Линейная регрессия

Линейные модели…

Вадим Хайтов, Марина Варфоломеева

Мы рассмотрим

- Базовые идеи корреляционного анализа
- Проблему двух статистических подходов: "Тестирование гипотез vs. построение моделей"
- Разнообразие статистических моделей
- Основы регрессионного анализа

Вы сможете

- Оценить взаимосвязь между измеренными величинами
- Объяснить что такое линейная модель
- Формализовать запись модели в виде уравнения
- Подобрать модель линейной регрессии
- Протестировать гипотезы о наличии зависимости при помощи t-критерия или F-критерия
- Оценить предсказательную силу модели

Знакомимся с даными

Пример: IQ и размеры мозга

Зависит ли уровень интеллекта от размера головного мозга? (Willerman et al. 1991)

- Было исследовано 20 девушек и 20 молодых людей

Пример взят из работы: Willerman, L., Schultz, R., Rutledge, J. N., and Bigler, E. (1991), "In Vivo Brain Size and Intelligence," Intelligence, 15, 223-228.
Данные представлены в библиотеке *"The Data and Story Library"* http://lib.stat.cmu.edu/DASL/

Знакомство с данными

Посмотрим на датасет

```
brain <- read.csv("data/IQ_brain.csv", header = TRUE)
head(brain)
```

```
##   Gender FSIQ VIQ PIQ Weight Height MRINACount
## 1 Female  133 132 124    118   64.5     816932
## 2   Male  140 150 124     NA   72.5    1001121
## 3   Male  139 123 150    143   73.3    1038437
## 4   Male  133 129 128    172   68.8     965353
## 5 Female  137 132 134    147   65.0     951545
## 6 Female   99  90 110    146   69.0     928799
```

Есть ли пропущенные значения?

```
sum(!complete.cases(brain))
```

```
## [1] 2
```

Где пропущенные значения?

Где именно?

```
sapply(brain, function(x) sum(is.na(x)))
```

```
##     Gender       FSIQ        VIQ        PIQ     Weight     Height
##          0          0          0          0          2          1
## MRINACount
##          0
```

Что это за случаи?

```
brain[!complete.cases(brain), ]
```

```
##    Gender FSIQ VIQ PIQ Weight Height MRINACount
## 2    Male  140 150 124     NA   72.5    1001121
## 21   Male   83  83  86     NA     NA     892420
```

Каков объем выборки

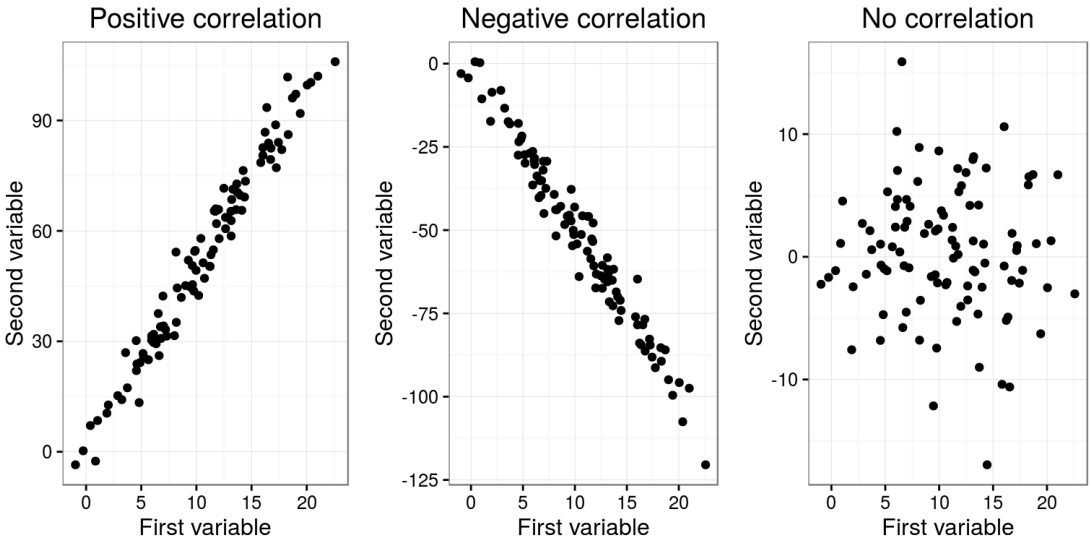
```
nrow(brain) ## Это без учета пропущенных значений
```

```
## [1] 40
```

- поиск взаимосвязи величин и создание базы для предсказания неизвестного на основе имеющихся данных

Корреляционный анализ

Вспомним:


Positive correlation
Second variable
First variable
0
30
60
90
0
5
10
15
20
Negative correlation
Second variable
First variable
0
-25
-50
-75
-100
-125
0
5
10
15
20
No correlation
Second variable
First variable
10
0
-10
0
5
10
15
20

Коэффициенты корреляции и условия их применимости

Коэффициент	Функция	Особенности применения
Коэф. Пирсона	`cor(x,y,method="pearson")`	Оценивает связь двух нормально распределенных величин. Выявляет только линейную составляющую взаимосвязи.
Ранговые коэффициенты (коэф. Спирмена, Кендалла)	`cor(x,y,method="spirman")` `cor(x,y,method="kendall")`	Не зависят от формы распределения. Могут оценивать связь для любых монотонных зависимостей.

Оценка достоверности коэффициентов корреляции

- Коэффициент корреляции - это статистика, значение которой описывает степень взаимосвязи двух сопряженных переменных. Следовательно применима логика статистического критерия.
- Нулевая гипотеза $H_0 : r = 0$
- Бывают двусторонние $H_a : r \neq 0$ и односторонние критерии $H_a : r > 0$ или $H_a : r < 0$
- Ошибка коэффициента Пирсона: $SE_r = \sqrt{\frac{1-r^2}{n-2}}$
- Стандартизованная величина $t = \frac{r}{SE_r}$ подчиняется распределению Стьюдента с параметром $df = n - 2$
- Для ранговых коэффициентов существует проблема "совпадающих рангов" (tied ranks), что приводит к приблизительной оценке r и приблизительной оценке уровня значимости.
- Достоверность коэффициента корреляции можно оценить пермутационным методом

Задание

- Определите силу и направление связи между всеми парами исследованных признаков
- Постройте точечную диаграмму, отражающую взаимосвязь между результатами IQ-теста (PIQ) и размером головного мозга (MRINACount)
- Оцените достоверность значения коэффициента корреляции Пирсона между этими двумя переменными

Hint 1: Обратите внимание на то, что в датафрейме есть пропущенные значения. Изучите, как работают с `NA` функции, вычисляющие коэффициенты корреляции.

