Корреляция и регрессия

Храмов Д.А.

27.01.2020

Содержание

- Связь между двумя переменными. Корреляция.
- Связанность и причинно-следственная связь.
- Постановка задачи регрессии.
- Простая линейная регрессия.
- Построение прогностической модели.
- Оценка точности модели.

Ковариация

Ковариация (*covariance = co + variance* — совместное изменение) — это мера того, как изменения одной переменной связаны с изменениями другой переменной.

Для пары случайных величин (X, Y), принимающих дискретные значения (x_i, y_i) $(i = 1, 2, \ldots, n)$ ковариация равна

$$\operatorname{cov}(X, Y) = \frac{1}{n^2} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$\bar{x}$, $\bar{y}$ — средние значения X и Y.

Величина $(x_i - \bar{x})(y_i - \bar{y})$ положительна тогда и только тогда, когда x_i и y_i лежат на одной стороне от соответствующих средних.

То есть: ковариация положительна, если x_i и y_i имеют тенденцию быть одновременно больше или одновременно меньше, чем их соответствующие средние значения. Если x_i и y_i находятся на противоположных сторонах от соответствующих средних значений, то ковариация отрицательна.

Но: на ковариацию влияет величина разброса (variation).

```
x = [1, 2, 3]
y = [4, 6, 10]

cov(x,y) = 2
```

Умножим обе переменные на 10. “Сила связи” между ними не изменится, а ковариация вырастет

```
x = [10, 20, 30]
y = [40, 60, 100]

cov(x,y) = 200
```

Коэффициент корреляции Пирсона

Нормируем ковариацию и получим

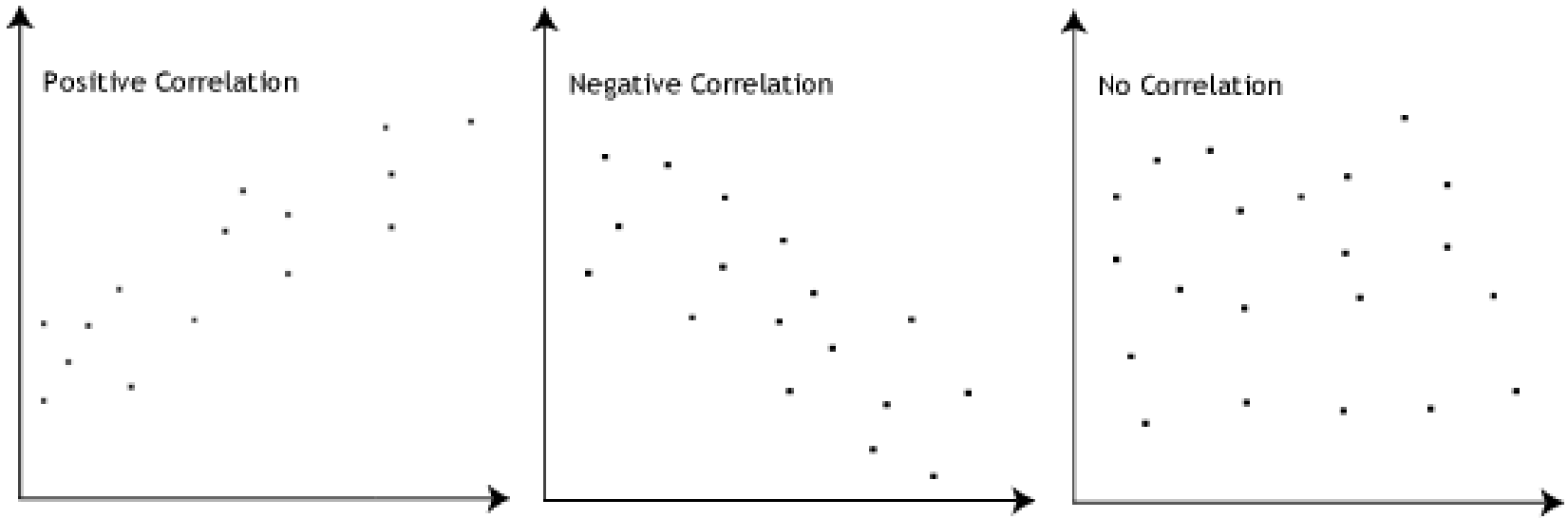
$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}},$$

— коэффициент корреляции Пирсона, который изменяется в пределах от -1 до 1.

Чем сильнее тенденция x_i и y_i быть одновременно больше или одновременно меньше соответствующих средних значений, тем больше абсолютное значение коэффициента корреляции.

Коэффициент корреляции Пирсона r распространен весьма широко. Часто его называют просто “коэффициентом корреляции” или даже “корреляцией”, хотя есть и другие подобные коэффициенты.

Направления связи: положительная, отрицательная, нулевая



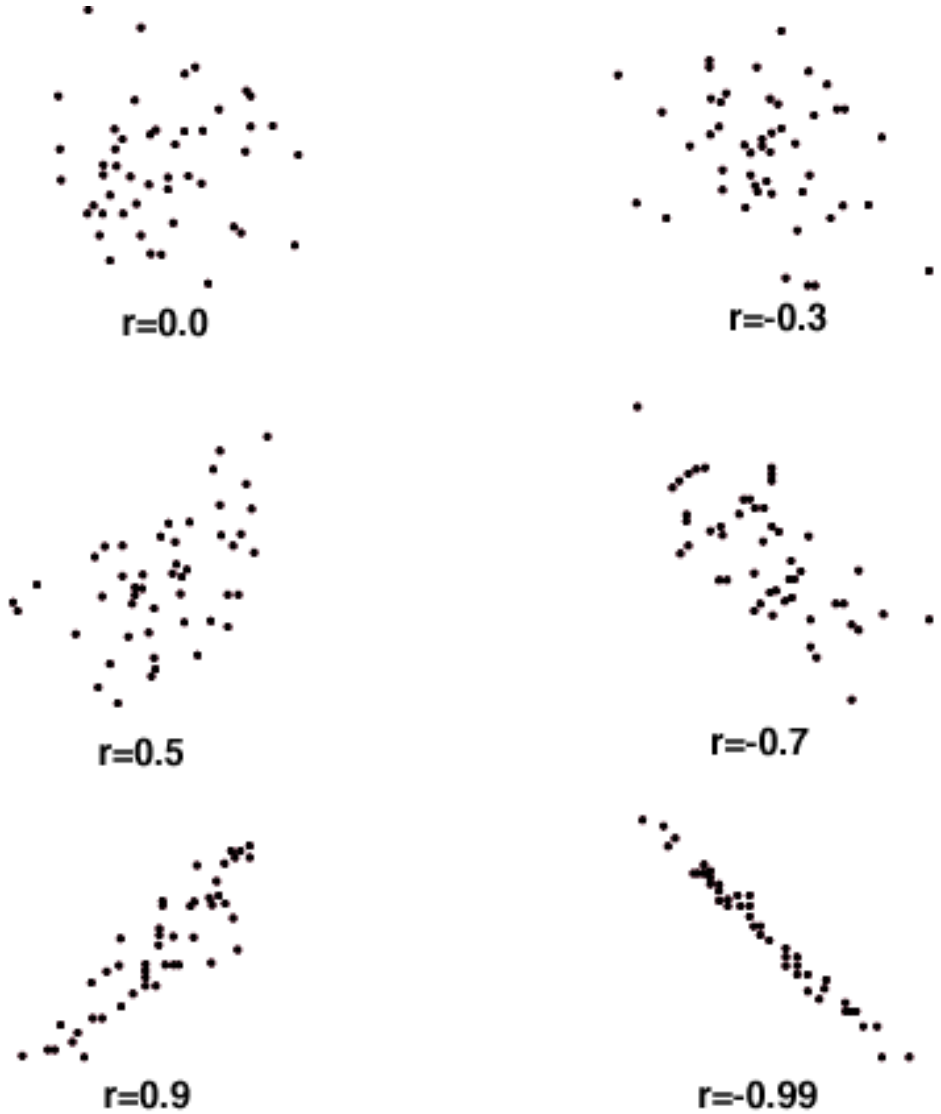
Источник:
https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

Сила корреляции

Интервал значений r	Интерпретация
0 – 0,2	Очень слабая корреляция
0,2 - 0,5	Слабая корреляция
0,5 – 0,7	Средняя корреляция
0,7 – 0,9	Высокая корреляция
0,9 - 1	Очень высокая корреляция

Зависит от области исследований. Следите за литературой по теме.

