



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Национальная олимпиада по анализу данных

РЕГРЕССИЯ

Автор: Слаболицкий Илья Сергеевич,
преподаватель Департамента прикладной экономики Факультета
экономических наук, НИУ ВШЭ

Электронная почта для связи: islabolitskiy@hse.ru

Москва, 2021



ОТ ВЫБОРКИ К ДАННЫМ



ОТ ВЫБОРКИ К ДАННЫМ

Как связаны между собой выборка и данные?



ОТ ВЫБОРКИ К ДАННЫМ

Как связаны между собой выборка и данные?

- Предположим, что мы при помощи некоторого способа получили выборку из индивидов и на ее основе провели опрос. Результаты опроса – это наши **данные** (data). Что же представляют собой полученные данные?



ОТ ВЫБОРКИ К ДАННЫМ

Как связаны между собой выборка и данные?

- Предположим, что мы при помощи некоторого способа получили выборку из индивидов и на ее основе провели опрос. Результаты опроса – это наши **данные** (data). Что же представляют собой полученные данные?
- Чаще всего, это таблица. Строки таблицы – это опрошенные индивиды (их ***n*** человек, ровно столько элементов в выборке), а столбцы таблицы – это те вопросы, которые индивидам задавали в ходе опроса. Соответственно, ячейки внутри таблицы – это ответ определенного индивида на определенный вопрос.

Номер индивида	Какой у Вас пол?	Сколько Вам полных лет?	...	Какой ежемесячный доход (тыс. руб.) удовлетворил бы Ваши ежемесячные потребности?
Индивид 1	Мужской	39	...	73.5
Индивид 2	Женский	21	...	93.0
...	...	...	...	...
Индивид <i>n</i>	Мужской	48	...	48.5



ОТ ДАННЫМ К МОДЕЛИ



ОТ ДАННЫМ К МОДЕЛИ

Для чего нам нужны полученные данные?



ОТ ДАННЫМ К МОДЕЛИ

Для чего нам нужны полученные данные?

Исследователи используют данные, чтобы строить модели. **Модель** (model) – это некоторая формула, которая связывает зависимую (объясняемую) переменную с одной или несколькими независимыми (объясняющими) переменными.



ОТ ДАННЫМ К МОДЕЛИ

Для чего нам нужны полученные данные?

Исследователи используют данные, чтобы строить модели. **Модель** (model) – это некоторая формула, которая связывает зависимую (объясняемую) переменную с одной или несколькими независимыми (объясняющими) переменными.

Продажа мороженого в летние дни

Номер дня	Доход с продаж за день	Температура воздуха
День 1	53.48	24.57
День 2	62.52	26.01
...	...	...
День 500	65.57	28.96

Данные по доходам с продажи мороженого. *Источник: Kaggle*



ОТ ДАННЫМ К МОДЕЛИ

Для чего нам нужны полученные данные?

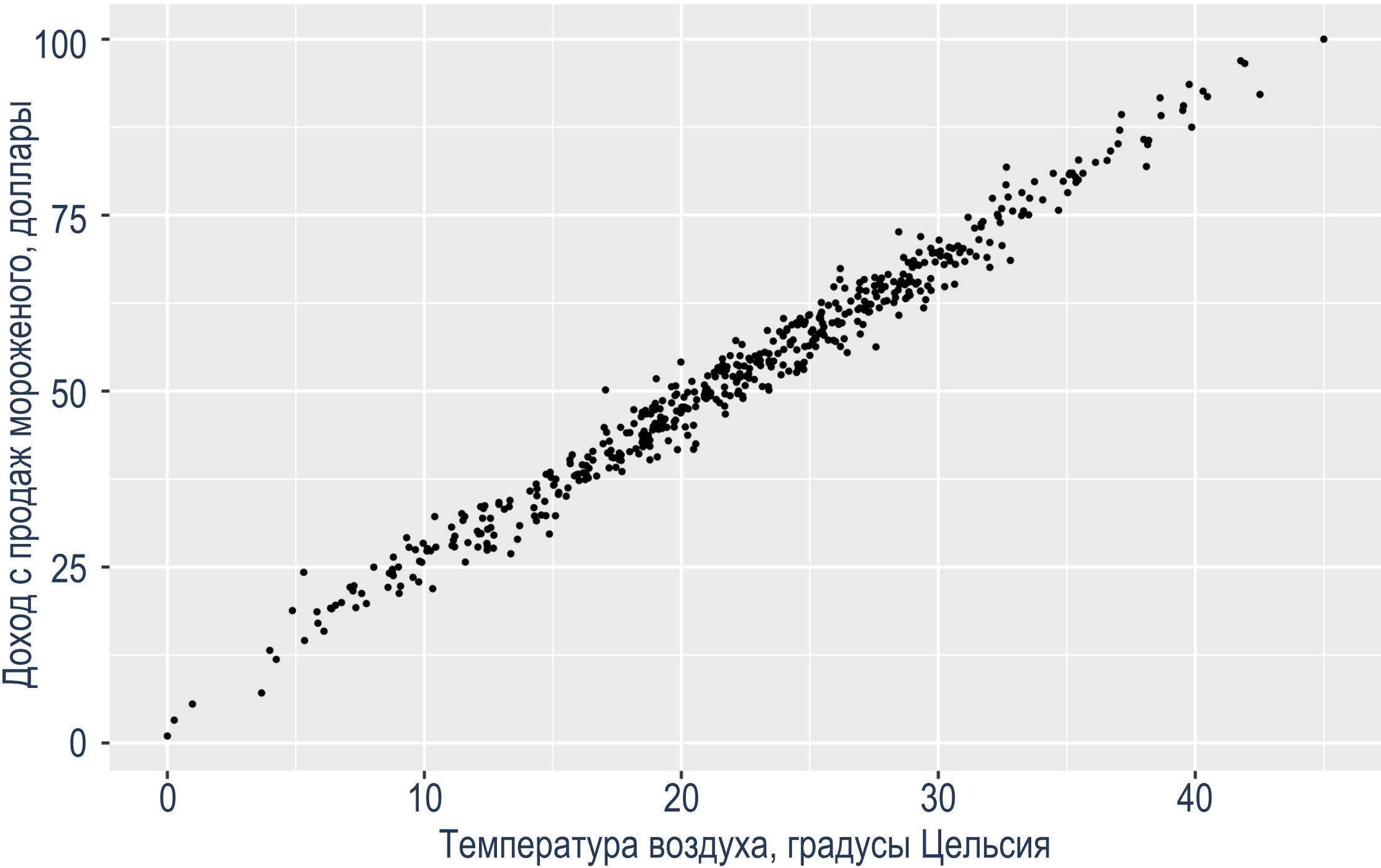
Исследователи используют данные, чтобы строить модели. **Модель** (model) – это некоторая формула, которая связывает зависимую (объясняемую) переменную с одной или несколькими независимыми (объясняющими) переменными.

Продажа мороженого в летние дни

Номер дня	Доход с продаж за день	Температура воздуха
День 1	53.48	24.57
День 2	62.52	26.01
...	...	...
День 500	65.57	28.96

Данные по доходам с продажи мороженого. *Источник: Kaggle*

График зависимости между температурой воздуха и доходом





ВИЗУАЛИЗАЦИЯ ДАННЫХ

Крайне удобный способ отображения наших данных – это точечный график или **диаграмма рассеяния** (scatter plot).

По горизонтальной оси отложены наблюдения, относящиеся к независимой переменной, по вертикальной – к зависимой.

Точки на графике – это пара объектов (независимая переменная и зависимая переменная) по каждому элементу выборки.



ВИЗУАЛИЗАЦИЯ ДАННЫХ

Крайне удобный способ отображения наших данных – это точечный график или **диаграмма рассеяния** (scatter plot).

По горизонтальной оси отложены наблюдения, относящиеся к независимой переменной, по вертикальной – к зависимой.

Точки на графике – это пара объектов (независимая переменная и зависимая переменная) по каждому элементу выборки.

