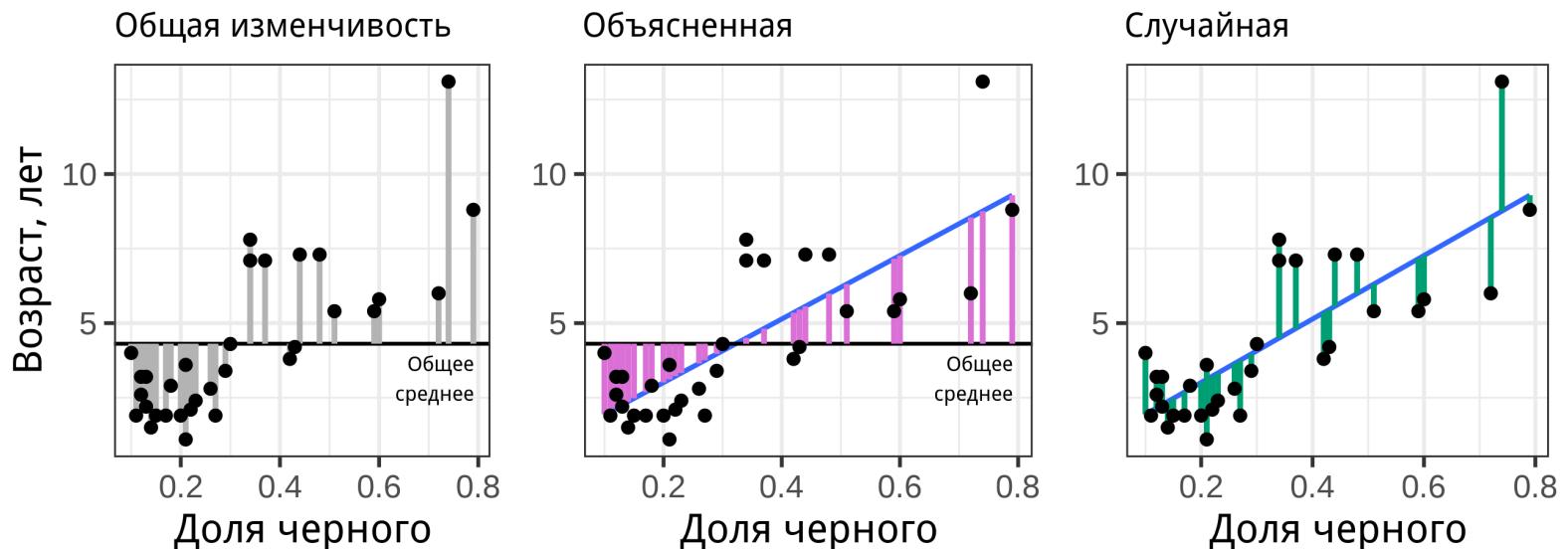


Структура общей изменчивости

$$SS_t = SS_r + SS_e$$



От изменчивостей к дисперсиям



MS_t , полная дисперсия

$$MS_t = \frac{SS_t}{df_t}$$

$$SS_t = \sum (y_i - \bar{y})^2$$

$$df_t = n - 1$$

MS_r , дисперсия, объясненная регрессией

$$MS_r = \frac{SS_r}{df_r}$$

$$SS_r = \sum (\hat{y} - \bar{y})^2$$

$$df_r = 1$$

MS_e , остаточная дисперсия

$$MS_e = \frac{SS_e}{df_e}$$

$$SS_e = \sum (y_i - \hat{y})^2$$

$$df_e = n - 2$$

С помощью MS_r и MS_e можно тестировать значимость коэффициентов

Если зависимости нет, то $MS_r \approx MS_e$

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$

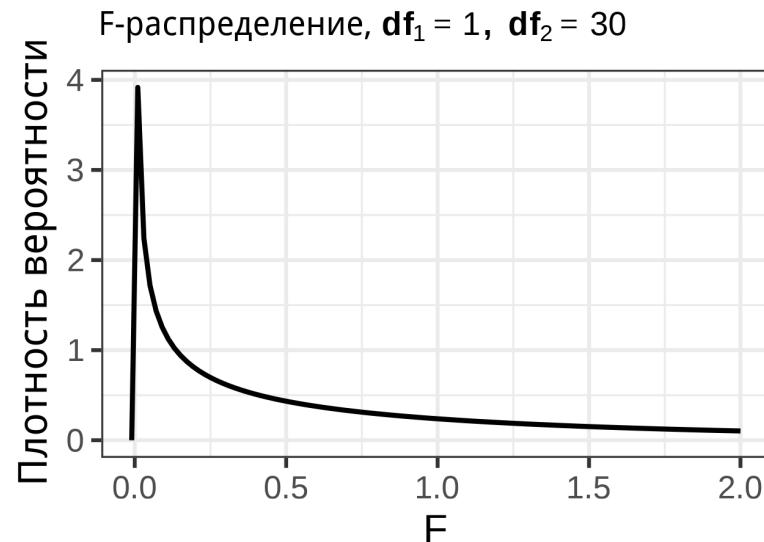
$$F_{df_r, df_e} = \frac{MS_r}{MS_e}$$

Тестирование значимости коэффициентов регрессии при помощи F-критерия

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$

$$F_{df_r, df_e} = \frac{MS_r}{MS_e}$$

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



Тестирование значимости коэффициентов регрессии при помощи F-критерия

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$

F-тест будет односторонним, т.к. соотношение дисперсий может быть только положительным.

