Введение в регрессионный анализ

Гарик Мороз

5 августа 2022, ДокМед ЛШ

План

Обо мне

Основы регрессии

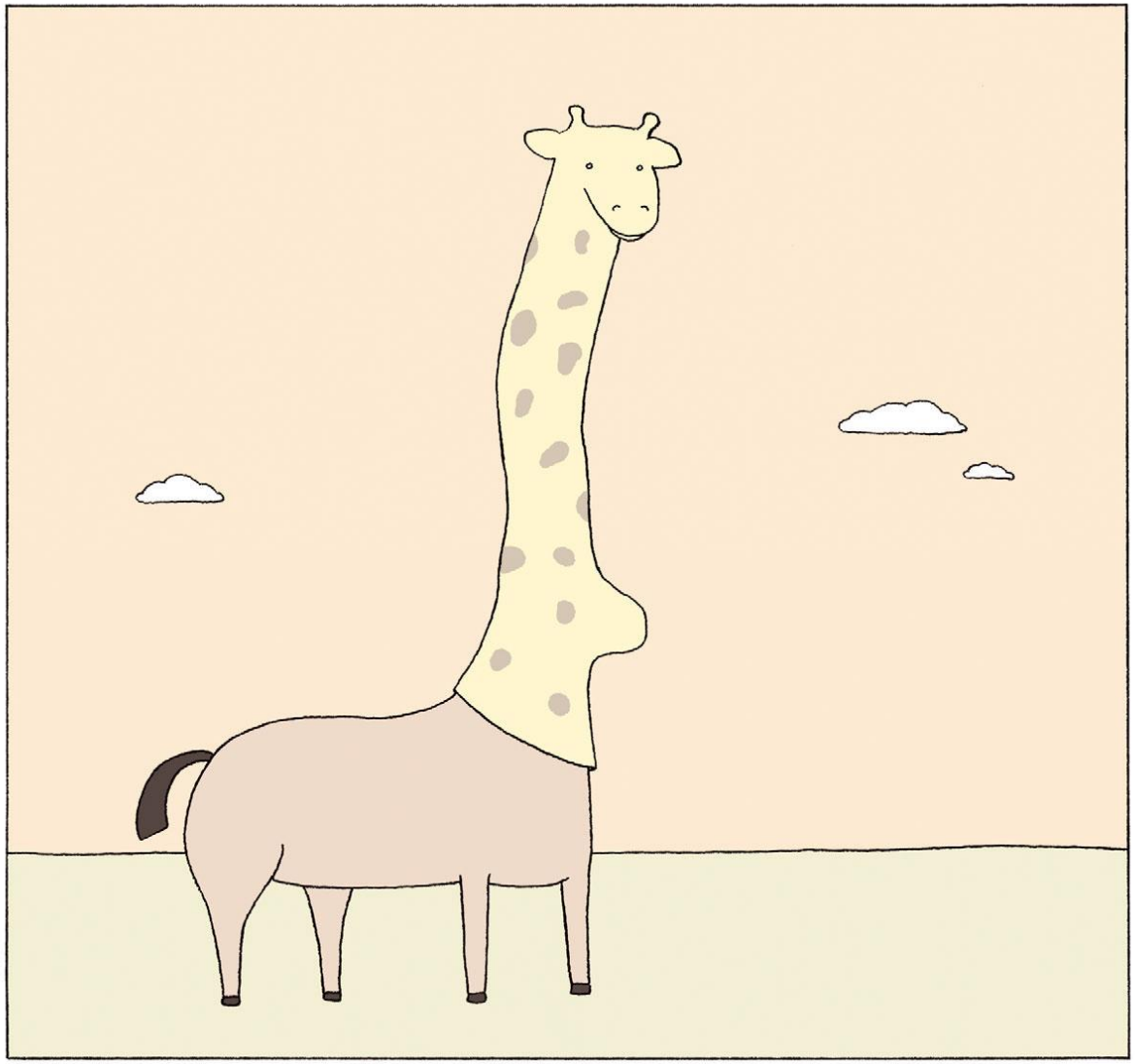
Усложнение регрессионой модели

На какие вопросы отвечает регрессия

Ограничения на применение регрессии

Не имею никакого отношения к медицине

Не имею никакого отношения к медицине



Картинка из [Farazmand 2017: 24]

Не имею никакого отношения к медицине

- теоретический и может быть даже компьютерный лингвист
- занимаюсь анализом данных, записал для лингвистов онлайн-курс, много лет веду курсы анализа данных в НИУ ВШЭ
- вел продвинутые треки мастерской АнДан на ЛШ

Не имею никакого отношения к медицине

- теоретический и может быть даже компьютерный лингвист
- занимаюсь анализом данных, записал для лингвистов онлайн-курс, много лет веду курсы анализа данных в НИУ ВШЭ
- вел продвинутые треки мастерской АнДан на ЛШ
- специально подготовился к вашей лекции и полистал статью и книгу с назвнанием *Regression analysis in medical research*
  - [Faguet and Davis 1984]
  - [Cleophas and Zwinderman 2021]

План

Обо мне

Основы регрессии

Усложнение регрессионой модели

На какие вопросы отвечает регрессия

Ограничения на применение регрессии

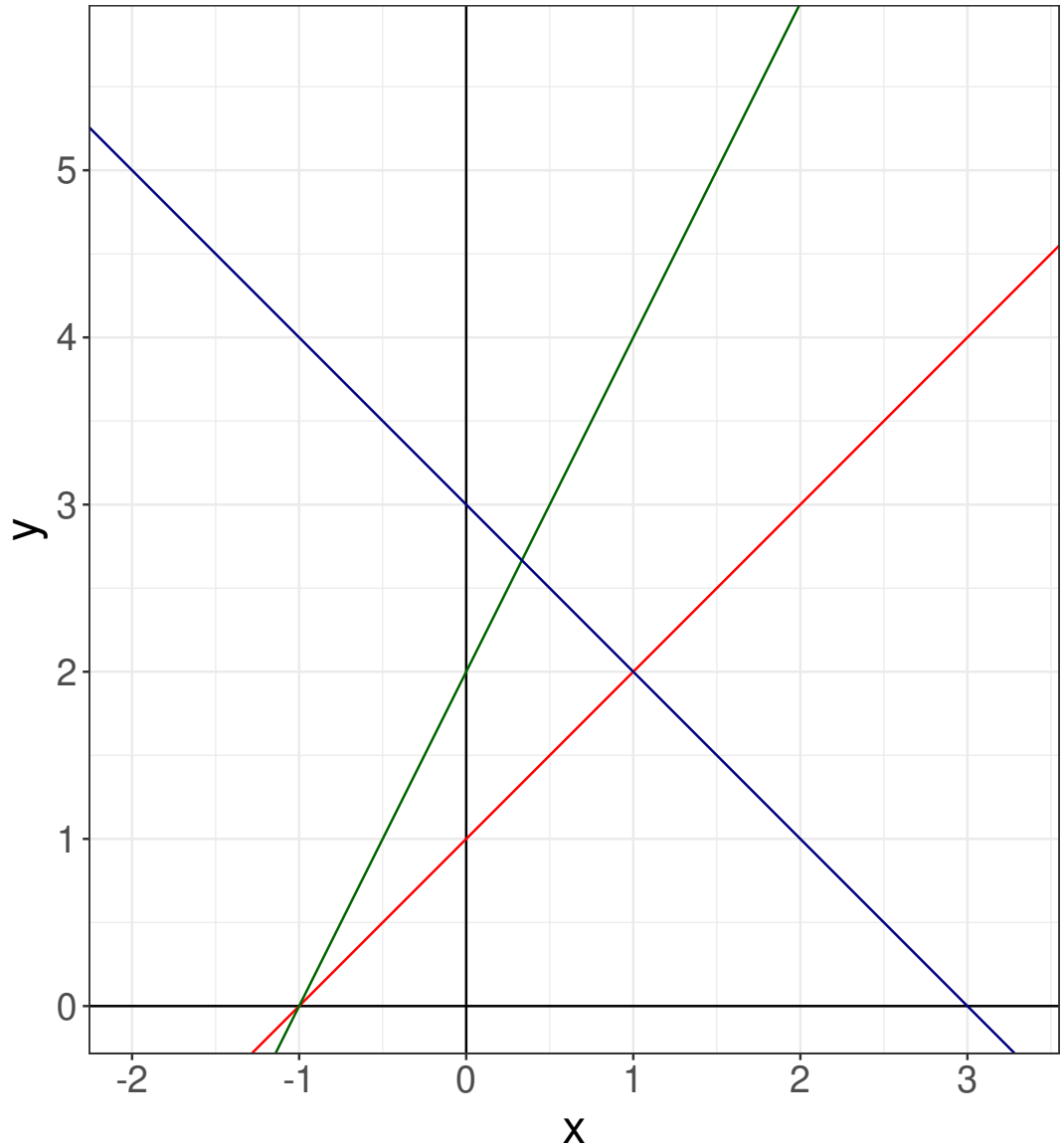
Основы регрессии

Суть регрессионного анализа в моделировании связи между двумя и более переменными при помощи прямой на плоскости. Формула прямой зависит от двух параметров: свободного члена (intercept) и углового коэффициента (slope).

