

связи
переменных
корреляции
aggregate()

регрессии

линейная
регрессия

dummy-
переменные

множественная
регрессия

сравнение
моделей

перебор
моделей

другие
регрессии

Корреляции и другие связи переменных. Регрессионный анализ

Г. Мороз

Связаны ли одни переменные с другими?

связи
переменных

корреляции
aggregate()

регрессии

линейная
регрессия

dummy-
переменные

множественная
регрессия

сравнение
моделей

перебор
моделей

другие
регрессии

- две количественные переменные как связаны?
⇒ коэффициенты корреляции Пирсона, Спирмана, Кенделла

- две качественные переменные
⇒ aggregate(), χ^2 , тест Фишера связи нет?

- одна количественная и одна качественная переменная
⇒ ANOVA связи нет?

К сожалению, слово *"корреляция"* и его однокоренные в языке используется куда шире (примеры из НКРЯ):

"...существует четкая корреляция между континентом проведения соревнований и результатом..."

"...существует прямая корреляция между владением азиатскими «тональными» языками и хорошим музыкальным слухом..."

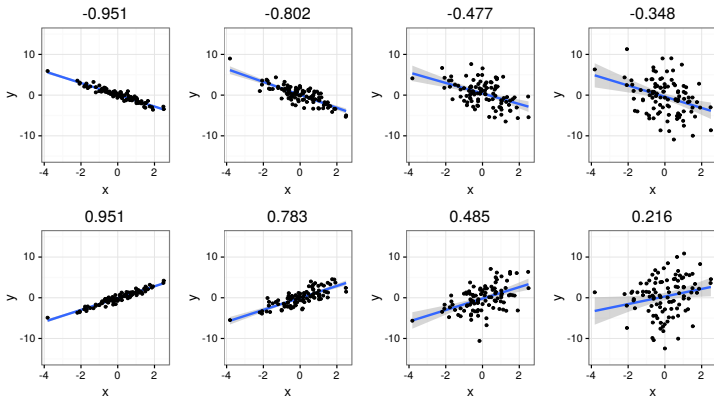
"...прямая корреляция количества крыс в доме зависит от наличия мусоропровода и его состояния..."

В статистике **корреляция** — это отношение между **числовыми переменными**.

презентация доступна: <http://goo.gl/AdqRQI>

Коэффициент корреляции Пирсона

Коэффициент корреляции позволяет показать степень взаимосвязи между двумя величинами. Коэффициент корреляции изменяется от -1 до 1, где 0 обозначает отсутствие взаимосвязи, положительное значение коэффициента указывает на прямую взаимосвязь (чем больше x , тем больше y), а отрицательное — на обратную (чем больше x , тем меньше y).



Коэффициент корреляции Пирсона

Значение коэффициента корреляции Пирсона зависит от удаленности точек от регрессионной прямой и **никак** не зависит от наклона данной прямой. См. примеры из [Википедии](#).

```
x <- c(2, 8, 3, 7, 11, 3)
```

```
y <- c(12, 7, 10, 8, 5, 11)
```

```
cor(x, y) # по умолчанию считается коэффициент корреляции Пирсона  
cor.test(x, y) # H0: коэффициент равен нулю
```

Pearson's product-moment correlation

data: x and y

t = -12.03, df = 4, p-value = 0.0002737

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.9985831 -0.8770124 доверительный интервал для коэффициента

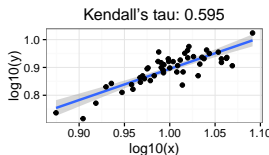
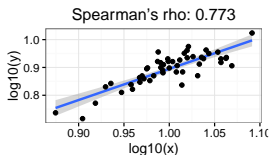
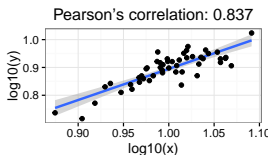
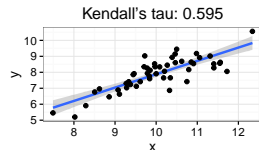
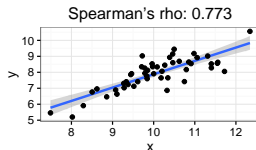
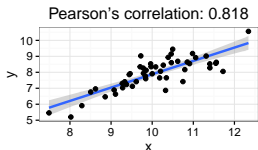
sample estimates:

-0.9864609 коэффициент корреляции

тип данных: числовой

параметрический (требует линейности связи)

Коэффициенты корреляции Спирмана, Кенделла



```
x <- c(2, 8, 3, 7, 11, 3)
y <- c(12, 7, 10, 8, 5, 11)
```

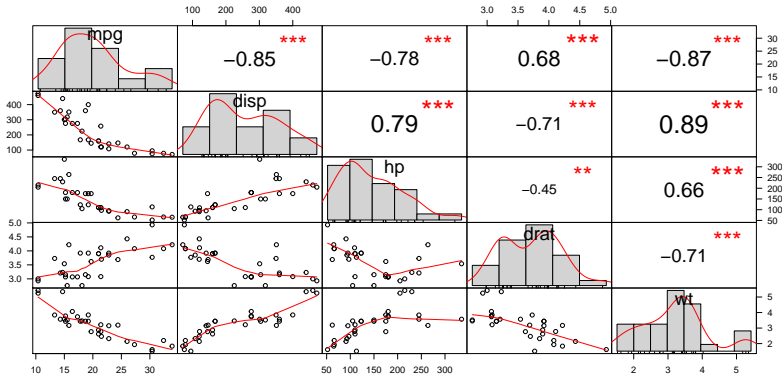
```
cor(x, y, method = "spearman") # коэффициент корреляции Спирмана
cor.test(x, y, method = "spearman")
cor(x, y, method = "kendall") # коэффициент корреляции Кенделла
cor.test(x, y, method = "kendall")
```

тип данных: числовой
непараметрический

презентация доступна: <http://goo.gl/AdqRQl>

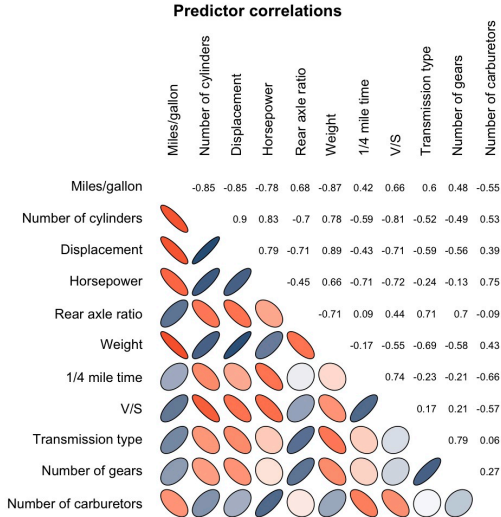
Корреляционная матрица: PerformanceAnalytics

```
library(PerformanceAnalytics)
chart.Correlation(mydata)
chart.Correlation(mydata, method = "spearman")
chart.Correlation(mydata, method = "kendall")
```



Корреляционная матрица: ellipse

Здесь пример использованием пакета ellipse.



aggregate()

В работе [Cedergren 1973] ис следовались отпадение согласного [s] в конце слова в речи жителей Панама. При каких условиях выпадение [s] происходит чаще всего?

tail(dat)

	s.deletion	gramm.cat	phon.cont	social
8841	not deleted	noun	pause	2
8842	not deleted	noun	pause	2
8843	not deleted	noun	pause	3
8844	not deleted	noun	pause	3
8845	not deleted	noun	pause	4
8846	not deleted	noun	pause	4

supply(dat, table)