$$F_{df_r, df_e} = \frac{MS_r}{MS_e}$$

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

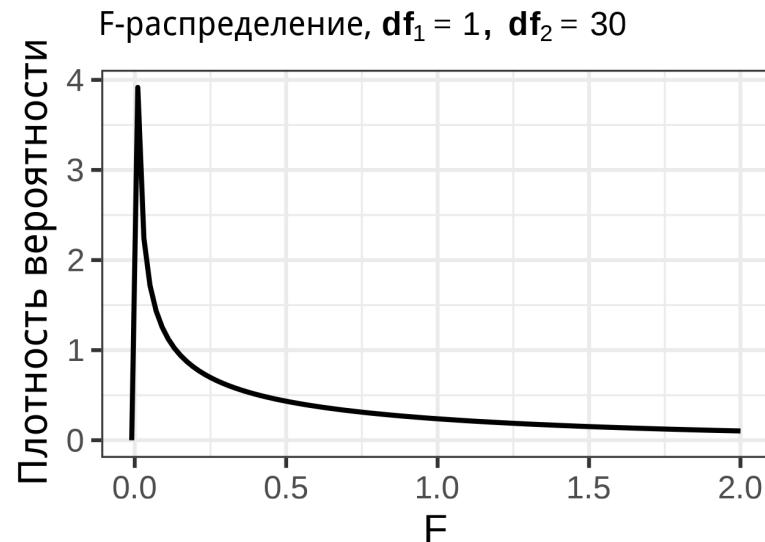


Таблица результатов дисперсионного анализа

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

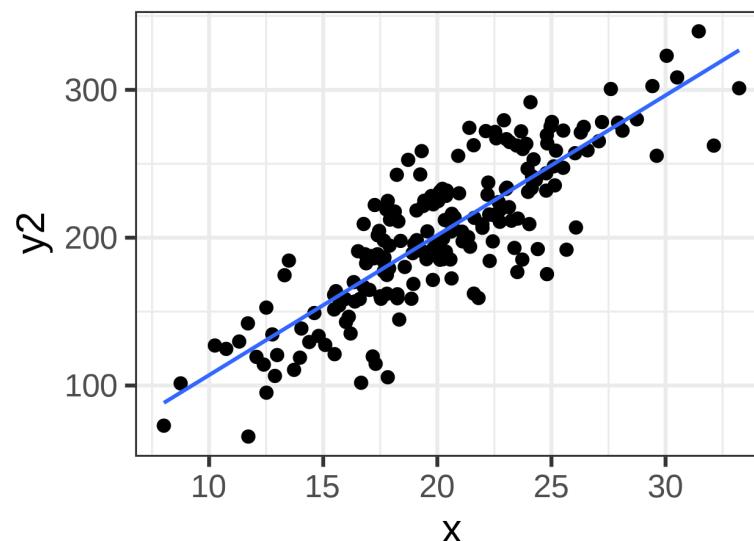
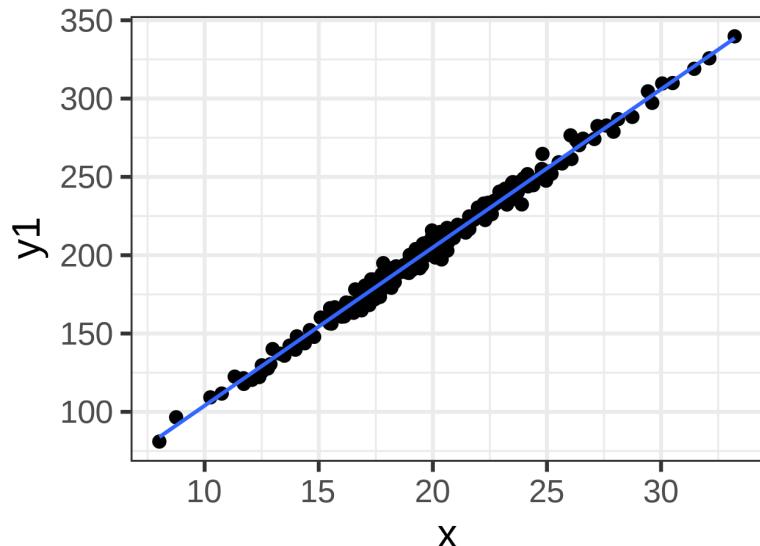
	Sum Sq	Df	F value	Pr(>F)
proportionBlack	138.5	1	49.8	0
Residuals	83.5	30		

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

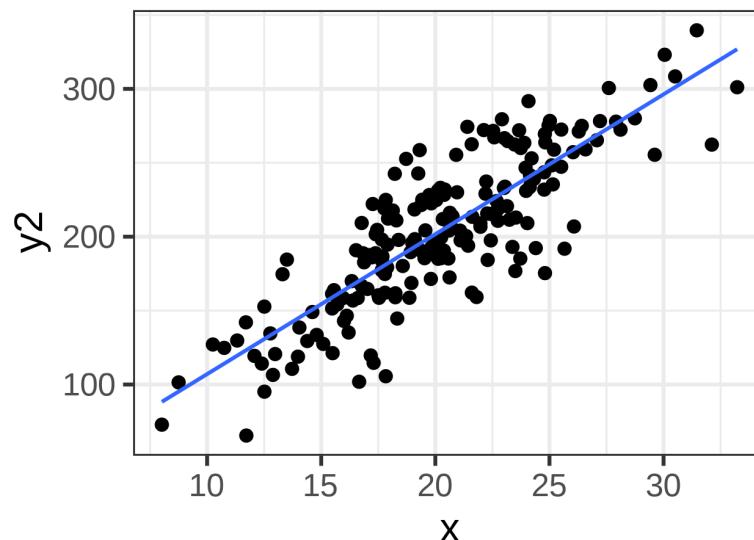
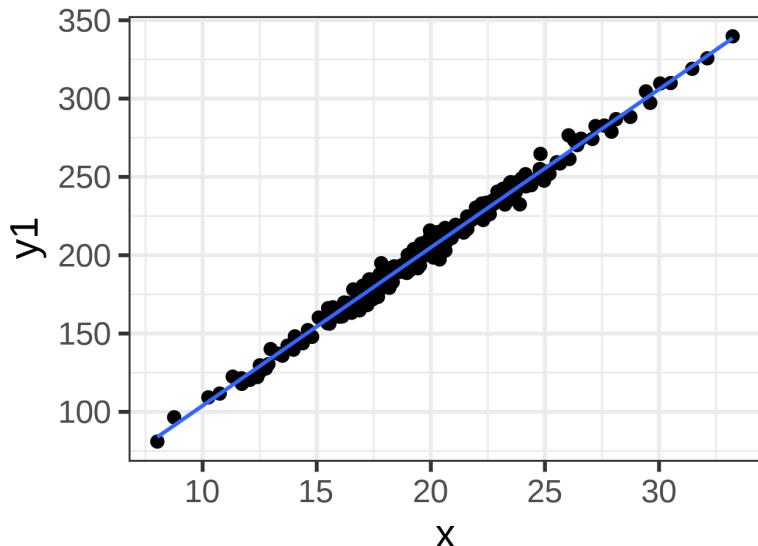
- Возраст льва статистически значимо зависит от количества черного пигмента на носу ($F_{1,30} = 49.75, p < 0.001$).

Оценка качества подгонки модели

В чем различие между этими двумя моделями?



В чем различие между этими двумя моделями?



У этих моделей разный разброс остатков:

- Модель слева объясняет практически всю изменчивость
- Модель справа объясняет не очень много изменчивости

Коэффициент детерминации – мера качества подгонки модели

Коэффициент детерминации описывает какую долю дисперсии зависимой переменной объясняет модель

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

- $0 < R^2 < 1$
- $R^2 = r^2$ — для простой линейной регрессии коэффициент детерминации равен квадрату коэффициента Пирсоновской корреляции

Если в модели много предикторов, нужно внести поправку

Скорректированный коэффициент детерминации (adjusted R-squared)

Применяется если необходимо сравнить две модели с разным количеством параметров

$$R_{adj}^2 = 1 - (1 - R^2) \frac{n - 1}{n - p}$$

p - количество параметров в модели

Вводится штраф за каждый новый параметр

Условия применимости простой линейной регрессии

Условия применимости простой линейной регрессии

Условия применимости линейной регрессии должны выполняться, чтобы тестировать гипотезы

1. Независимость наблюдений
2. Линейность связи
3. Нормальное распределение остатков
4. Равенство дисперсий остатков