Как это выглядит



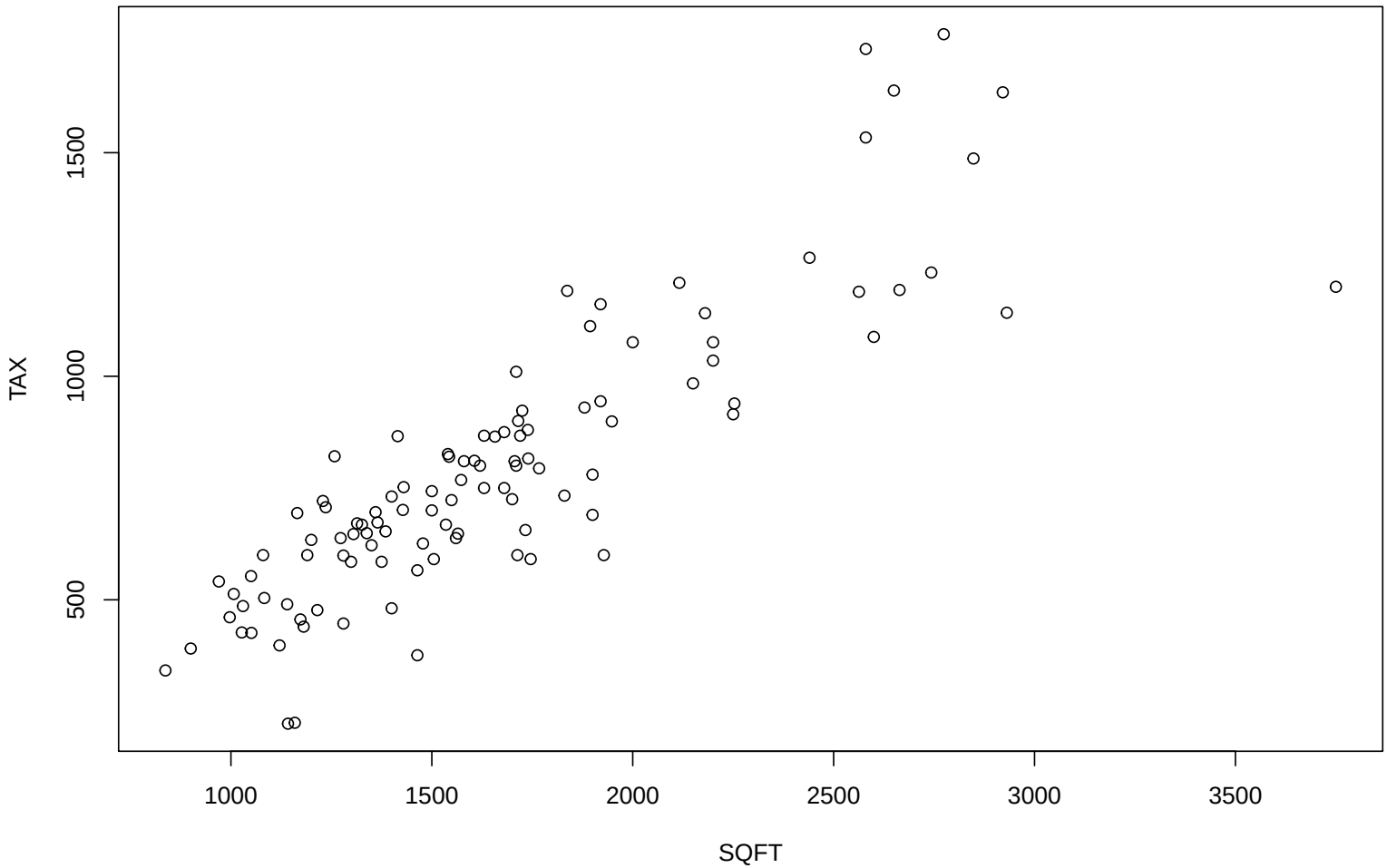
Диаграммы рассеяния при различных значениях коэффициента корреляции r Пирсона

Пример: недвижимость в Альбукерке

Строим зависимость налогов (TAX) от площади дома (SQFT)

```
x <- read.table("Albuquerque_Home_Prices_data.txt",
                header=T, na.strings="-9999")
# Чтобы не писать каждый раз 'x'
attach(x)
# Рассмотрим зависимость налогов от площади дома
plot(SQFT,TAX)
# Вычислим корреляцию между величинами
(r = cor(SQFT,TAX))
```

Зависимость налогов (TAX) от площади дома (SQFT)




```
## [1] 0.8585828
```

Сила линейной связи

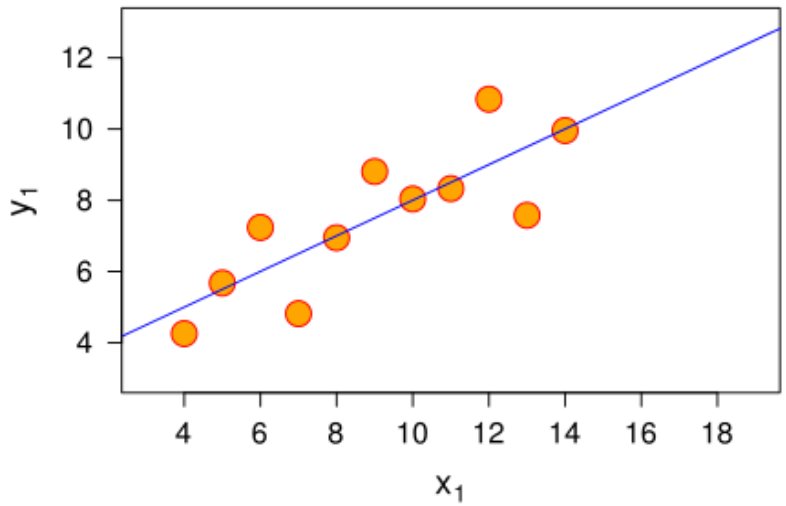
Коэффициент корреляции Пирсона отражает **линейную** взаимосвязь между переменными.

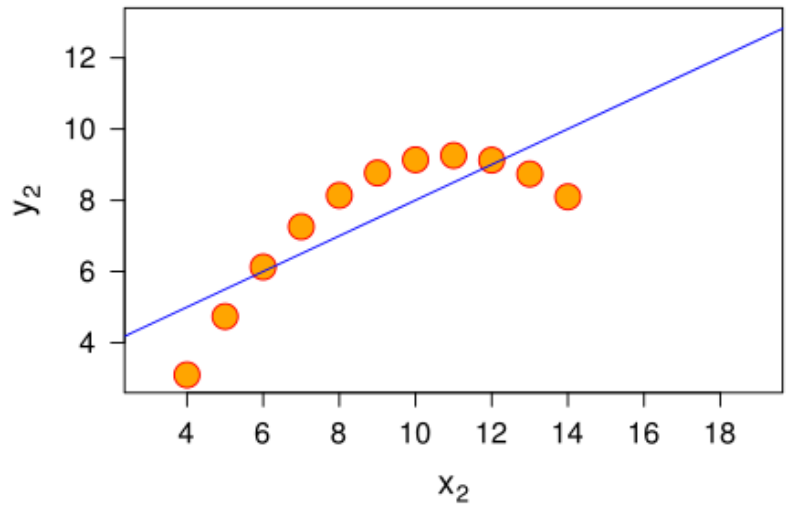
Предположим $Y = bX$, тогда

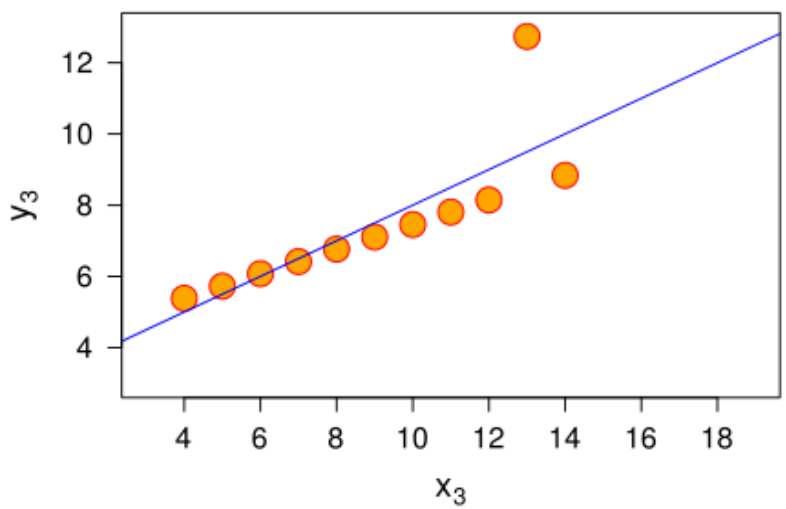
$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{b\sum(x_i - \bar{x})(x_i - \bar{x})}{b\sqrt{\sum(x_i - \bar{x})^2 \sum(x_i - \bar{x})^2}} = 1.$$

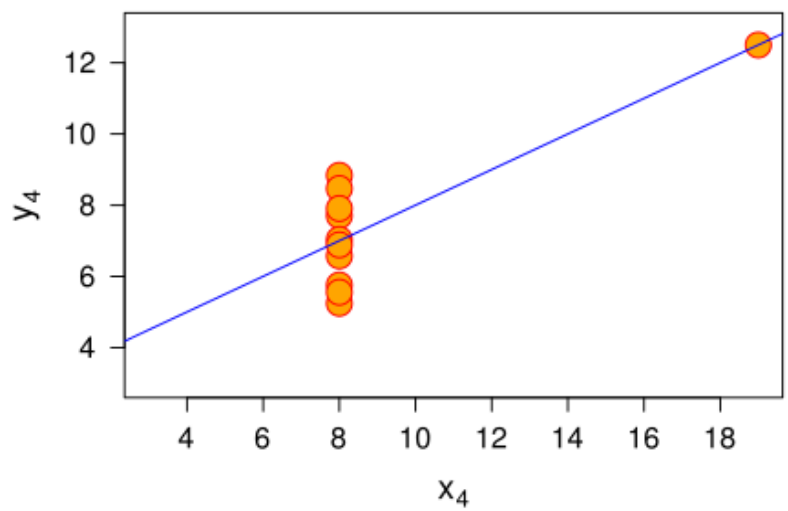
Малый коэффициент корреляции может означать как то, что линейная связь слаба, так и то что эта связь нелинейна.