$$y = \beta_0 + \beta_1 \times x$$

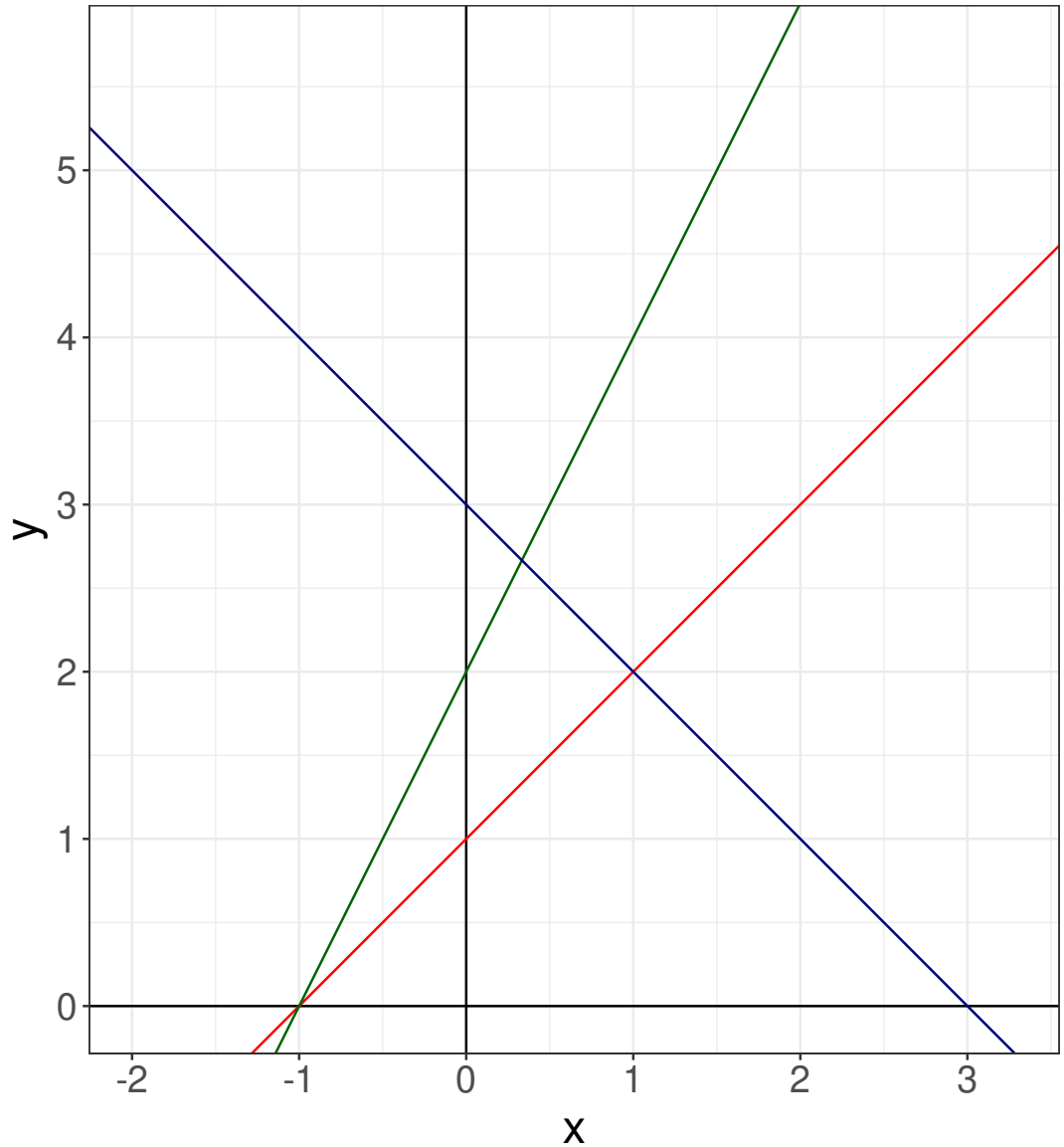
- β_0 – свободный член (intercept)
- β_1 – угловой коэффициента (slope)

Как провести линию на плоскости?



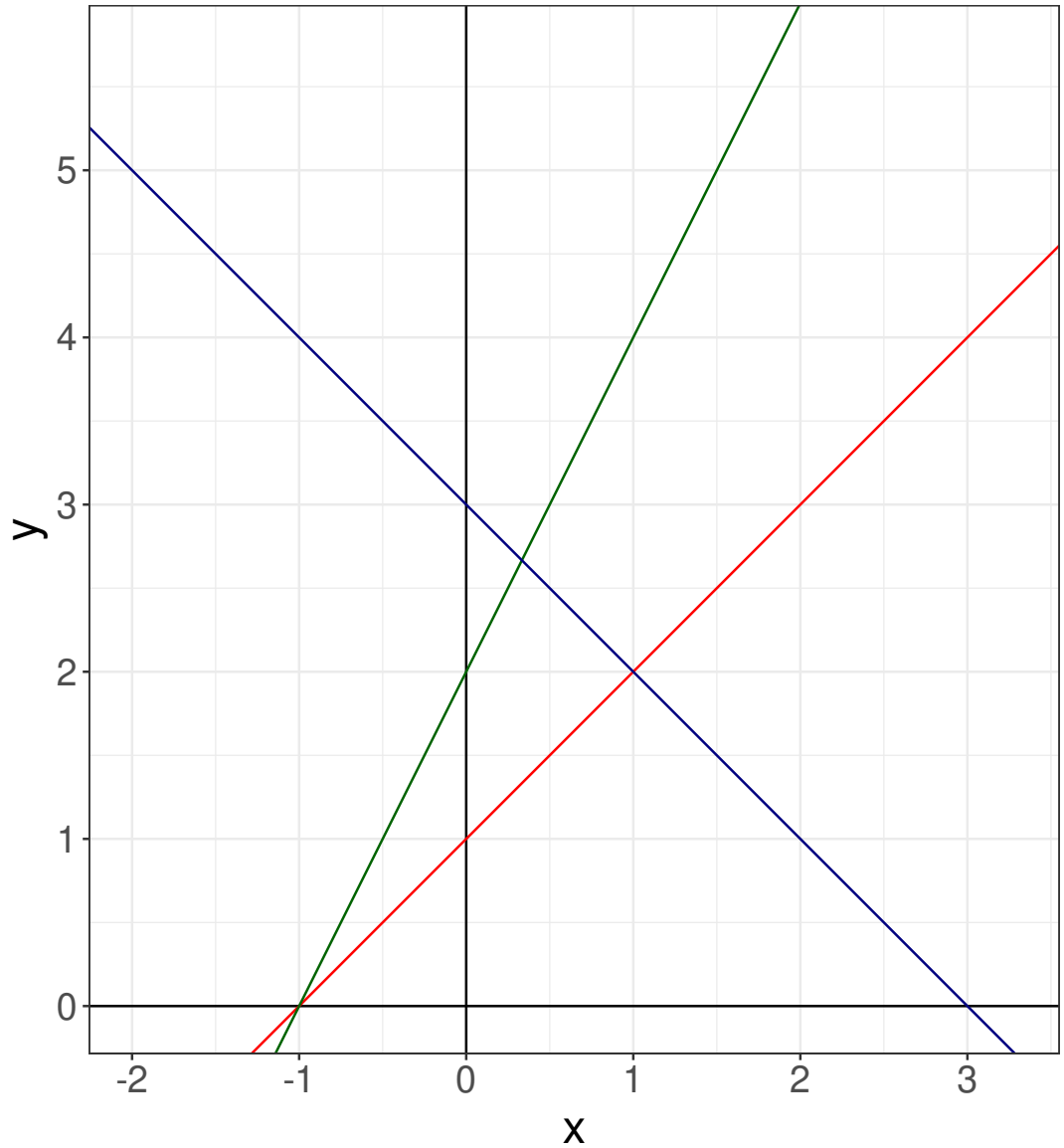
Какое значение свободного члена у красной прямой?

Как провести линию на плоскости?



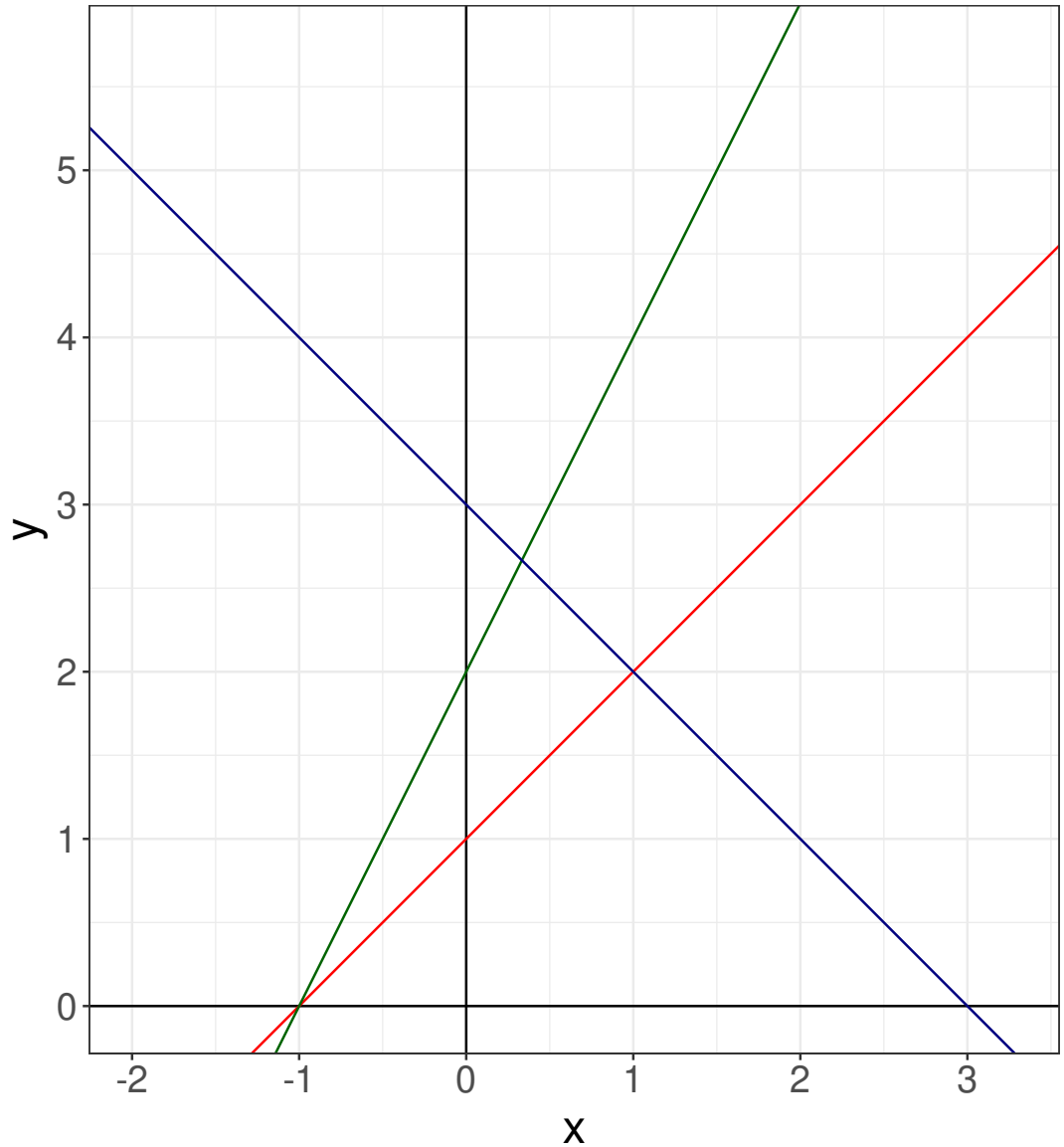
Какое значение свободного члена у зеленой прямой?

Как провести линию на плоскости?