\$s.deletion

deleted	not deleted
5091	3755

\$gramm.cat

adjective	determiner	noun	separate morpheme	verb
609	1393	2268	4460	116

\$phon.cont

consonant	pause	vowel
5600	1304	1942

\$social

1	2	3	4
579	2547	2385	3335


```
aggregate(s.deletion~gramm.cat,
data = a[a$s.deletion == "deleted",],
length)
```

```
# формула
# данные
# функция
```

	gramm.cat	s.deletion
1	adjective	298
2	determiner	604
3	noun	1594
4	separate morpheme	2556
5	verb	39

```
aggregate(s.deletion~gramm.cat + phon.cont,
data = a[a$s.deletion == "deleted",],
length)
```

```
# формула
# данные
# функция
```

	gramm.cat	phon.cont	s.deletion
1	adjective	consonant	205
2	determiner	consonant	524
3	noun	consonant	733
4	separate morpheme	consonant	1355
5	verb	consonant	23
6	adjective	pause	48
7	noun	pause	485
8	separate morpheme	pause	471
9	adjective	vowel	45
10	determiner	vowel	80
11	noun	vowel	376
12	separate morpheme	vowel	730
13	verb	vowel	16

Correlation does not imply causation!

Это говорят на всех курсах по статистике. Примером могут служить **сайт** и сделанная на его основе **книга** *Spurious correlations*.

Если есть корреляция между двумя переменными a и b , то может быть один из следующих вариантов:

- a вызывает b
- b вызывает a
- a вызывает b , а b вызывает a
- a вызывает c , а c вызывает b
- c вызывает a и b , но a и b не связаны
- a и b не связаны

Однако часто приводят примеры лишь на последнее. Кстати, вы знали, что **количество пиратов влияет на глобальное потепление**? А еще... чем больше пожарников посылают тушить пожар, тем больше ущерба он наносит.

Статистическая модель

Когда мы работаем с данными и находим отношения между какими-то из переменных, мы создаем упрощенное представление некоторой системы, которые мы в дальнейшем будем называть *моделью*. Получившаяся модель позволяет нам с некоторой точностью предсказывать некоторый результат на основе той или иной конфигурации параметров модели. Чаще всего в статистические модели закладывается стохастический элемент, т. е. даже в самой простой модели будет случайная переменная, которую еще называют **остатками модели**:

$$y = 4 + \epsilon_i$$

Таким образом любое статистическое моделирование — это поиск наилучшей аппроксимация закона распределения исследуемой переменной, так чтобы обеспечить минимум **средней ошибки ϵ_i** .

Линейная регрессия

В работе исследовалась зависимость средней значения частоты основного тона от возраста (мужчины и женщины следует считать отдельно). Какие коэффициенты получит регрессионные линии и сколько процентов дисперсии объяснят наши модели, если мы предположим линейную зависимость между переменными?

связи
переменных

корреляции
aggregate()

регрессии

линейная
регрессия

dummy-
переменные

множественная
регрессия

сравнение
моделей

перебор
моделей

другие
регрессии

Линейная регрессия

В работе исследовалась зависимость средней значения частоты основного тона от возраста (мужчины и женщины следует считать отдельно). Какие коэффициенты получит регрессионные линии и сколько процентов дисперсии объяснят наши модели, если мы предположим линейную зависимость между переменными?

$$y = \beta_0 + \beta_1 \cdot x + \varepsilon_i,$$

где x — предиктор, β_0 — свободный член, β_1 — угловой коэффициент, ε_i — средняя ошибка.

Линейная регрессия: строим модель

summary(df)

sex	age	pitch
f:20	Min. :23.00	Min. :139.3
m:20	1st Qu. :46.50	1st Qu. :154.3
	Median :62.50	Median :158.8
	Mean :59.12	Mean :159.3
	3rd Qu. :71.25	3rd Qu. :162.4
	Max. :83.00	Max. :176.4

```
dfm <- subset(df, sex=="m")
```

сгруппируем по полу

```
dff <- subset(df, sex=="f")
```

сгруппируем по полу

```
fit.f <- lm(pitch~age, dff)
```

строим регрессию

```
fit.m <- lm(pitch~age, dfm)
```

строим регрессию

В случае, если в задаче требуется исключить свободный член, то в формулу нужно добавить -1:

```
fit.m2 <- lm(pitch~age - 1, dfm)
```