Hint 2 Для построения точечной диаграммы вам понадобится `geom_point()`

Решение

```
cor(brain[, 2:6], use = "pairwise.complete.obs")
```

```
##           FSIQ     VIQ      PIQ   Weight  Height
## FSIQ    1.0000  0.9466  0.93413 -0.05148 -0.0860
## VIQ     0.9466  1.0000  0.77814 -0.07609 -0.0711
## PIQ     0.9341  0.7781  1.00000  0.00251 -0.0767
## Weight -0.0515 -0.0761  0.00251  1.00000  0.6996
## Height -0.0860 -0.0711 -0.07672  0.69961  1.0000
```

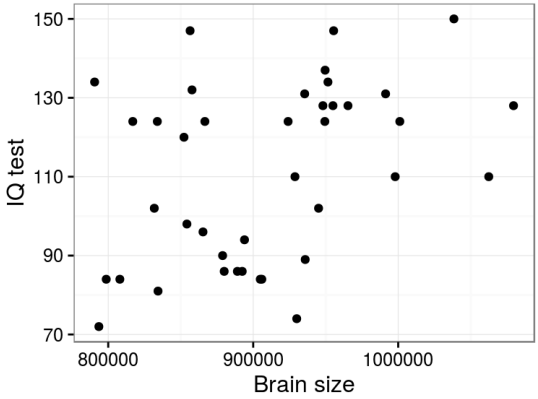
Решение

```
cor.test(brain$PIQ, brain$MRINACount, method = "pearson", alternative = "two.side
```

```
##
##  Pearson's product-moment correlation
##
## data:  brain$PIQ and brain$MRINACount
## t = 3, df = 40, p-value = 0.01
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0856 0.6232
## sample estimates:
##   cor
## 0.387
```

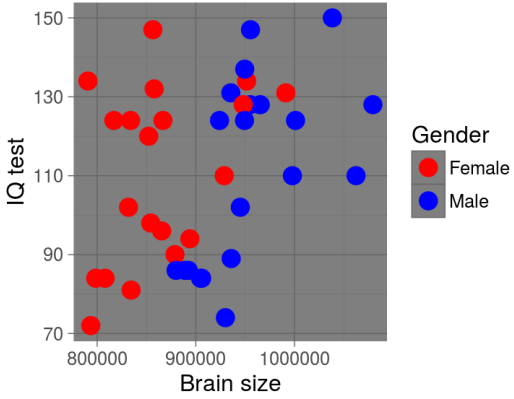
Решение

```
pl_brain <- ggplot(brain,
                   aes(x = MRINACount, y = PIQ)) +
  geom_point() +
  xlab("Brain size") +
  ylab("IQ test")
pl_brain
```



Решение

```
pl_brain + theme_dark() +
  geom_point(aes(color = Gender), size = 4) +
  scale_color_manual(values =  c("red", "blue"))
```



Два подхода к исследованию:
Тестирование гипотезы
VS
Построение модели

- Проведя корреляционный анализ, мы лишь ответили на вопрос "Существует ли статистически значимая связь между величинами?"
- Сможем ли мы, используя это знание, *предсказать* значения одной величины, исходя из знаний другой?

Тестирование гипотезы VS построение модели

- Простейший пример
- Между путем, пройденным автомобилем, и временем, проведенным в движении, несомненно есть связь. Хватает ли нам этого знания?
- Для расчета величины пути в зависимости от времени необходимо построить модель: $S = Vt$, где S - зависимая величина, t - независимая переменная, V - параметр модели.
- Зная параметр модели (скорость) и значение независимой переменной (время), мы можем рассчитать (*смоделировать*) величину пройденного пути

Какие бывают модели?

Линейные и нелинейные модели

Линейные модели

$$y = b_0 + b_1 x$$

$$y = b_0 + b_1 x_1 + b_2 x_2$$

Нелинейные модели

$$y = b_0 + b_1^x$$

$$y = b_0^{b_1 x_1 + b_2 x_2}$$

Простые и многокомпонентные (множественные) модели

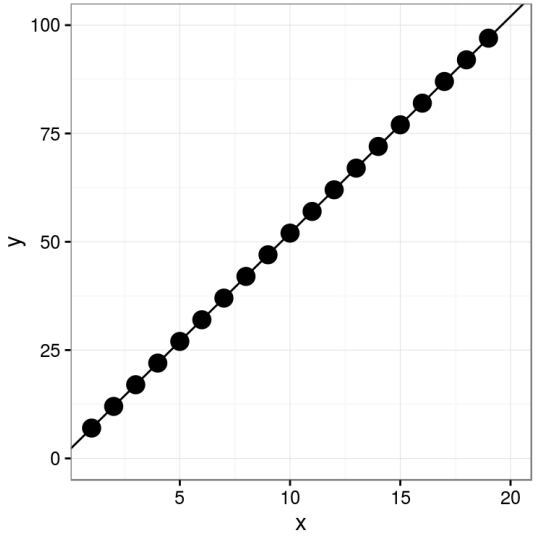
- Простая модель

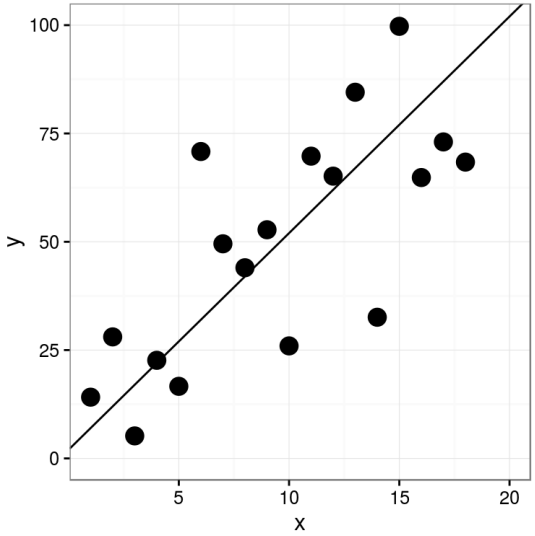
$$y = b_0 + b_1 x$$

- Множественная модель

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n$$

Детерминистские и стохастические модели





Модель: $y_i = 2 + 5x_i$

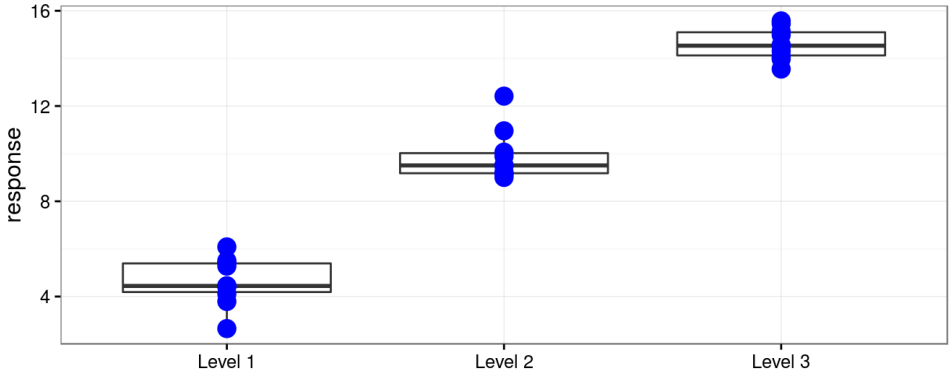
Два параметра: угловой коэффициент (slope) $b_1 = 5$; свободный член (intercept) $b_0 = 2$

Чему равен y при $x = 10$?

