Введение в регрессионный анализ. Часть 1

Линейные модели на R, осень 2014

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
* Проверить состоятельность модели при помощи t-критерия или F-критерия
* Оценить предсказательную силу модели

--- .segue

# Знакомимся с даными

---- &twocol

## Зависит ли уровень интеллекта от размера головного мозга?

\*\*\* =left

* Было исследовано 20 девушек и 20 молодых людей (праворуких, англоговорящих, не склонных к алкоголизму, наркомании и прочим смещающим воздействиям)
* У каждого индивида определяли биометрические параметры: вес, рост, размер головного мозга (количество пикселей на изображении ЯМР сканера)
* Интеллект был протестирован с помощью IQ тестов

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/>

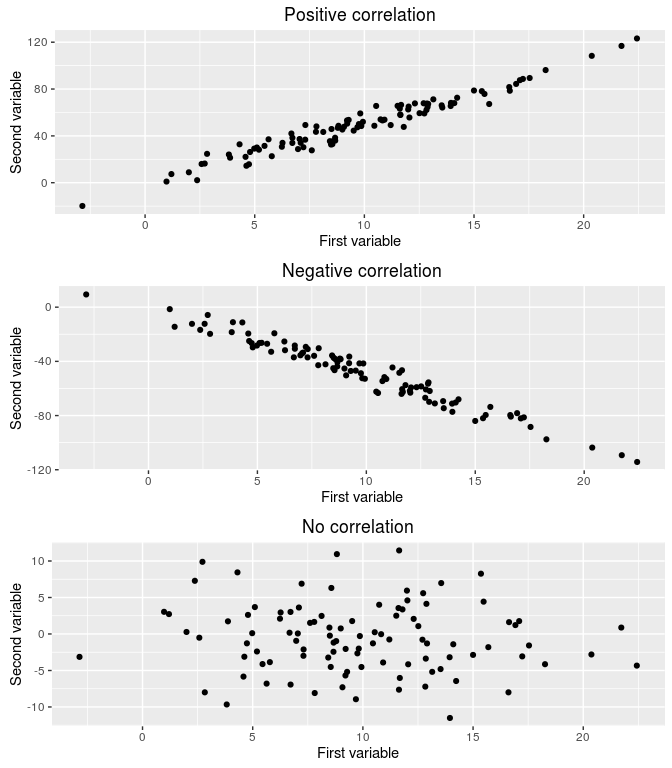
\*\*\* =right

## Посмотрим на датасет

brain <- read.csv("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

## Вспомним: *Сила и направление связи между величинами*



--- .segue

# Основы корреляционного анализа

## Коэффициенты корреляции и условия их применимости

|  |  |  |
| --- | --- | --- |
| Коэффициент | Фукция | Особенности примененения |
| Коэф. Пирсона | cor(x,y,method="pearson") | Оценивает связь двух нормально распределенных величин. Выявляет только лиейную составляющую взамосвязи. |
| Ранговые коэффициенты (коэф. Спирмена, Кэндалла) | 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
* Бывают двусторонние и односторонние критерии или
* Ошибка коэффициента Пирсона:
* Стандартизованная величина подчиняется распределению Стьюдента с парметром
* Для ранговых коэффициентов существует проблема "совпадающих рангов" (tied ranks), что приводит к приблизительной оценке и приблизительной оценке уровня значимости.
* Достоверность коэффициента кореляции можно оценить пермутационным методом

--- .prompt

## 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 перменными

*Hint 1*: Обратите внимание на то, что в датафрейме есть пропущенные значения. Изучите, как работают c NA функуции, вычисляющие коэффициенты корреляции.

*Hint 2* Для построения точечной диаграммы вам понадобится geom\_point()