исключаем свободный член

Линейная регрессия: анализ результатов

```
fit.f <- lm(pitch~age, dff)          # строим регрессию по данным женщин
summary(fit.f)
```

```
Call:
lm(formula = pitch ~age, data = dff)      # формула, вдруг забыли
```

```
Residuals:                                # распределение остатков
      Min       1Q       Median       3Q      Max
-2.33997 -0.62471 -0.06519  0.70728  1.66992
```

```
Coefficients:                                # коэффициенты модели
      Estimate Std. Error t value Pr(>|t|)      *** #  $\beta_0$ 
(Intercept)  189.87935   1.00978   188.04 < 2e-16      ***
age          -0.39820    0.01605   -24.81 2.27e-15      *** #  $\beta_1$ 
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 1.134 on 18 degrees of freedom

Multiple R-squared: 0.9716, **Adjusted R-squared: 0.97**

F-statistic: 615.5 on 1 and 18 DF, p-value: 2.267e-15

Что означают p-values?

презентация доступна: <http://goo.gl/AdqRQl>

Линейная регрессия: анализ результатов

```
fit.m <- lm(pitch~age, dfm)      # строим регрессию по данным женщин
summary(fit.m)
```

```
Call:
lm(formula = pitch ~age, data = dfm)      # формула, вдруг забыли
```

```
Residuals:                                # распределение остатков
      Min       1Q     Median       3Q      Max
-1.87771   -0.54867    0.05222    0.88251    1.79433
```

```
Coefficients:                                # коэффициенты модели
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  130.07015   0.90750   143.33 < 2e-16 *** #  $\beta_0$ 
age          0.39790    0.01534   25.94 1.04e-15 *** #  $\beta_1$ 
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.9985 on 18 degrees of freedom

Multiple R-squared: 0.974, Adjusted R-squared: 0.9725

F-statistic: 673 on 1 and 18 DF, p-value: 1.036e-15

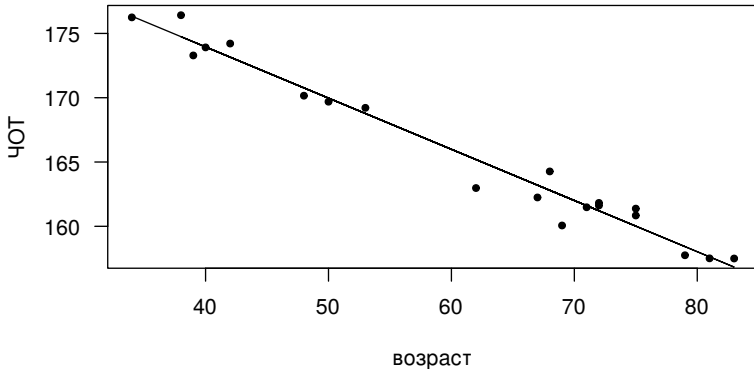
Что означают p-values?

презентация доступна: <http://goo.gl/AdqRQl>

Линейная регрессия: визуализация, R-base

связи
переменных
корреляции
aggregate()
регрессии
линейная
регрессия
dummy-
переменные
множественная
регрессия
сравнение
моделей
перебор
моделей
другие
регрессии

ЧОТ vs. возраст: женщины

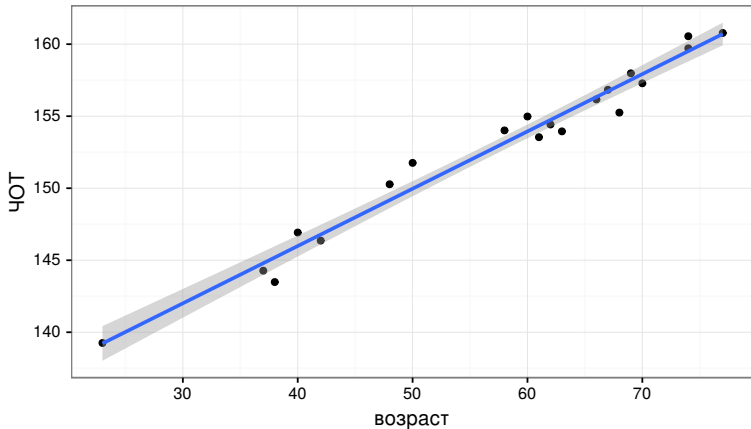


```
fit.f <- lm(pitch~age, dff)
plot(dff$age, dff$pitch)
lines(dff$age, fit.f$fitted.values)
```