Модель: $= 2 + 5 \quad +$

Появляется дополнительный член Он вводит в модель влияние неучтенных моделью факторов. Обычно считают, что $\epsilon \in N(0, \quad)$

Модели с дискретными предикторами



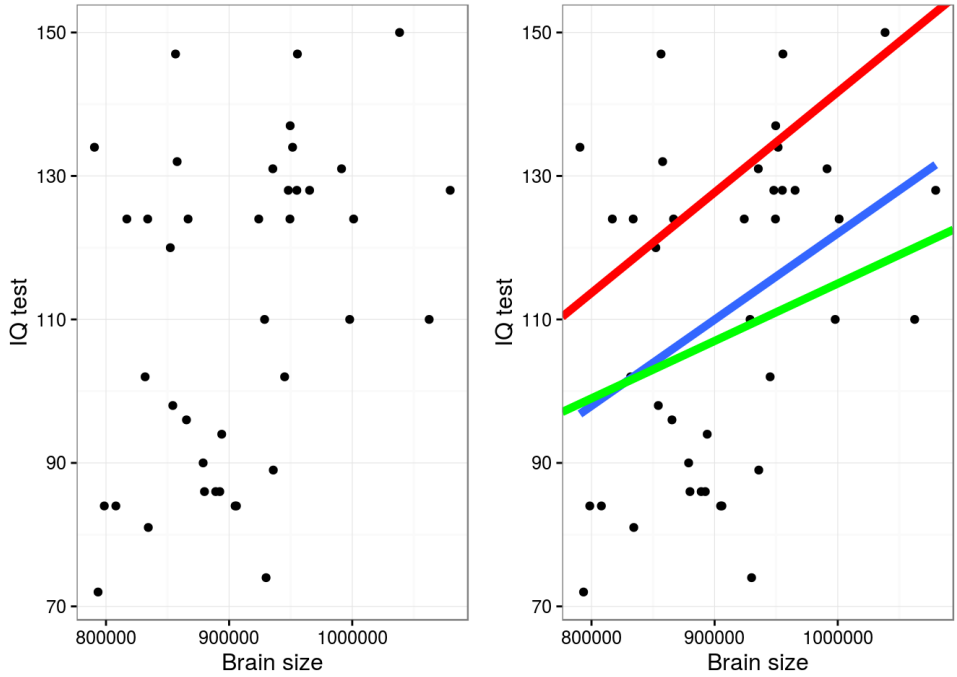
Модель для данного примера имеет такой вид

$$response = 4.6 + 5.3I_{Level2} + 9.9I_{Level3}$$

I_i - dummy variable

Модель для зависимости величины IQ от размера головного мозга

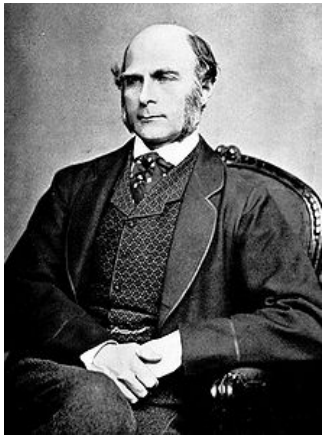
Какая из линий "лучше" описывает облако точек?



"Essentially, all models are wrong,
but some are useful"
(Georg E. P. Box)

Найти оптимальную модель позволяет регрессионный анализ

Происхождение термина "регрессия"



Френсис Галтон (Francis Galton)

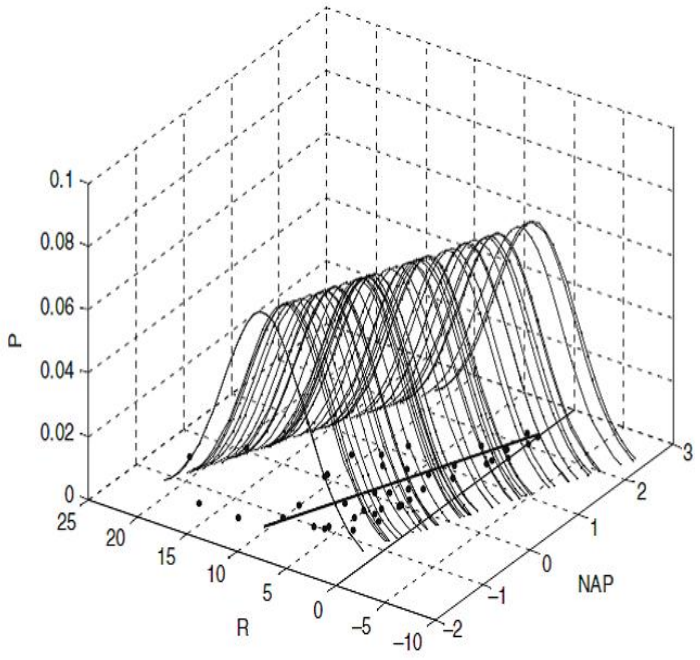
"the Stature of the adult offspring … [is] … more mediocre than the stature of their Parents" (цит. по `Legendre & Legendre, 1998`)

Рост *регрессирует* (возвращается) к популяционной средней
Угловой коэффициент в зависимости роста потомков от роста родителей- *коэффициент регресси*

Подбор линии регрессии проводится с помощью двух методов

- С помощью метода наименьших квадратов (Ordinary Least Squares) - используется для простых линейных моделей
- Через подбор функции максимального правдоподобия (Maximum Likelihood) - используется для подгонки сложных линейных и нелинейных моделей.