Диаграмма рассеяния с положительной линейной связью
между двумя переменными

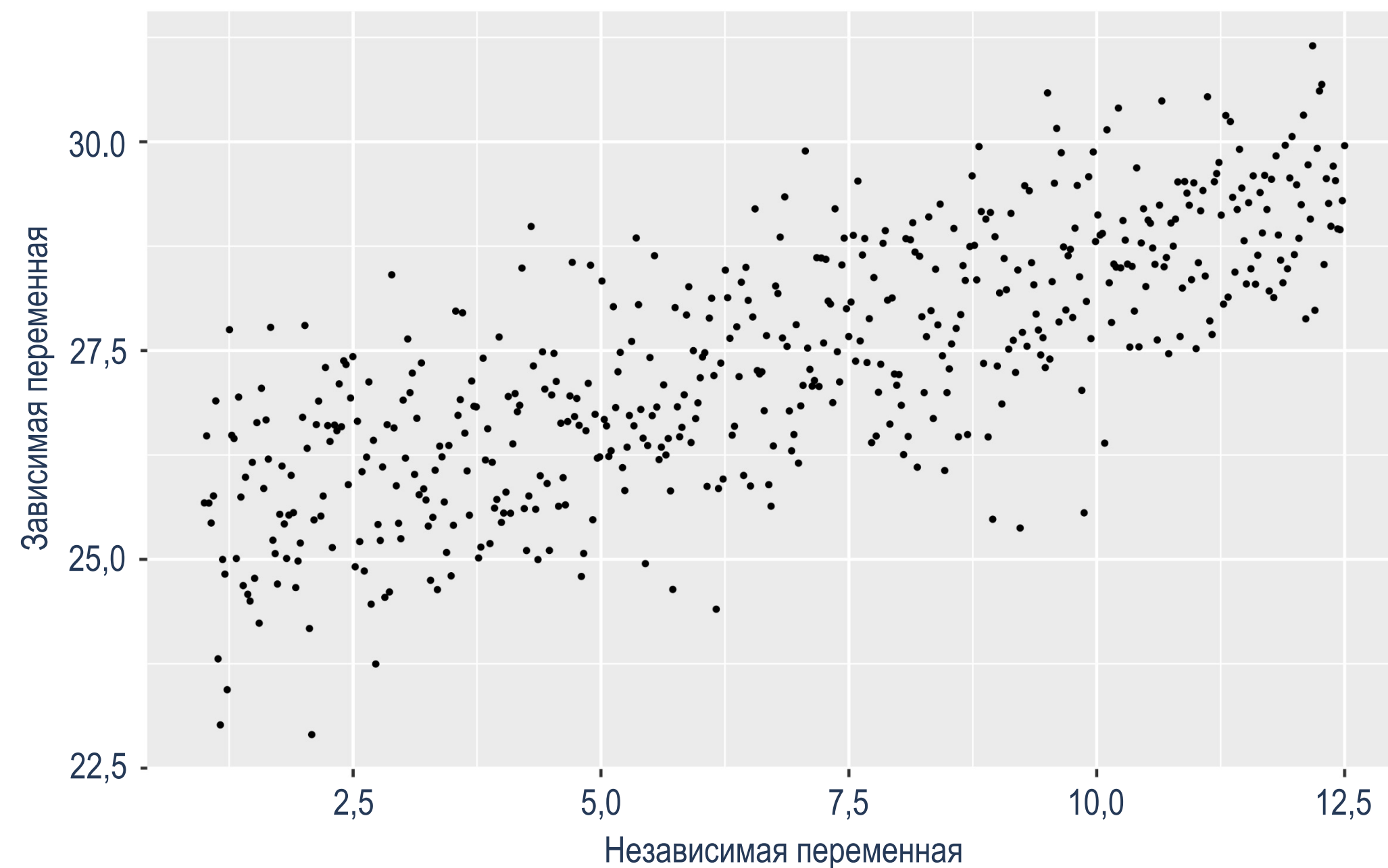
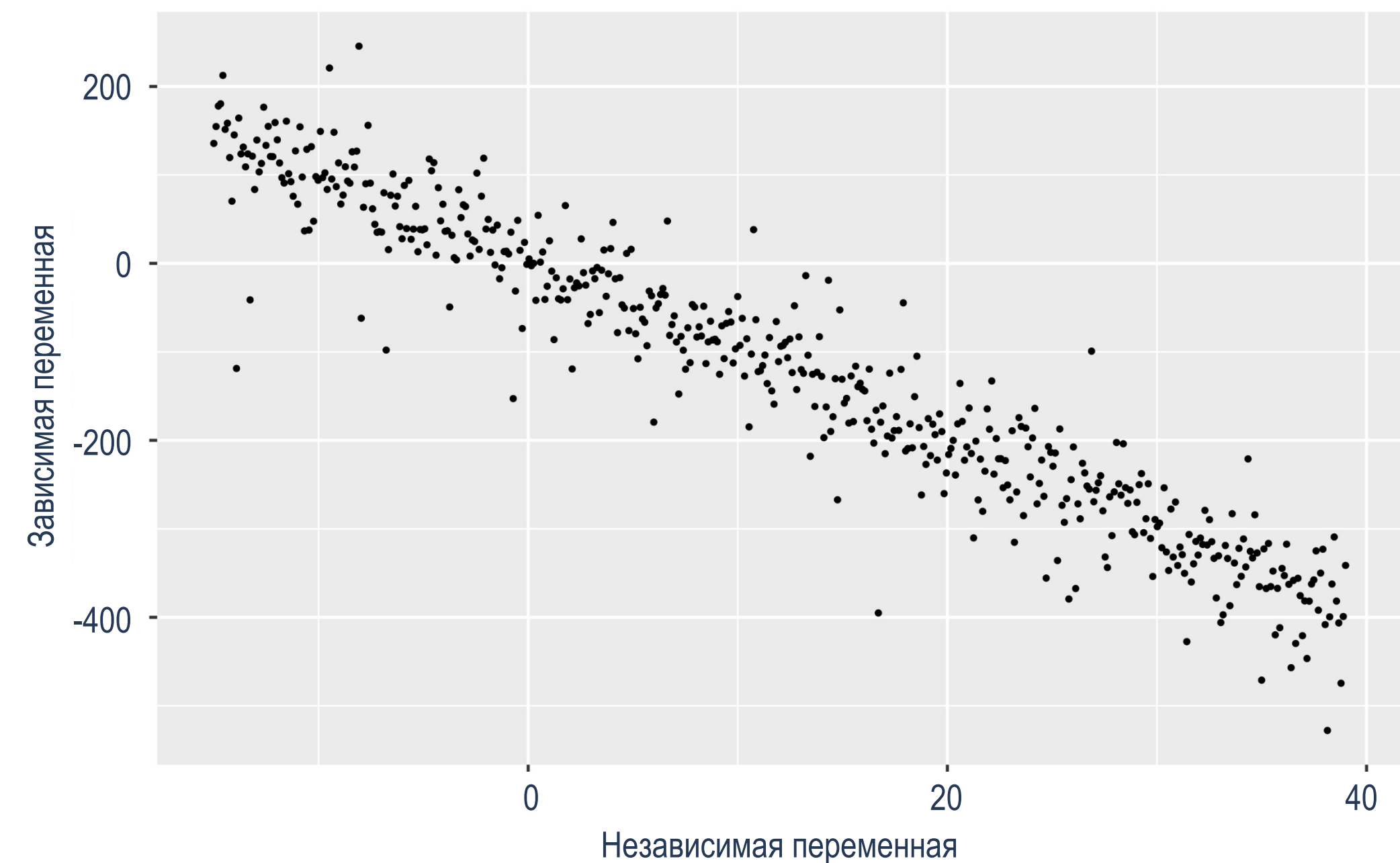


Диаграмма рассеяния с отрицательной линейной связью
между двумя переменными





ВИЗУАЛИЗАЦИЯ ДАННЫХ

Крайне удобный способ отображения наших данных – это точечный график или **диаграмма рассеяния** (scatter plot).

По горизонтальной оси отложены наблюдения, относящиеся к независимой переменной, по вертикальной – к зависимой.

Точки на графике – это пара объектов (независимая переменная и зависимая переменная) по каждому элементу выборки.

Диаграмма рассеяния с положительной линейной связью
между двумя переменными

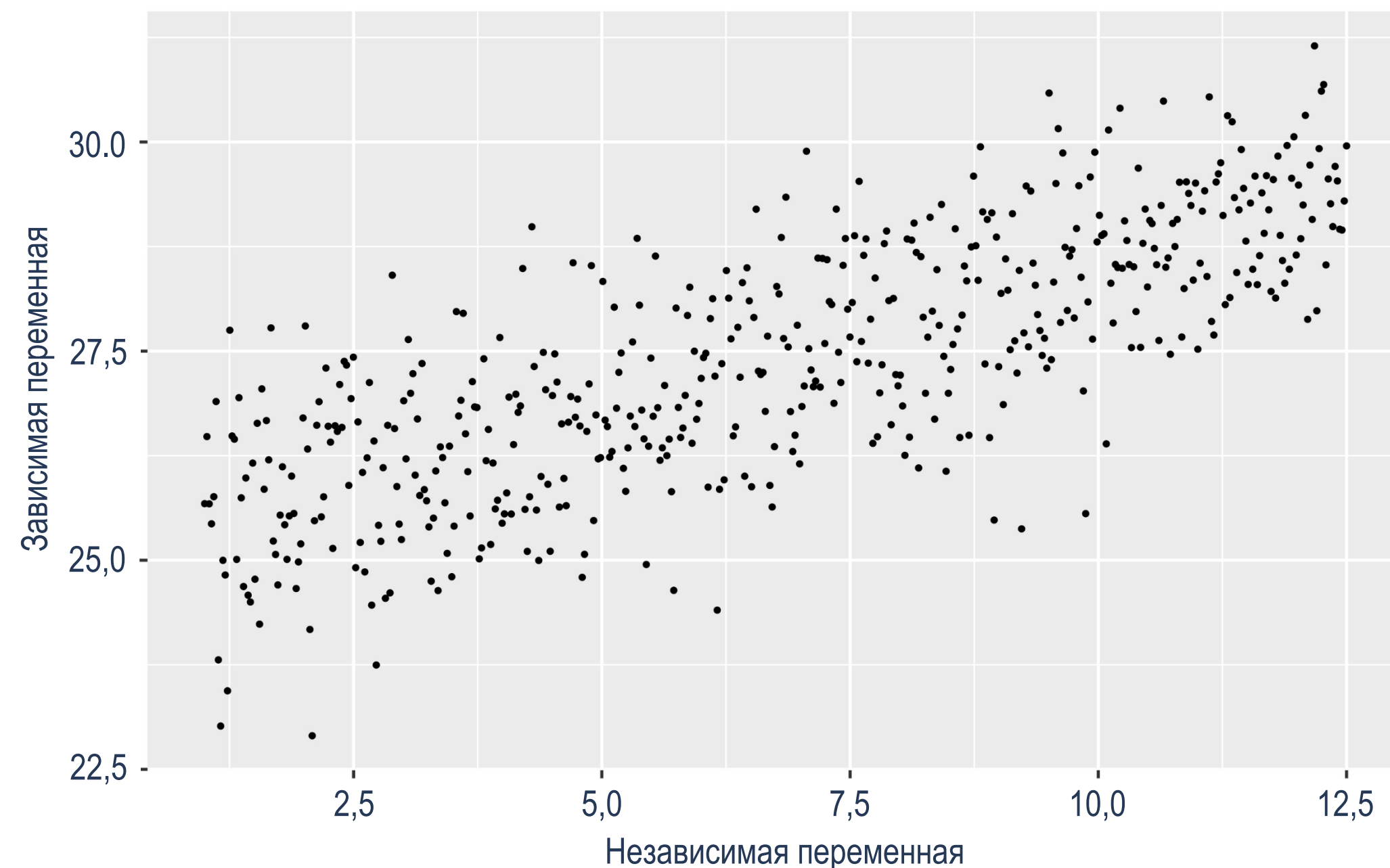
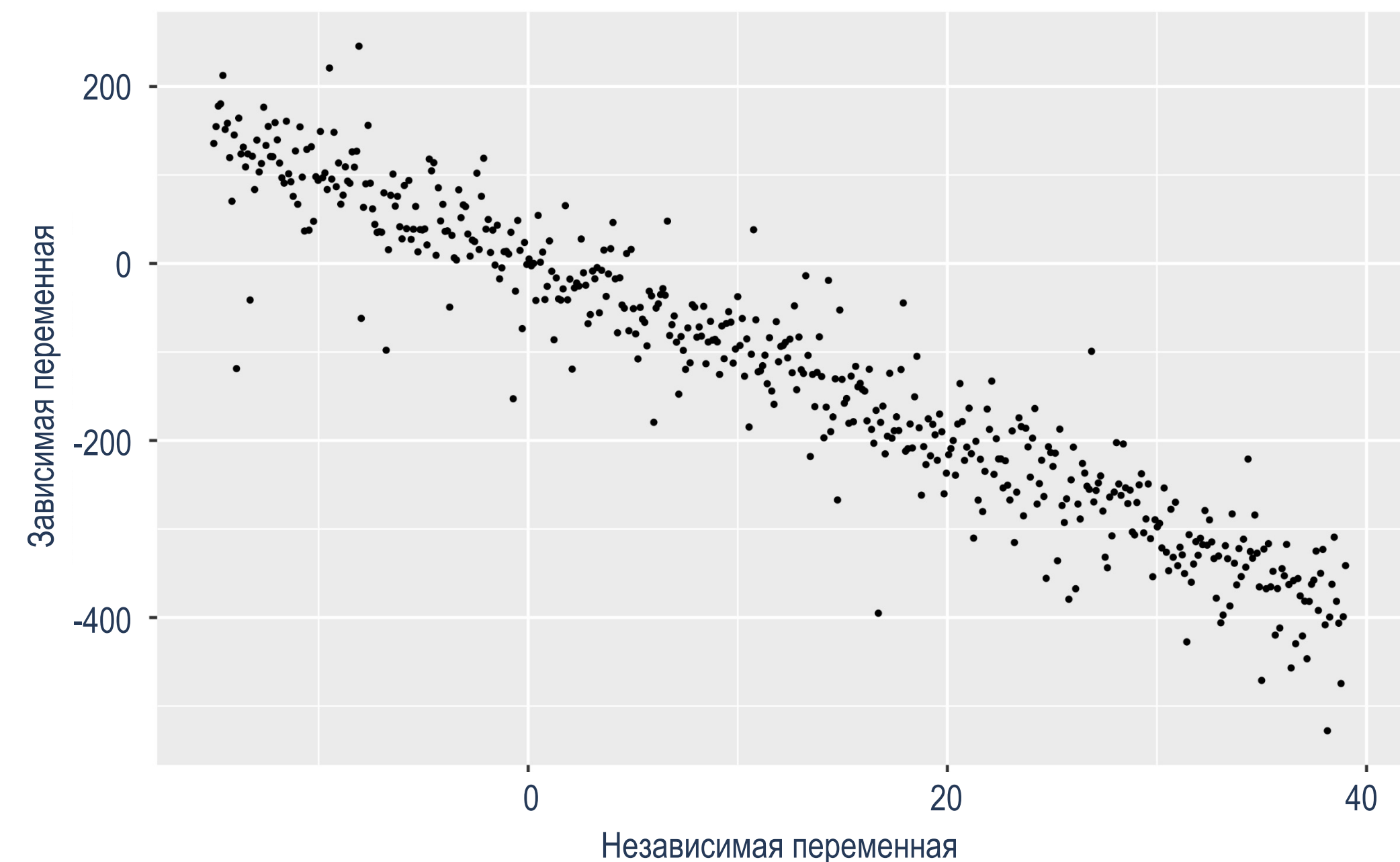