Условия применимости простой линейной регрессии

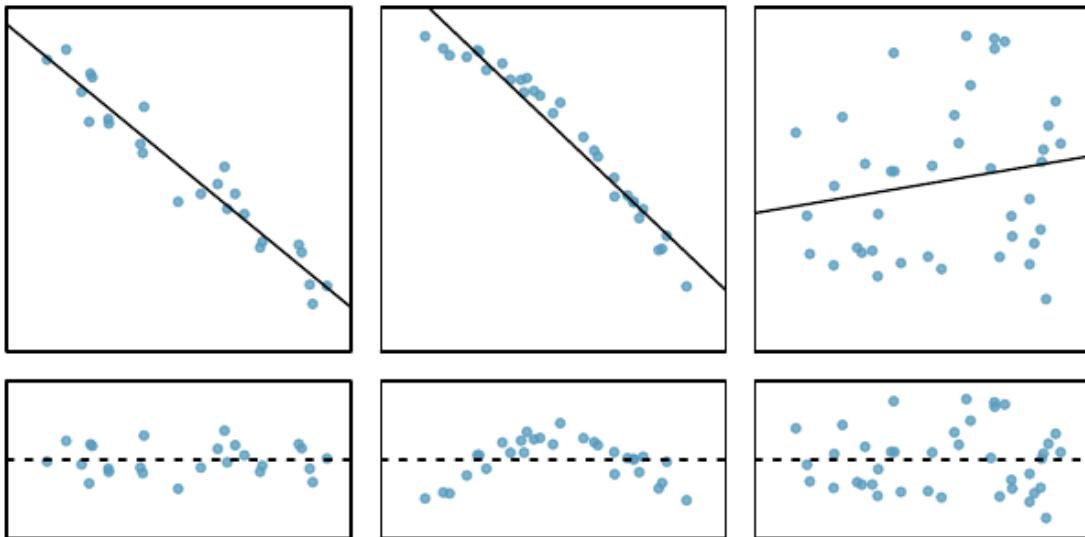
Условия применимости линейной регрессии должны выполняться, чтобы тестировать гипотезы

1. Независимость наблюдений
2. Линейность связи
3. Нормальное распределение остатков
4. Равенство дисперсий остатков

Для множественной линейной регрессии добавляется требование независимости предикторов друг от друга (отсутствие мультиколлинеарности).

1. Независимость наблюдений

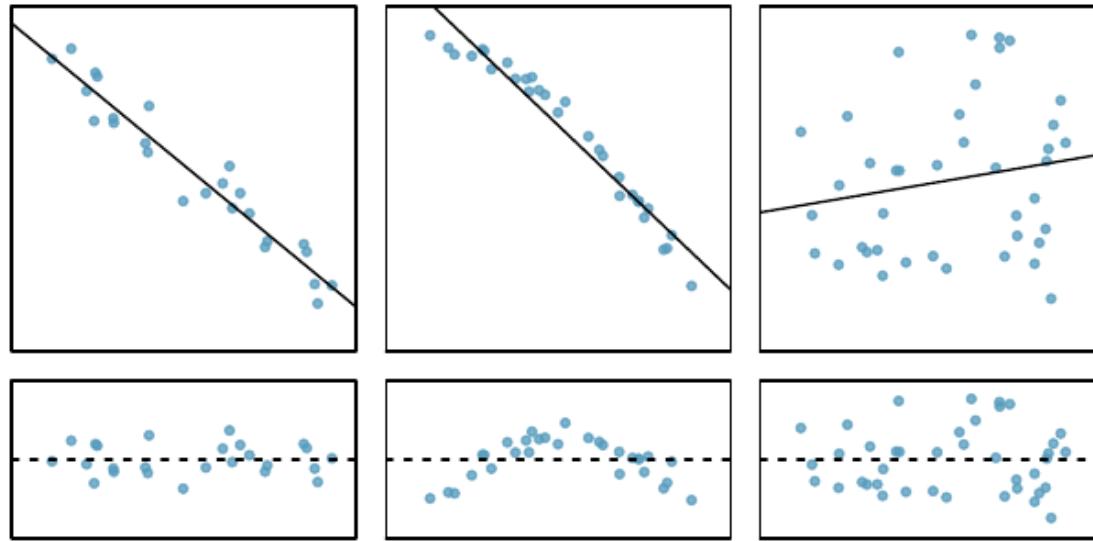
- Значения y_i должны быть независимы друг от друга
- Берегитесь псевдоповторностей и автокорреляций (например, временных)
- Контролируется на этапе планирования
- Проверяют на графике зависимости остатков от предсказанных значений



Из кн. Diez et al., 2010, стр. 332, рис. 7.8

2. Линейность связи

- Проверяем на графике зависимости остатков от предсказанных значений



Из кн. Diez et al., 2010, стр. 332, рис. 7.8

3. Нормальное распределение остатков

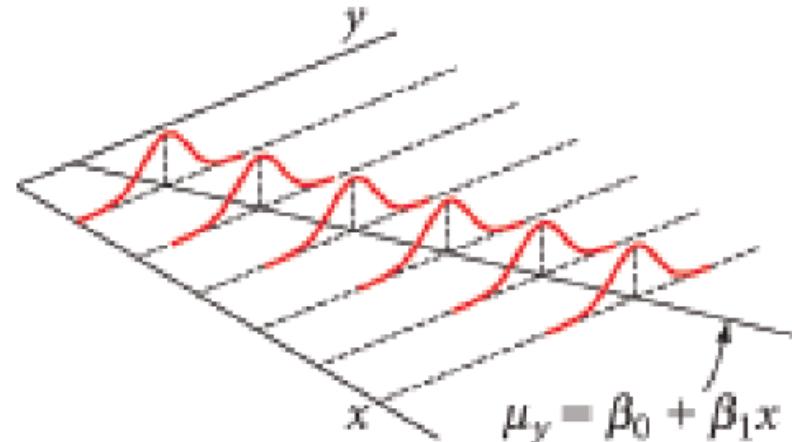
Нужно, т.к.

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

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

$$\text{а значит } \epsilon_i \sim N(0, \sigma^2)$$

- Нужно для тестов параметров, а не для подбора коэффициентов
- Нарушение не страшно — тесты устойчивы к небольшим отклонениям
- Проверяем на квантильном графике остатков



Из кн. Watkins et al., 2008, стр. 743, рис. 11.4

4. Гомогенность дисперсий

Нужно, т.к.