Квартет Анскомба: $r = 0.816$









Источник: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

$r = 0$ и независимость переменных

Если две переменные независимы, их корреляция равна 0. Но наличие нулевой корреляции не означает, что переменные независимы.

Источник: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Функции для расчета корреляции в R

```
cor(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))

cor.test(x, y,
         alternative = c("two.sided", "less", "greater"),
         method = c("pearson", "kendall", "spearman"),
         exact = NULL, conf.level = 0.95,
         continuity = FALSE, ...)
```

- `use = "everything"` — результат `cor()` будет равен NA, если хотя бы одно из наблюдений имеет пробел (NA).
- `use = "all.obs"` — наличие пробелов в наблюдениях вызовет ошибку.
- `use = "complete.obs"` — наблюдения с пробелами удаляются. Рассматриваются только полные наблюдения, а если таких нет возникает ошибка.

Мы рассматривали коэффициент корреляции, относящийся к генеральной совокупности. На практике он чаще относится к выборке и является статистикой критерия связи между двумя переменными.

- *Хили Дж. Статистика: социологические и маркетинговые исследования. ДиасофтЮП, 2005.* — главы 12—15 посвящены проверке гипотез о связи между переменными.

Три вопроса про статистические связи

1. Существует ли связь?
2. Если связь существует, то насколько на сильна?
3. Каково направление связи?

Связь между переменными $\neq$ Причинная обусловленность

В идеале желательно выявить причинно-следственную связь между переменными (переменная X является причиной изменения переменной Y), измерить ее силу и направление.

Но установить причинно-следственный характер связи между переменными статистика не может. Переменная X связана с переменной Y, но обе они могут зависеть от третьей переменной Z.

Впрочем, для прогнозирования причинно-следственная связь не обязательна.

Скрытая переменная. Бернард Шоу “Доктор на распутье”

“Даже опытные статистики часто оказываются не в состоянии оценить, до какой степени смысл статистических данных искажается молчаливыми предположениями их интерпретаторов... Легко доказать, что ношение цилиндров и зонтиков расширяет грудную клетку, удлиняет жизнь и дает относительный иммунитет от болезней... Университетский диплом, ежедневная ванна, обладание тридцатью парами брюк, знание музыки Вагнера, скамья в церкви короче, все, что подразумевает ...”

Что по мнению Шоу объясняет все эти зависимости?

О вреде огурцов

Огурцы вас погубят! Каждый съеденный огурец приближает вас к смерти. Удивительно, как думающие люди до сих пор не распознали смертоносности этого растительного продукта и даже прибегают к его названию для сравнения в положительном смысле («как огурчик!»). И несмотря ни на что, производство консервированных огурцов растет.

С огурцами связаны все главные телесные недуги и все вообще людские несчастья.

1. Практически все люди, страдающие хроническими заболеваниями, ели огурцы. Эффект явно кумулятивен.
2. 99,9% всех людей, умерших от рака, при жизни ели огурцы.
3. 100% всех солдат ели огурцы.
4. 99,7% всех лиц, ставших жертвами автомобильных и авиационных катастроф, употребляли огурцы в пищу в течение двух недель, предшествовавших фатальному несчастному случаю.
5. 93,1% всех малолетних преступников происходят из семей, где огурцы потребляли постоянно.

Единственный способ избежать вредного действия огурцов — изменить диету. Ешьте, например, суп из болотных орхидей. От него, насколько нам известно, еще никто не умирал.

Источник: Физики продолжают шутить. — М.: “Мир”, 1968.

Зоопарк коэффициентов корреляции

Коэффициент корреляции Пирсона используется для оценки тесноты линейной связи между переменными в метрических шкалах (переменные ведут себя как действительные числа).

Для оценки корреляции между переменными в ранговых (порядковых) шкалах используются коэффициенты корреляции Спирмана (Spearman) или Кендала (Kendall).

Для оценки корреляции номинативных переменных используется коэффициент корреляции Крамера (Cramér).

Пример связи между номинативными переменными: влияние аккредитации на трудоустройство социальных работников

Случайная выборка объемом 100 выпускников колледжа, получивших диплом социального работника, разбита на категории по двум признакам: 1) обучался ли данный студент по программе, аккредитованной Советом по образованию социальных работников (независимая переменная X, категории которой соответствуют столбцам таблицы); 2) принят ли данный студент в течение 3-х месяцев после выпуска на должность социального работника (зависимая переменная Y, категории которой соответствуют строкам таблицы).

Таблица сопряженности:

	Аккредитован	Не аккредитован	Итого
Работает	30	10	40
Не работает	25	35	60
Итого	55	45	100

Существует ли зависимость между X и Y?

- Мы имеем дело с номинативными (номинальными) переменными.
- Зависимая и независимая переменные назначены произвольно. Можно поменять.
- По расположению переменных (столбцы, строки): думайте о таблице как о графике.
- Столбцы представляют собой условные распределения зависимой переменной Y — т.е. значения зависимой переменной, соответствующие определенному значению (категории) независимой переменной.

Критерий χ^2 (хи-квадрат)

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_o — наблюдаемые частоты ячеек, взятые из таблицы сопряженности;

f_e — ожидаемые частоты ячеек, которые бы имели место в случае независимости переменных.

Чем больше различие между наблюдаемыми и ожидаемыми частотами, тем меньше вероятность того, что переменные являются независимыми.

Расчет ожидаемых частот

Исходная таблица сопряженности:

	Аккредитован	Не аккредитован	Итого
Работает	30	10	40
Не работает	25	35	60
Итого	55	45	100

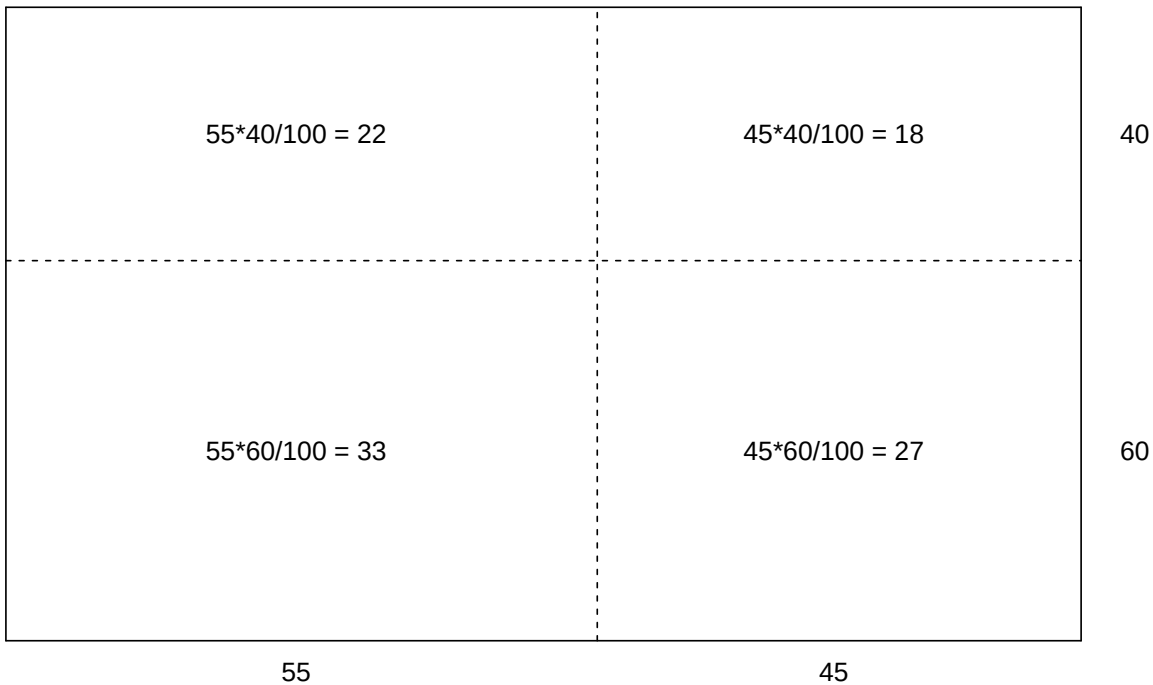
$f_{e,11}$ = 55x40/100 = 22; $f_{e,12}$ = 45x40/100 = 18;

$f_{e,21}$ = 55x60/100 = 33; $f_{e,22}$ = 45x60/100 = 27.