Диаграмма рассеяния с отрицательной линейной связью
между двумя переменными



Теперь мы готовы строить модель!



ЛИНЕЙНАЯ РЕГРЕССИОННАЯ МОДЕЛЬ



ЛИНЕЙНАЯ РЕГРЕССИОННАЯ МОДЕЛЬ

Линейная регрессионная модель

- Самая простая функция, которую только можно себе представить, – это линейная функция.

Предположим, что наши данные можно описать следующей формулой:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

- Такой простейший вид модели называется парной линейной регрессией или линейной регрессией с одной объясняющей переменной.



ЛИНЕЙНАЯ РЕГРЕССИОННАЯ МОДЕЛЬ

Линейная регрессионная модель

- Самая простая функция, которую только можно себе представить, – это линейная функция.

Предположим, что наши данные можно описать следующей формулой:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

- Такой простейший вид модели называется парной линейной регрессией или линейной регрессией с одной объясняющей переменной.
- Что это за красивая греческая буква ε (читается «эпсилон»)?
 ε – это такой набор факторов, который как-то влияет на y , но который мы не включили в модель.



ЛИНЕЙНАЯ РЕГРЕССИОННАЯ МОДЕЛЬ

Линейная регрессионная модель

- Самая простая функция, которую только можно себе представить, – это линейная функция.

Предположим, что наши данные можно описать следующей формулой:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

- Такой простейший вид модели называется парной линейной регрессией или линейной регрессией с одной объясняющей переменной.
- Что это за красивая греческая буква ε (читается «эпсилон»)?
 ε – это такой набор факторов, который как-то влияет на y , но который мы не включили в модель.
- β_0 показывает усредненный вклад в величину зависимой переменной y не включенных в модель переменных ε .
- β_1 позволяют нам интерпретировать влияние переменной x на переменную y : при увеличении переменной x на 1 у.е. переменная y увеличится на β_1 у.е.



ЛИНЕЙНАЯ РЕГРЕССИОННАЯ МОДЕЛЬ

Линейная регрессионная модель

- Самая простая функция, которую только можно себе представить, – это линейная функция.

Предположим, что наши данные можно описать следующей формулой:

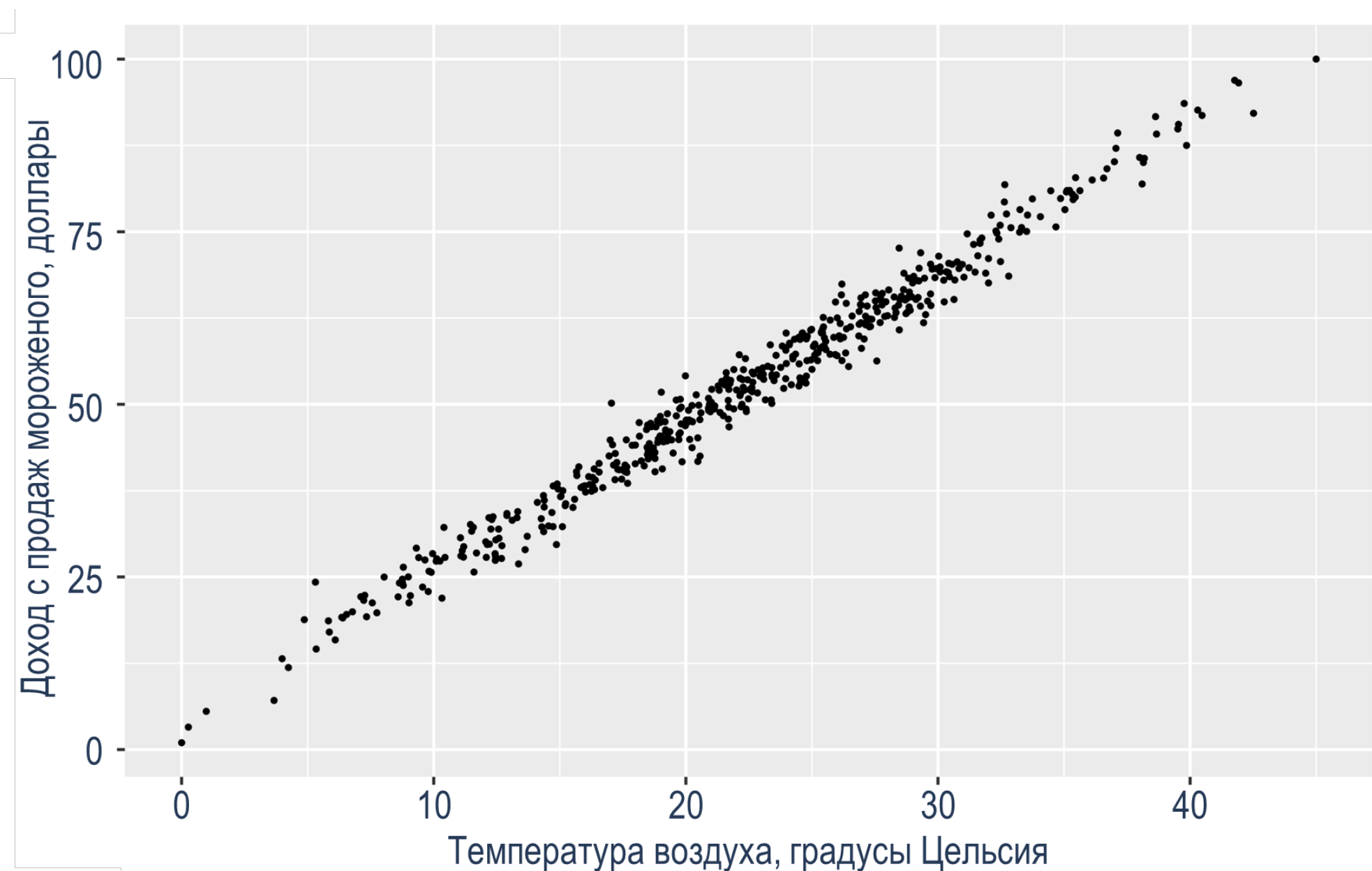
$$y = \beta_0 + \beta_1 x + \varepsilon.$$

- Такой простейший вид модели называется парной линейной регрессией или линейной регрессией с одной объясняющей переменной.
- Что это за красивая греческая буква ε (читается «эпсилон»)?
 ε – это такой набор факторов, который как-то влияет на y , но который мы не включили в модель.
- β_0 показывает усредненный вклад в величину зависимой переменной y не включенных в модель переменных ε .
- β_1 позволяют нам интерпретировать влияние переменной x на переменную y : при увеличении переменной x на 1 у.е. переменная y увеличится на β_1 у.е.
- Линейную модель можно также использовать для прогнозирования!