## Решение

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

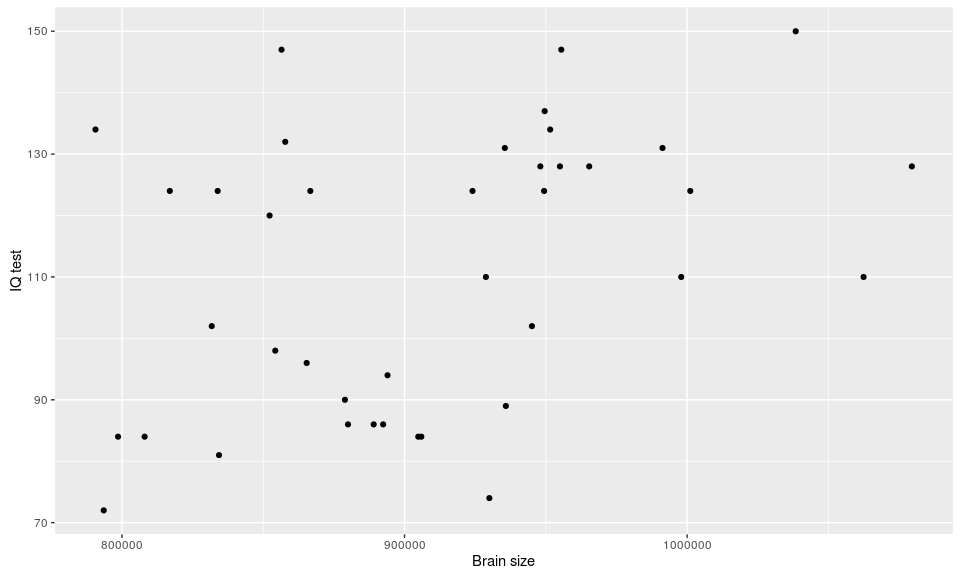
## FSIQ VIQ PIQ Weight Height  
## FSIQ 1.00000 0.94664 0.934125 -0.051483 -0.08600  
## VIQ 0.94664 1.00000 0.778135 -0.076088 -0.07107  
## PIQ 0.93413 0.77814 1.000000 0.002512 -0.07672  
## Weight -0.05148 -0.07609 0.002512 1.000000 0.69961  
## Height -0.08600 -0.07107 -0.076723 0.699614 1.00000

cor.test(brain$PIQ,   
 brain$MRINACount,   
 method = "pearson",   
 alternative = "two.sided")

##   
## Pearson's product-moment correlation  
##   
## data: brain$PIQ and brain$MRINACount  
## t = 2.6, df = 38, p-value = 0.01  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.08563 0.62323  
## sample estimates:  
## cor   
## 0.3868

## Решение

pl\_brain <- ggplot(brain, aes(x = MRINACount, y = PIQ)) + geom\_point() + xlab("Brain size") + ylab("IQ test")  
pl\_brain



## Частные корреляции

Частная корреляция - описывает связь между двумя переменными при условии, что влияние других переменных удалено.

Мы удаляем из и ту часть зависимости, которая вызвана влиянием

library(ppcor)  
brain\_complete <- brain[complete.cases(brain),]  
pcor.test(brain\_complete$PIQ, brain\_complete$MRINACount, brain\_complete$Height, )

## estimate p.value statistic n gp Method  
## 1 0.5373 0.0006051 3.769 38 1 pearson

--- .segue

# Два подхода к исследованию: Тестирование гипотезы и Построение модели

## Тестирование гипотезы VS построение модели

* Проведя корреляционный анализ, мы лишь ответили на вопрос "Существет ли достоверная связь между величинами?"
* Сможем ли мы, используя это знание, *предсказть* значения одной величины, исходя из знаний другой?