Ожидаемые частоты:

	Аккредитован	Не аккредитован	Итого
Работает	22	18	40
Не работает	33	27	60
Итого	55	45	100

Расчет ожидаемых частот 2



Расчет хи-квадрат

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

```
# chi_2 =
(30-22)^2/22 + (10-18)^2/18 + (25-33)^2/33 + (35-27)^2/27
```

```
## [1] 10.77441
```

Оценка силы связи

Один из вариантов — коэффициент ϕ

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

$$\phi = \sqrt{\frac{10.78}{100}} = 0.33$$

Для таблицы 2x2 значения ϕ находятся в пределах от 0 (отсутствие связи) до 1 (полная связь). Для таблиц большего размера есть коэффициент **V Крамера**.

Линейная регрессия

Если две переменные связаны, можно прогнозировать значения одной из них, пользуясь значениями другой.

Регрессия — прогнозирование значения одной *непрерывной* величины по наблюдениям других величин (как непрерывных, так и дискретных).

Линейная регрессия или линейный регрессионный анализ предполагает, что наблюдаемые величины (предикторы или регрессоры) связаны с прогнозируемой величиной (откликом) **линейно**.

Например, строится линейное уравнение, описывающее статистическую зависимость переменной Y (уровня продаж) от переменной X (расходов на рекламу). В результате аналитик может прогнозировать значение переменной Y.

Случай двух переменных X и Y

Дано

Наблюдения, то есть пары чисел (x_i, y_i).
Гипотеза, что имеется линейная статистическая зависимость между переменными X и Y

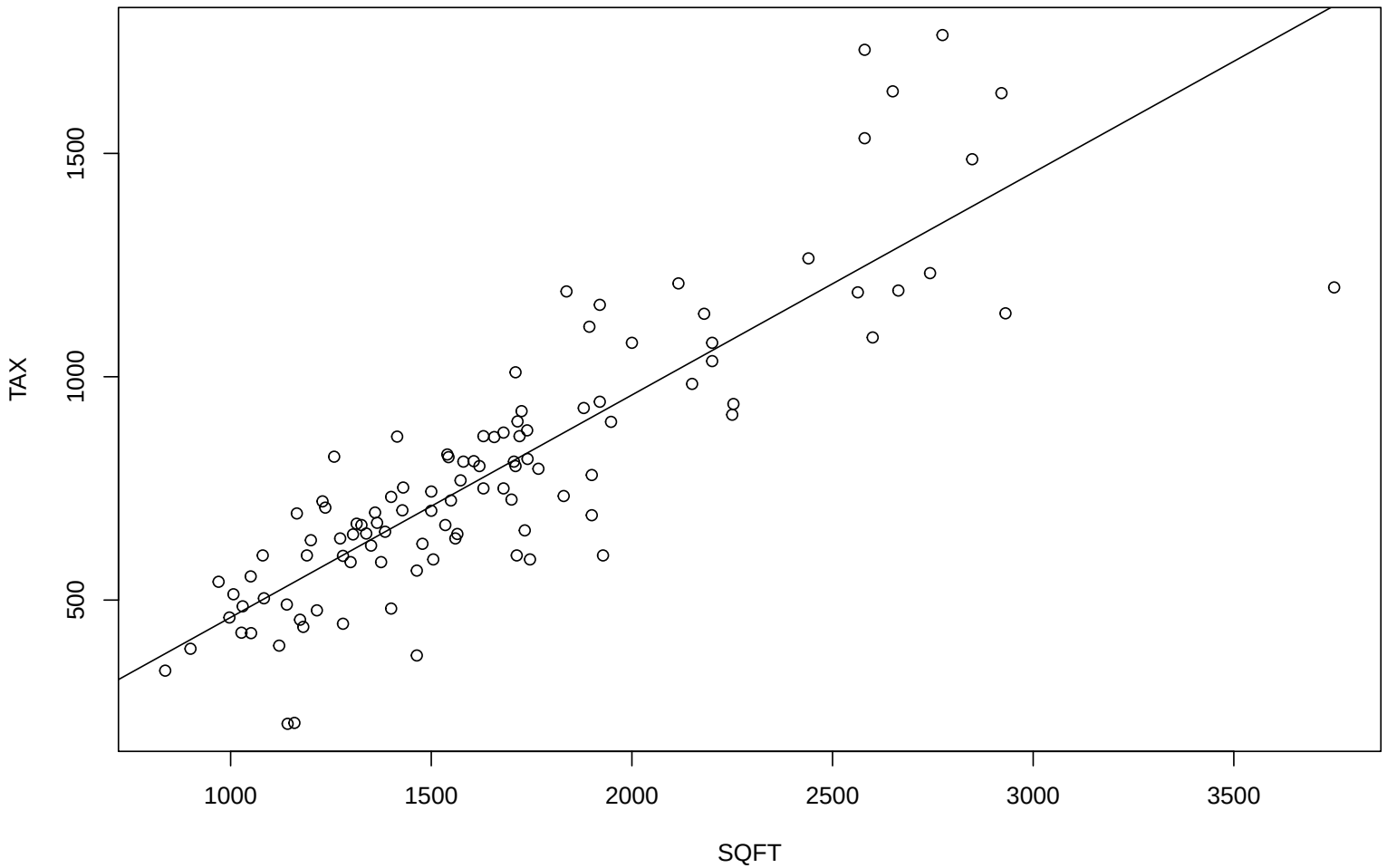
$$Y = a + bX. \tag{1}$$

Найти

Оценки коэффициентов a и b уравнения регрессии (1).
Ключевое допущение: модель линейна по параметрам a и b.

Геометрическая идея решения

Уравнение регрессии определяет прямую, наиболее близко проходящую ко всем точкам с координатами (x_i, y_i).



“Наиболее близко” означает “с наименьшим суммарным отклонением от...”. То есть прямая проходит через средние арифметические.

Вопросы

1. Как считается расстояние между прямой и точкой наблюдений?
2. Чем отличается график регрессии (scatter plot, диаграмма рассеяния) от обычного графика функции?

Функция потерь

Линия регрессии в нашем случае — прямая:

$$Y = f(X) = a + bX$$

Подбираем a и b так, чтобы сумма квадратов отклонений точек линии регрессии Y от наблюдаемых значений y_i была минимальной:

$$F(a, b) = \sum_{i=1}^{N} (y_i - f(x_i, a, b))^2 \to \min$$

F — **функция потерь**. Потери стремятся свести к минимуму.

Ищем минимум функции потерь

Найдем значения a и b, обращающие $F(a, b)$ в минимум

$$\sum_{i=1}^{N}(Y_i - (a + bX_i))\frac{\partial f}{\partial a} = 0,$$

$$\sum_{i=1}^{N}(Y_i - (a + bX_i))\frac{\partial f}{\partial b} = 0.$$

Вычислим производные

$$\frac{\partial f}{\partial a} = 1, \quad \frac{\partial f}{\partial b} = X.$$

Получим

$$\sum_{i=1}^{N}(Y_i - (a + bX_i)) = 0,$$

$$\sum_{i=1}^{N}(Y_i - (a + bX_i))X_i = 0.$$

Коэффициенты регрессии

Из двух уравнений выражаем значения двух неизвестных: b (коэффициента наклона) и a (свободного члена).

Коэффициент наклона характеризует силу влияния X на Y:

$$b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2},$$

При каждом изменении X на 1, Y изменяется на b.

$$a = \bar{Y} - b\bar{X}.$$

Итог: мы получили формулу линейной связи между величинами Y и X ($Y = a + bX$) с конкретными числовыми значениями a и b.

Если отклик Y линейно зависит от единственной переменной, мы имеем дело с **простой линейной регрессией** (simple linear regression).

Простая линейная регрессия

```
data("cars")
```

На основе 50 наблюдений построим уравнение регрессии

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X.$$

которое является оценкой для

$$Y = b_0 + b_1 X + \varepsilon.$$

- X — предиктор (speed = скорость, миль/час)
- Y — отклик (dist = длина тормозного пути, фут)
- b_0, b_1 — коэффициенты регрессии

$\varepsilon = Y - \hat{Y}$ — ошибка (остаток, погрешность, невязка) равна разности между наблюдаемым и прогнозируемым значениями отклика.

```
Residual = Observed - Predicted
```