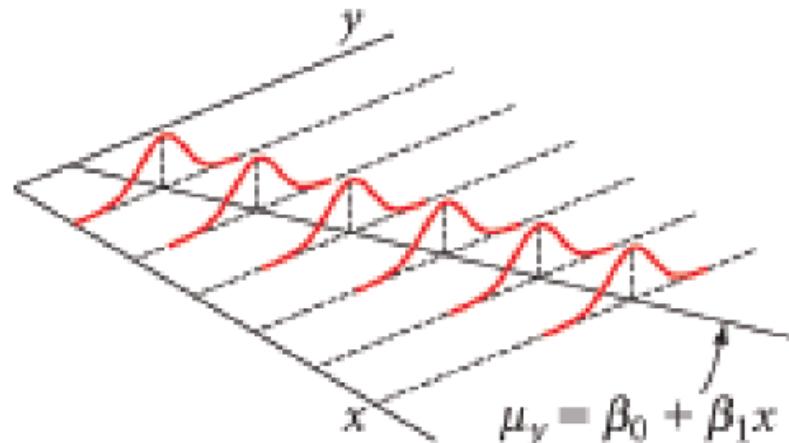
$$Y_i = \beta_0 + \beta x_i + \epsilon_i$$

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

$$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_i^2 \text{ для каждого } Y_i$$

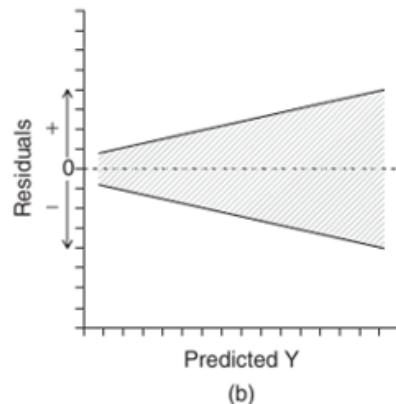
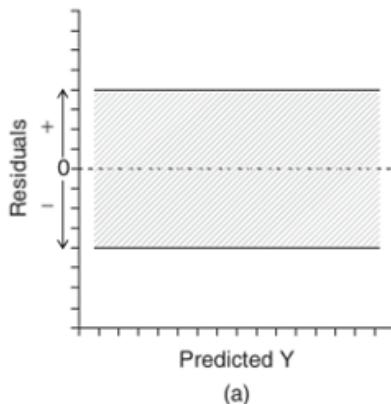
Но, поскольку $\epsilon_i \sim N(0, \sigma^2)$, можно проверить равенство дисперсий остатков ϵ_i

- Нужно и важно для тестов параметров
- Проверяем на графике остатков по отношению к предсказанным значениям
- Формальные тесты слишком чувствительны (тест Брайша-Пагана, тест Кокрана)

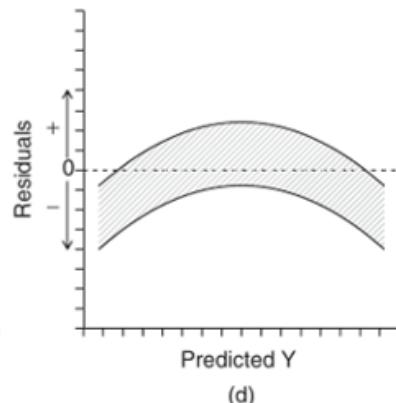
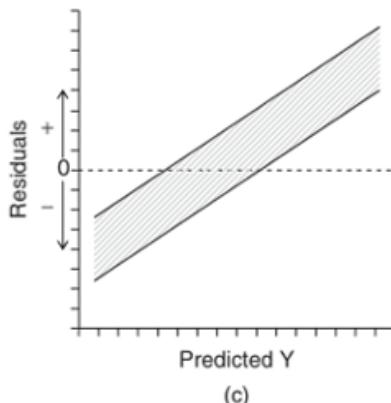


Из кн. Watkins et al., 2008, стр. 743, рис. 11.4

Диагностика регрессии по графикам остатков



- (a) все условия выполнены
- (b) разброс остатков разный (wedge-shaped pattern)
- (c) разброс остатков одинаковый, но нужны дополнительные предикторы
- (d) к нелинейной зависимости применили линейную регрессию



Из кн. Logan, 2010, стр. 174, рис. 8.5 д

Ловушки при использовании корреляции и регрессии

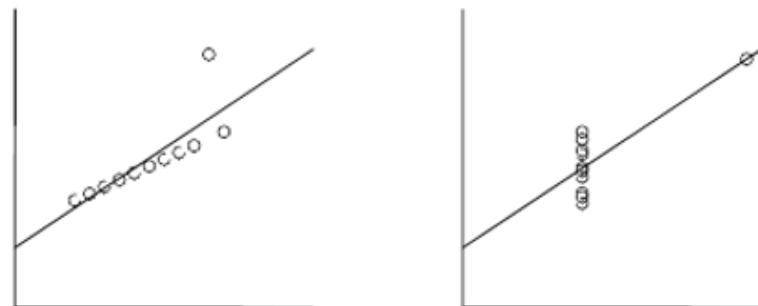
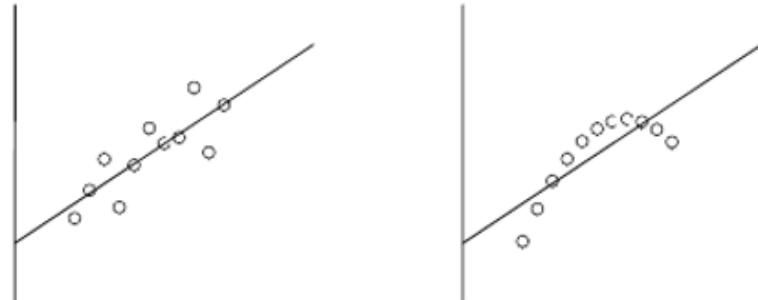
Последствия необдуманного применения линейной регрессии