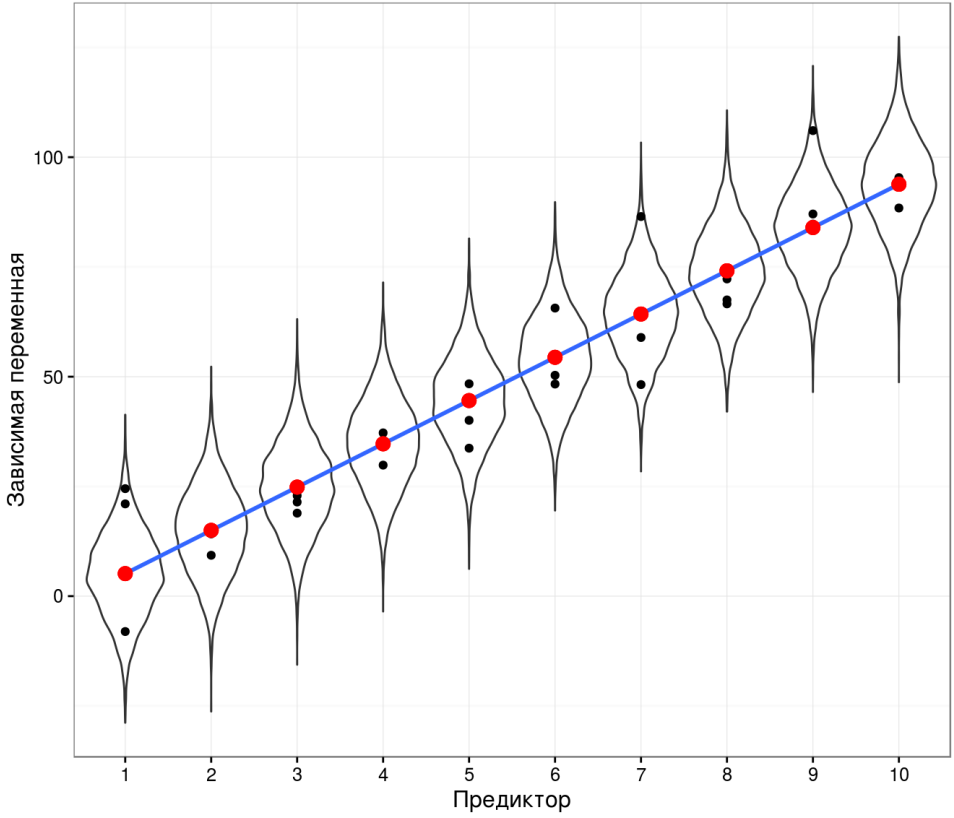
Кратко о методе максимального правдоподобия



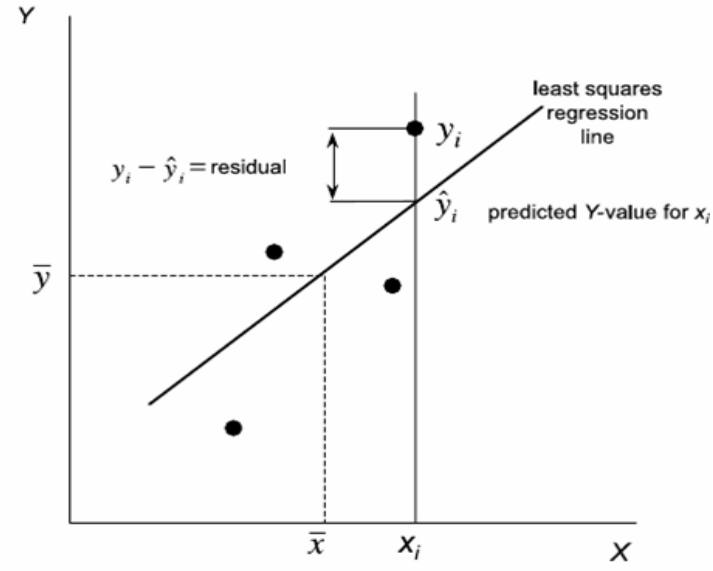
(из кн. Zuur et al., 2009, стр. 19)

Кратко о методе максимального правдоподобия

Симулированный пример с использованием `geom_violin()`



Метод наименьших квадратов



Остатки (Residuals):

$$= \quad -$$

Линия регрессии (подобранная модель) - это та линия, у которой $\sum$ минимальна.

(из кн. Quinn, Keough, 2002, стр. 85)

Подбор модели методом наименьших квадратов с помощью функци `lm()`

```
fit <- lm(formula, data)
```

Модель записывается в виде формулы

Модель	Формула

Простая линейная регрессия $\hat{y_i} = b_0 + b_1 x_i$	`Y ~ X`
	`Y ~ 1 + X`
	`Y ~ X + 1`

Простая линейная регрессия (без b_0, "no intercept") $\hat{y_i} = b_1 x_i$	`Y ~ -1 + X` `Y ~ X - 1`
Уменьшенная простая линейная регрессия $\hat{y_i} = b_0$	`Y ~ 1` `Y ~ 1 - X`
Множественная линейная регрессия $\hat{y_i} = b_0 + b_1 x_i + b_2 x_2$	`Y ~ X1 + X2`

Подбор модели методом наименьших квадратов с помощью функци `lm()`

```
fit <- lm(formula, data)
```

Элементы формул для записи множественных моделей

Элемент формулы	Значение
`:`	Взаимодействие предикторов `Y ~ X1 + X2 + X1:X2`

`*`	Обозначает полную схему взаимодействий
	`Y ~ X1 * X2 * X3`
	аналогично
	`Y ~ X1 + X2 + X3+ X1:X2 + X1:X3 + X2:X3 + X1:X2:X3`

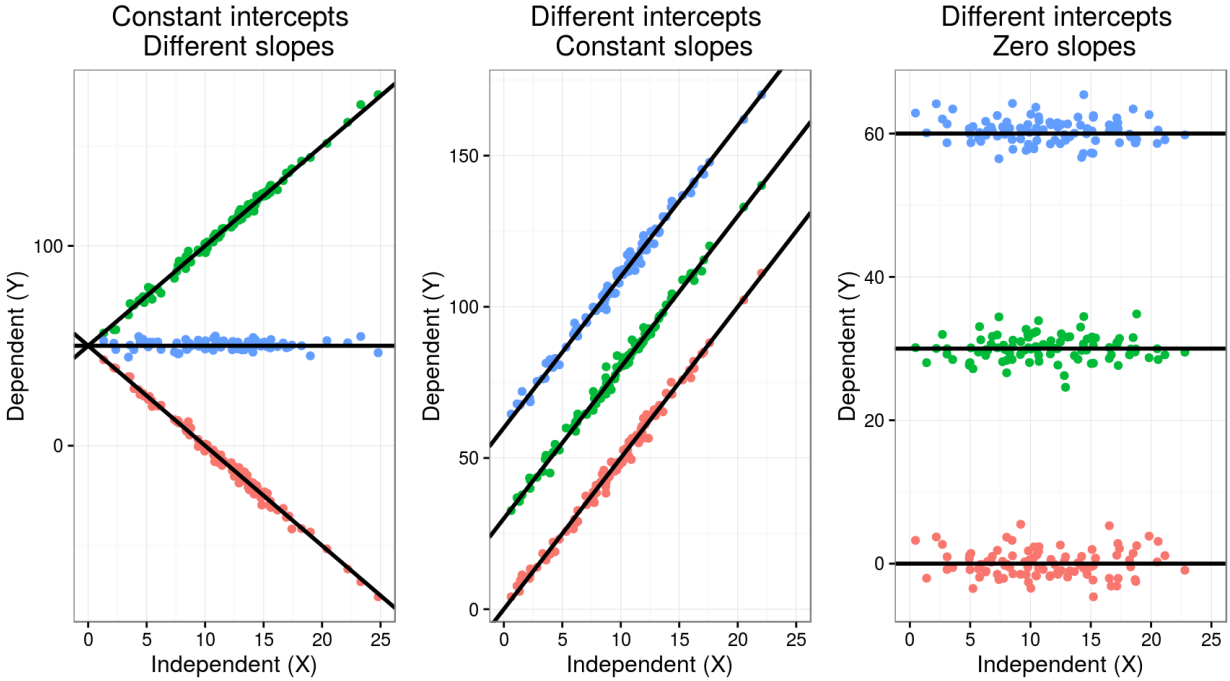
`.`	`Y ~ .` В правой части формулы записываются все переменные из датафрейма, кроме `Y`