## Тестирование гипотезы VS построение модели

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

--- .segue

# Какие бывают модели?

## Линейные и нелинейные модели

### Линейные модели

### Нелинейные модели

## Простые и многокомпонентные (множественные) модели

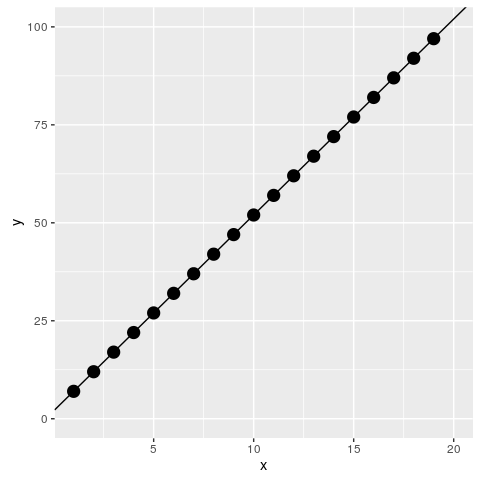
### Простая модель

### Множественная модель

---- &twocol

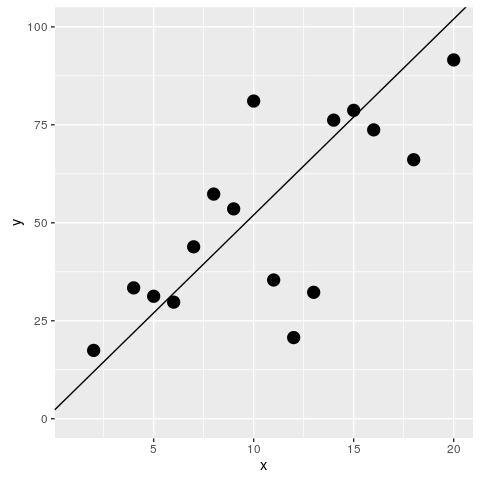
## Детерминистские и стохастические модели

\*\*\* =left



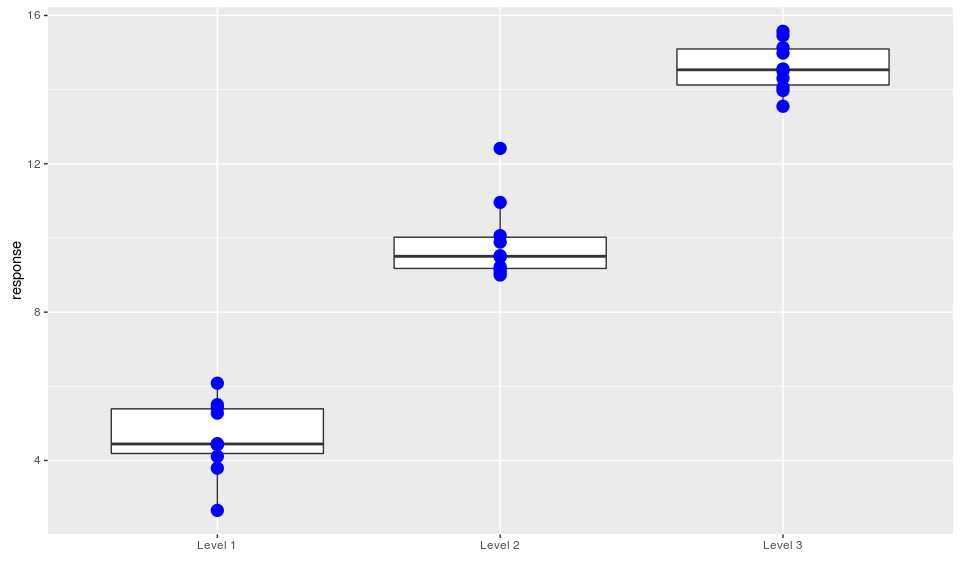
модель Два парметра: угловой коэффициент (slope) ; свободный член (intercept) Чему равен при ?

\*\*\* =right



модель Появляется дополнительный член Он вводит в модель влияние неучтенных моделью факторов. Обычно считают, что

## Модели с дискретными предикторами

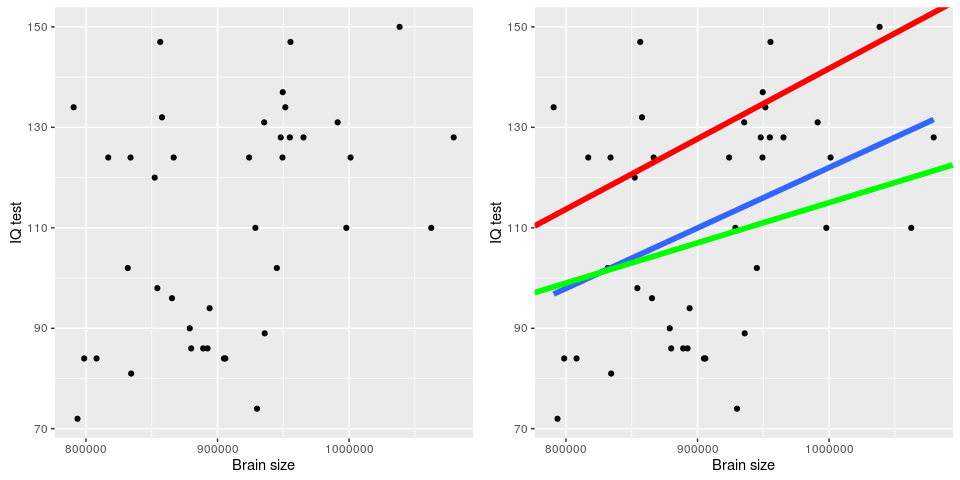


Модель для данного примера имеет такой вид + 9.9

- 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)