Данные cars

```
str(cars)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

```
head(cars)
```

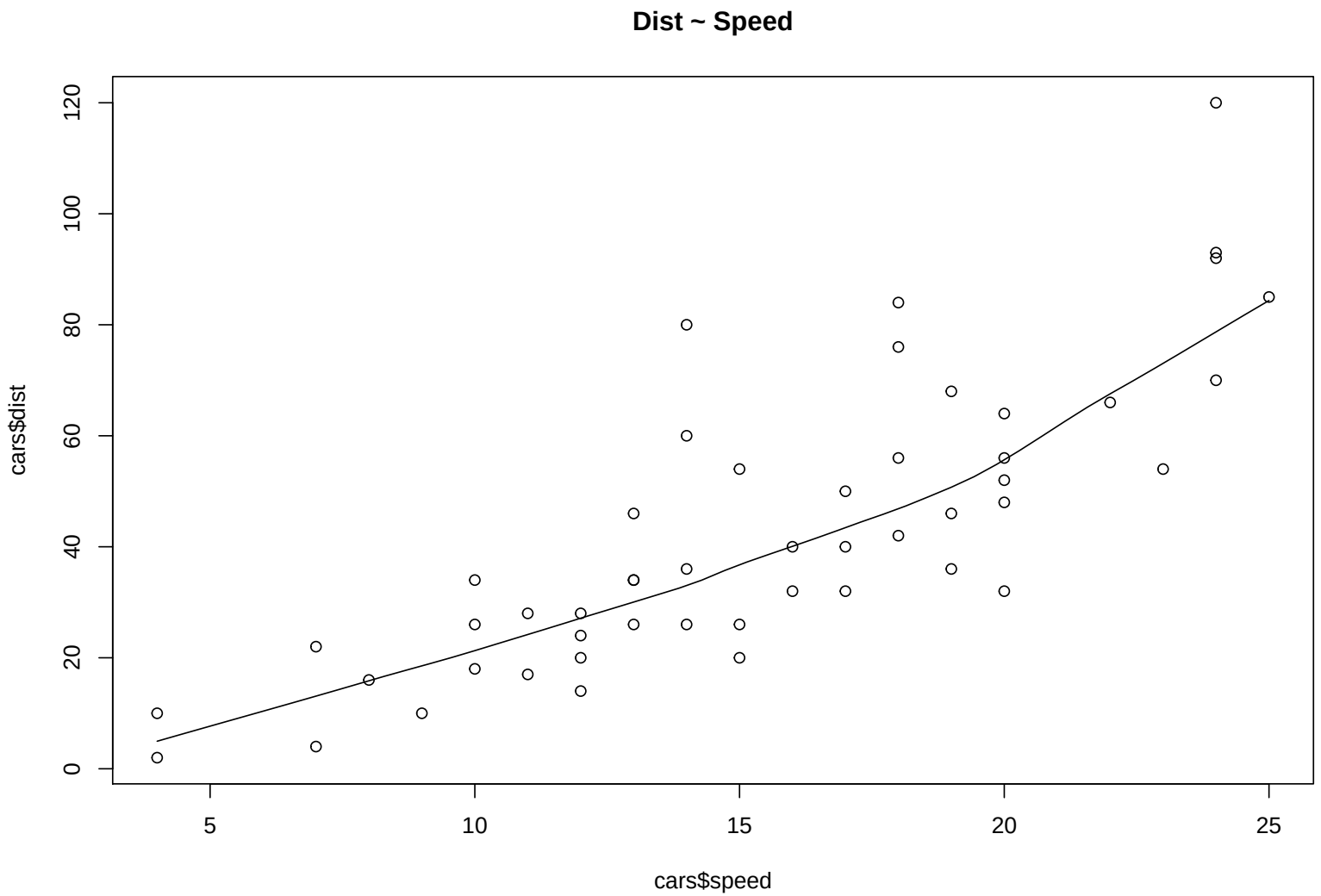
```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

Встроенный в R набор данных.

Источник: *Ezekiel, M.* (1930) Methods of Correlation Analysis. Wiley.

График

```
scatter.smooth(x=cars$speed, y=cars$dist,
                                main="Dist ~ Speed")
```



Корреляция

```
# Вычислим корреляцию между скоростью и
# длиной тормозного пути
cor(cars$speed, cars$dist)
```

```
## [1] 0.8068949
```

Построим линейную модель

```r
# Построим модель для полного набора данных
linear_mod <- lm(dist ~ speed, data=cars)
print(linear_mod)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Coefficients:
## (Intercept)        speed
##     -17.579        3.932
```

Получим формулу:

```
dist = -17.579 + 3.932*speed
```

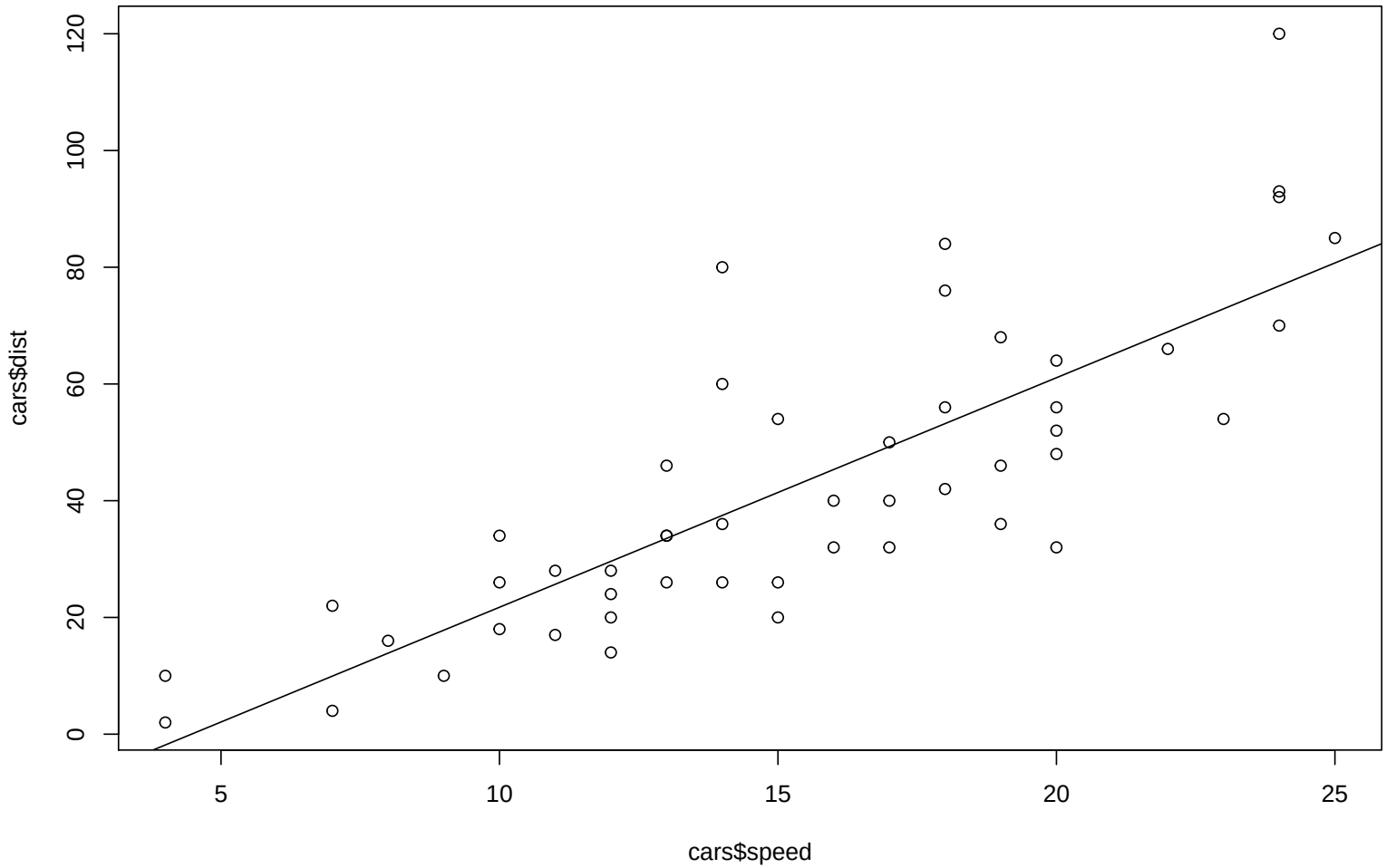
lm() — функция для подгонки линейных моделей

Функция `lm` принимает на вход обучающие данные и описание связи между откликом и предикторами, а возвращает линейную регрессионную модель.

```
lm(Y ~ X, data)
```

- `Y ~ X` — объект класса `formula`.
- `data` — обучающие данные: таблица с колонками `X`, `Y`, ...

```r
plot(cars$speed, cars$dist)
abline(linear_mod)
```



Что находится внутри линейной модели

```
str(linear_mod)
```

```
## List of 12
##  $ coefficients : Named num [1:2] -17.58 3.93
##   ..- attr(*, "names")= chr [1:2] "(Intercept)" "speed"
##  $ residuals    : Named num [1:50] 3.85 11.85 -5.95 12.05 2.12 ...
##   ..- attr(*, "names")= chr [1:50] "1" "2" "3" "4" ...
##  $ effects      : Named num [1:50] -303.914 145.552 -8.115 9.885 0.194 ...
##   ..- attr(*, "names")= chr [1:50] "(Intercept)" "speed" "" "" ...
##  $ rank         : int 2
##  $ fitted.values: Named num [1:50] -1.85 -1.85 9.95 9.95 13.88 ...
##   ..- attr(*, "names")= chr [1:50] "1" "2" "3" "4" ...
##  $ assign       : int [1:2] 0 1
##  $ qr           :List of 5
##   ..$ qr   : num [1:50, 1:2] -7.071 0.141 0.141 0.141 0.141 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   .. .. ..$ : chr [1:50] "1" "2" "3" "4" ...
##   .. .. ..$ : chr [1:2] "(Intercept)" "speed"
##   .. ..- attr(*, "assign")= int [1:2] 0 1
##   ..$ qraux: num [1:2] 1.14 1.27
##   ..$ pivot: int [1:2] 1 2
##   ..$ tol  : num 1e-07
##   ..$ rank : int 2
##   ..- attr(*, "class")= chr "qr"
##  $ df.residual  : int 48
##  $ xlevels      : Named list()
```