ЛИНЕЙНАЯ РЕГРЕССИОННАЯ МОДЕЛЬ

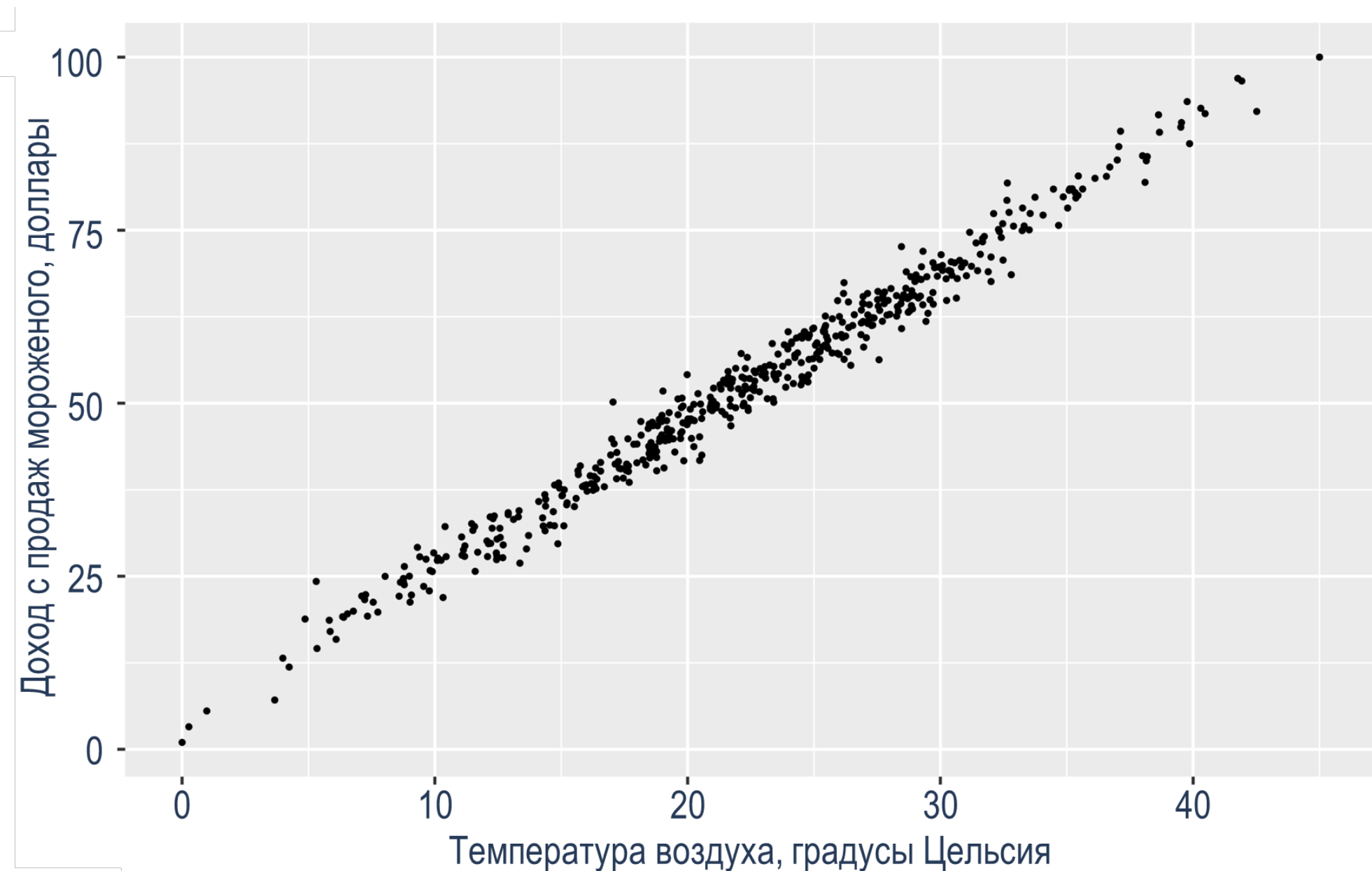
Так что там с мороженым?



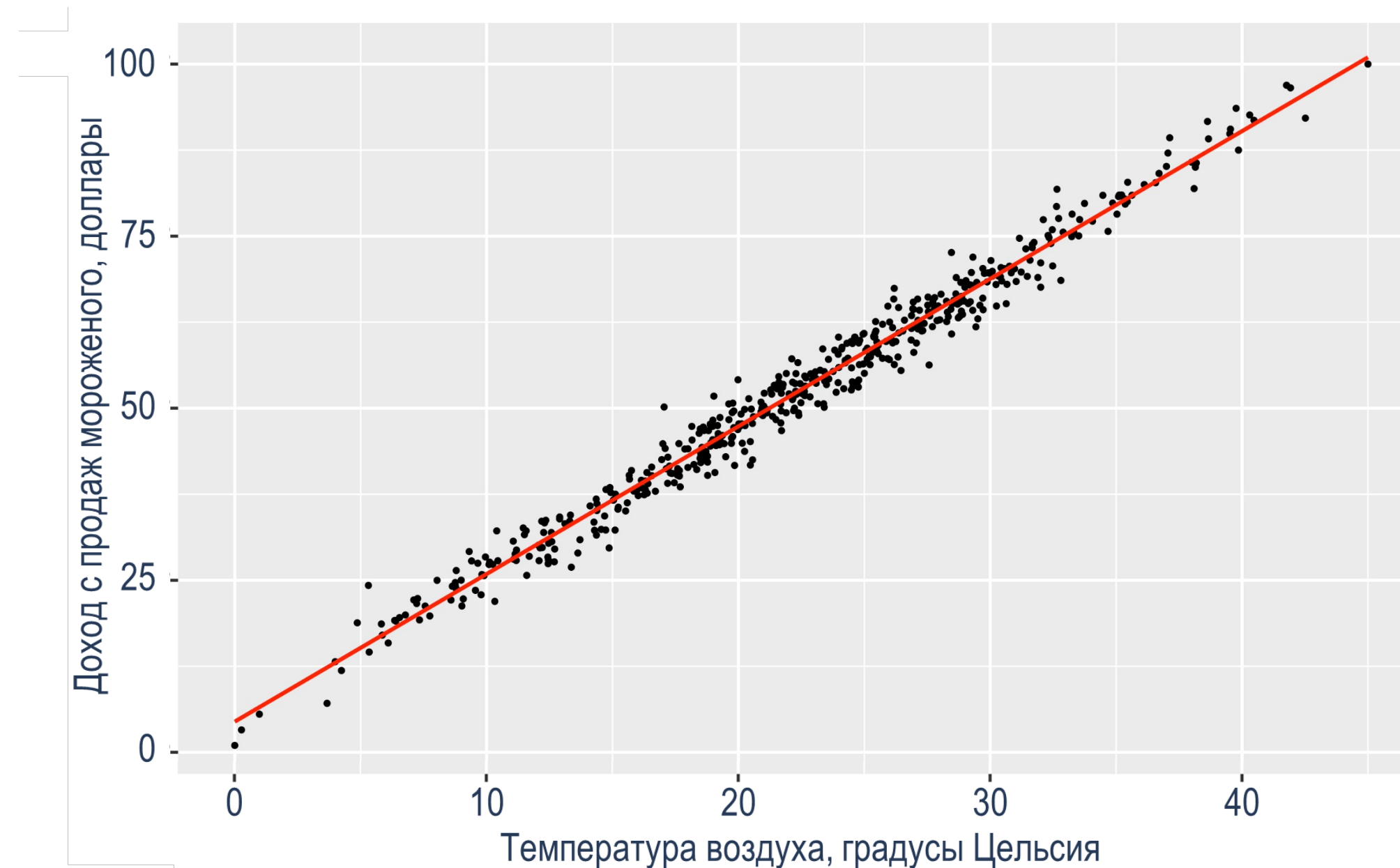


ЛИНЕЙНАЯ РЕГРЕССИОННАЯ МОДЕЛЬ

Так что там с мороженым?



линия
регрессии



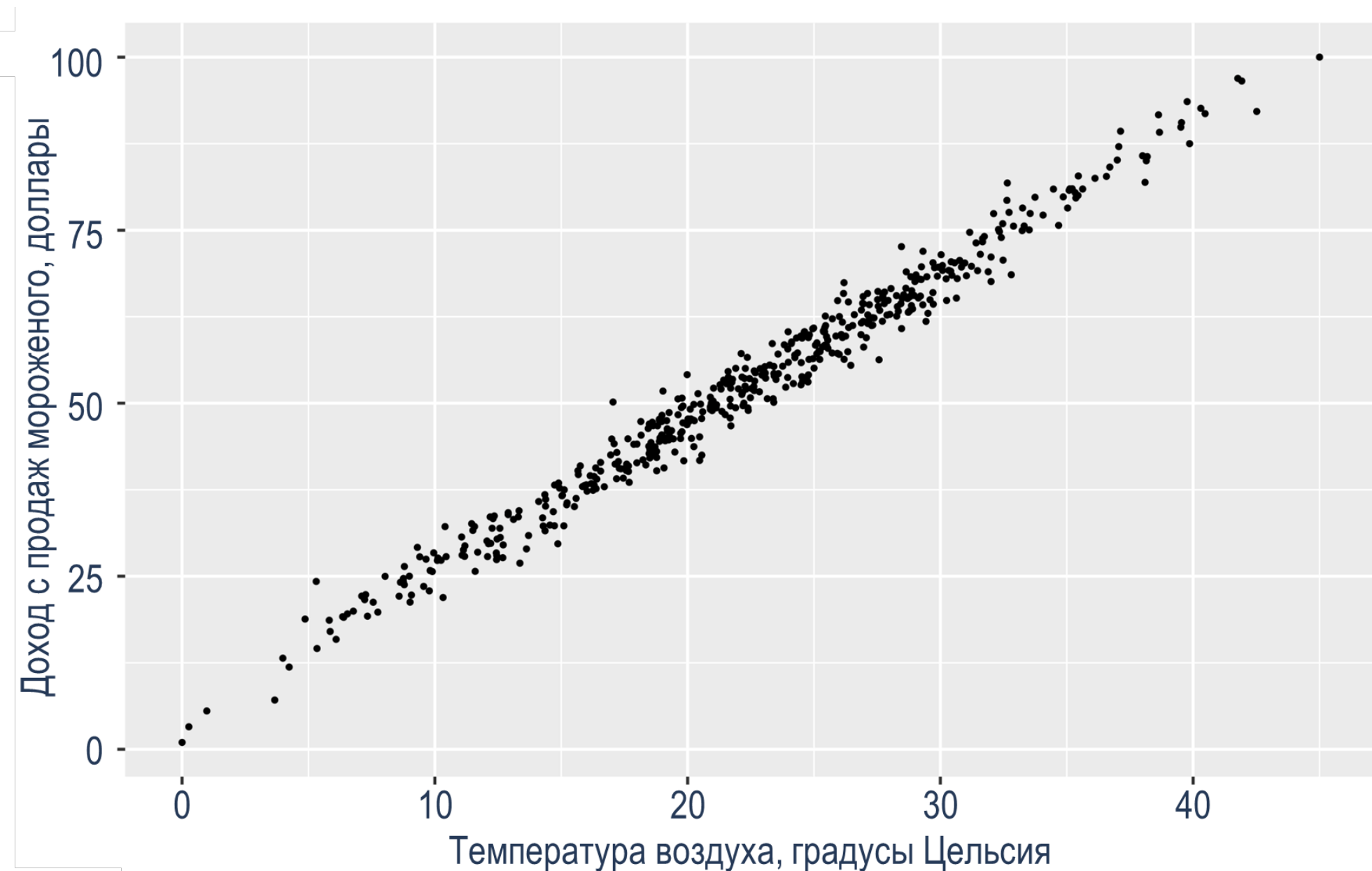
В данном примере уравнение регрессии имеет следующий вид:

$$income = 44.83 + 22.44 \cdot temperature .$$

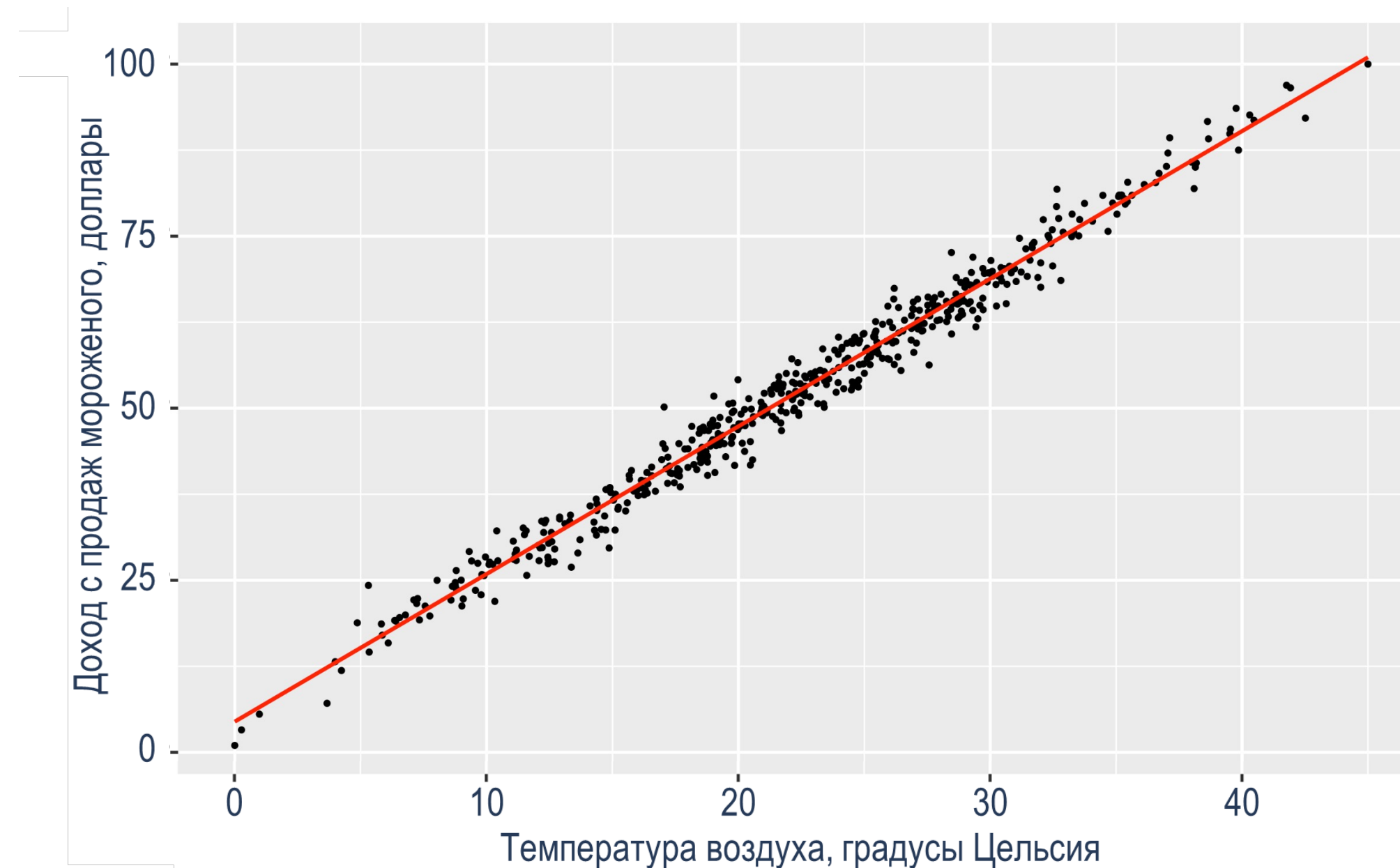


ЛИНЕЙНАЯ РЕГРЕССИОННАЯ МОДЕЛЬ

Так что там с мороженым?



линия
регрессии



В данном примере уравнение регрессии имеет следующий вид:

$$income = 44.83 + 22.44 \cdot temperature .$$

Как мы построили прямую?



МЕТОД НАИМЕНЬШИХ КВАДРАТОВ



МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Метод, при помощи которого мы получили уравнение прямой (которая, кстати, называется линией тренда), именуется **методом наименьших квадратов** или **МНК** (least squares).

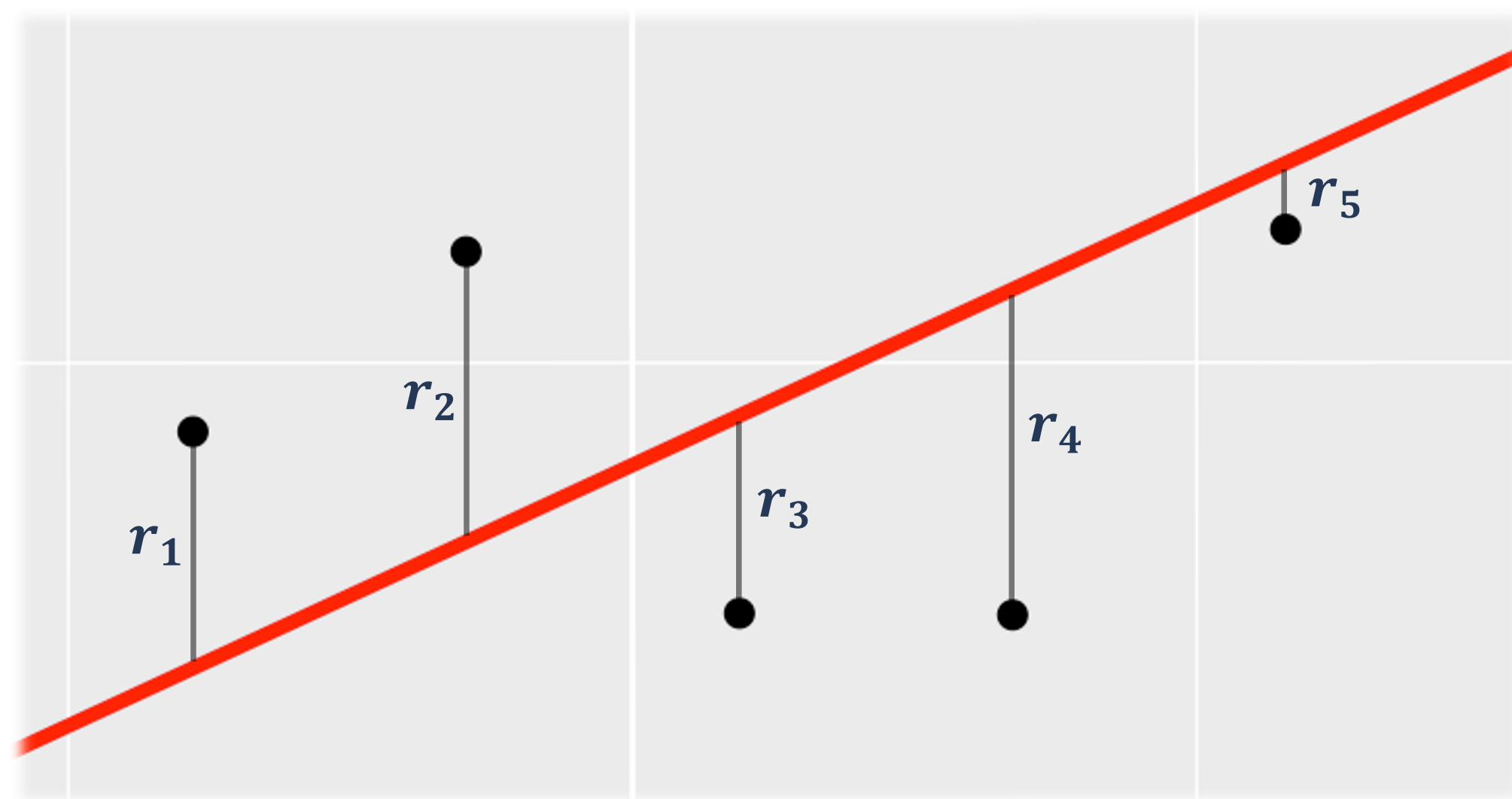
Идея метода: провести прямую линию через набор точек так, чтобы суммарное отклонение от этой линии до каждой из точек было минимальным. Но не все так просто.



МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Метод, при помощи которого мы получили уравнение прямой (которая, кстати, называется линией тренда), именуется **методом наименьших квадратов** или **МНК** (least squares).

Идея метода: провести прямую линию через набор точек так, чтобы суммарное отклонение от этой линии до каждой из точек было минимальным. Но не все так просто.