--- .segue

# Найти оптимальную модель позволяет регрессионный анализ

--- &twocol

## Происхождение термина "регрессия"

\*\*\* =left

Френсис Галтон (Francis Galton)

\*\*\* =right

"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- *коэффциент регресси*

## Подбор лиии 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 макcимального правдоподобия

(из кн. Zuur et al., 2009, стр. 19)

--- &twocol

## Метод наименьших квадратов

\*\*\* =left (из кн. Quinn, Keough, 2002, стр. 85)

\*\*\* =right

Остатки (Residuals):

Линия регрессии (подобраная модель) - это та линия, у которой минимальна.

## Подбор модели методом наменьших квадратов с помощью функци lm()

fit <- lm(formula, data)

Модель записывается в виде формулы

|  |  |
| --- | --- |
| Mодель | Формула |
| Простая линейная регресся | Y ~ X Y ~ 1 + X Y ~ X + 1 |
| Простая линейная регрессия (без , "no intercept") | Y ~ -1 + X Y ~ X - 1 |
| Уменьшенная простая линейная регрессия | Y ~ 1 Y ~ 1 - X |
| Множественная линейная регрессия | Y ~ X1 + X2 |

## Подбор модели методом наменьших квадратов с помощью функци lm()

fit <- lm(formula, data)

Элементы формул для записи множественных моделей

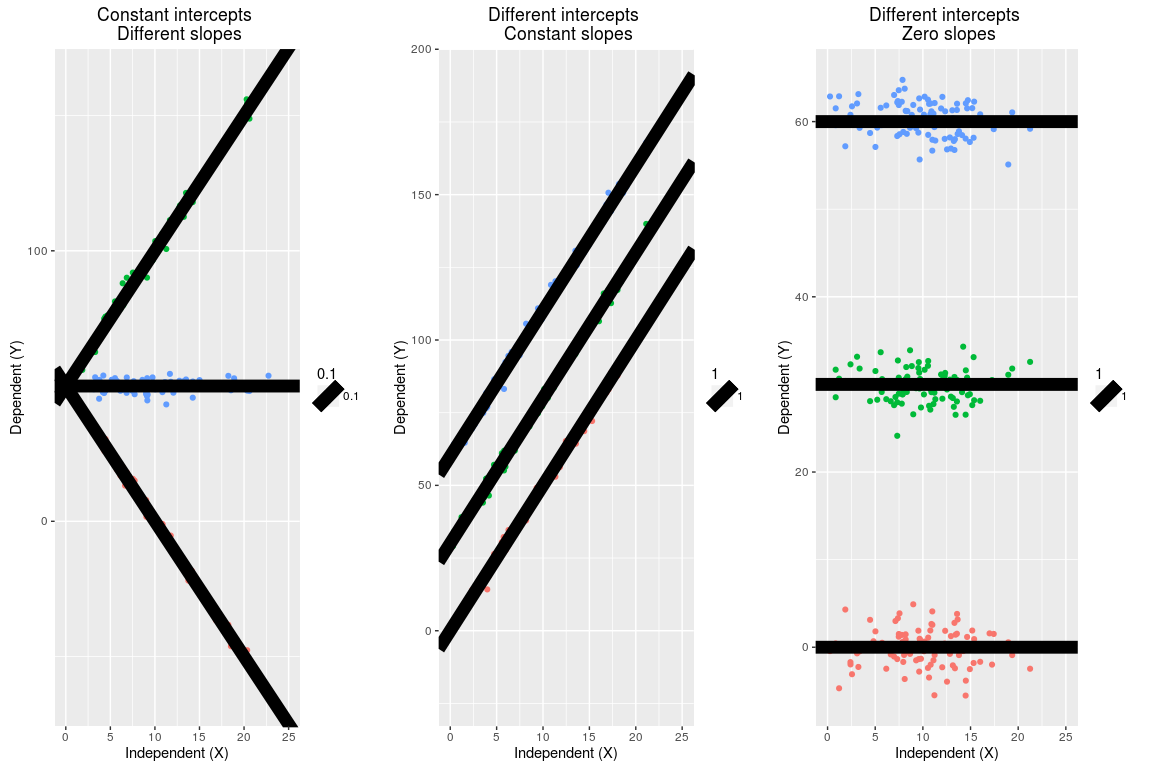
|  |  |
| --- | --- |
| Элемент формулы | Значение |
| : | Взаимодействие предикторов Y ~ X1 + X2 + X1:X2 |
| \* | Обзначает полную схему взаимодействий Y ~ X1 \* X2 \* X3 аналогично Y ~ X1 + X2 + X3+ X1:X2 + X1:X3 + X2:X3 + X1:X2:X3 |
| . | Y ~ . В правой части формулы записываются все переменные из 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 при изменении на *одну единицу* значения предиктора ()
* Свободный член (*intercept*) - величина во многих случаях не имеющая "смысла", просто поправочный коэффициент, без которого нельзя вычислить . *NB!* В некоторых линейных моделях он имеет смысл, например, значения при .
* Остатки (*residuals*) - характеризуют влияние неучтенных моделью факторов.