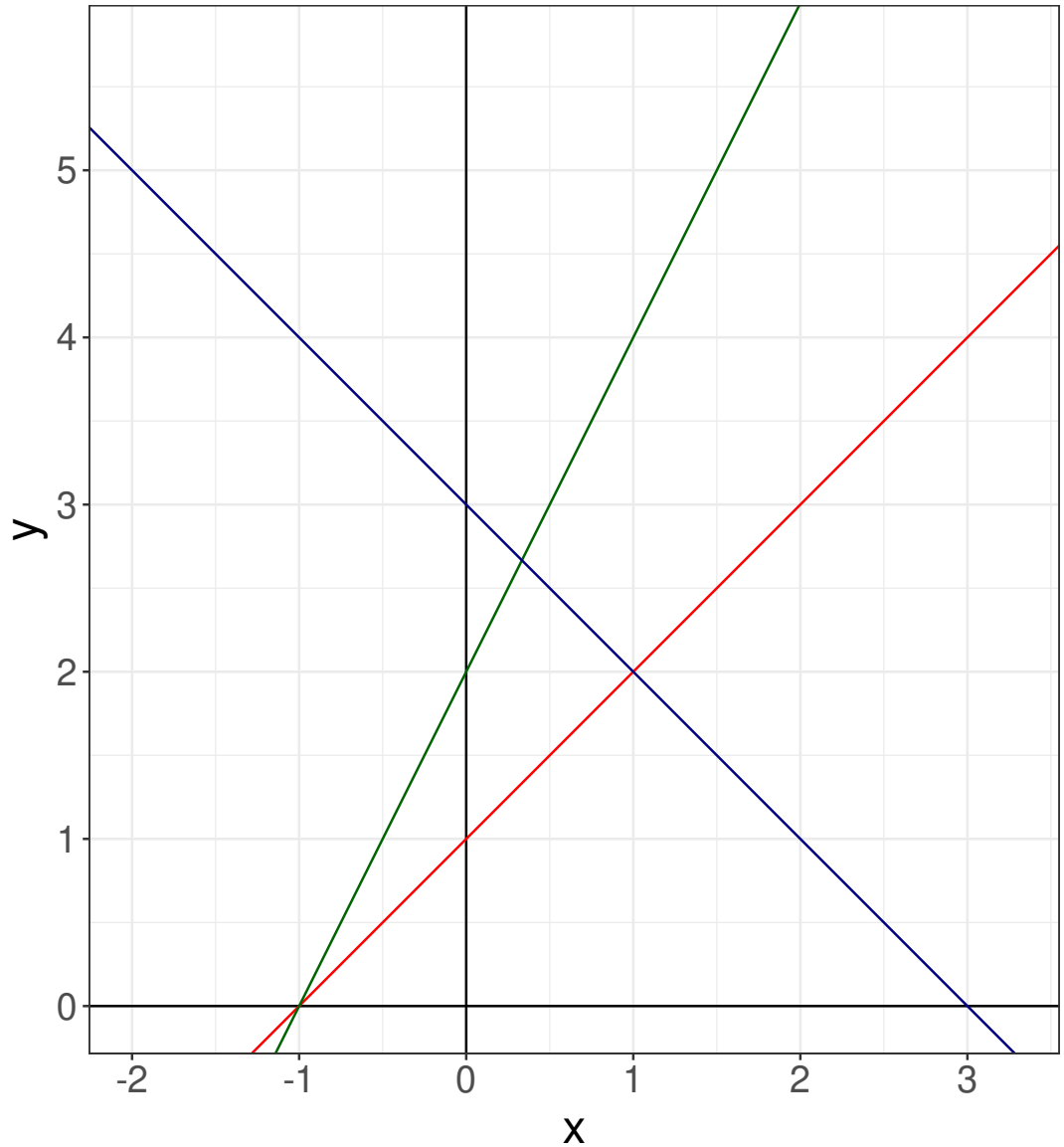
Какое значение свободного члена у синей прямой?

Как провести линию на плоскости?



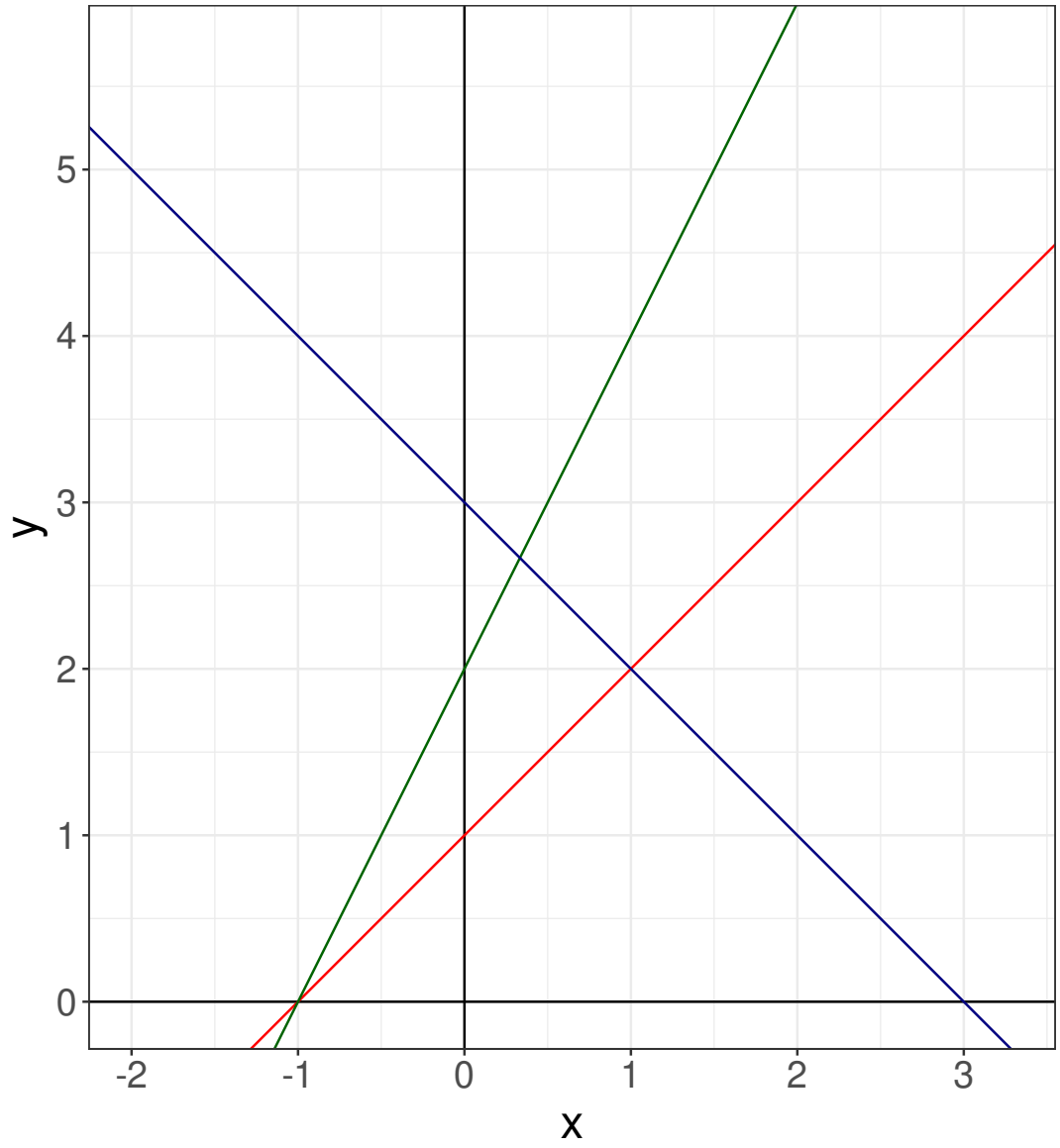
Какое значение углового коэффициента у красной прямой?

Как провести линию на плоскости?



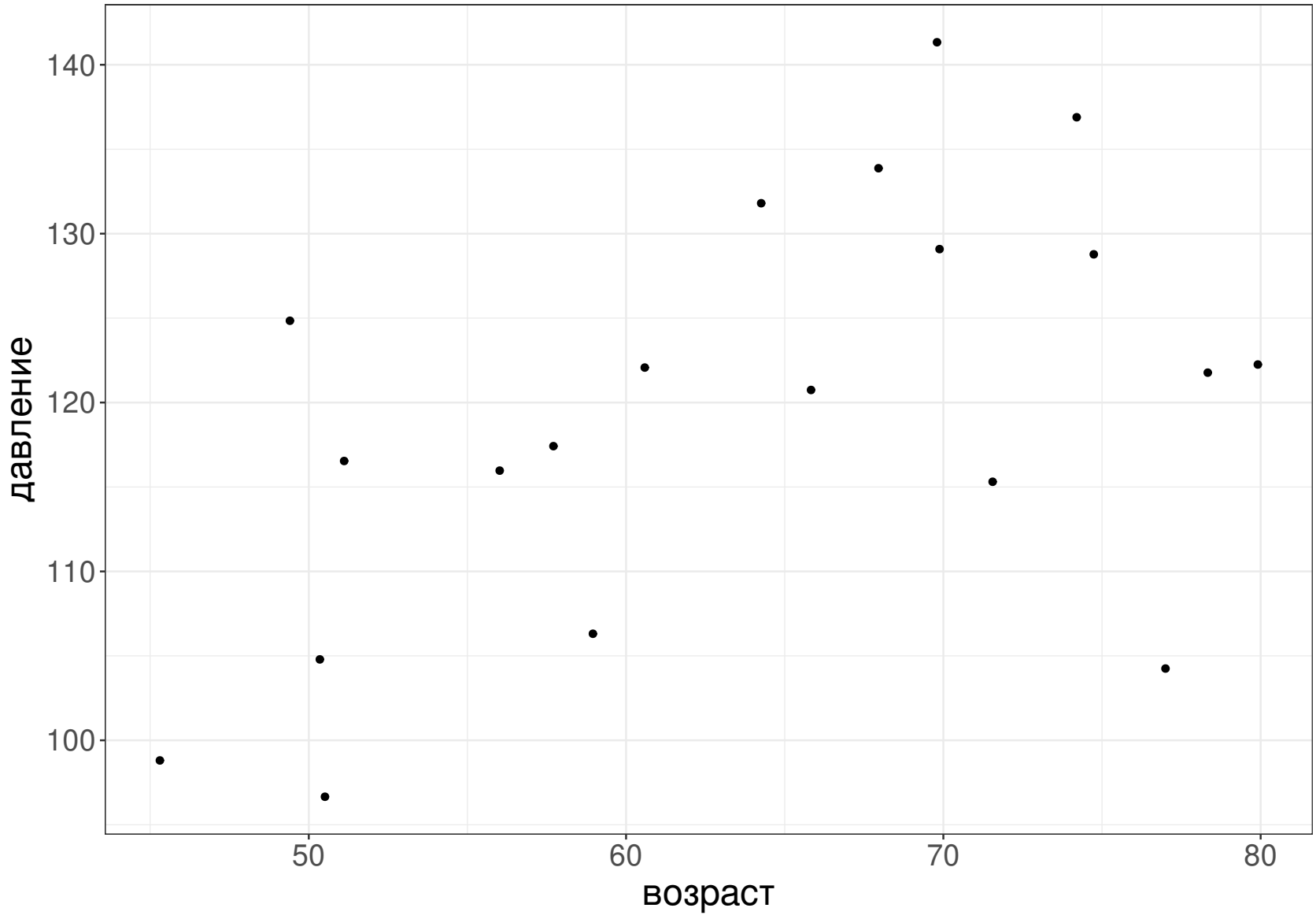
Какое значение углового коэффициента у зеленой прямой?

Как провести линию на плоскости?



Какое значение углового коэффициента у синей прямой?