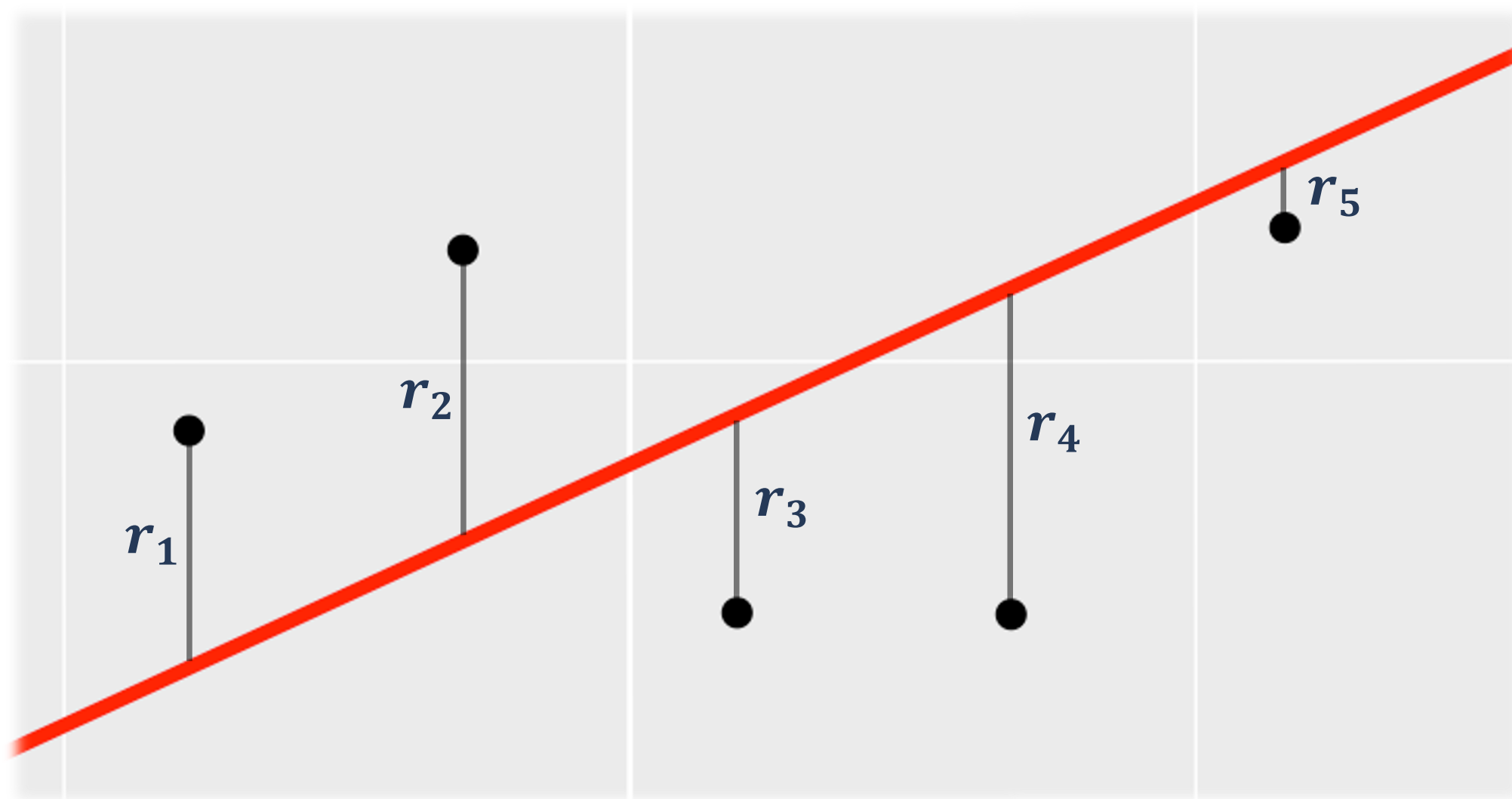




МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Метод, при помощи которого мы получили уравнение прямой (которая, кстати, называется линией тренда), именуется **методом наименьших квадратов** или **МНК** (least squares).

Идея метода: провести прямую линию через набор точек так, чтобы суммарное отклонение от этой линии до каждой из точек было минимальным. Но не все так просто.



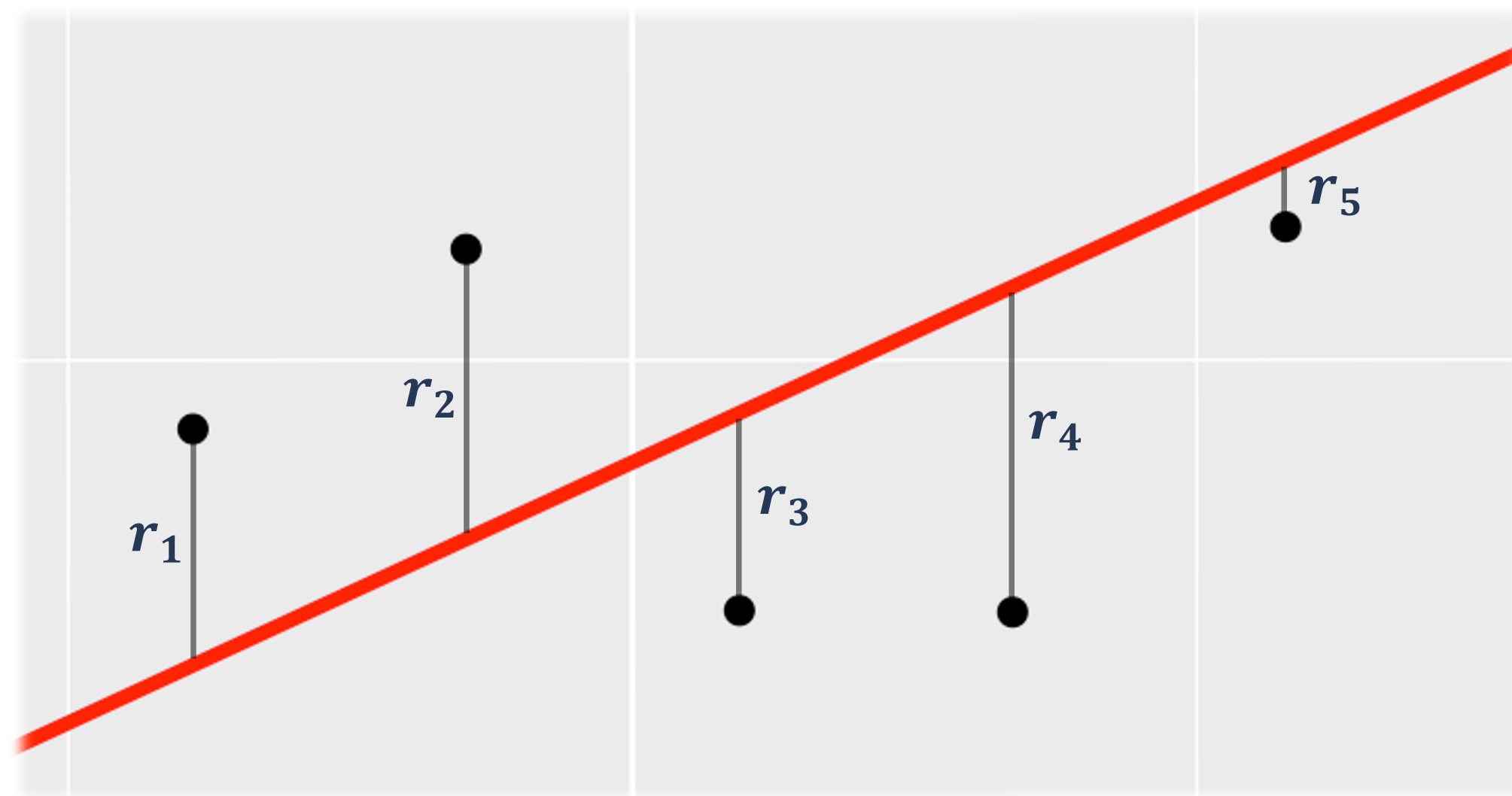
- Минимизация суммы $r_1 + r_2 + r_3 + r_4 + r_5$ — это плохой вариант.



МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Метод, при помощи которого мы получили уравнение прямой (которая, кстати, называется линией тренда), именуется **методом наименьших квадратов** или **МНК** (least squares).

Идея метода: провести прямую линию через набор точек так, чтобы суммарное отклонение от этой линии до каждой из точек было минимальным. Но не все так просто.



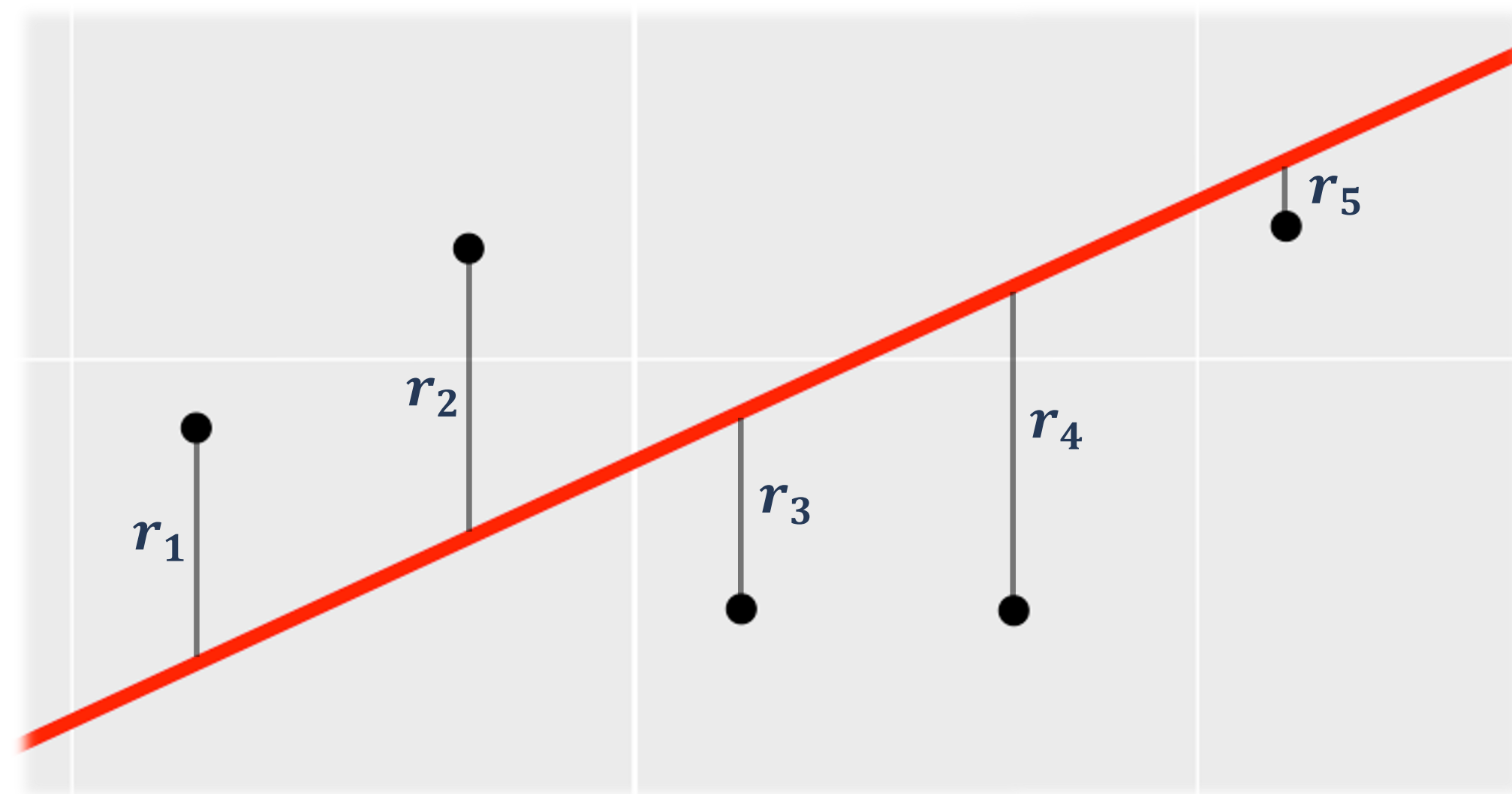
- Минимизация суммы $r_1 + r_2 + r_3 + r_4 + r_5$ — это плохой вариант.
- Минимизация суммы $|r_1| + |r_2| + |r_3| + |r_4| + |r_5|$ — это хороший с точки зрения здравого смысла вариант, но плохой с точки зрения математики.



МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Метод, при помощи которого мы получили уравнение прямой (которая, кстати, называется линией тренда), именуется **методом наименьших квадратов** или **МНК** (least squares).

Идея метода: провести прямую линию через набор точек так, чтобы суммарное отклонение от этой линии до каждой из точек было минимальным. Но не все так просто.