--- .prompt

## Вопросы:

1. Чему равны угловой коэффициент и свободный член полученной модели brain\_model?
2. Какое значеие 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)   
## 1.744

coefficients(brain\_model) [2]

## MRINACount   
## 0.0001203

coefficients(brain\_model) [1] + coefficients(brain\_model) [2] \* 900000

## (Intercept)   
## 110

brain$PIQ[10] - fitted(brain\_model)[10]

## 10   
## 30.36

residuals(brain\_model)[10]

## 10   
## 30.36

## 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

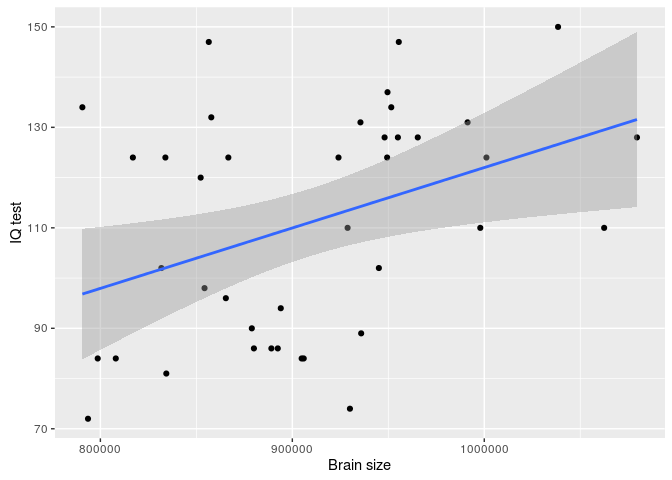
Параметр | Оценка | Стандартная ошибка  
|-------------|--------------------|-------------|  
 | или проще |   
 | |   
 | |   
 Для чего нужны стандартные ошибки? >- Они нужны, поскольку мы *оцениваем* параметры по *выборке* >- Они позволяют построить доверительные интервалы для параметров >- Их используют в статистических тестах

---- &twocol

## Графическое представление результатов

\*\*\* =left

pl\_brain + geom\_smooth(method="lm")