Подберем модель, наилучшим образом описывающую зависимость результатов IQ-теста от размера головного мозга

```
brain_model <- lm(PIQ ~ MRINACount, data = brain)
brain_model
```

```
##
## Call:
## lm(formula = PIQ ~ MRINACount, data = brain)
##
## Coefficients:
## (Intercept)   MRINACount
##     1.74376      0.00012
```

Как трактовать значения параметров регрессионной модели?



Как трактовать значения параметров регрессионной модели?

- Угловой коэффициент (*slope*) показывает на сколько *единиц* изменяется предсказанное значение $\hat{y}$ при изменении на *одну единицу* значения предиктора (x)

- Свободный член (*intercept*) - величина во многих случаях не имеющая "смысла", просто поправочный коэффициент, без которого нельзя вычислить $\hat{y}$. *NB!* В некоторых линейных моделях он имеет смысл, например, значения $\hat{y}$ при $x = 0$.

- Остатки (*residuals*) - характеризуют влияние неучтенных моделью факторов.

Вопросы:

1. Чему равны угловой коэффициент и свободный член полученной модели `brain_model`?
2. Какое значение IQ-теста предсказывает модель для человека с объемом мозга равным 900000
3. Чему равно значение остатка от модели для человека с порядковым номером 10?

Ответы

```
coefficients(brain_model) [1]
```

```
## (Intercept)
##        1.74
```

```
coefficients(brain_model) [2]
```

```
## MRINACount
##    0.00012
```

Ответы

```
as.numeric(coefficients(brain_model) [1] + coefficients(brain_model) [2] * 900000
```

```
## [1] 110
```

Ответы

```
brain$PIQ[10] - fitted(brain_model)[10]
```

```
##   10
## 30.4
```

```
residuals(brain_model)[10]
```

```
##   10
## 30.4
```

Углубляемся в анализ модели: функция `summary()`

```
summary(brain_model)
```

```
##
## Call:
## lm(formula = PIQ ~ MRINACount, data = brain)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -39.6  -17.9   -1.6   17.0   42.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.7437570 42.3923825    0.04    0.967
## MRINACount   0.0001203  0.0000465    2.59    0.014 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21 on 38 degrees of freedom
## Multiple R-squared:  0.15,   Adjusted R-squared:  0.127
## F-statistic: 6.69 on 1 and 38 DF,  p-value: 0.0137
```

Что означают следующие величины?

```
Estimate
Std. Error
t value
Pr(>|t|)
```

Оценки параметров регрессионной модели

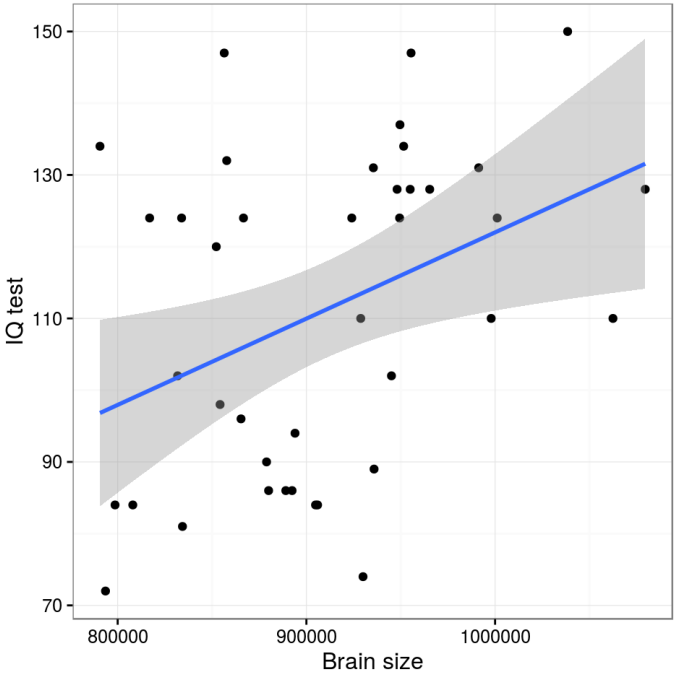
Параметр	Оценка	Стандартная ошибка
β_1 **Slope**	$b_1 = \frac{\sum_{i=1}^{n}[(x_i-\bar{x})(y_i-\bar{y})]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ или проще $b_0 = r\frac{sd_y}{sd_x}$	$SE_{b_1} = \sqrt{\frac{MS_e}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$
β_0 **Intercept**	$b_0 = \bar{y} - b_1\bar{x}$	$SE_{b_0} = \sqrt{MS_e[\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}]}$
ϵ_i	$e_i = y_i - \hat{y_i}$	$\approx \sqrt{MS_e}$

Для чего нужны стандартные ошибки?

- Они нужны, поскольку мы *оцениваем* параметры по *выборке*
- Они позволяют построить доверительные интервалы для параметров
- Их используют в статистических тестах