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

$$y_i = 3.0 + 0.5x_i$$

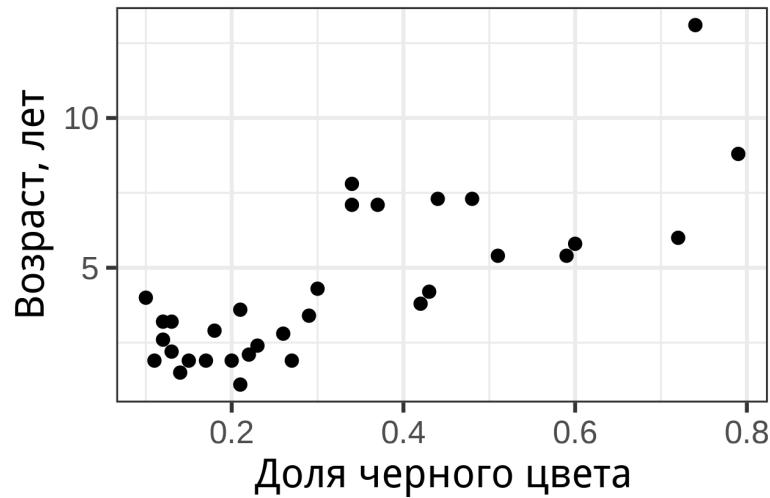
$$r^2 = 0.68$$

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

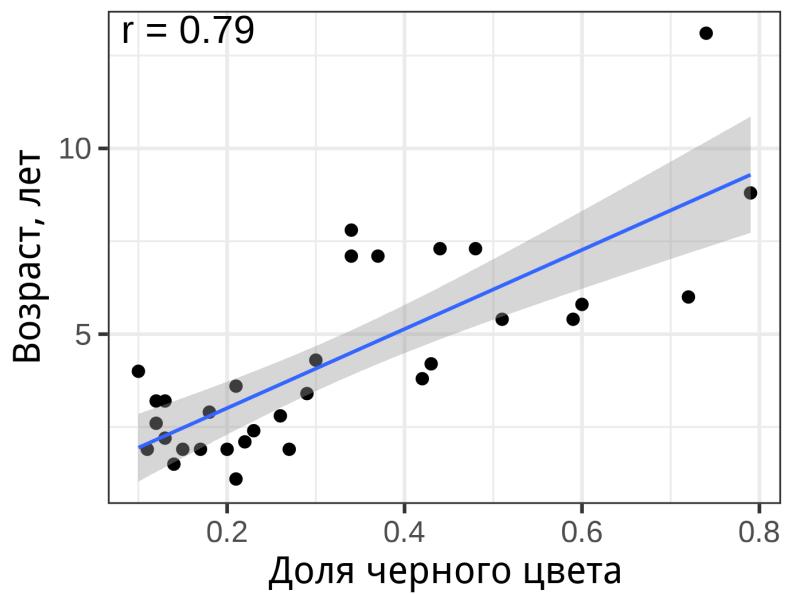
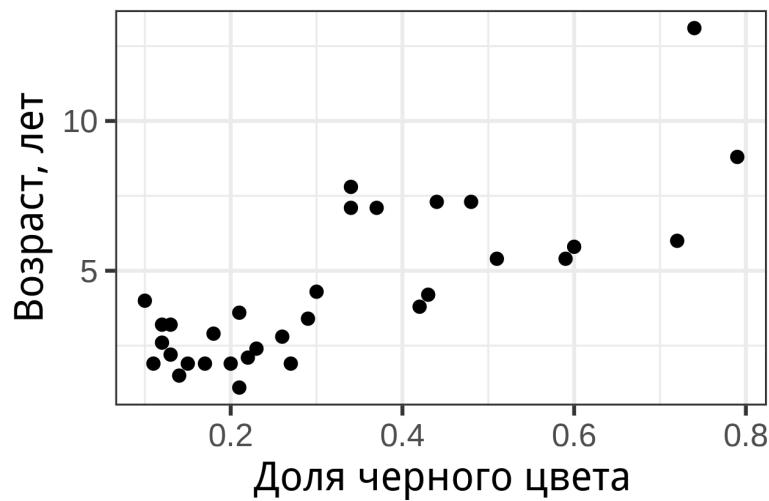


Из кн. Quinn, Keough, 2002, стр. 97, рис. 5.9

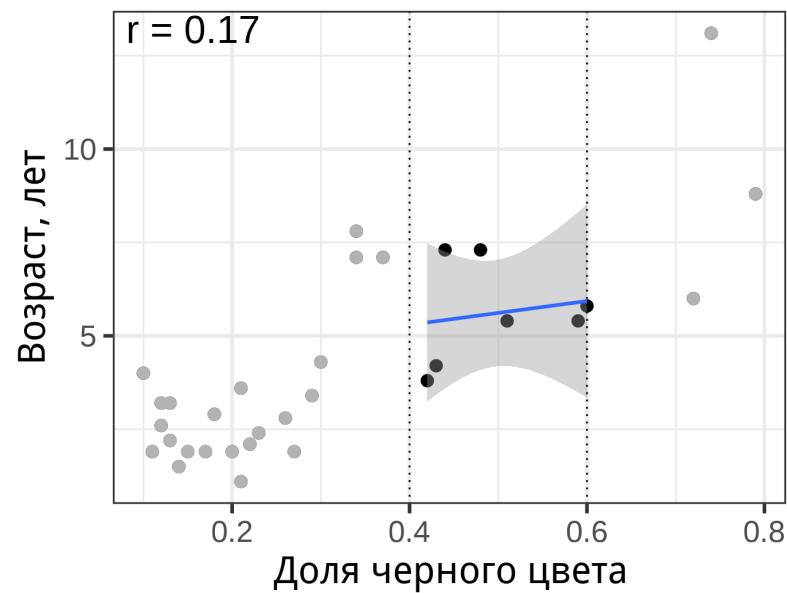
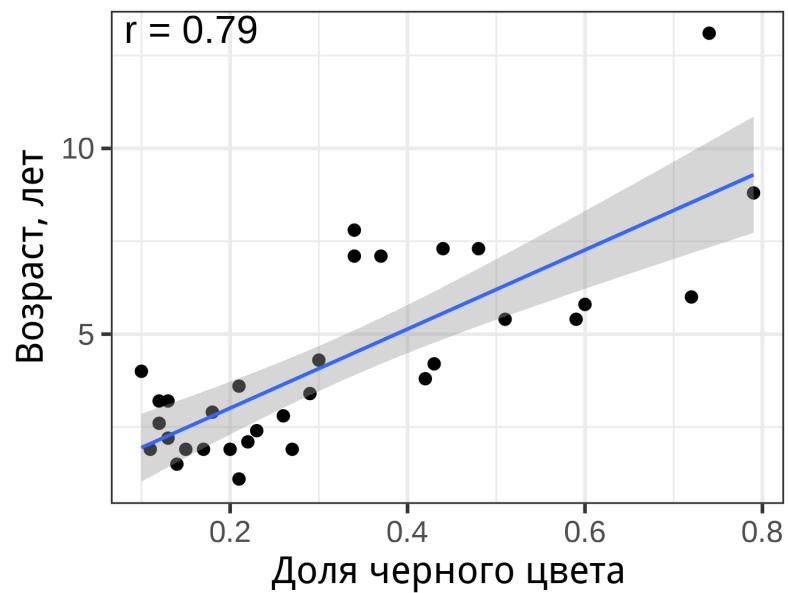
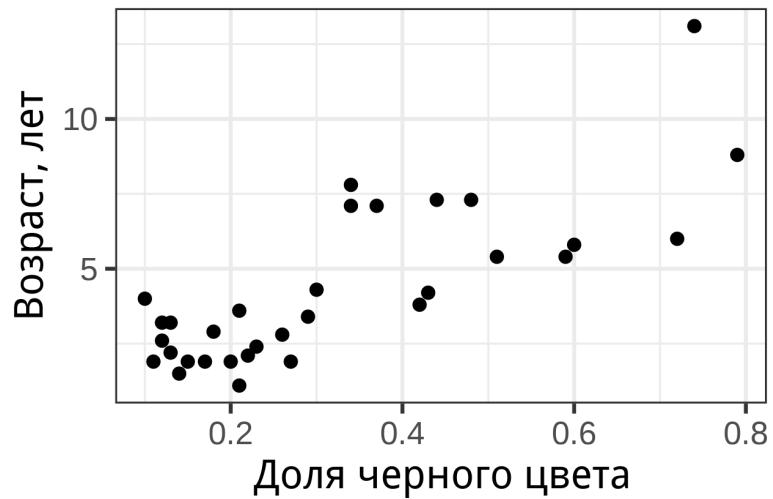
Обратите внимание на диапазон значений