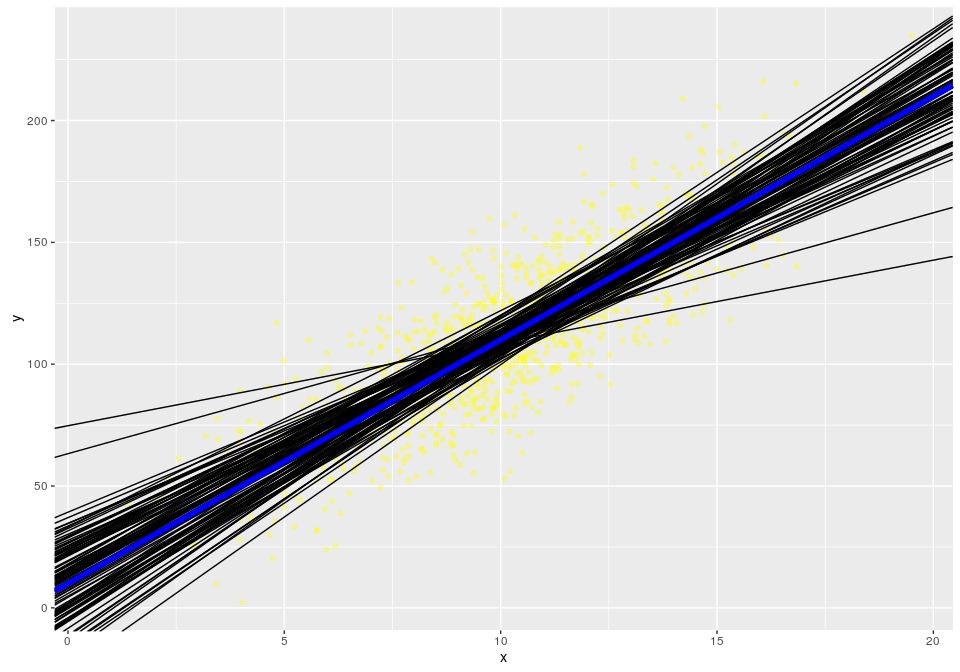


\*\*\* =right

Что это за серая область?

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

## Симулированный пример

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

## Доверительные интервалы для коэффициентов уравнения регрессии

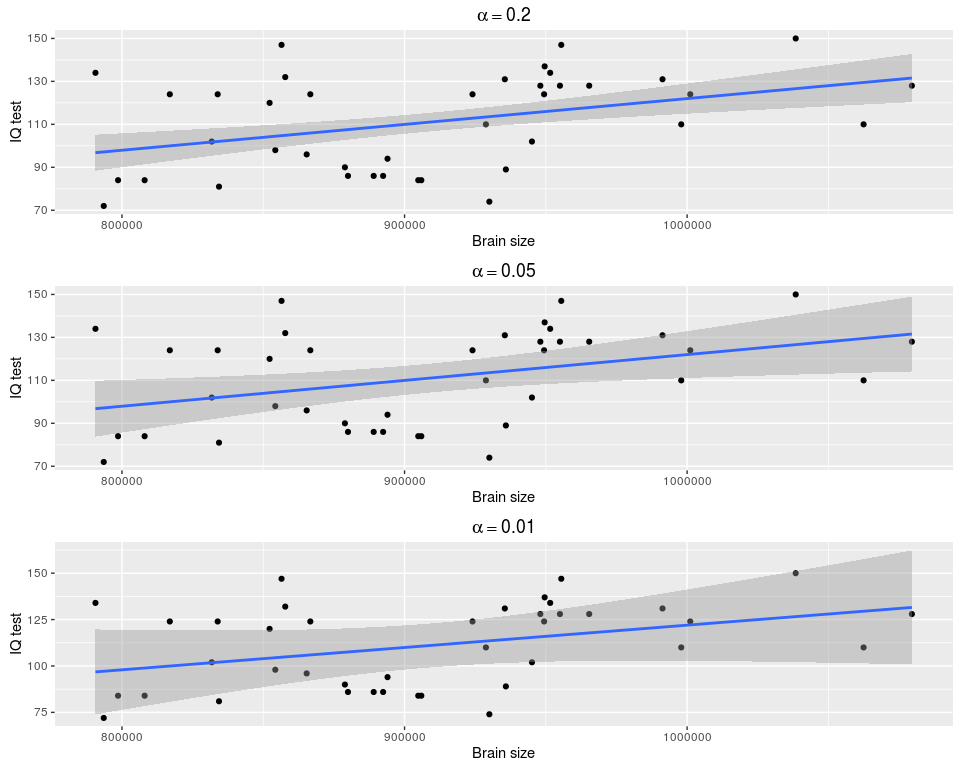
coef(brain\_model)

## (Intercept) MRINACount   
## 1.7437570 0.0001203

confint(brain\_model)

## 2.5 % 97.5 %  
## (Intercept) -84.07513478 87.5626489  
## MRINACount 0.00002611 0.0002144

## Для разных можно построить разные доверительные интервалы



*Важно!* Если коэффициенты уравнения регресси - лишь приблизительные оценки параметров, то предсказать значения зависимой переменной можно только *с нeкоторой вероятностью*.