Графическое представление результатов

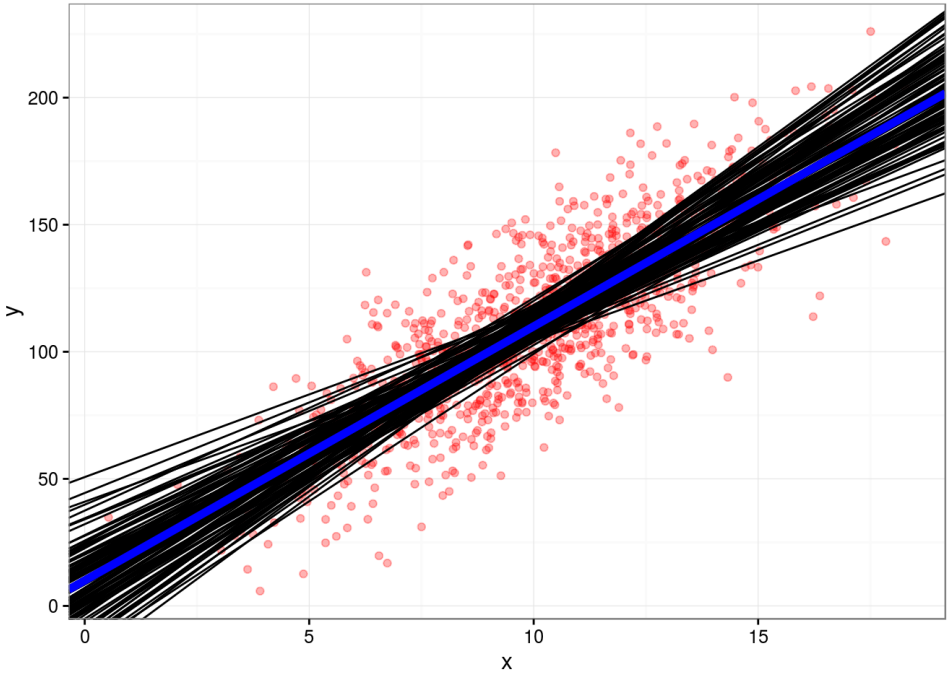
```
pl_brain + geom_smooth(method="lm")
```



Доверительная зона регрессии. В ней с 95% вероятностью лежит регрессионная прямая, описывающая связь в генеральной совокупности.
Возникает из-за неопределенности оценок коэффициентов модели, вследствие выборочного характера оценок.

Симулированный пример

Линии регрессии, полученные для 100 выборок (по 20 объектов в каждой), взятых из одной и той же генеральной совокупности



Доверительные интервалы для коэффициентов уравнения регрессии

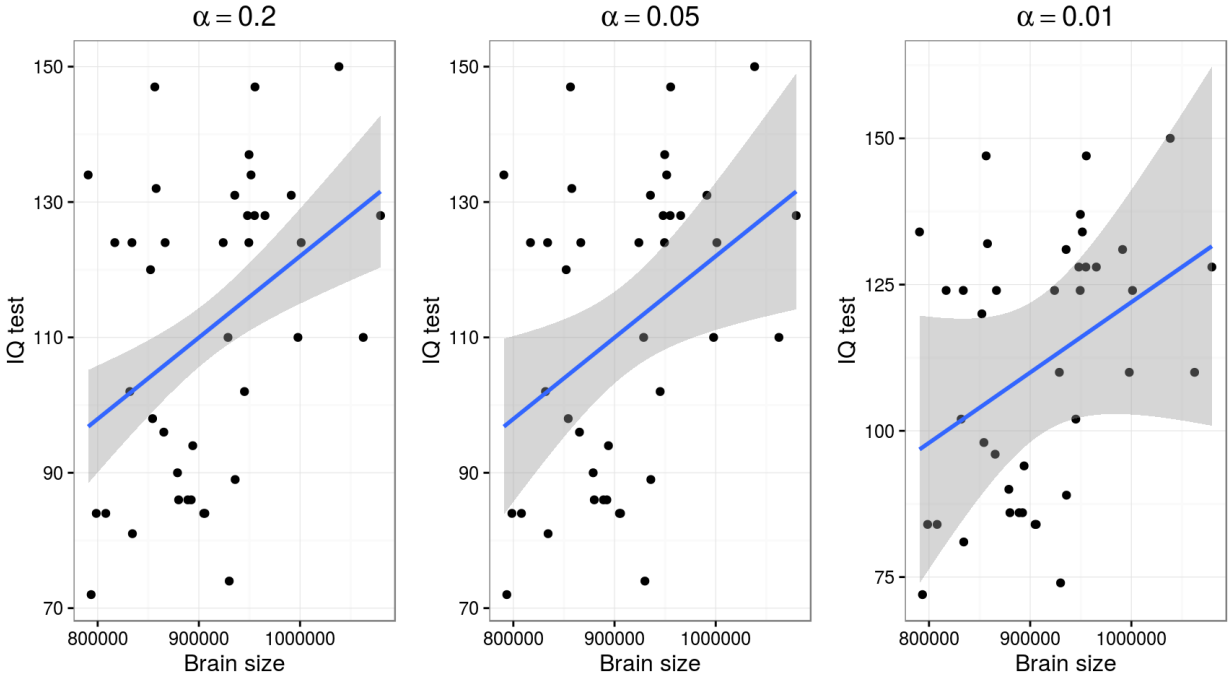
```
coef(brain_model)
```

```
## (Intercept)  MRINACount
##     1.74376     0.00012
```

```
confint(brain_model)
```

```
##                   2.5 %    97.5 %
## (Intercept) -84.0751348 87.562649
## MRINACount    0.0000261  0.000214
```

Для разных α можно построить разные доверительные интервалы



Важно!

Если коэффициенты уравнения регрессии - лишь приблизительные оценки параметров, то предсказать значения зависимой переменной можно только *с некоторой вероятностью*.

Какое значение IQ можно ожидать у человека с размером головного мозга 900000?

```
newdata <- data.frame(MRINACount = 900000)

predict(brain_model, newdata, interval = "prediction", level = 0.95, se = TRUE)$f

##   fit  lwr upr
## 1 110 66.9 153
```

- При размере мозга 900000 среднее значение IQ будет, с вероятностью 95%, находиться в интервале от 67 до 153.

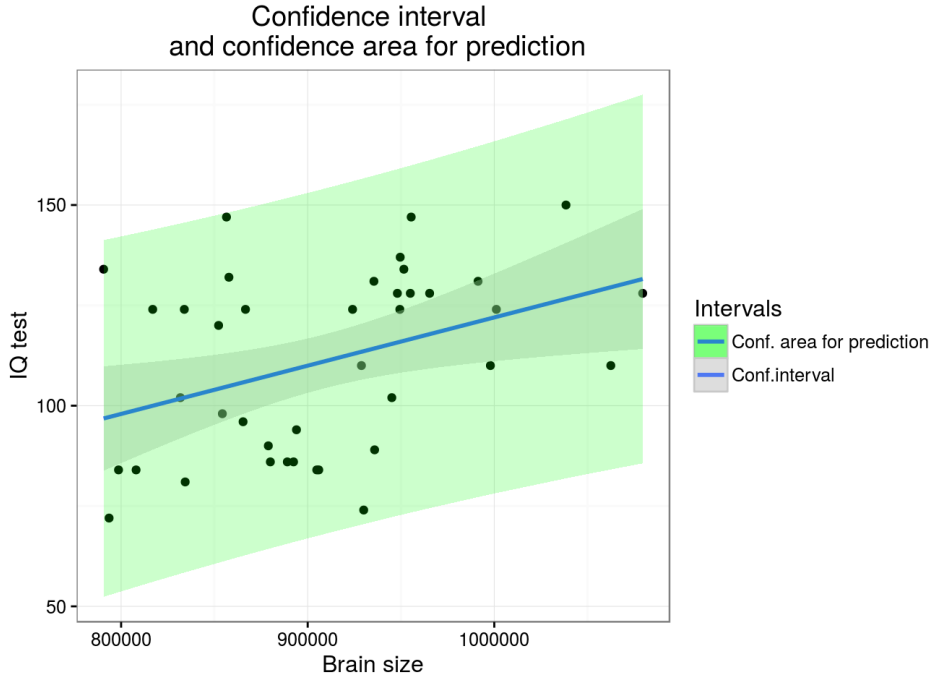
Отражаем на графике область значений, в которую попадут 95% предсказанных величин IQ