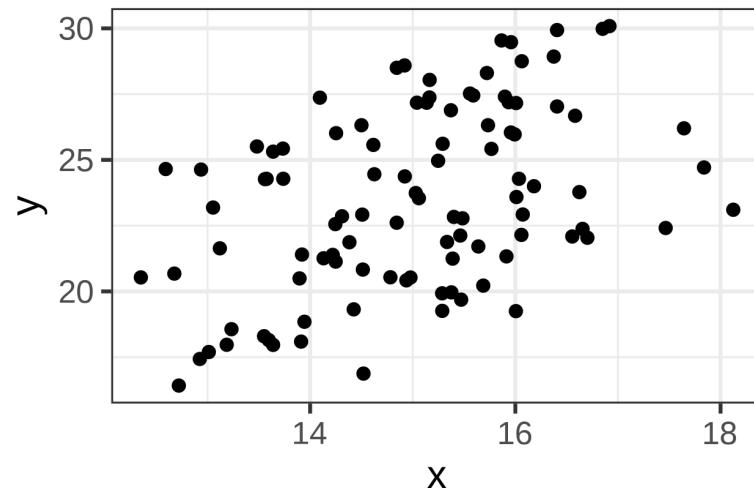
Обратите внимание на диапазон значений



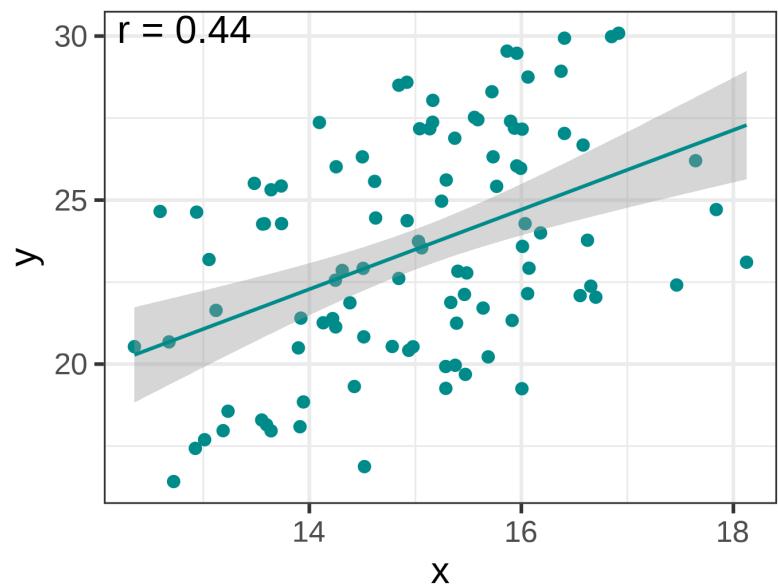
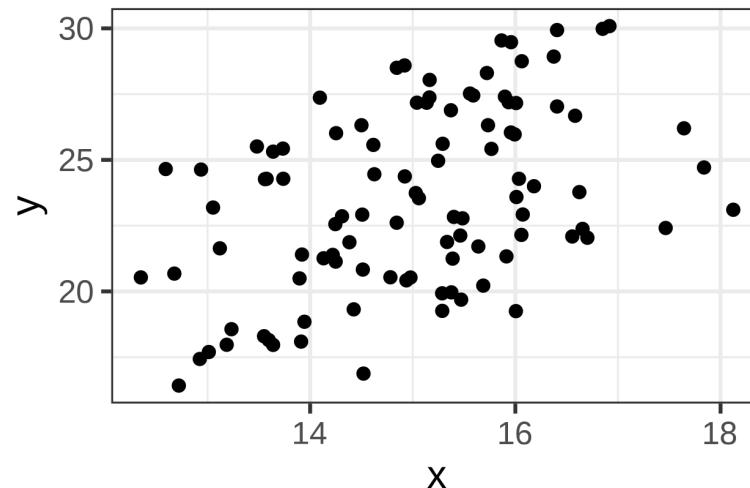
Обратите внимание на диапазон значений



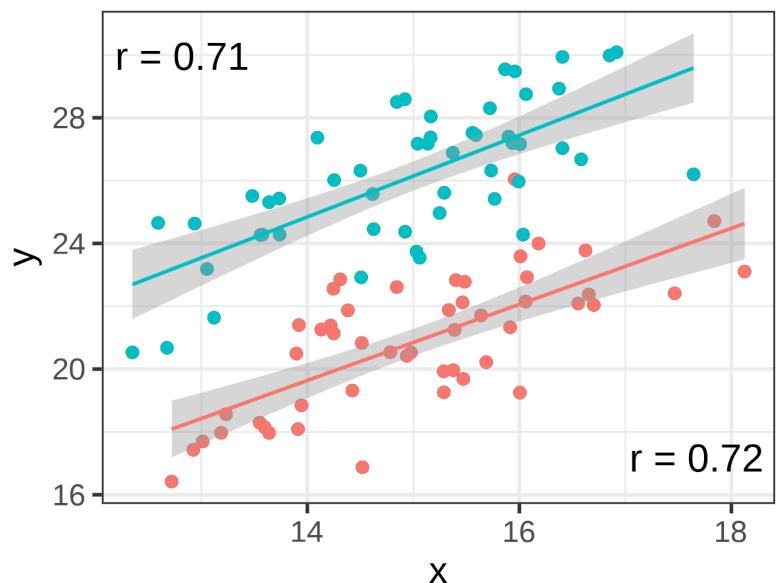
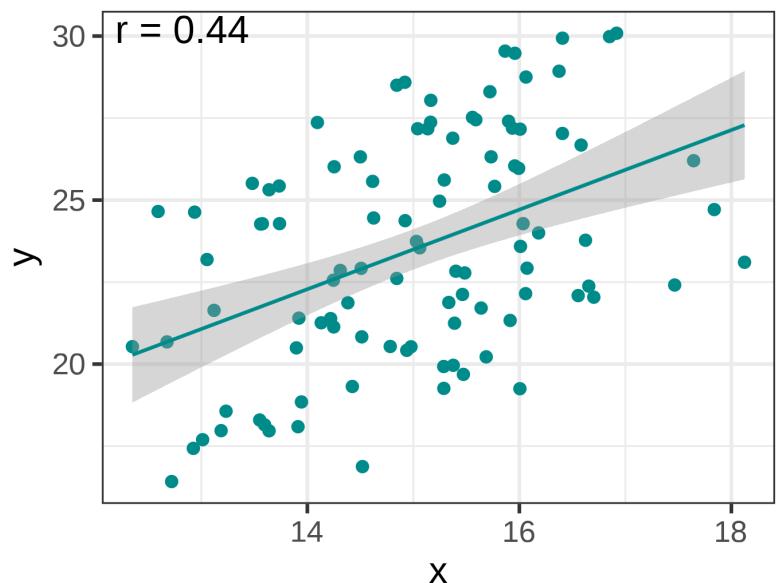
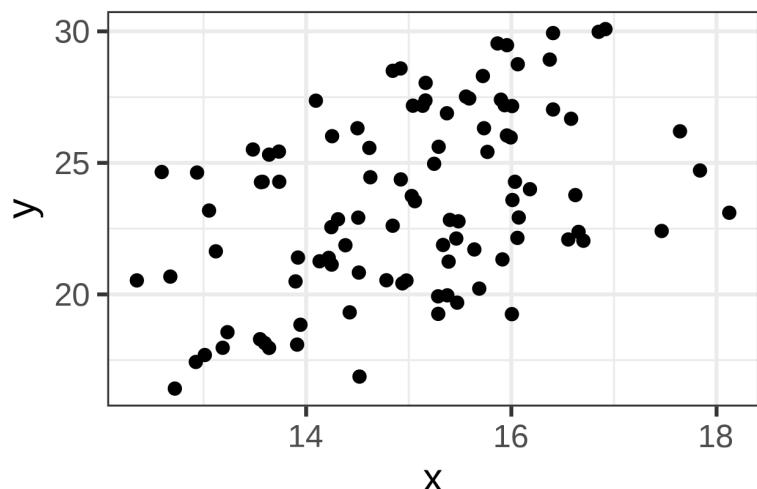
Парadox Симпсона