презентация доступна: <http://goo.gl/AdqRQl>

Линейная регрессия: визуализация, ggplot2

ЧОТ vs. возраст: мужчины



```
library(ggplot2)
ggplot(dfm, aes(x=age, y = pitch))+
  geom_point()+
  geom_smooth(method = "lm")
```

уже встроена линейная регрессия

презентация доступна: <http://goo.gl/AdqRQl>

Линейная регрессия: доверительный интервал

Так как регрессия строится по выборочным данным, мы не знаем, как бы проходила линия, если бы мы взяли другую выборку. Т. е. значения коэффициентов β_0 и β_1 , вычисленные функцией `lm()`, являются лишь некоторым приближением к коэффициентам регрессии, которая бы описывала параметры генеральной совокупности.

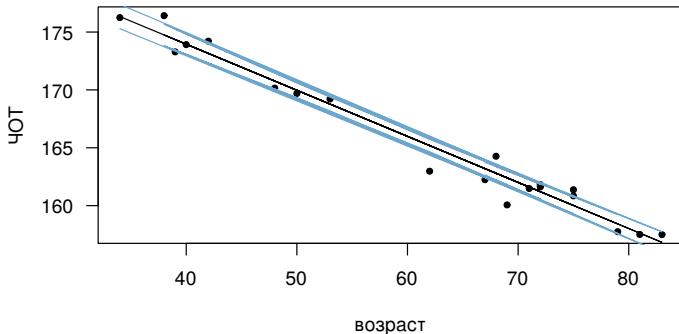
В связи с этим стоит строить доверительный интервал регрессии, что позволяет делать команда `predict()`.

```
head(predict(fit.f, interval = "conf"))
```

	fit	lwr	upr
1	158.4218	157.6118	159.2319
2	162.8020	162.2180	163.3859
3	162.4038	161.8052	163.0023
4	163.2002	162.6292	163.7711
5	161.6074	160.9752	162.2396
6	161.2092	160.5582	161.8602

Линейная регрессия: CI, R-base

ЧОТ vs. возраст: женщины



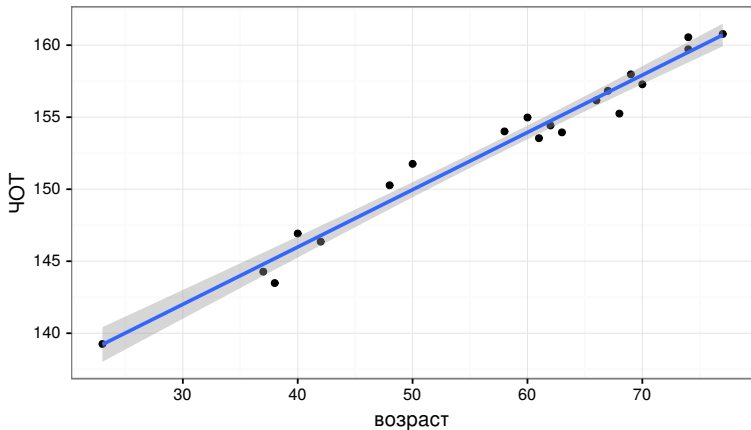
```
fit.f <- lm(pitch~age, dff)
pred.f <- predict(fit.f, interval = "conf")
plot(dff$age, dff$pitch)
lines(dff$age, pred.f[,1])
lines(dff$age, pred.f[,2], col = "skyblue3")
lines(dff$age, pred.f[,3], col = "skyblue3")
```

строим модель
строим границы

линия регрессии
нижняя гр. CI
верхняя гр. CI

Линейная регрессия: CI, ggplot2

ЧОТ vs. возраст: мужчины



```
library(ggplot2)
ggplot(dfm, aes(x=age, y = pitch))+
  geom_point()+
  geom_smooth(method = "lm")
```

уже встроен CI

презентация доступна: <http://goo.gl/AdqRQl>

dummy-переменные

В регрессионные модели можно включить и категориальные предикторы. Для этого вводятся фиктивные переменные (dummy variables), принимающие значение либо 1, либо 0. При этом фиктивных переменных должно быть на одну меньше, чем значений, которые принимает категориальные переменные.