Подготавливаем данные

```
brain_predicted <- predict(brain_model, interval="prediction")
brain_predicted <- data.frame(brain, brain_predicted)
head(brain_predicted)
```

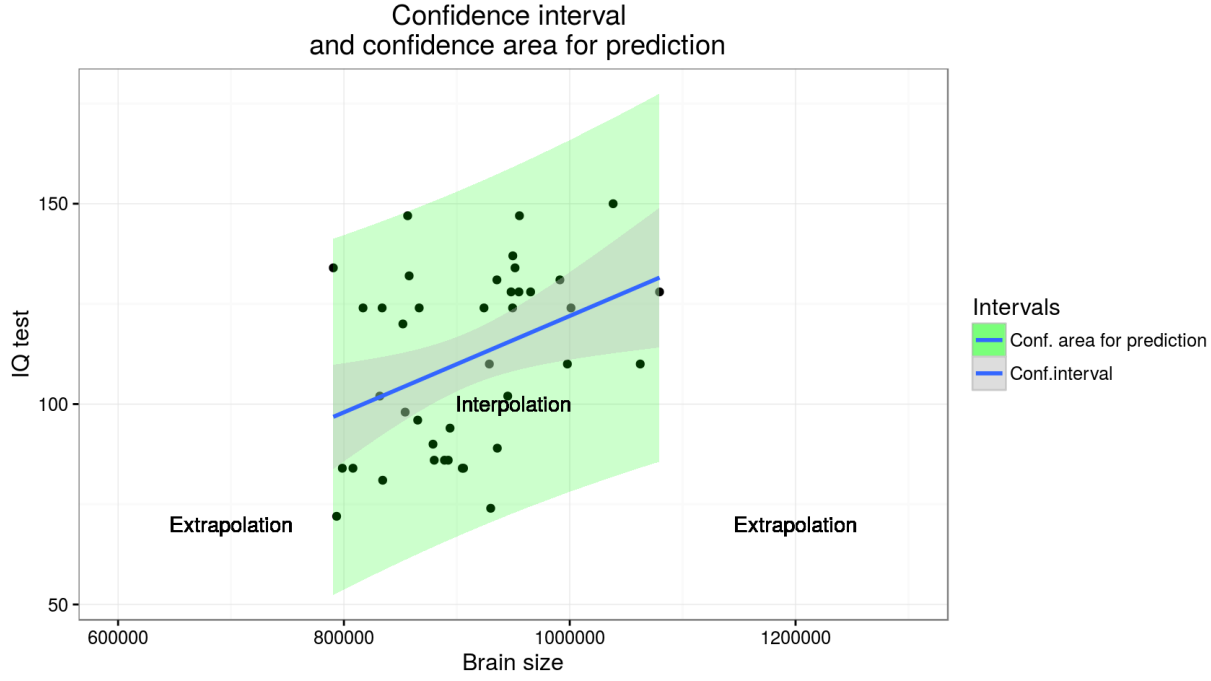
```
##   Gender FSIQ VIQ PIQ Weight Height MRINACount fit  lwr upr
## 1 Female  133 132 124    118   64.5     816932 100 56.1 144
## 2   Male  140 150 124     NA   72.5    1001121 122 78.2 166
## 3   Male  139 123 150    143   73.3    1038437 127 81.9 171
## 4   Male  133 129 128    172   68.8     965353 118 74.5 161
## 5 Female  137 132 134    147   65.0     951545 116 73.0 159
## 6 Female   99  90 110    146   69.0     928799 113 70.4 157
```

Отражаем на графике область значений, в которую попадут 95% предсказанных величин IQ



Важно!

Модель "работает" только в том диапазоне значений независимой переменной (x), для которой она построена (интерполяция). Экстраполяцию надо применять с большой осторожностью.



Итак, что означают следующие величины?

- `Estimate`
- Оценки праметров регрессионной модели
- `Std. Error`
- Стандартная ошибка для оценок
- Осталось решить, что такое `t value`, `Pr(>|t|)`

Тестирование гипотез с помощью линейных моделей

Два равноправных способа

- Проверка достоверности оценок коэффициента b_1 (t-критерий).
- Оценка соотношения описанной и остаточной дисперсии (F-критерий).

Тестирование гипотез с помощью t-критерия

Зависимость есть, если $\beta_1 \neq 0$

Нулевая гипотеза $H_0 : \beta = 0$

Тестируем гипотезу

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

Число степеней свободы: $df = n - 2$
>- Итак,
>- `t value` - Значение t-критерия
>- `Pr(>|t|)` - Уровень значимости

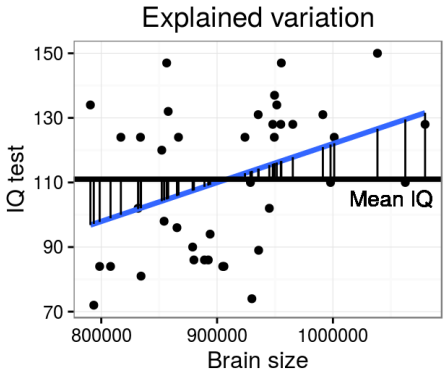
Зависит ли IQ от размера головного мозга?