У нас есть вот такие данные



Первый подход

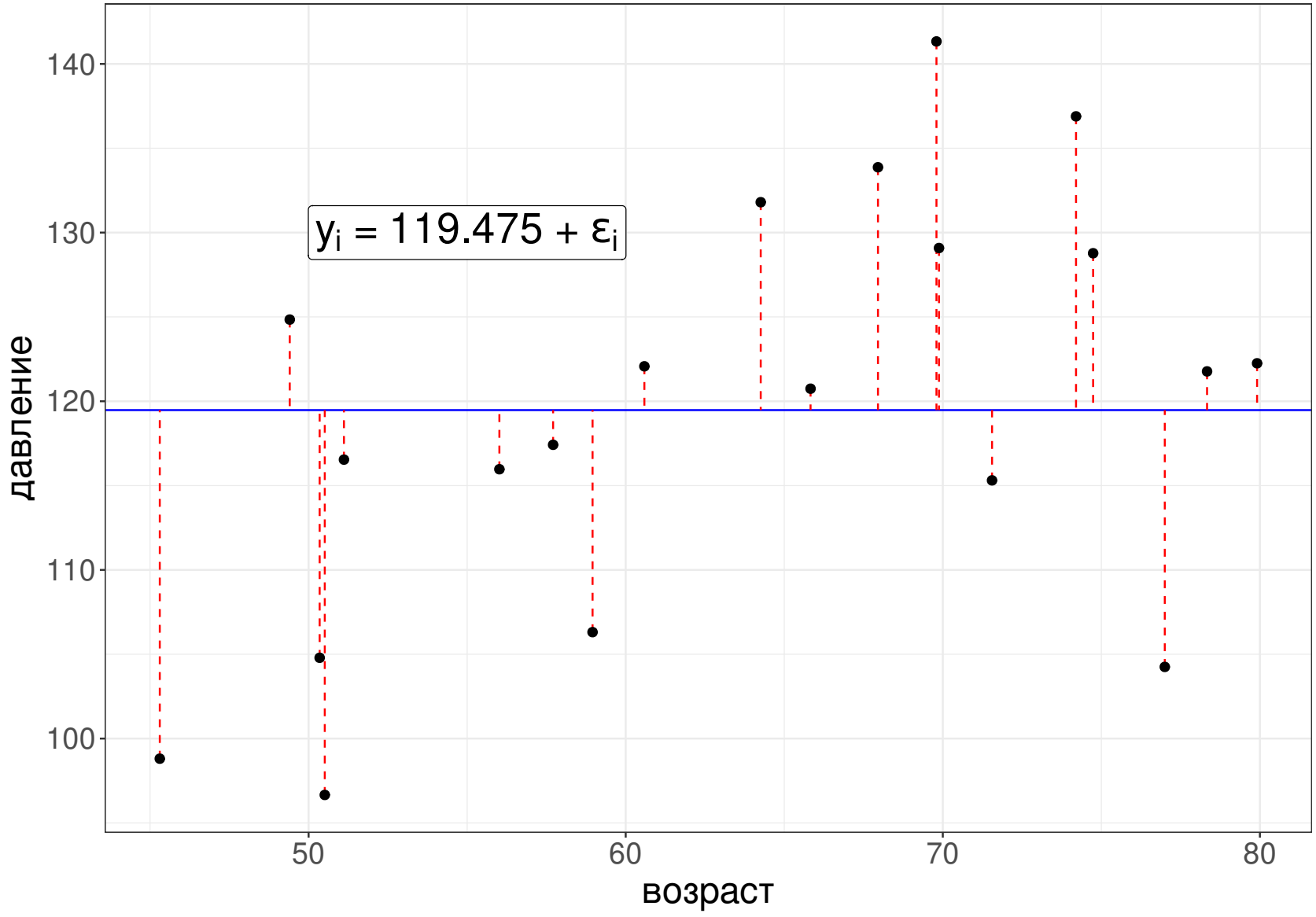
Представим, что мы пытаемся научиться предсказывать данные переменной Y, не используя других переменных. Какую меру можно выбрать?

$y_i = \hat{\beta}_0 + \epsilon_i$

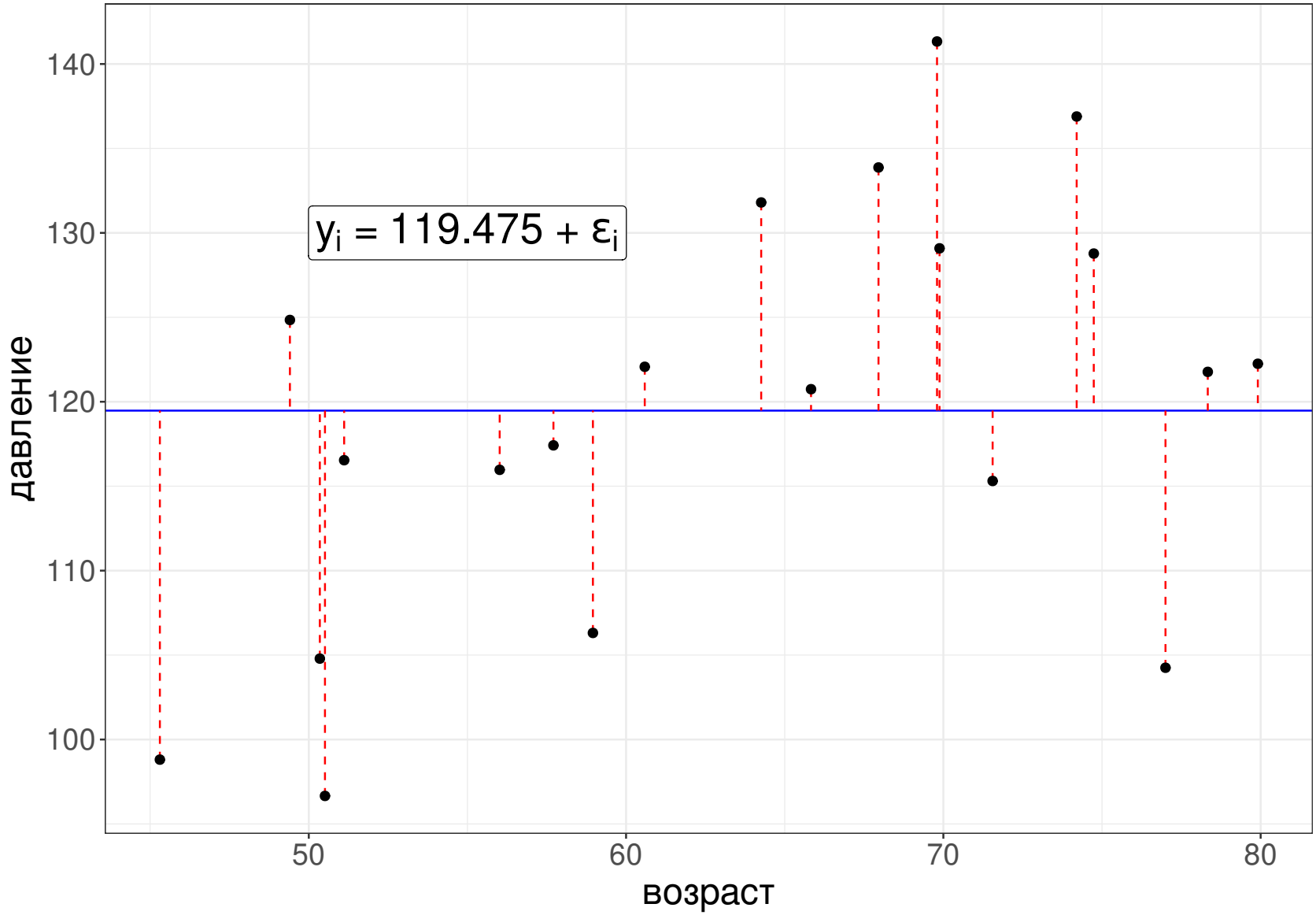
Представим, что мы пытаемся научиться предсказывать данные переменной Y, не используя других переменных. Тогда мы будем использовать формулу в заголовке

- y_i — i-ый элемент вектора значений Y (предсказываемая переменная);
- $\hat{\beta}_0$ — оценка случайного члена (intercept);
- ϵ_i — i-ый остаток, разница между оценкой модели ($\hat{\beta}_0$) и реальным значением y_i; весь вектор остатков иногда называют случайным шумом (на графике выделены красным);
- i — номер наблюдения.

$$y_i = \hat{\beta}_0 + \epsilon_i$$

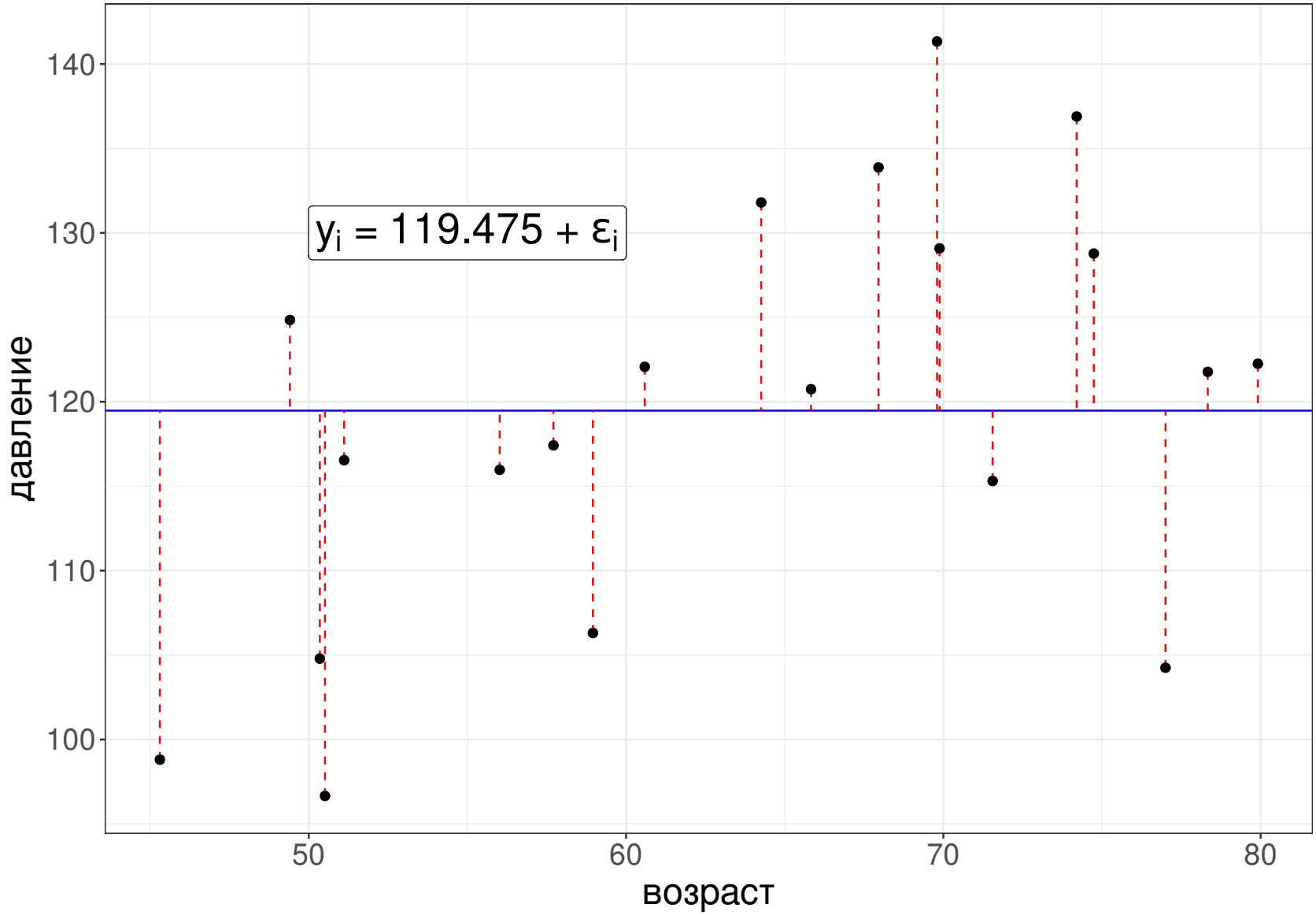


$$y_i = \hat{\beta}_0 + \epsilon_i$$



Все, теперь вы знаете основу регрессионного анализа.