связи
переменных

корреляции
aggregate()

регрессии

линейная
регрессия

dummy-
переменные

множественная
регрессия

сравнение
моделей

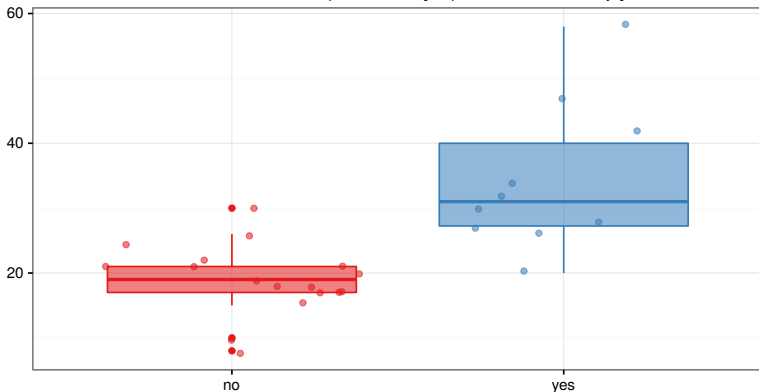
перебор
моделей

другие
регрессии

dummy-переменные

Проанализируем **данные**, содержащих выборку языков с указанием количества согласных и наличия в данном языке абруптивных согласных. На графике представлен результат (можно посмотреть **более интерактивный вариант**):

Количество согласных (по LAPSyD) ~ наличие абруптивных



dummy-переменные

```
m <- lm(n.cons.lapsyd ~ ejectives, data = ejectives)
summary(m)
```

Call:

```
lm(formula = n.cons.lapsyd ~ ejectives, data = ejectives)
```

Residuals:

Min	1Q	Median	3Q	Max
-14.400	-4.229	-1.059	2.441	23.600

Coefficients:

	Estimate	Std. Error	t	value	Pr(> t)
(Intercept)	19.059	1.953	9.758	5.25e-10	***
<u>ejectivesyes</u>	<u>15.341</u>	<u>3.209</u>	<u>4.780</u>	<u>6.59e-05</u>	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.053 on 25 degrees of freedom

Multiple R-squared: 0.4775, Adjusted R-squared: 0.4566

F-statistic: 22.85 on 1 and 25 DF, p-value: 6.588e-05

dummy-переменные

Естественно в модели переменная может принимать больше двух значений.

Регрессионное моделирование, в котором предсказываемая переменная — количественная, а все предикторы — категориальные, называют **ANOVA (Analysis of Variance)**.

А что если количество предикторов больше двух?

- попарное сравнение pairs()
- несколько регрессий
- множественная регрессия
- ...

связи
переменных

корреляции
aggregate()