Полезные функции

```
coef(linear_mod)          # коэффициенты модели
```

```
## (Intercept)       speed
##  -17.579095    3.932409
```

```
fitted(linear_mod)[1:5] # подогнанные значения Y
```

```
##         1         2         3         4         5
## -1.849460 -1.849460  9.947766  9.947766 13.880175
```

```
resid(linear_mod)[1:5]   # остатки
```

```
##         1         2         3         4         5
##  3.849460 11.849460 -5.947766 12.052234  2.119825
```

Прогнозирование и доверительный интервал

Только прогноз для новых данных:

```
predict(model, data.frame(новые данные))
```

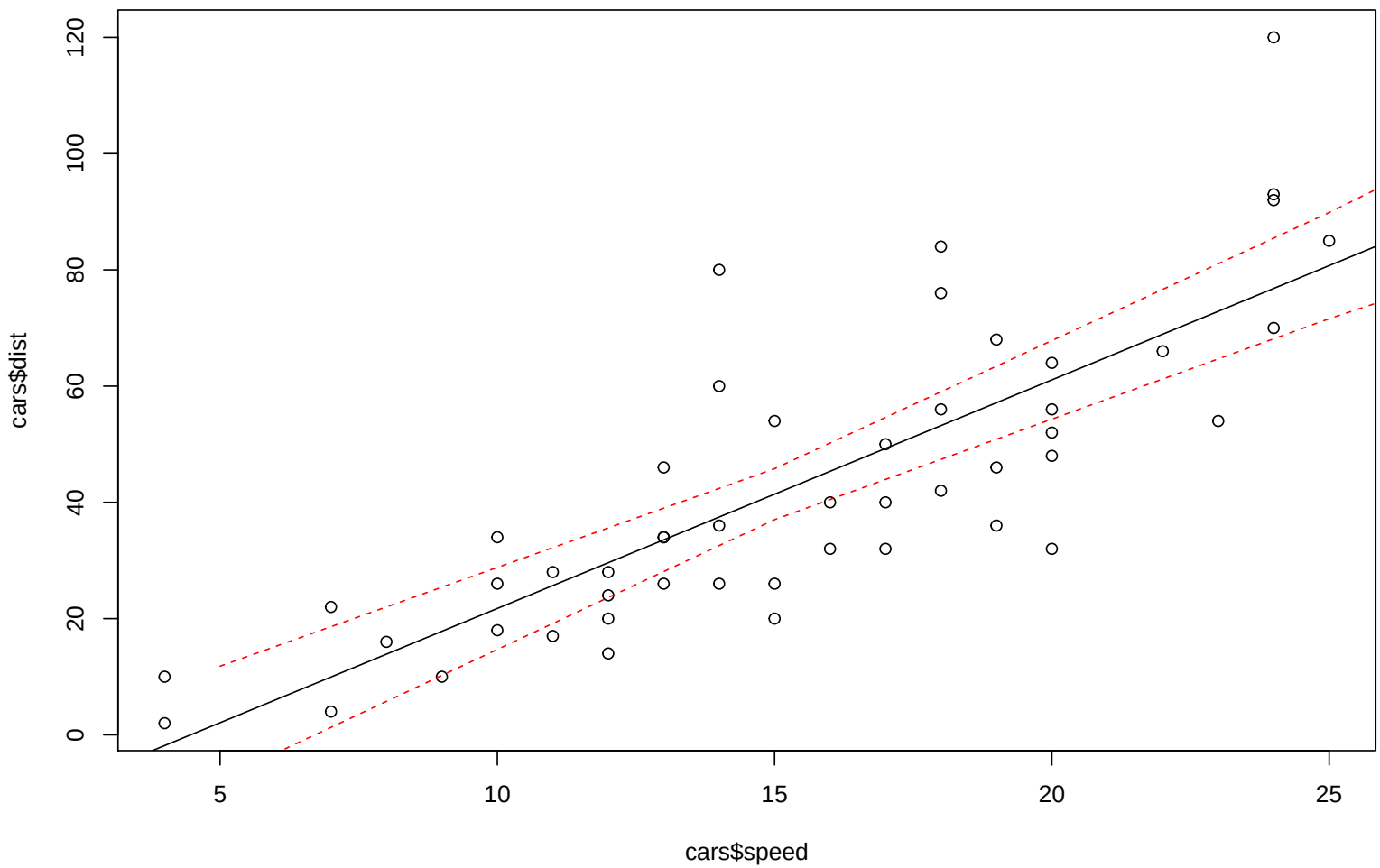
Прогноз с расчетом доверительного интервала:

```
predict(model, data.frame(новые данные),
                level = 0.95, interval = "confidence")
```

```
speed.new <- c(5,15,25,35,45)
preds <- predict(linear_mod, data.frame(speed = speed.new),
                 level = 0.95, interval = "confidence")
```

```
##          fit       lwr       upr
## 1   2.082949  -7.64415  11.81005
## 2  41.407036  37.02115  45.79292
## 3  80.731124  71.59608  89.86617
## 4 120.055212 103.10660 137.00382
## 5 159.379299 134.26645 184.49215
```

```r
plot(cars$speed, cars$dist)
abline(linear_mod)
lines(speed.new, preds[ ,3], lty = 'dashed', col = 'red')
lines(speed.new, preds[ ,2], lty = 'dashed', col = 'red')
```



Этапы линейной регрессии в R

1. Собрать числовые данные о связи предикторов с откликом.
2. Записать формулу, предположительно связывающую предикторы и отклик. Поместить эту формулу в функцию `lm()`.
3. Проанализировать качество регрессионной модели при помощи `summary()`.
4. По результатам выполнения `lm()` найти коэффициенты регрессии. Записать с их помощью уравнение регрессии $Y = b_0 + b_1 X$.
5. Для прогнозирования новых значений отклика использовать функцию `predict()`. Определить доверительный интервал прогноза.

Измеряем качество регрессионной модели: среднеквадратичная ошибка

$$\text{Mean Square Error (MSE)} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (a + bX_i))^2.$$

$$\text{Root Mean Square Error (RMSE)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - (a + bX_i))^2}.$$

```
MSE <- mean(linear_mod$residuals^2)
(RMSE <- sqrt(MSE))
```

```
## [1] 15.06886
```

Измеряем качество регрессионной модели: коэффициент детерминации

Насколько построенная нами модель лучше описывает данные по сравнению с некой базовой характеристикой?

В качестве базовой характеристики выступает среднее арифметическое $\bar{Y}$.

- Измеряем сумму квадратов отклонений для регрессионной модели

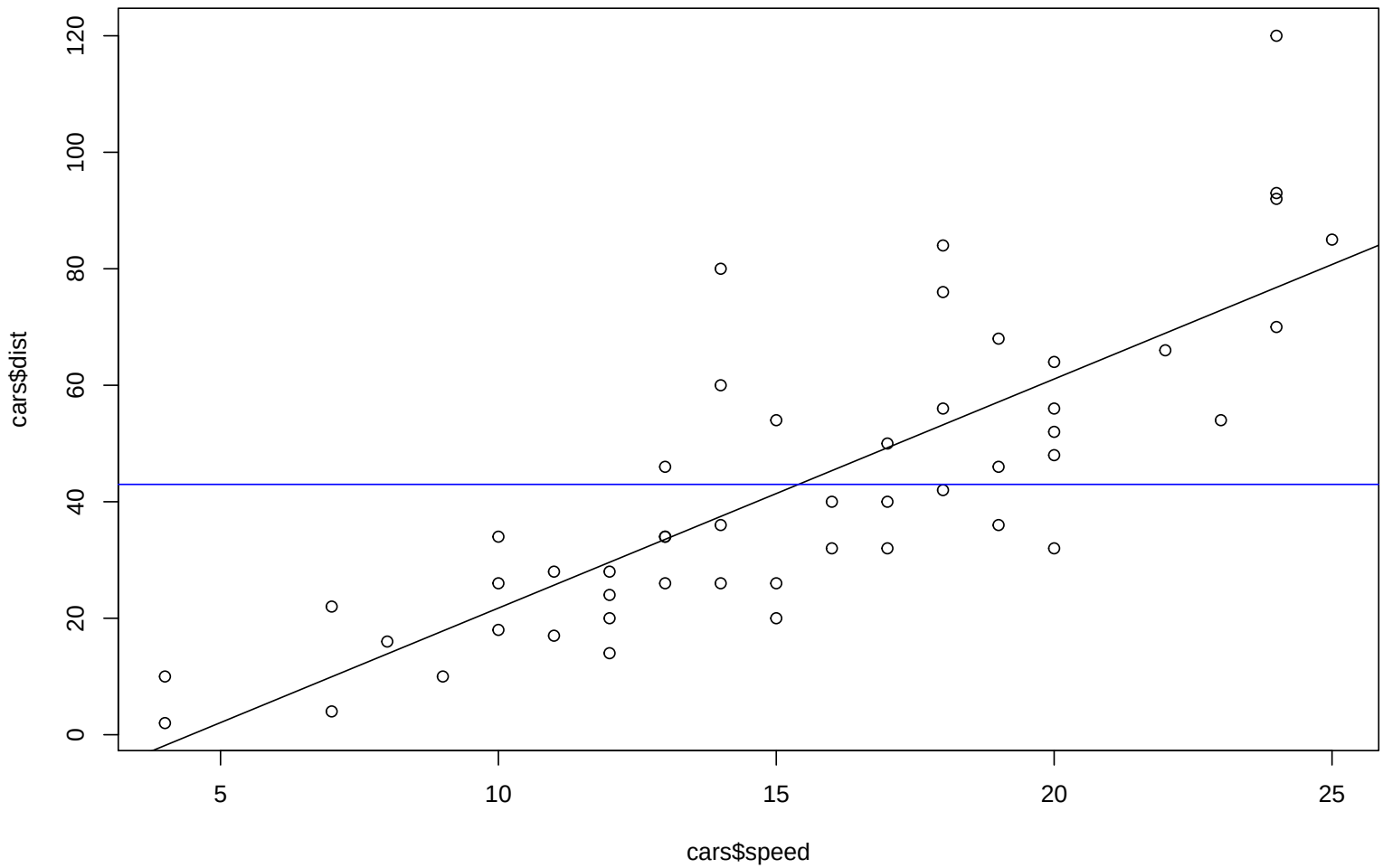
$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - (a + bX_i))^2$$