$$y_i = \hat{\beta}_0 + \epsilon_i$$



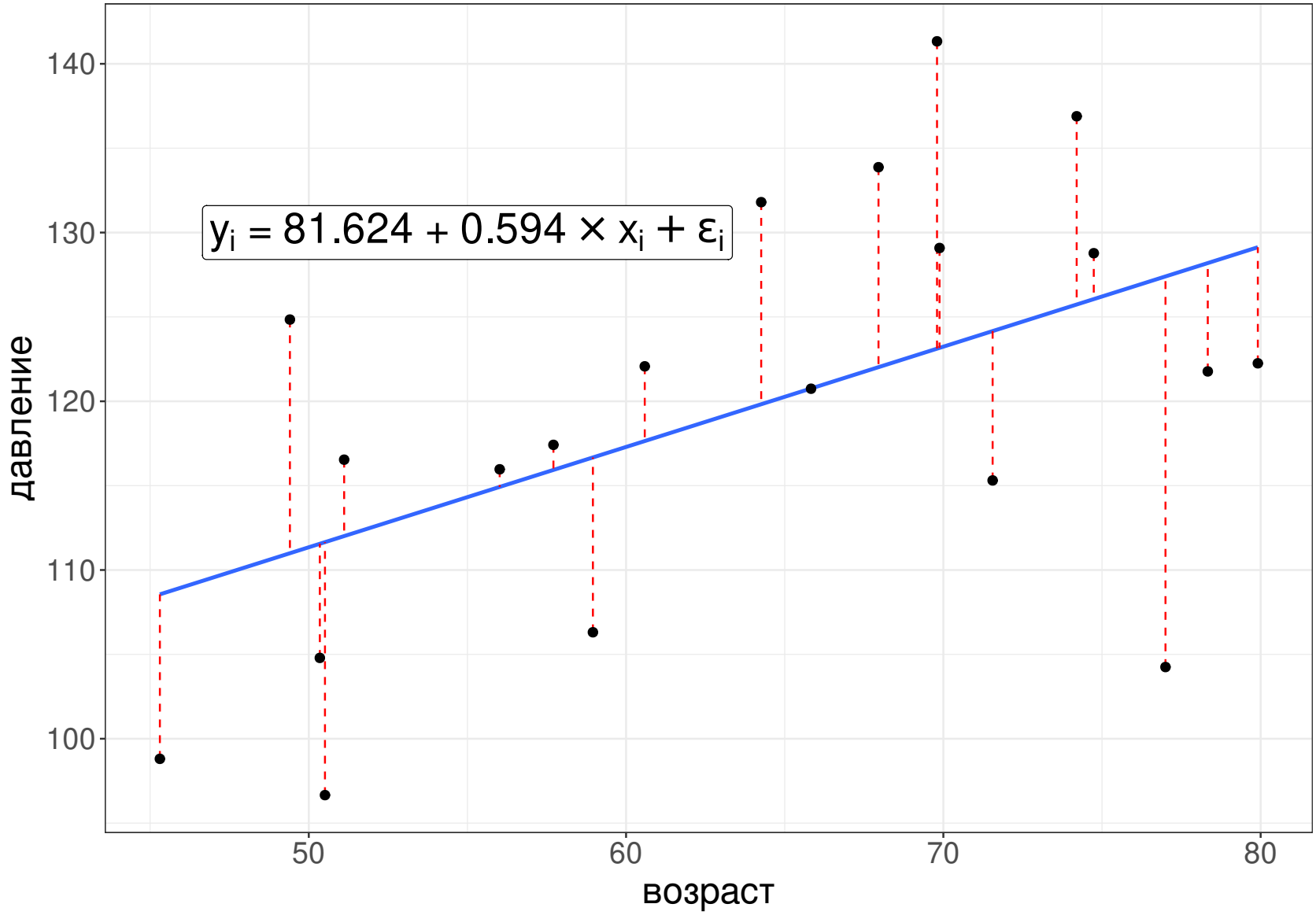
Все, теперь вы знаете основу регрессионного анализа. Почти.

$y_i = \hat{\beta}_0 + \hat{\beta}_1 \times x_i + \epsilon_i$

Когда мы пытаемся научиться предсказывать данные одной переменной Y при помощи другой переменной X, мы получаем формулу в заголовке, где

- x_i — i-ый элемент вектора значений X (предиктор);
- y_i — i-ый элемент вектора значений Y (предсказываемая переменная);
- $\hat{\beta}_0$ — оценка случайного члена (intercept);
- $\hat{\beta}_1$ — оценка углового коэффициента (slope);
- ϵ_i — i-ый остаток, разница между оценкой модели ($\hat{\beta}_0 + \hat{\beta}_1 \times x_i$) и реальным значением y_i; весь вектор остатков иногда называют случайным шумом (на графике выделены красным);
- i — номер наблюдения.

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \times x_i + \epsilon_i$$



$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \times x_i + \epsilon_i$$

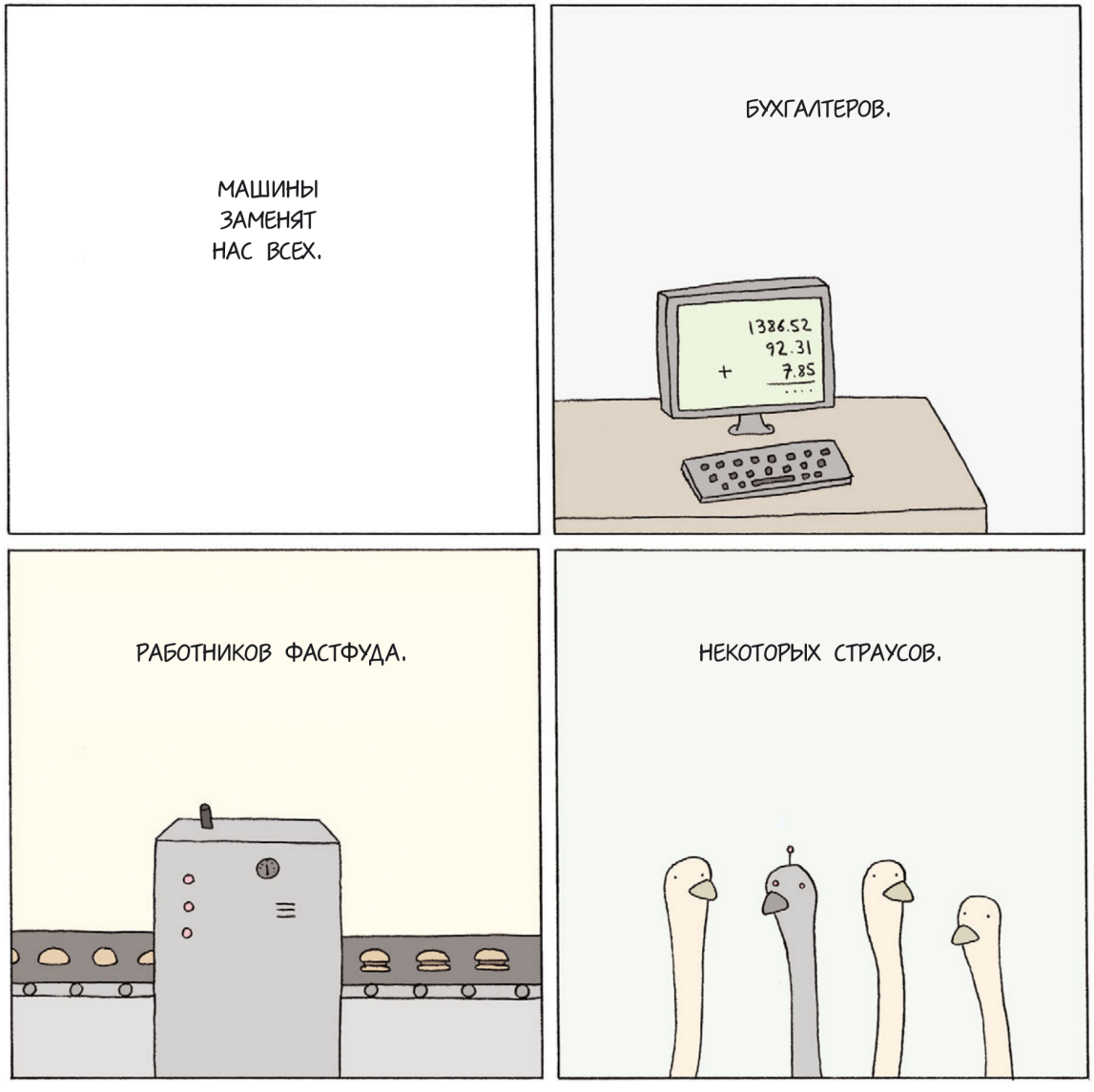
Таким образом, задача регрессии — оценить параметры $\hat{\beta}_0$ и $\hat{\beta}_1$, если нам известны все значения x_i и y_i и мы пытаемся минимизировать значния ϵ_i. В данном конкретном случае, задачу можно решить аналитически и получить следующие формулы:

$$\hat{\beta}_1 = \frac{(\sum_{i=1}^{n} x_i \times y_i) - n \times \bar{x} \times \bar{y}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \times \bar{x}$$

Не волнуйтесь, софт посчитает все за вас.

Не волнуйтесь, софт посчитает все за вас.



Картинка из [Farazmand 2017: 48]

Проба руки

Вот данные:

```
x: 21, 21, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 17.8, 16.4, 17.3, 15.2
y: 2.62, 2.875, 2.32, 3.215, 3.44, 3.46, 3.57, 3.19, 3.15, 3.44, 3.44, 4.07, 3.73, 3.78
```

Попробуйте посчитать коэффициенты регрессии вот здесь.