регрессии

линейная
регрессия

dumpty-
переменные

множественная
регрессия

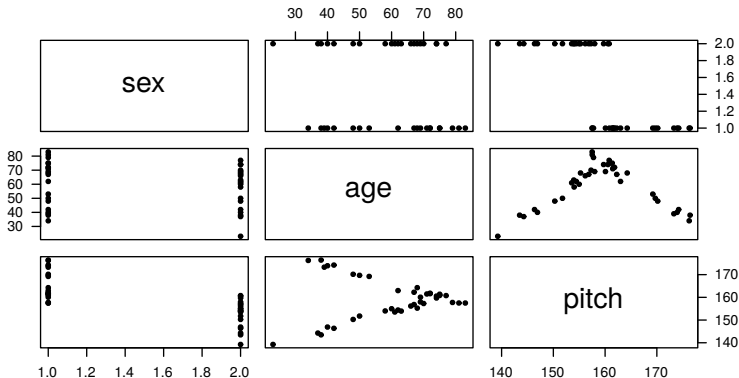
сравнение
моделей

перебор
моделей

другие
регрессии

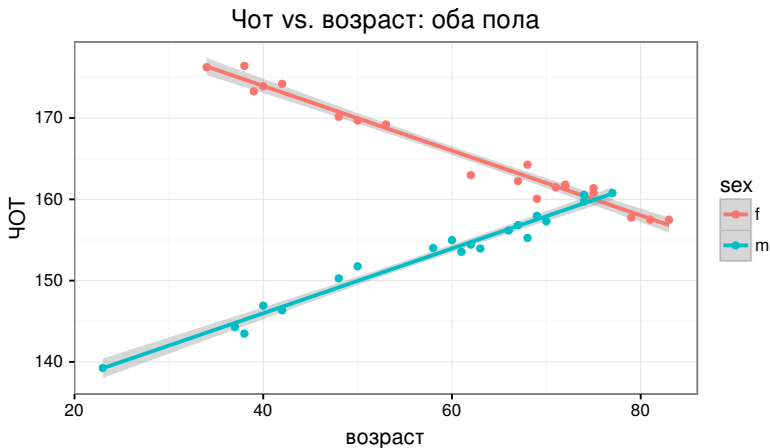
А что если количество предикторов больше двух?

- попарное сравнение `pairs()`



А что если количество предикторов больше двух?

- несколько регрессий



Множественная регрессия

Естественным обобщением линейной регрессии является множественная регрессия, в которой имеется не один предиктор, а несколько:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_k \cdot x_k + \varepsilon_i,$$

где x_1, x_2, \dots, x_k — предиктор, β_0 — свободный член, $\beta_1, \beta_2, \dots, \beta_k$ — коэффициенты регрессии, ε_i — средняя ошибка.

Как сравнить модели?

Для сравнения моделей используют несколько параметров:

- p-value модели, и p-value коэффициентов регрессии
- R^2 и adjusted R^2 — доля дисперсии, объясняемая моделью
- AIC — информационный критерий Акаике (чем меньше, тем лучше)
- BIC — байесовский информационный критерий Шварца (чем меньше, тем лучше)
- результаты перекрестной проверки (cross-validation) — существует много разных техник

В работе [Stone 1977], видимо, показано, что AIC и некоторые методы перекрестной проверки асимптотически эквивалентны.

Как сравнить модели? p-value

Для сравнения моделей используют несколько параметров:

- p-value модели, и p-value коэффициентов регрессии.

Call:

lm(formula = pitch ~ age, data = dfm) # формула, вдруг забыли

Residuals: # распределение остатков

Min	1Q	Median	3Q	Max
-1.87771	-0.54867	0.05222	0.88251	1.79433

Coefficients: # коэффициенты модели

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	130.07015	0.90750	143.33	< 2e-16	*** # β_0
age	0.39790	0.01534	25.94	1.04e-15	*** # β_1

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9985 on 18 degrees of freedom

Multiple R-squared: 0.974, Adjusted R-squared: 0.9725

F-statistic: 673 on 1 and 18 DF, p-value: 1.036e-15

Как сравнить модели? R^2 и adjusted R^2

Для сравнения моделей используют несколько параметров:

- R^2 и adjusted R^2 — доля дисперсии, объясняемая моделью.