- Измеряем сумму квадратов отклонений для базовой модели

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Коэффициент детерминации R^2

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - (a + bX_i))^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \quad 0 \leq R^2 \leq 1$$



Коэффициент детерминации для зависимости тормозного пути от скорости

```
summary(linear_mod)$r.squared
```

```
## [1] 0.6510794
```

Наша модель хороша, если она дает большой выигрыш по сравнению с базовой моделью.

В данном случае это так.

Свойства R^2

- Коэффициент детерминации, в отличие от MSE, — величина безразмерная.
- Коэффициент детерминации работает для зависимостей вида

$$Y = a + b_1 X_1 + b_2 X_2 + ... + b_n X_n.$$

Интерпретации коэффициента детерминации

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \frac{1}{n}\sum_{i=1}^{n}(Y_i - (a + bX_i))^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \cdot 100\%$$

- На сколько процентов улучшилась модель по сравнению с базовой.
- Какой процент вариации Y объясняется влиянием всех независимых переменных (предикторов).

Если X_i всего один, то R^2 равен квадрату коэффициента корреляции между этим X_i и Y.

Диагностика модели: summary()

```
summary(linear_mod)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

summary()

- `Call` — как вызывалась функция `lm`.
- `Residuals` — описание остатков модели.
- `Coefficients` — оценка качества подбора коэффициентов модели.
  - `Estimate` — оценки коэффициентов модели,
  - `Std. Error` — их стандартные отклонения,
  - `t value` и `Pr(>|t|)` — t-значения и вероятности нулевой гипотезы, что коэффициент равен нулю.
- стандартное отклонение регрессии (Residual standard error) — RMSE.
- коэффициенты детерминации — обычный (Multiple R-squared) и скорректированный (Adjusted R-squared).
- `F-statistic` — результаты F-теста нулевой гипотезы об одновременном равенстве нулю всех коэффициентов регрессионной модели.

Шаг 1. Создадим обучающую и проверочную выборки данных

До сих пор мы строили модель линейной регрессии, используя весь набор данных. В этом случае невозможно оценить точность работы модели на новых данных.

Новых данных у нас нет, но мы сделаем их из имеющихся.

Разделим набор данных в соотношении 80:20 (обучение:тестирование). Построим модель на обучающей выборке из 80% данных и затем оценим точность ее прогноза на тестовой (проверочной) выборке из оставшихся 20% данных.

```r
# Разделим данные на обучающую и тестовую выборки

# Зерно генератора случайных чисел нужно
# для воспроизводимости результата
set.seed(100)
# Номера строк, которые попадут в обучающую выборку
train_ind <- sample(1:nrow(cars), 0.8*nrow(cars))
training <- cars[ train_ind, ] # обучающая выборка
test     <- cars[-train_ind, ] # тестовая выборка
```

Шаг 2. Построим модель на тренировочных данных и используем ее для прогноза на проверочных данных

```
# Построим модель, обученную на обучающей выборке
lm_mod <- lm(dist ~ speed, data=training)
# Спрогнозируем тормозной путь на тестовой выборке
predicted <- predict(lm_mod, test)
```

Шаг 3. Проанализируем качество модели по summary

```
summary(lm_mod)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.350 -10.771  -2.137   9.255  42.231
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -22.657      7.999  -2.833  0.00735 **
## speed          4.316      0.487   8.863 8.73e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 15.84 on 38 degrees of freedom
## Multiple R-squared:  0.674,  Adjusted R-squared:  0.6654
## F-statistic: 78.56 on 1 and 38 DF,  p-value: 8.734e-11
```

Шаг 4. Оценка точности прогноза

Коэффициент корреляции между реальными и прогнозными значениями можно использовать как простейшую меру точности прогноза. Высокие значения корреляции показывают, что реальные и прогнозные значения изменяются сонаправленно (одновременно увеличиваются или уменьшаются).

```
actuals_preds <- data.frame(actual=test$dist, predicted)
(correlation_accuracy <- cor(actuals_preds))
```

```
##                actual predicted
## actual    1.0000000 0.8277535
## predicted 0.8277535 1.0000000
```

```
actuals_preds[1:4,]
```

```
##    actual predicted
## 1       2 -5.392776
## 4      22  7.555787
## 8      26 20.504349
## 20     26 37.769100
```

Дополнительные меры точности: MinMax accuracy и MAPE

$$\text{MinMax Accuracy} = mean\left(\frac{min(actual, predicted)}{max(actual, predicted)}\right),$$

$$\text{Mean Absolute Percentage Error (MAPE)} =$$

$$= mean\left(\frac{|predicted - actual|}{actual}\right).$$

MinMax accuracy и MAPE

```
min_max_accuracy <- mean(apply(actuals_preds, 1, min) /
                         apply(actuals_preds, 1, max))
mape <- mean(abs((actuals_preds$predicted -
        actuals_preds$actual))/actuals_preds$actual)

min_max_accuracy
```

```
## [1] 0.3800489
```

```
mape
```

```
## [1] 0.6995032
```

Необходимость перекрестной проверки

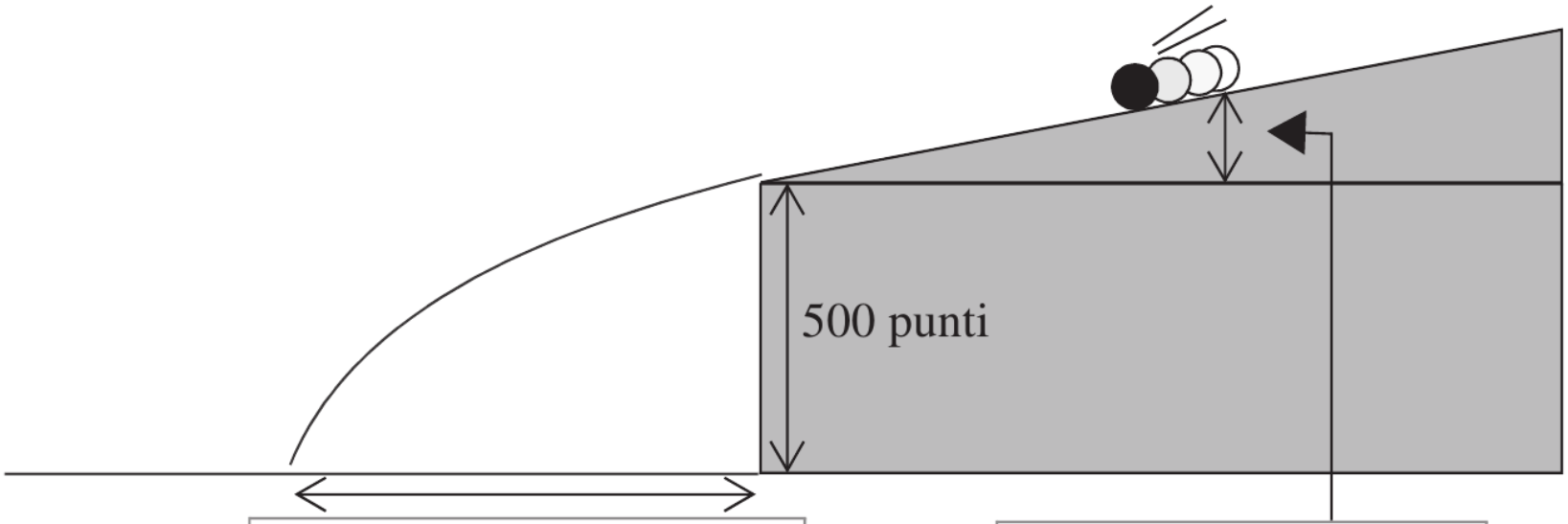
Предположим, модель удовлетворительно предсказывает на 20%-ной тестовой выборке. Достаточно ли этого, чтобы утверждать, что модель всегда будет предсказывать столь же точно? Возможно, наша тестовая выборка всего лишь счастливое исключение.

Нужно показать, что построенная модель хорошо работает на любой подобной обучающей выборке.

Разделим наши данные на $k = 5$ непересекающихся частей. Сохраняя каждую часть в качестве тестовой выборки, построим модель на оставшихся $(k - 1)$ частях данных и рассчитаем среднеквадратичную ошибку прогноза.

Проделаем эту операцию для каждой из k частей. Затем вычислим среднее значение полученных среднеквадратичных ошибок. Эта метрика является более устойчивым показателем качества прогноза чем среднеквадратичная ошибка, построенная по единственной тестовой выборке.

Эксперимент Галилея



Horizontal Distance (punti)	Initial Height (punti)
253	100
337	200
395	300
451	450
495	600
534	800
573	1,000

Ramsey F., Schafer D. The Statistical Sleuth: A Course in Methods of Data Analysis, 3rd Edition, 2013

Квадратичная аппроксимация