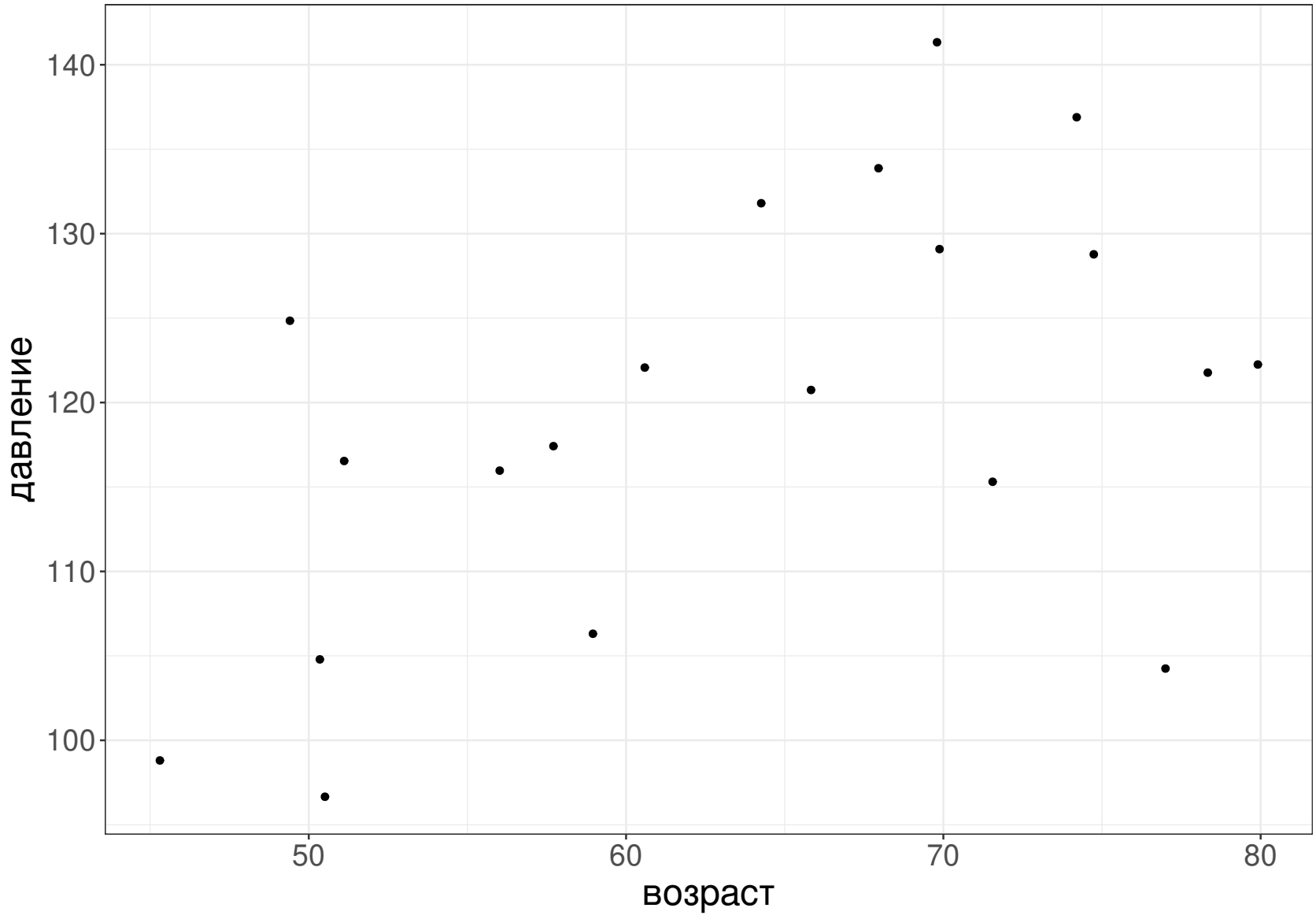
Проба руки

Вот те же данные, но предиктор и предсказываемая переменная поменяны местами:

```
x: 2.62, 2.875, 2.32, 3.215, 3.44, 3.46, 3.57, 3.19, 3.15, 3.44, 3.44, 4.07, 3.73, 3.78
y: 21, 21, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 17.8, 16.4, 17.3, 15.2
```

Попробуйте посчитать коэффициенты регрессии вот здесь.

Снова рассмотрим наш эксперимент



Снова рассмотрим наш эксперимент

```
Call:
lm(formula = bp ~ age, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-23.149  -7.371   1.265   7.229  18.215

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  81.6241    15.0994   5.406 3.89e-05 ***
age           0.5945     0.2340   2.541   0.0205 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.04 on 18 degrees of freedom
Multiple R-squared:  0.264, Adjusted R-squared:  0.2231
F-statistic: 6.456 on 1 and 18 DF,  p-value: 0.02048
```

Снова рассмотрим наш эксперимент

```
Call: код для вызова регрессии
lm(formula = bp ~ age, data = df)

Residuals: распределение остатков (должно быть вокруг нуля)
    Min      1Q  Median      3Q     Max
-23.149  -7.371   1.265   7.229  18.215

Coefficients: оценка коэффициентов
            Estimate Std. Error      t value  Pr(>|t|)
              оценка ст. ошибка t-статистика  p-value  ст. знач.
(Intercept)  81.6241    15.0994        5.406 3.89e-05  ***
age           0.5945     0.2340        2.541   0.0205  *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.04 on 18 degrees of freedom
Multiple R-squared:  0.264, Adjusted R-squared:  0.2231
r-квадрат -- коэффициент корреляции Пирсона в квадрате
F-statistic: 6.456 on 1 and 18 DF,  p-value: 0.02048
p-value в этом месте совпадает с результатом ANOVA
```

План

Обо мне

Основы регрессии

Усложнение регрессионой модели

На какие вопросы отвечает регрессия

Ограничения на применение регрессии

Множественная регрессионая модель

Вообще-то можно инкорпорировать много предикторов в одну регрессию:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \times x_i^1 + \hat{\beta}_2 \times x_i^2 ... + \hat{\beta}_k \times x_i^k + \epsilon_i$$

Множественная регрессионая модель

Вообще-то можно инкорпорировать много предикторов в одну регрессию:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \times x_i^1 + \hat{\beta}_2 \times x_i^2 ... + \hat{\beta}_k \times x_i^k + \epsilon_i$$

В таком случае все предикторы просто становятся весами для нашего углового коэффициента. Т. е. регрессия в каком-то смысле ранжирует предикторы.

$$bp_i = 45.0604 + 0.7728 \times age_i + 0.242 \times Na_i + \epsilon_i$$

```
Call:
lm(formula = bp ~ age + na_plus, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-32.936  -7.698   2.057  10.129  25.399

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.0604    29.4028   1.533    0.133
age           0.7728     0.1769   4.368 7.52e-05 ***
na_plus       0.2426     0.1839   1.319    0.194
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.27 on 44 degrees of freedom
Multiple R-squared:  0.3083,    Adjusted R-squared:  0.2769
F-statistic: 9.805 on 2 and 44 DF,  p-value: 0.0003008
```

Категориальная переменная

Когда в данных есть категориальная переменная с n возможных значений, то принято ее превращать в $n - 1$ фиктивную переменную (dummy variable). Например:

цвет глаз	→	голубые	карие	серые
голубые	→	1	0	0
карие	→	0	1	0
серые	→	0	0	1
зеленые	→	0	0	0

Категориальная переменная

Когда в данных есть категориальная переменная с n возможных значений, то принято ее превращать в $n - 1$ фиктивную переменную (dummy variable). Например:

цвет глаз	→	голубые	карие	серые
голубые	→	1	0	0
карие	→	0	1	0
серые	→	0	0	1
зеленые	→	0	0	0

Какие значения примут переменные в случае карих глаз?

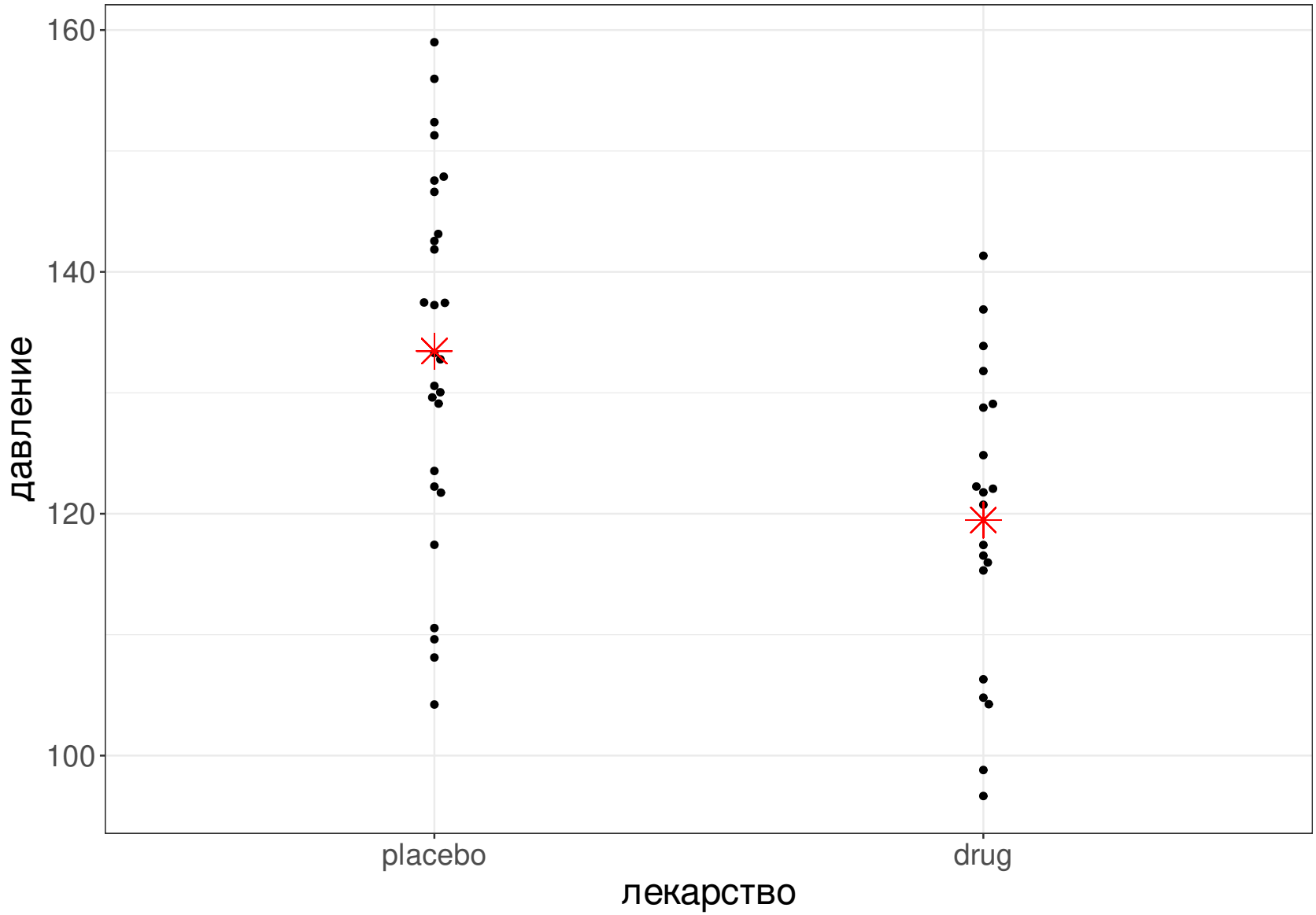
Категориальная переменная

Когда в данных есть категориальная переменная с n возможных значений, то принято ее превращать в $n - 1$ фиктивную переменную (dummy variable). Например:

цвет глаз	→	голубые	карие	серые
голубые	→	1	0	0
карие	→	0	1	0
серые	→	0	0	1
зеленые	→	0	0	0

Какие значения примут переменные в случае карих глаз?
Какие значения примут переменные в случае зеленых глаз?