- Минимизация суммы $r_1 + r_2 + r_3 + r_4 + r_5$ — это плохой вариант.
- Минимизация суммы $|r_1| + |r_2| + |r_3| + |r_4| + |r_5|$ — это хороший с точки зрения здравого смысла вариант, но плохой с точки зрения математики.
- Минимизация суммы $r_1^2 + r_2^2 + r_3^2 + r_4^2 + r_5^2$ — это отличный вариант!



ПРИМЕР НА РЕАЛЬНЫХ ДАННЫХ



ПРИМЕР НА РЕАЛЬНЫХ ДАННЫХ

Данные представляют собой выборку домов в одном из штатов США в 2018 году (Kaggle). Данные содержат большое количество переменных, включая цену дома, расстояния до ближайшей автобусной станции, информацию о соседях и так далее.

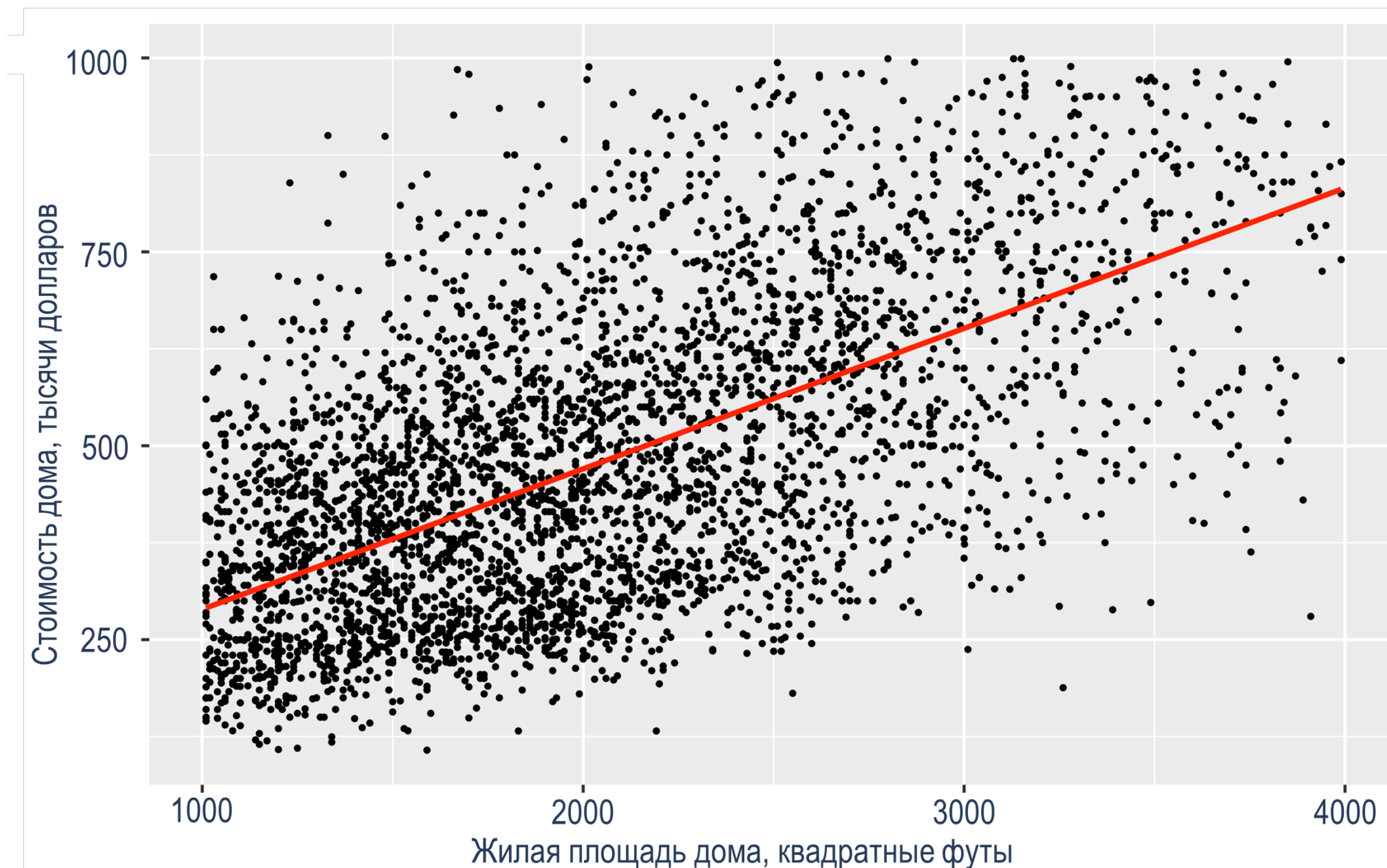


ПРИМЕР НА РЕАЛЬНЫХ ДАННЫХ

Данные представляют собой выборку домов в одном из штатов США в 2018 году (Kaggle). Данные содержат большое количество переменных, включая цену дома, расстояния до ближайшей автобусной станции, информацию о соседях и так далее.

Рассмотрим зависимость между стоимостью дома и жилой площадью дома.

График зависимости между температурой воздуха и доходом с линией тренда





ПРИМЕР НА РЕАЛЬНЫХ ДАННЫХ

В данном примере в соответствии с МНК уравнение прямой имеет вид:

$$price = 144.20 + 0.17 \cdot living_square.$$

Какую информацию мы можем отсюда достать?

График зависимости между температурой воздуха и доходом с линией тренда





ПРИМЕР НА РЕАЛЬНЫХ ДАННЫХ

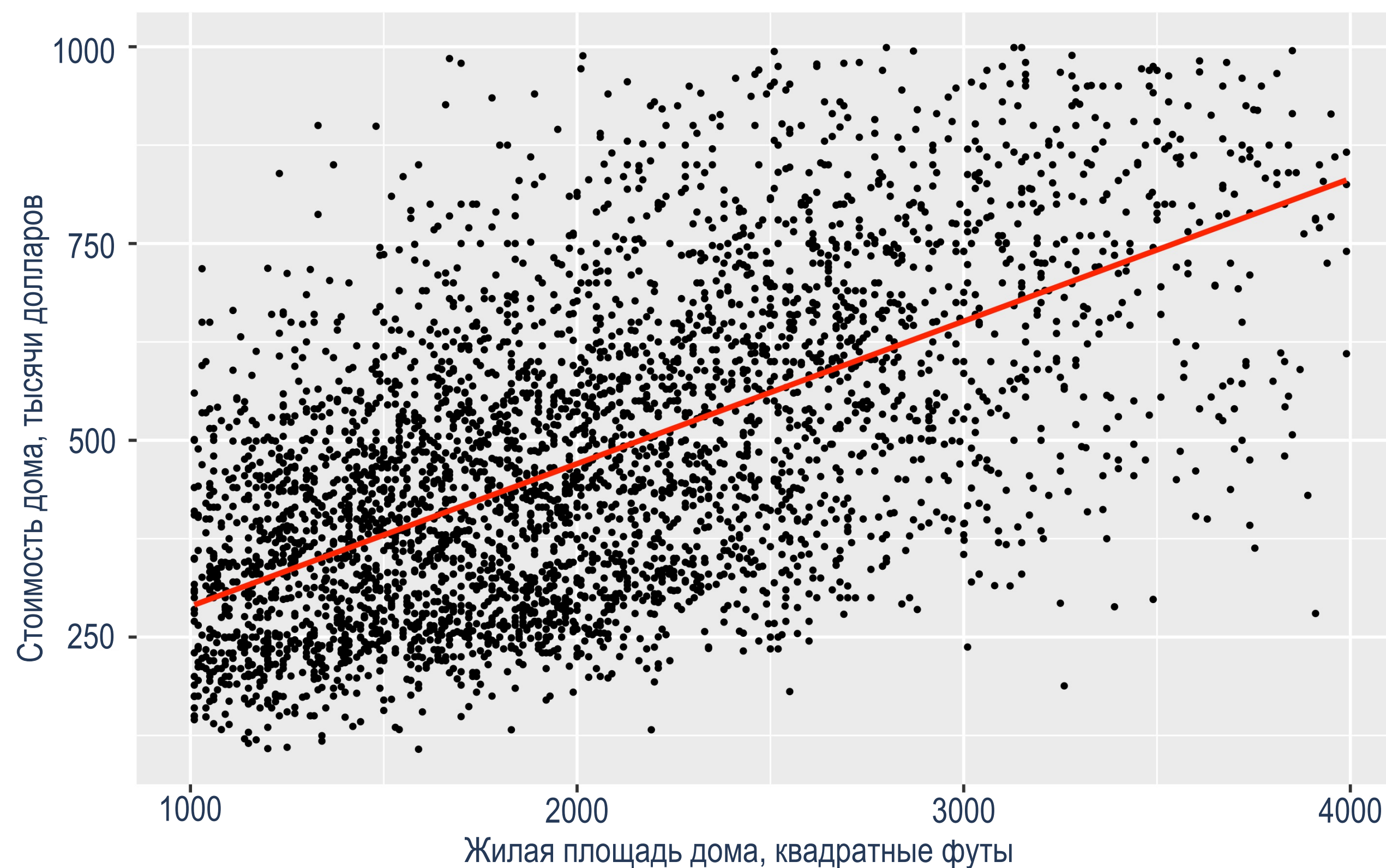
В данном примере в соответствии с МНК уравнение прямой имеет вид:

$$price = 144.20 + 0.17 \cdot living_square.$$

Какую информацию мы можем отсюда достать?

- Коэффициент β_1 равен **0.17**, что говорит о положительном влиянии жилой площади на цену дома: при увеличении жилой площади на 1 квадратный фут цена вырастет в среднем на 170 долларов.