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

newdata <- data.frame(MRINACount = 900000)  
  
predict(brain\_model, newdata, interval = "confidence", level = 0.95, se = TRUE)

## $fit  
## fit lwr upr  
## 1 110 103.2 116.7  
##   
## $se.fit  
## [1] 3.344  
##   
## $df  
## [1] 38  
##   
## $residual.scale  
## [1] 20.99

* При размере мозга 900000 среднее значение IQ будет, с вероятностью 95%, находиться в интервале от 103 до 117 (110 7).

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

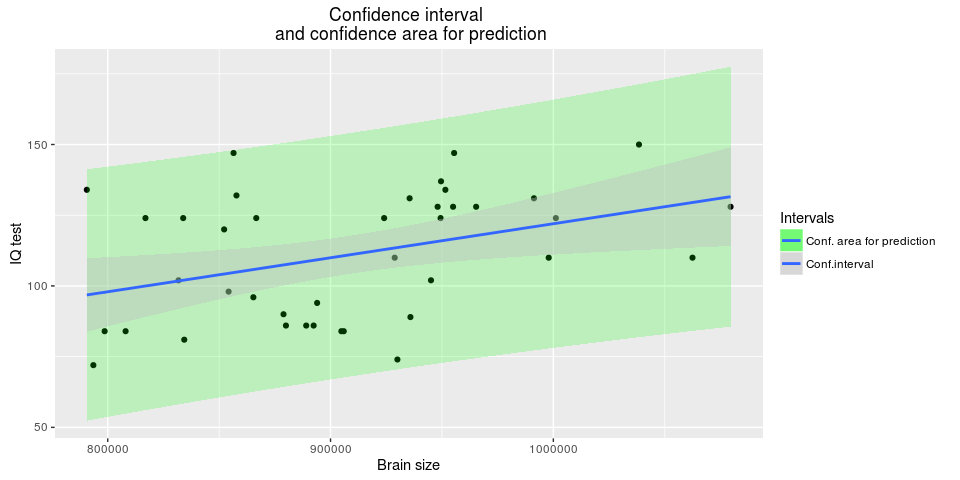
Подготавливаем данные

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 99.98 56.10 143.9  
## 2 Male 140 150 124 NA 72.5 1001121 122.13 78.24 166.0  
## 3 Male 139 123 150 143 73.3 1038437 126.62 81.90 171.3  
## 4 Male 133 129 128 172 68.8 965353 117.83 74.48 161.2  
## 5 Female 137 132 134 147 65.0 951545 116.17 72.96 159.4  
## 6 Female 99 90 110 146 69.0 928799 113.44 70.37 156.5

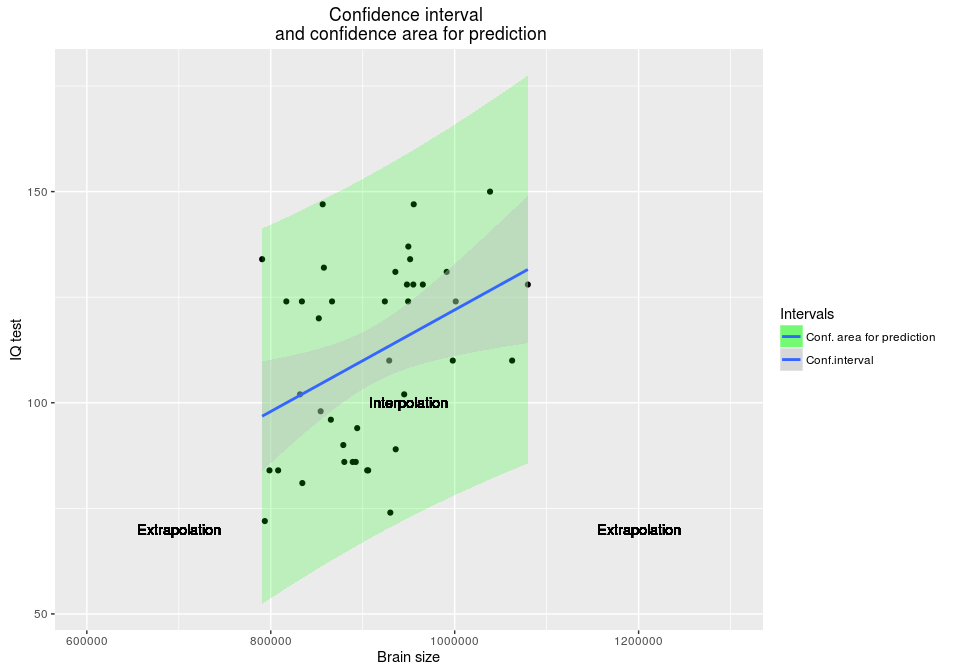
## Отражаем на графике область значений, в которую попадут 95% предсказанных величин IQ