Категориальная переменная



$y_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{dummy treatment}_i + \epsilon_i$

treatment	→	dummy treatment
drug	→	1
placebo	→	0

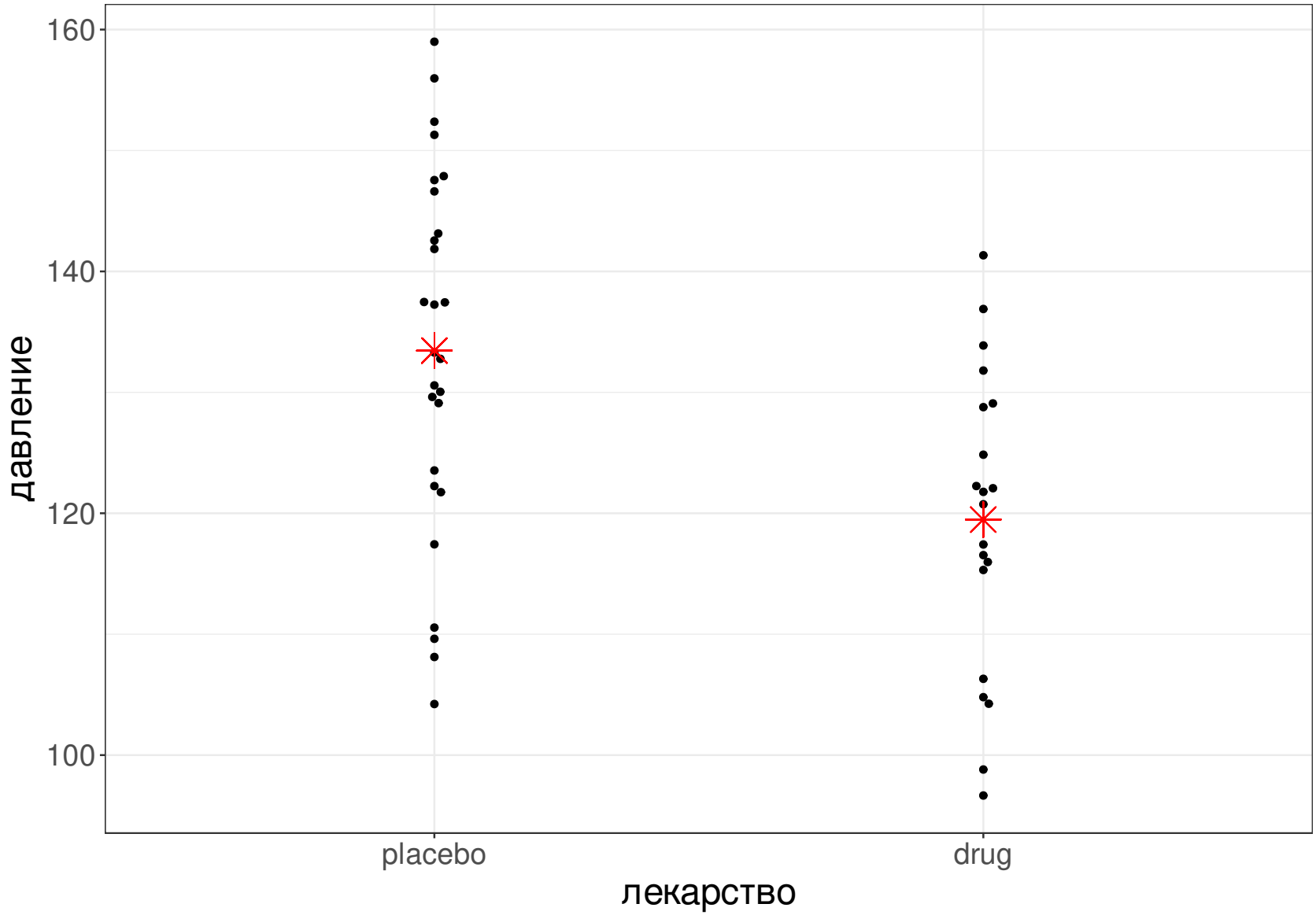
Так как `dummy_treatment` принимает либо значение 1, либо значение 0, то получается, что модель предсказывает лишь два значения:

$$y_i = \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 \times 1 + \epsilon_i = \hat{\beta}_0 + \hat{\beta}_1 + \epsilon_i, \text{ если лекарство} \\ \hat{\beta}_0 + \hat{\beta}_1 \times 0 + \epsilon_i = \hat{\beta}_0 + \epsilon_i, \text{ если плацебо} \end{cases}$$

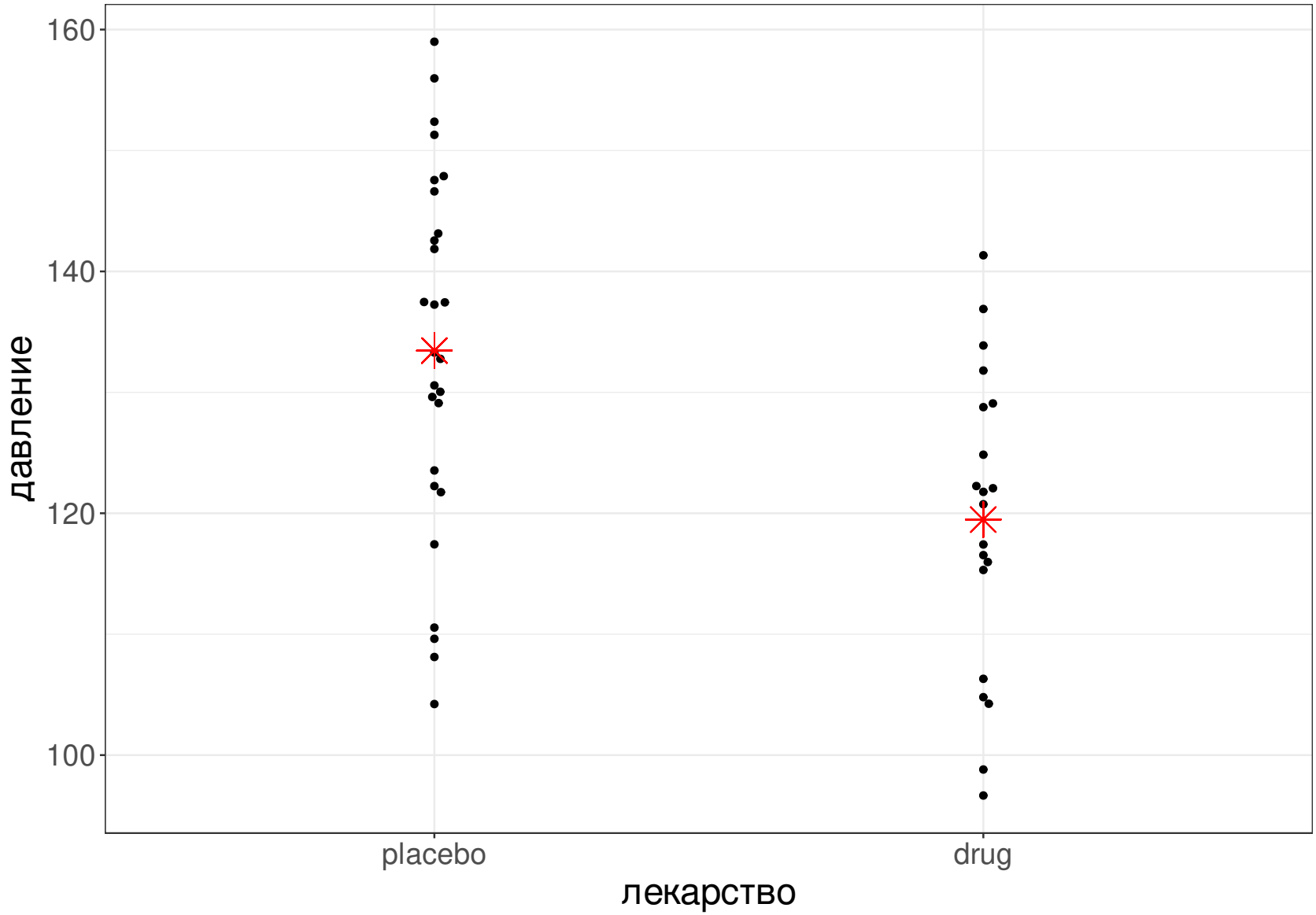
$y_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{dummy treatment}_i + \epsilon_i$

```
Call:
lm(formula = bp ~ treatment, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-29.231 -10.562   1.270   9.648  25.540

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   133.454      2.714   49.17   <2e-16 ***
treatmentdrug -13.980      4.161   -3.36   0.0016 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.1 on 45 degrees of freedom
Multiple R-squared:  0.2006,    Adjusted R-squared:  0.1828
F-statistic: 11.29 on 1 and 45 DF,  p-value: 0.001597
```

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{dummy treatment}_i + \epsilon_i$$



$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{dummy treatment}_i + \epsilon_i$$