```
# Данные из: Ramsey F., Schafer D. The Statistical
# Sleuth: A Course in Methods of Data Analysis,
# 3rd Edition, 2013

height = c(100, 200, 300, 450, 600, 800, 1000)
distance = c(253, 337, 395, 451, 495, 534, 574)

# Модель в форме квадратичного полинома
lm.r = lm(distance ~ height + I(height^2))
```

```r
summary(lm.r)
```

```
##
## Call:
## lm(formula = distance ~ height + I(height^2))
##
## Residuals:
##       1       2       3       4       5       6       7
## -14.420   9.192  13.624   2.060  -6.158 -12.912   8.614
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.002e+02  1.695e+01  11.811 0.000294 ***
## height       7.062e-01  7.568e-02   9.332 0.000734 ***
## I(height^2) -3.410e-04  6.754e-05  -5.049 0.007237 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 13.79 on 4 degrees of freedom
## Multiple R-squared:  0.9902, Adjusted R-squared:  0.9852
## F-statistic: 201.1 on 2 and 4 DF,  p-value: 9.696e-05
```

Закон движения тела

Галилей теоретически обосновал и экспериментально доказал, что тело, брошенное горизонтально или под углом к горизонту, будет двигаться по параболе.

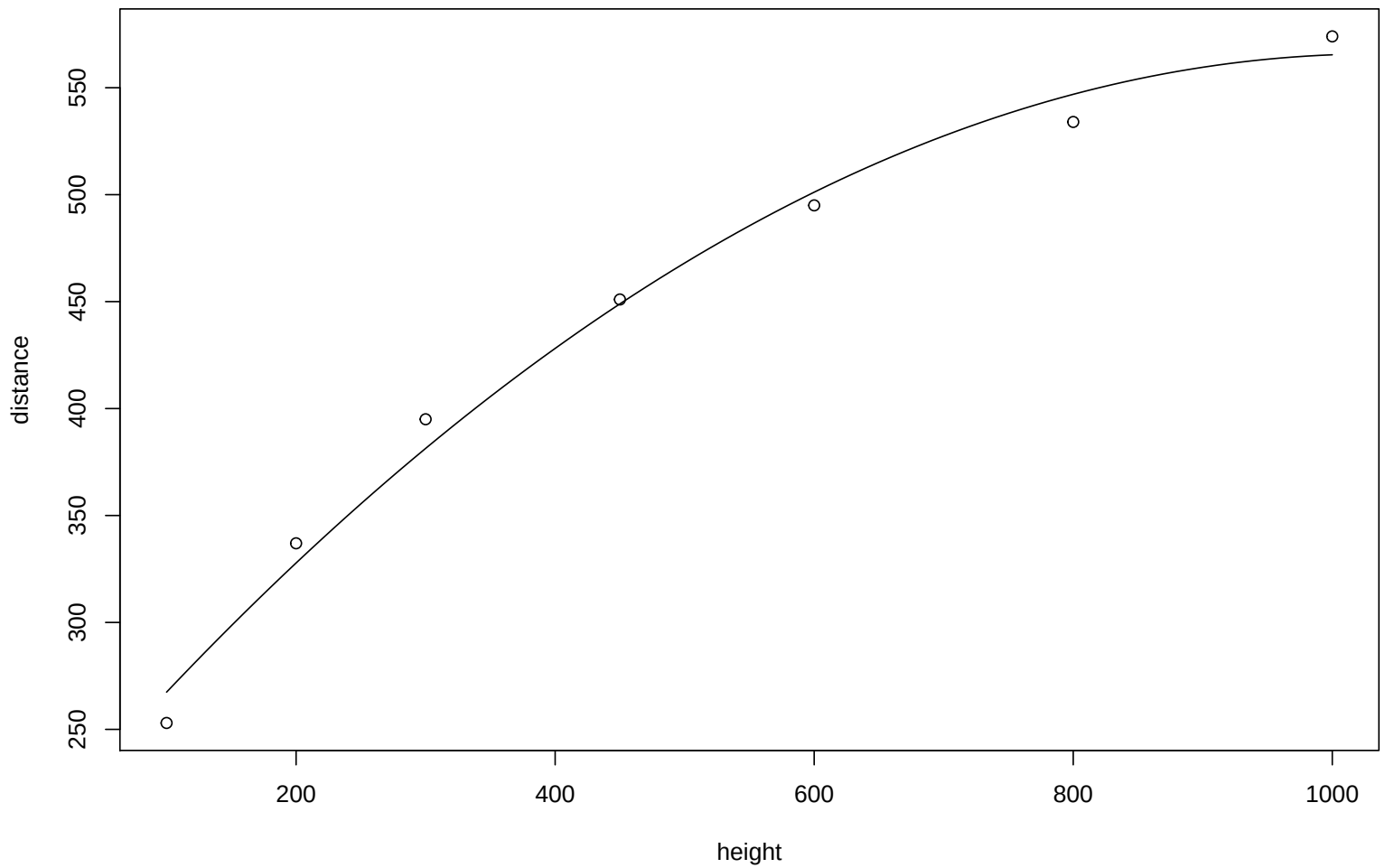
Данные эксперимента дают нам формулу:

```
distance = 200.211950 + 0.706182*height - 0.000341*height^2
```

```
# Создадим высоты для прогноза
h = seq(100, 1000, 10)
# Вычислим расстояния для каждой из новых высот
dist = 200.211950 + 0.706182*h - 0.000341*h^2
```

Зависимость длины полета от начальной высоты

```
plot(height, distance)   # исходные данные
lines(h, dist, lty=1)    # результаты подгонки
```



Формулы

“Линейная модель” означает линейность относительно коэффициентов регрессии b_i

- $y = b_0 + b_1 x + b_2 x^2$ — линейная модель.
- $y = b_0 x^{b_1}$ — нелинейная модель.

Примеры

- `Y ~ A` — прямая со свободным членом b_0, заданным неявно

$$Y = b_0 + b_1 A$$

- `Y ~ -1 + A` — прямая без свободного члена; будет проходить через (0,0)

$$Y = b_1 A$$

- `Y ~ A + I(A^2)` — полином; внутри функции `I()` можно задавать операции, которые трактуются в обычном для математики смысле

$$Y = b_0 + b_1 A + b_2 A^2$$

Примеры формул

- `Y ~ A + B` — модель 1-го порядка, в которой A и B не взаимодействуют

$$Y = b_0 + b_1 A + b_2 B$$

- `Y ~ A:B` — модель, содержащая только взаимодействие между A и B (1-го порядка)

$$Y = b_0 + b_1 AB$$

- `Y ~ A*B` — полная модель 1-го порядка, учитывающая как A и B, так и взаимодействие между ними (эквивалент: `Y ~ A + B + A:B`)

$$Y = b_0 + b_1 A + b_2 B + b_3 AB$$

- `Y ~ (A + B + C)^2` — модель включает все эффекты 1-го порядка и все взаимодействия вплоть до n-го порядка, где n задается показателем в `(...)^n` (эквивалент: `Y ~ A*B*C - A:B:C`)

$$Y = b_0 + b_1 A + b_2 B + b_3 C + b_4 AB + b_5 AC + b_6 AC$$

Дополнительные материалы

- MachineLearning.ru: Коэффициент детерминации — о скорректированном коэффициенте детерминации.
- Prabhakaran S. Linear Regression With R
- Ramsey F., Schafer D. The Statistical Sleuth: A Course in Methods of Data Analysis, 3rd Edition, 2013. Данные Галилея находятся в пакете Sleuth3.