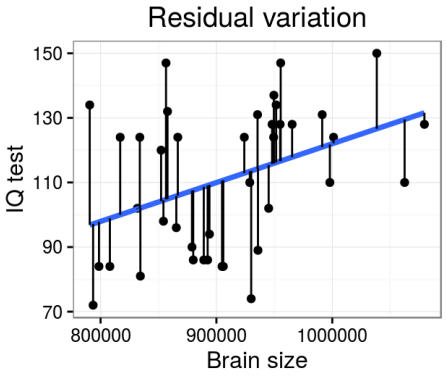
$$PIQ = 1.744 + 0.0001202 MRINACount$$

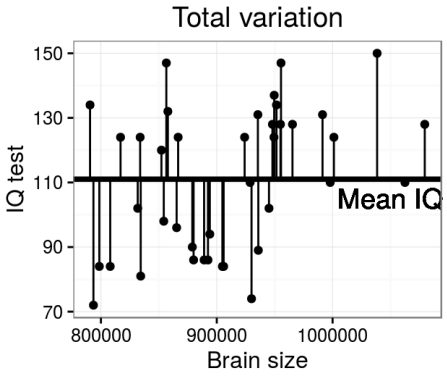
```
summary(brain_model)
```

```
##
## Call:
## lm(formula = PIQ ~ MRINACount, data = brain)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -39.6  -17.9   -1.6   17.0   42.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.7437570 42.3923825    0.04    0.967
## MRINACount   0.0001203  0.0000465    2.59    0.014 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21 on 38 degrees of freedom
## Multiple R-squared:  0.15,   Adjusted R-squared:  0.127
## F-statistic: 6.69 on 1 and 38 DF,  p-value: 0.0137
```

Тестирование гипотез с помощью F-критерия







Объясненная дисперсия

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

$$df_{Regression} = 1$$

$$MS_{Regression} = \frac{SS_{Regression}}{df}$$

Остаточная дисперсия

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

$$df_{Residual} = n - 2$$

$$MS_{Residual} = \frac{SS_{Residual}}{df_{Residual}}$$

Полная дисперсия

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

$$df_{Total} = n - 1$$

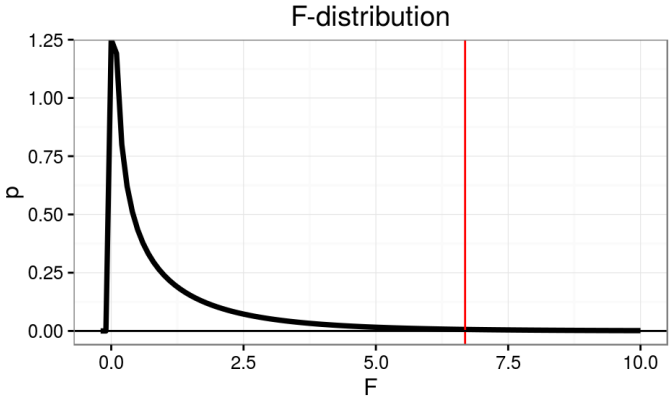
$$MS_{Total} = \frac{SS_{Total}}{df_{Total}}$$

F критерий

Если зависимости нет, то $MS_{Regression} = MS_{Residual}$

$$F = \frac{MS_{Regression}}{MS_{Residual}}$$

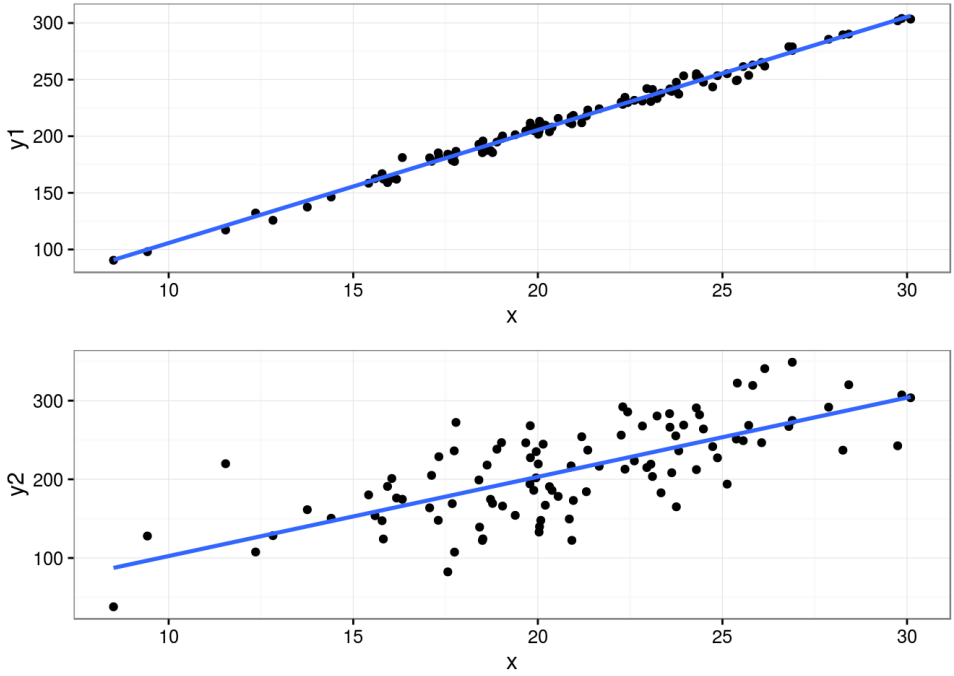
Логика та же, что и с t-критерием



Форма F-распределения зависит от двух параметров: $df_{Regression} = 1$ и $df_{Residual} = n - 2$

Оценка качества подгонки модели с помощью коэффициента детерминации

В чем различие между этми двумя моделями?



Оценка качества подгонки модели с помощью коэффициента детерминации

Коэффициент детерминации описывает какую долю дисперсии зависимой переменной объясняет модель

- $$R^2 = \frac{SS_{Regression}}{SS_{Total}}$$
- $$0 < R^2 < 1$$
- $$R^2 = r^2$$

Еще раз смотрим на результаты регрессионного анализа зависимости IQ от размеров мозга

```
summary(brain_model)
```

```
##
## Call:
## lm(formula = PIQ ~ MRINACount, data = brain)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -39.6  -17.9   -1.6   17.0   42.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.7437570 42.3923825    0.04    0.967
## MRINACount   0.0001203  0.0000465    2.59    0.014 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21 on 38 degrees of freedom
## Multiple R-squared:  0.15,   Adjusted R-squared:  0.127
## F-statistic: 6.69 on 1 and 38 DF,  p-value: 0.0137
```

Adjusted R-squared - скорректированный коэффициет детерминации

Применяется если необходимо сравнить две модели с разным количеством параметров

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k}$$

k - количество параметров в модели

Вводится штраф за каждый новый параметр

Как записываются результаты регрессионного анлиза в тексте статьи?

Мы показали, что связь между результатами теста на IQ описывается моделью вида IQ = 1.74 + 0.00012 MRINACount ($F_{1,38}$ = 6.686, p = 0.0136, R^2 = 0.149)

Summary

- Модель простой линейной регрессии $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- Параметры модели оцениваются на основе выборки
- В оценке коэффициентов регрессии и предсказанных значений существует неопределенность: необходимо вычислять доверительный интервал.
- Доверительные интервалы можно расчитать, зная стандартные ошибки.
- Гипотезы о наличии зависимости можно тестировать при помощи t- или F-теста. $(H_0 : \beta_1 = 0)$
- Качество подгонки модели можно оценить при помощи коэффициента детерминации (R^2)

Что почитать

- Гланц, 1999, стр. 221-244
- Open Intro to Statistics: Chapter 7. Introduction to linear regression, pp. 315-353.
- Quinn, Keough, 2002, pp. 78-110