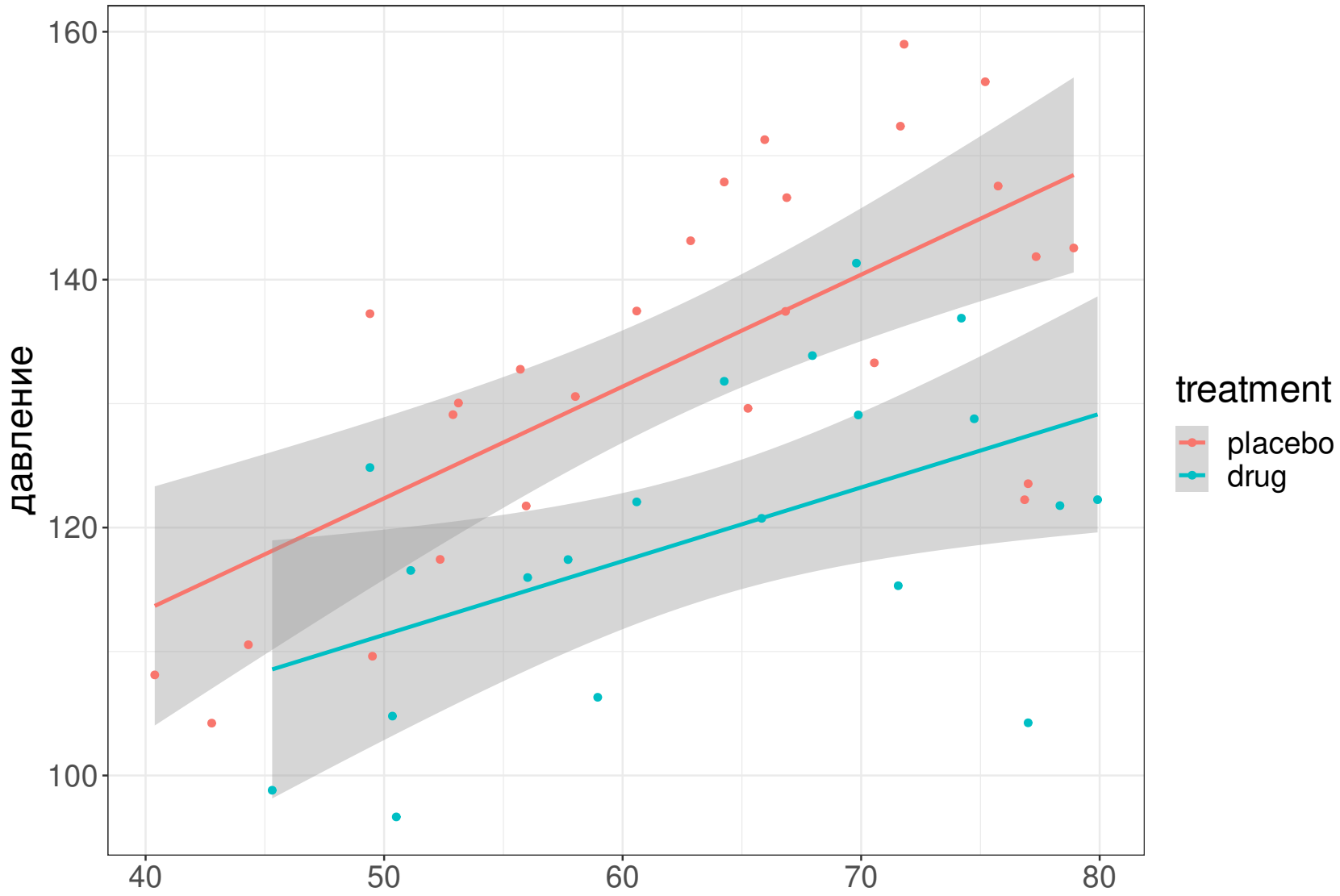


Да, мы запустили регрессию, чтобы посчитать два средних

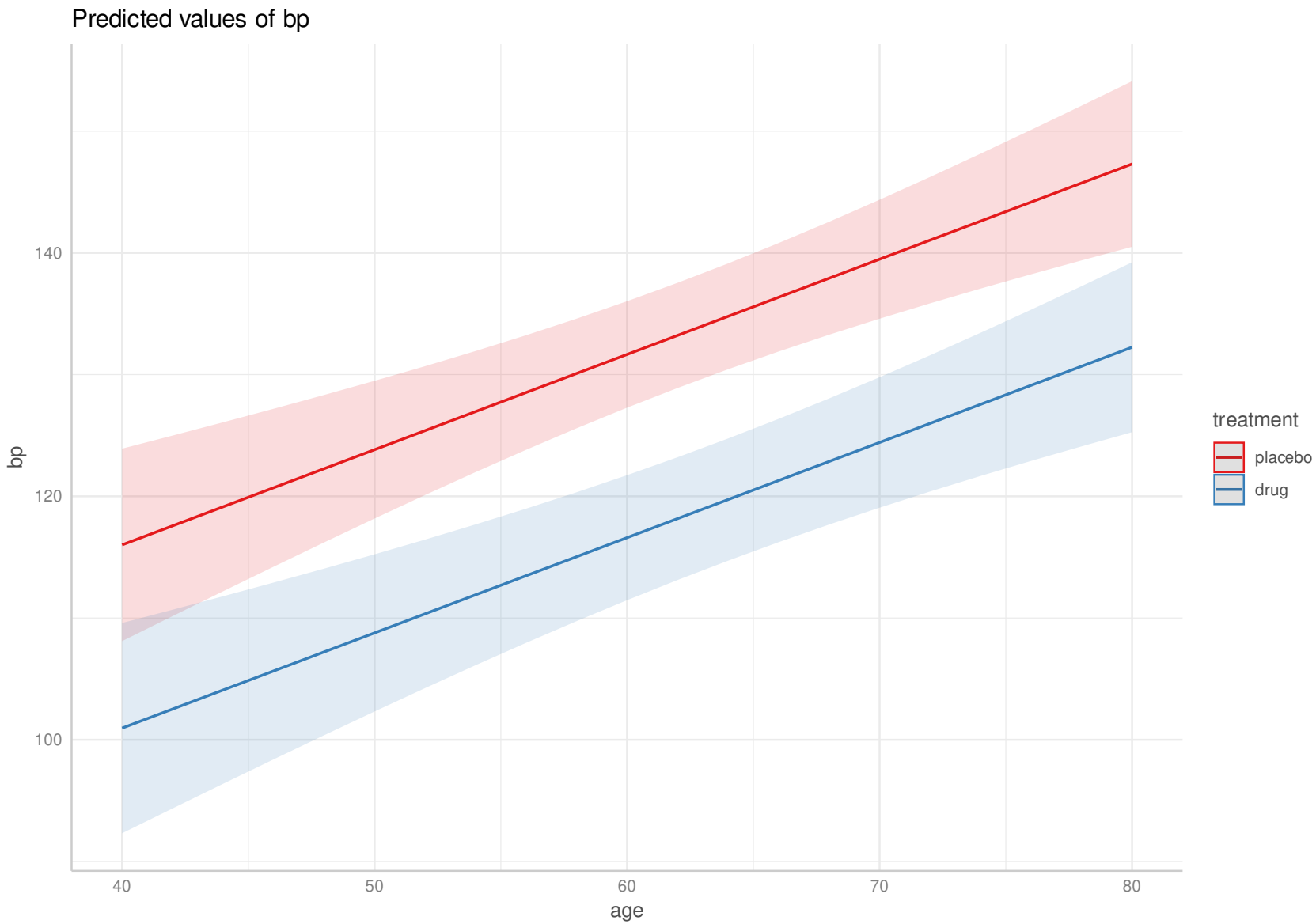
А насколько мы уверены в наших линиях?

А насколько мы уверены в наших линиях?

Благодаря станадртным ошибкам коэффициентов в выдаче регрессии, можно строить доверительные интервалы:



Я предпочитаю effect plots



А как сравнивать модели?

Есть модели, какая лучше?

- bp ~ age + treatment + Na+
- bp ~ age + treatment
- bp ~ age + Na+
- bp ~ treatment + Na+
- bp ~ age
- bp ~ treatment
- bp ~ Na+

А как сравнивать модели?

Есть модели, какая лучше?

- bp ~ age + treatment + Na+
- bp ~ age + treatment
- bp ~ age + Na+
- bp ~ treatment + Na+
- bp ~ age
- bp ~ treatment
- bp ~ Na+

Люди придумали некоторые методы:

- можно сравнивать статистическую значимость предикторов
- можно сравнивать R^2
- чаще всего используют так называемые информационные критерии, самый популярный – AIC (Akaike information criterion). Чем меньше значение, тем модель лучше.

План

Обо мне

Основы регрессии

Усложнение регрессионой модели

На какие вопросы отвечает регрессия

Ограничения на применение регрессии

На какие вопросы отвечает регрессия

- оценка коэффициентов
- проверка стат. значимости коэффициентов
- проверка стат. значимости модели
- интерполяция/экстраполяция значений

На какие вопросы отвечает регрессия

- оценка коэффициентов
- проверка стат. значимости коэффициентов
- проверка стат. значимости модели
- интерполяция/эктраполяция значений
- подбор модели, но это, видимо, не для медицины

План

Обо мне

Основы регрессии

Усложнение регрессионой модели

На какие вопросы отвечает регрессия

Ограничения на применение регрессии

Ограничения на применение

- связь между предсказываемой переменной и предикторами должна быть линейной
  - можно как-то трансформировать переменные (корень, логарифм)
  - нелинейные регресии (если связь между переменными нелинейна)

Ограничения на применение

- связь между предсказываемой переменной и предикторами должна быть линейной
  - можно как-то трансформировать переменные (корень, логарифм)
  - нелинейные регресии (если связь между переменными нелинейна)
- остатки должны быть нормально распределены

Ограничения на применение

- связь между предсказываемой переменной и предикторами должна быть линейной
  - можно как-то трансформировать переменные (корень, логарифм)
  - нелинейные регресии (если связь между переменными нелинейна)
- остатки должны быть нормально распределены
- дисперсия остатков вокруг регрессионной линии должно быть постоянно (гомоскидастично)

Ограничения на применение

- связь между предсказываемой переменной и предикторами должна быть линейной
  - можно как-то трансформировать переменные (корень, логарифм)
  - нелинейные регресии (если связь между переменными нелинейна)
- остатки должны быть нормально распределены
- дисперсия остатков вокруг регрессионной линии должно быть постоянно (гомоскидастично)
  - см. этот пост про это и предыдущую проблемы
- предикторы не должны коррелировать друг с другом

Ограничения на применение

- связь между предсказываемой переменной и предикторами должна быть линейной
  - можно как-то трансформировать переменные (корень, логарифм)
  - нелинейные регресии (если связь между переменными нелинейна)
- остатки должны быть нормально распределены
- дисперсия остатков вокруг регрессионной линии должно быть постоянно (гомоскидастично)
  - см. этот пост про это и предыдущую проблемы
- предикторы не должны коррелировать друг с другом
- все наблюдения в регрессии должны быть независимы друг от друга
  - регрессия со смешанными эффектами (если внутри данных есть группировки)

Ограничения на применение

- связь между предсказываемой переменной и предикторами должна быть линейной
  - можно как-то трансформировать переменные (корень, логарифм)
  - нелинейные регресии (если связь между переменными нелинейна)
- остатки должны быть нормально распределены
- дисперсия остатков вокруг регрессионной линии должно быть постоянно (гомоскидастично)
  - см. этот пост про это и предыдущую проблемы
- предикторы не должны коррелировать друг с другом
- все наблюдения в регрессии должны быть независимы друг от друга
  - регрессия со смешанными эффектами (если внутри данных есть группировки)
- предсказываемая переменная должна быть числовой переменной
  - логистическая (два возможных исхода)
  - мультиномиальная (больше двух дискретных исходов)

Список литературы

Ton J. Cleophas and Aeilko H. Zwinderman. *Regression Analysis In Medical Research: For Starters And 2nd Levelers*. Springer, 2nd ed. edition, 2021. ISBN 3030613933; 9783030613938; 9783030613945; 3030613941.

G. B. Faguet and H. C. Davis. Regression analysis in medical research. *Southern medical journal*, 77(6):722–5, 1984.

Reza Farazmand. *Comics for a Strange World: A Book of Poorly Drawn Lines*. Penguin, 2017.