Call:

`lm(formula = pitch ~ age, data = dfm)` # формула, вдруг забыли

Residuals: # распределение остатков

Min	1Q	Median	3Q	Max
-1.87771	-0.54867	0.05222	0.88251	1.79433

Coefficients: # коэффициенты модели

	Estimate	Std. Error	t value	Pr(> t)		
(Intercept)	130.07015	0.90750	143.33	< 2e-16	***	# β_0
age	0.39790	0.01534	25.94	1.04e-15	***	# β_1

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9985 on 18 degrees of freedom

Multiple R-squared: 0.974, Adjusted R-squared: 0.975

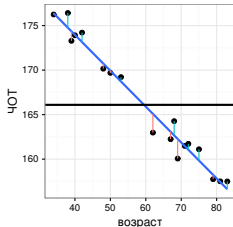
F-statistic: 673 on 1 and 18 DF, p-value: 1.036e-15

Как сравнить модели? R^2 и adjusted R^2

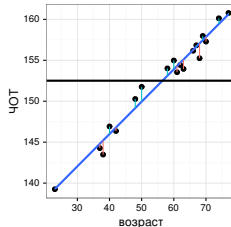
Для сравнения моделей используют несколько параметров:

- R^2 и adjusted R^2 — доля дисперсии, объясняемая моделью

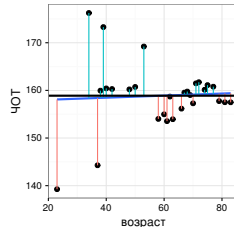
Чот vs. возраст: женщины



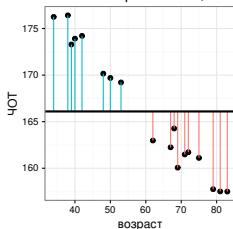
Чот vs. возраст: мужчины



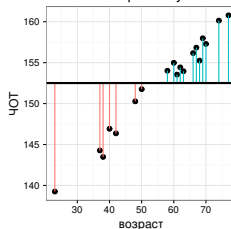
Чот vs. возраст: оба пола



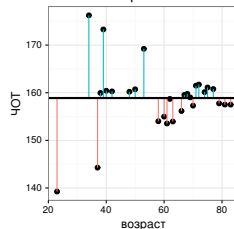
Чот vs. возраст: женщины



Чот vs. возраст: мужчины



Чот vs. возраст: оба пола



Как сравнить модели? AIC, BIC

Для сравнения моделей используют несколько параметров:

- AIC — информационный критерий Акаике (чем меньше, тем лучше)
AIC()
- BIC — байесовский информационный критерий Шварца (чем меньше, тем лучше)
BIC()

... надо сказать, разных вариантов этих критериев много, AIC и BIC — самые популярные.

Перебор моделей

Существует несколько стратегий перебора моделей:

- построить модель из всех предикторов, а потом "выкидывать" не значимые
- строить модель снизу вверх добавляя по одному предиктору, выясняя какие значимы, а какие нет

Если много предикторов, то возникает желание перебрать все возможные варианты и узнать, в какой модели лучше скорректированный R^2 , AIC и BIC. Для этого, естественно уже написаны готовые функции (см. функцию `regsubsets` в пакете `leaps`, `bestglm` в пакете `bestglm` или пакет `FWDselect` и его описание [Sestelo et al. 2016]). Однако многие высказывают недовольство такой стратегией выискивания лучшей из моделей, построенных на одних и тех же данных, приравнивая ее к *data fishing*.

В данной презентации не рассказано о...

- полиномиальной регрессии,
- нелинейной регрессии,
- логистической регрессии,
- гребневой и лассо-регрессии,
- и, наверное, о массе всего другого.

связи
переменных

корреляции
aggregate()

регрессии

линейная
регрессия

dummy-
переменные

множественная
регрессия

сравнение
моделей

перебор
моделей

другие
регрессии

связи
переменных
корреляции
aggregate()

регрессии

линейная
регрессия

dummy-
переменные

множественная
регрессия

сравнение
моделей

перебор
моделей

другие
регрессии

Спасибо за внимание

Пишите письма

agricolamz@gmail.com

Список литературы

- Cedergren, H. C. J. (1973). *The interplay of social and linguistic factors in Panama*. Cornell University.
- Sestelo, M., N. M. Villanueva, and J. Roca-Pardiñas (2016). FWDselect: An R package for selecting variables in regression models. *Discussion Papers in Statistics and Operation Research* 13(02).
- Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44--47.