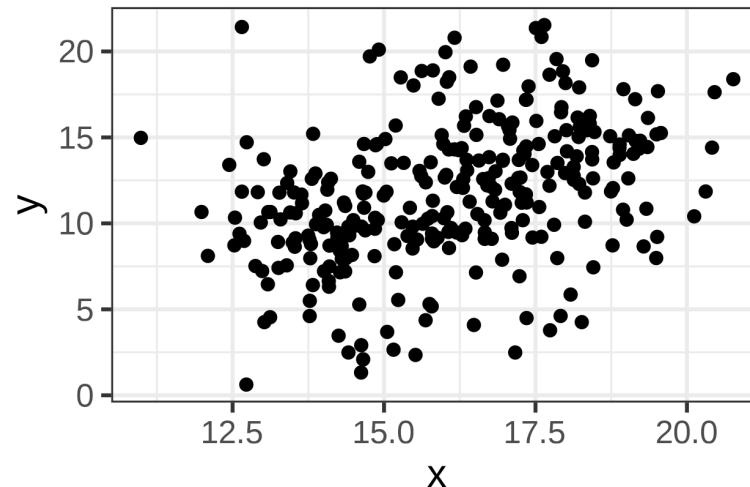
Парadox Симпсона



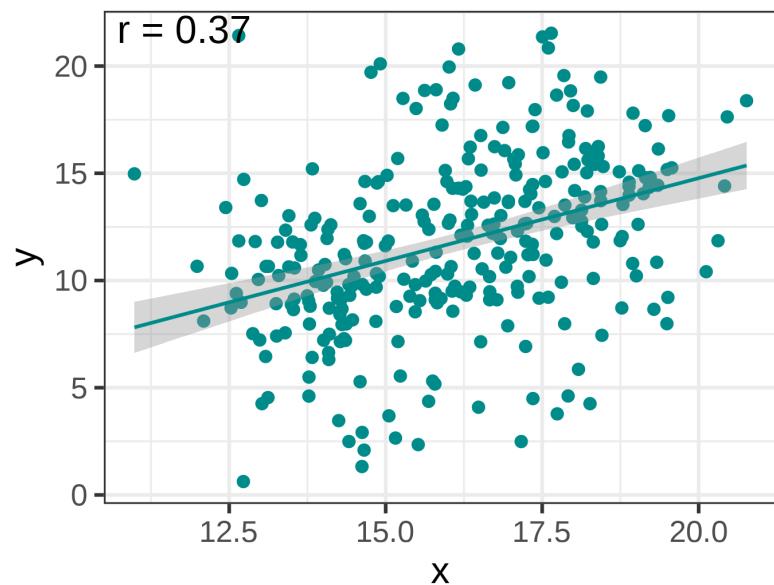
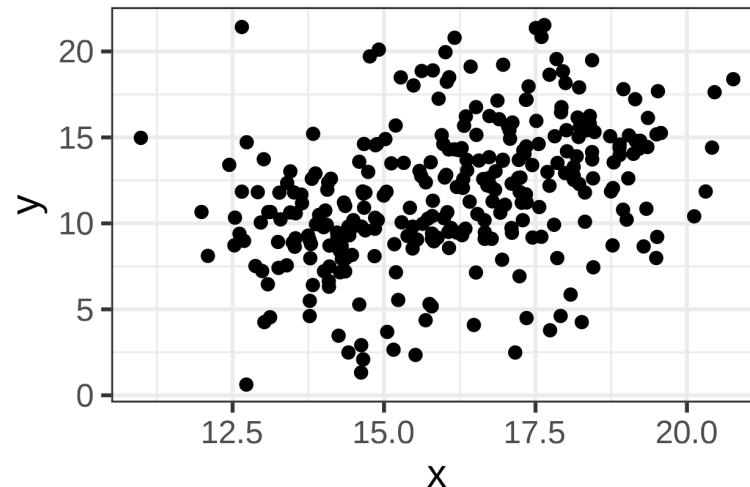
Парадокс Симпсона