График зависимости между температурой воздуха и доходом с линией тренда





ПРИМЕР НА РЕАЛЬНЫХ ДАННЫХ

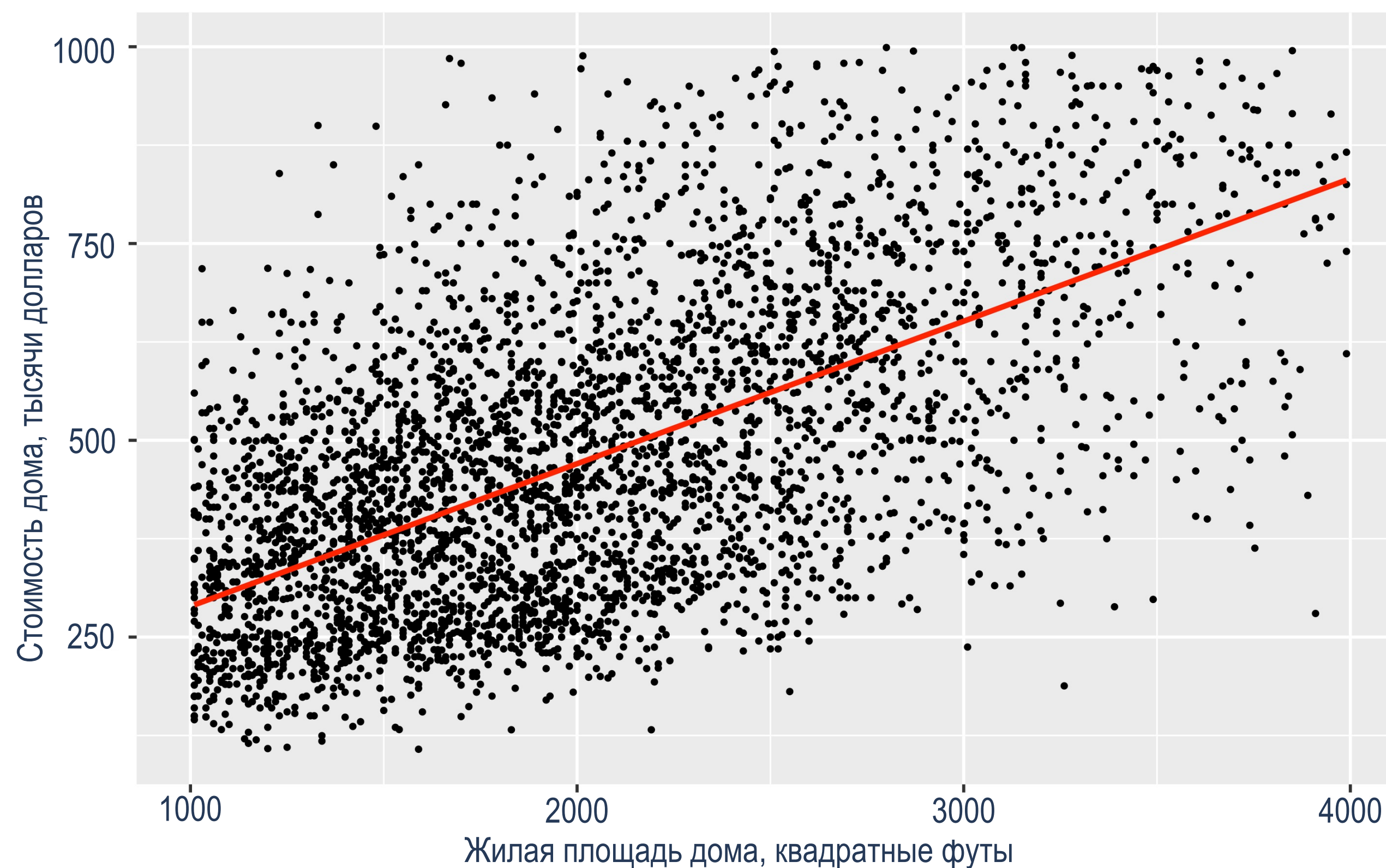
В данном примере в соответствии с МНК уравнение прямой имеет вид:

$$price = 144.20 + 0.17 \cdot living_square.$$

Какую информацию мы можем отсюда достать?

- Коэффициент β_1 равен **0.17**, что говорит о положительном влиянии жилой площади на цену дома: при увеличении жилой площади на 1 квадратный фут цена вырастет в среднем на 170 долларов.
- Средний вклад не включенных в модель переменных в стоимость дома равен 144.20.

График зависимости между температурой воздуха и доходом с линией тренда





ПРИМЕР НА РЕАЛЬНЫХ ДАННЫХ

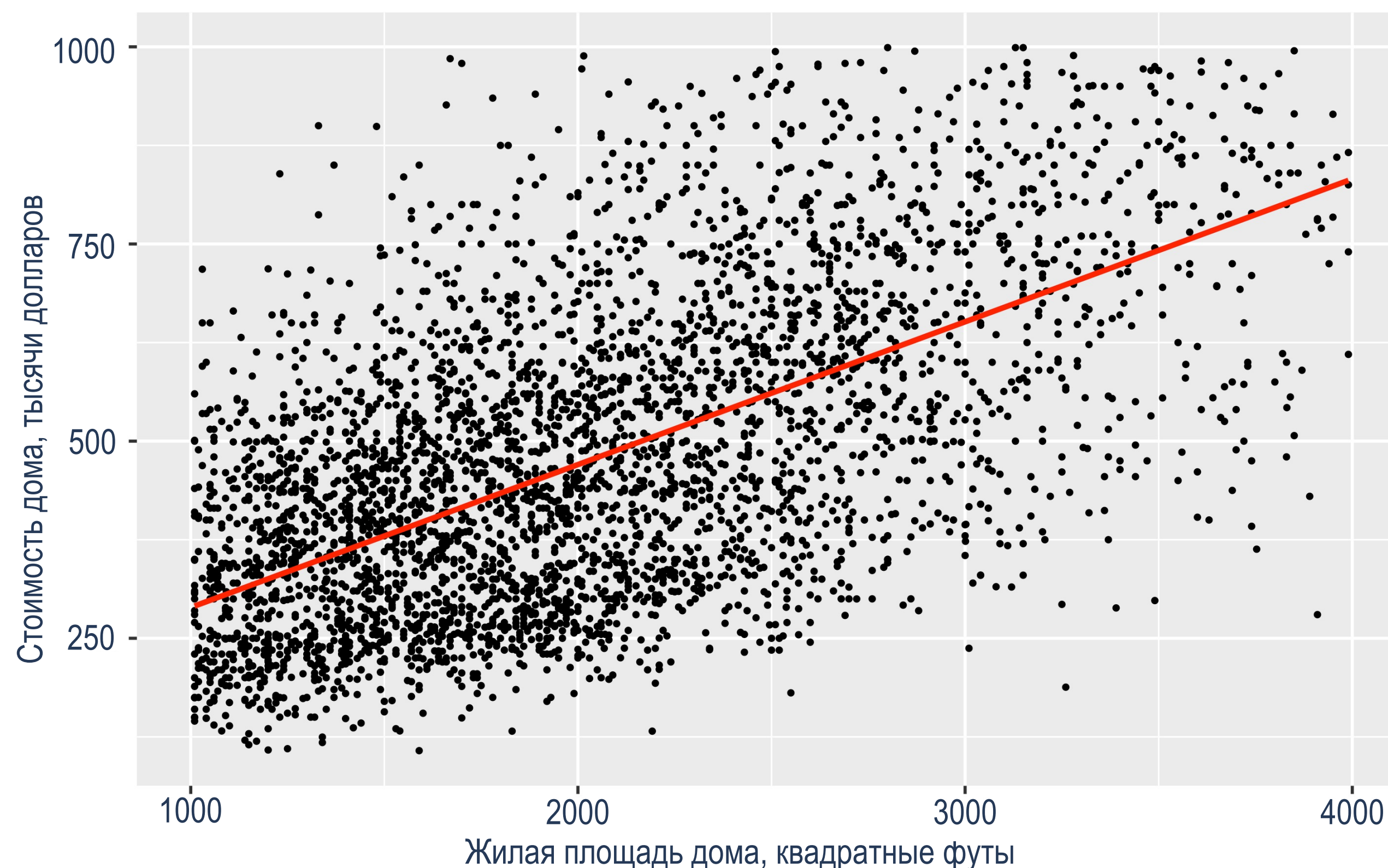
В данном примере в соответствии с МНК уравнение прямой имеет вид:

$$price = 144.20 + 0.17 \cdot living_square.$$

Какую информацию мы можем отсюда достать?

- Коэффициент β_1 равен **0.17**, что говорит о положительном влиянии жилой площади на цену дома: при увеличении жилой площади на 1 квадратный фут цена вырастет в среднем на 170 долларов.
- Средний вклад не включенных в модель переменных в стоимость дома равен 144.20.
- Прогноз цены дома. Илон Маск присматривает новый дом площадью 398.26 квадратных футов. Наша модель утверждает, что цена такого дома составит $144.20 + 0.17 \cdot 398.26 = 211.90$.
- Вывод: Илон Маск сделал правильный выбор :)

График зависимости между температурой воздуха и доходом с линией тренда





НЕУЖЕЛИ ВСЕ ТАК ПРОСТО?



НЕУЖЕЛИ ВСЕ ТАК ПРОСТО?

Линейная регрессионная модель кажется весьма простым инструментом. Так ли это?



НЕУЖЕЛИ ВСЕ ТАК ПРОСТО?

Линейная регрессионная модель кажется весьма простым инструментом. Так ли это?

С математической точки зрения, регрессия – это достаточно сложный инструмент.

При работе с регрессией исследователь сталкивается с большим количеством проблем:



НЕУЖЕЛИ ВСЕ ТАК ПРОСТО?

Линейная регрессионная модель кажется весьма простым инструментом. Так ли это?

С математической точки зрения, регрессия – это достаточно сложный инструмент.

При работе с регрессией исследователь сталкивается с большим количеством проблем:

- большой выбор методов, при помощи которых можно получить значения коэффициентов регрессионного уравнения



НЕУЖЕЛИ ВСЕ ТАК ПРОСТО?

Линейная регрессионная модель кажется весьма простым инструментом. Так ли это?

С математической точки зрения, регрессия – это достаточно сложный инструмент.

При работе с регрессией исследователь сталкивается с большим количеством проблем:

- большой выбор методов, при помощи которых можно получить значения коэффициентов регрессионного уравнения
- грамотный выбор функциональной формы модели



НЕУЖЕЛИ ВСЕ ТАК ПРОСТО?

Линейная регрессионная модель кажется весьма простым инструментом. Так ли это?

С математической точки зрения, регрессия – это достаточно сложный инструмент.

При работе с регрессией исследователь сталкивается с большим количеством проблем:

- большой выбор методов, при помощи которых можно получить значения коэффициентов регрессионного уравнения
- грамотный выбор функциональной формы модели
- пагубное влияние на результаты исследования не включенных в модель переменных



НЕУЖЕЛИ ВСЕ ТАК ПРОСТО?

Линейная регрессионная модель кажется весьма простым инструментом. Так ли это?

С математической точки зрения, регрессия – это достаточно сложный инструмент.

При работе с регрессией исследователь сталкивается с большим количеством проблем:

- большой выбор методов, при помощи которых можно получить значения коэффициентов регрессионного уравнения
- грамотный выбор функциональной формы модели
- пагубное влияние на результаты исследования не включенных в модель переменных

↓ ↓ ↓
ЭКОНОМЕТРИКА

СПАСИБО ЗА ВНИМАНИЕ!



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Электронная почта для связи: islabolitskiy@hse.ru