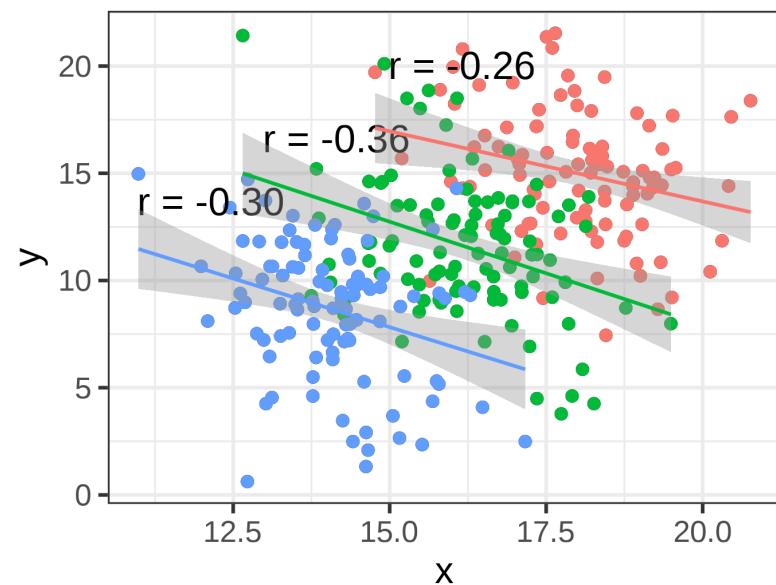
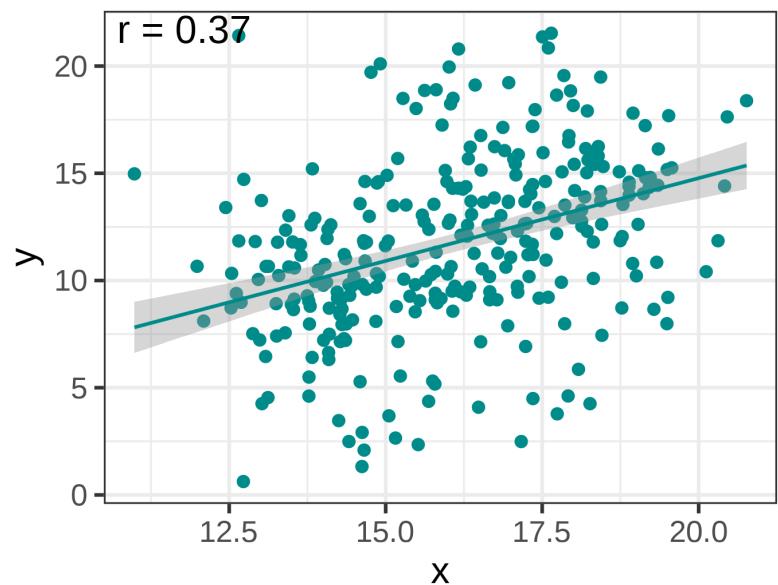
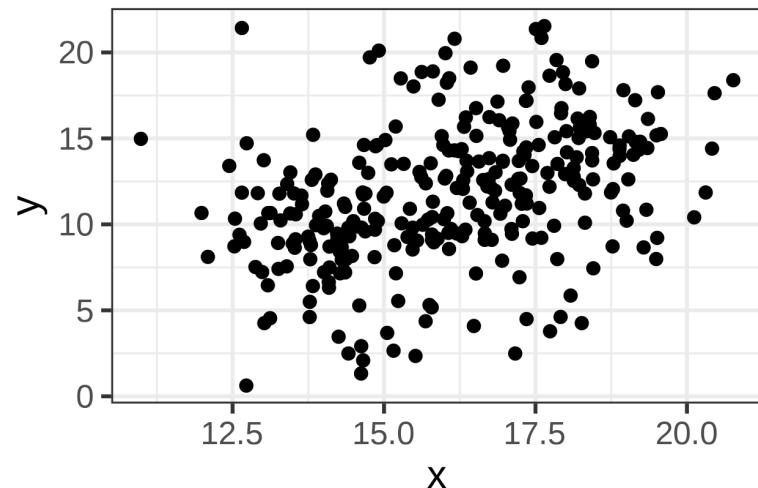
Парadox Симпсона



Парadox Симпсона



Парадокс Симпсона



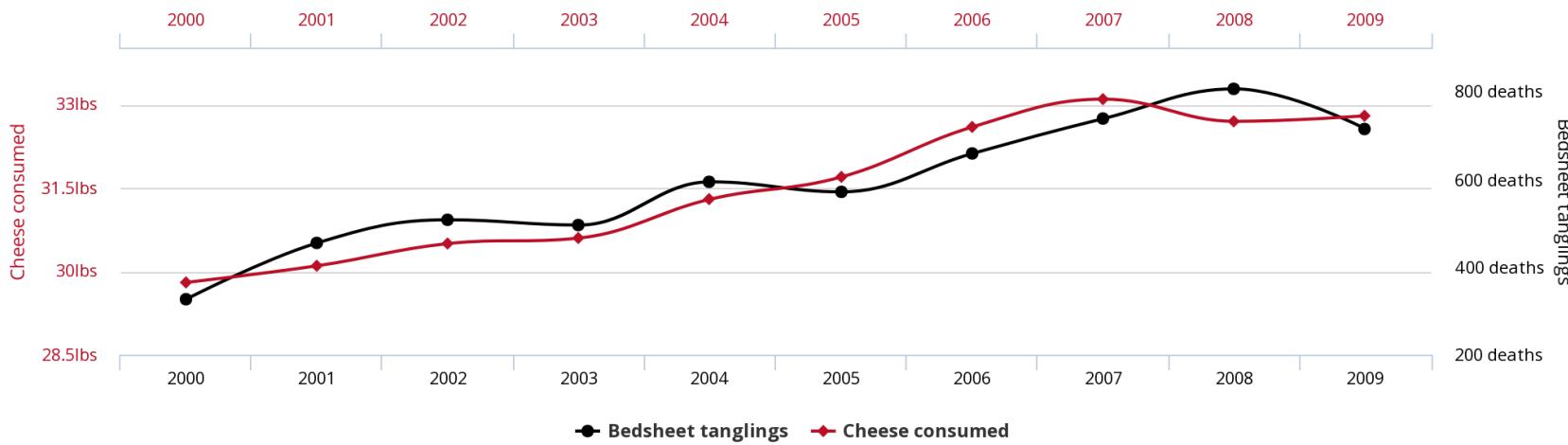
Наличие связи между переменными не означает причинно-следственных отношений

| Correlation does not imply causation

Per capita cheese consumption

correlates with

Number of people who died by becoming tangled in their bedsheets



<https://www.tylervigen.com/spurious-correlations>

Summary

Summary

- Коэффициент корреляции Пирсона r оценивает силу и направление связи между численными величинами (линейную составляющую)
- Корреляция не означает причинно-следственной связи между переменными.
- Условия применимости коэффициента корреляции Пирсона
 - Двумерное нормальное распределение переменных
 - Дисперсия одной переменной не должна зависеть от другой
 - Не должно быть выбросов (= outliers)
 - Связь должна быть линейной.
- Непараметрическая альтернатива — использование коэффициентов корреляции Спирмена или Кендалла.

Summary

- Модель простой линейной регрессии $y_i = b_0 + b_1x_i + e_i$
- y называется откликом, а x — предиктором, коэффициент b_0 — свободный член линейной модели кодирует отрезок, b_1 — это коэффициент угла наклона.
- Параметры модели оцениваются на основе выборки.
- В оценке коэффициентов регрессии, положения регрессионной прямой и предсказанных значений существует неопределенность.
- Доверительные интервалы (двух сортов) можно рассчитать, зная стандартные ошибки.
- Гипотезы о наличии взаимосвязи между откликом и предиктором можно тестировать при помощи t- или F-теста.
- Качество подгонки модели можно оценить при помощи коэффициента детерминации R^2 .

Summary

- Условия применимости линейных моделей
 - Независимость наблюдений
 - Линейность связи
 - Нормальное распределение остатков
 - Равенство дисперсий остатков
- Если условия применимости нарушены, то результатам тестов для этой модели нельзя верить (получаются заниженные доверительные вероятности, возрастает вероятность ошибок I рода).
- Анализ остатков дает разностороннюю информацию о валидности моделей.

ЧТО ПОЧИТАТЬ

- Гланц, С., 1998. Медико-биологическая статистика. М., Практика
- Кабаков Р.И. R в действии. Анализ и визуализация данных на языке R. М.: ДМК Пресс, 2014
- Diez, D.M., Barr, C.D. and Çetinkaya-Rundel, M., 2015. OpenIntro Statistics. OpenIntro.
- Zuur, A., Ieno, E.N. and Smith, G.M., 2007. Analyzing ecological data. Springer Science & Business Media.
- Quinn G.P., Keough M.J. 2002. Experimental design and data analysis for biologists
- Logan M. 2010. Biostatistical Design and Analysis Using R. A Practical Guide