pl\_brain +   
 geom\_ribbon(data=brain\_predicted, aes(y=fit, ymin=lwr, ymax=upr, fill = "Conf. area for prediction"), alpha=0.2) +   
 geom\_smooth(method="lm", aes(fill="Conf.interval"), alpha=0.4) +   
 scale\_fill\_manual("Intervals", values = c("green", "gray")) +   
 ggtitle("Confidence interval \n and confidence area for prediction")



## Важно!

*Модель "работает" только в том диапазоне значений независимой переменной (), 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|)

## Проверка состоятельности модели

### Существует два равноправных способа

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

## Проверка состоятельности модели с помощью t-критерия

Модель "работает" если в генеральной совокупности

Гипотеза: антигипотеза Тестируем гипотезу

Число степеней свободы: >- Итак, >- t value - Значение t-критерия >- Pr(>|t|) - Уровень значимости

## Состоятельна ли модель, описывающая связь IQ и размера головного 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

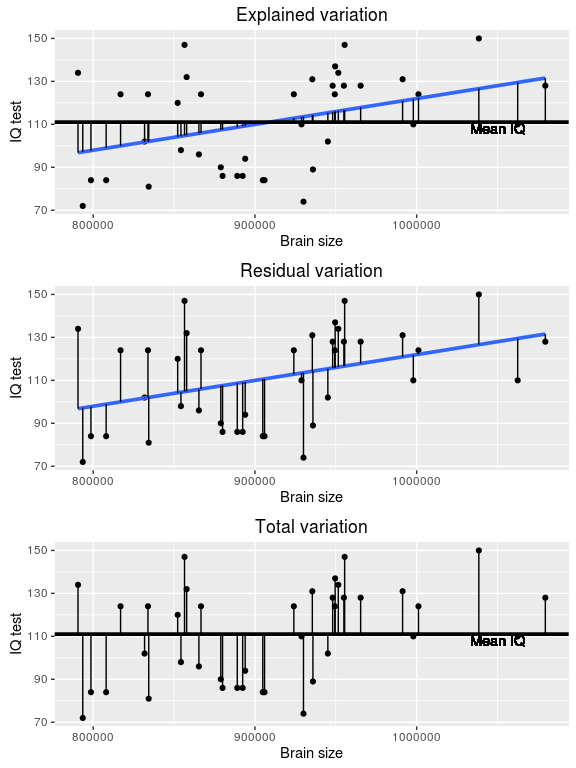
---- &twocol

## Проверка состоятельности модели с помощью F-критерия

\*\*\* =left

*Объясненная дисперсия зависимой перменной*  
  
  
 *Остаточная дисперсия завсимой переменной*  
  
  
  
 *Полная дисперсия зависимой переменной*

\*\*\* =right



---- &twocol

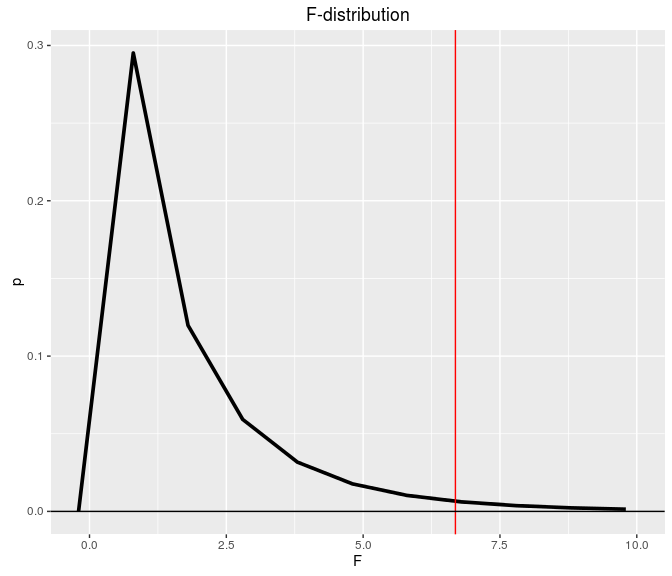
## F критерий

\*\*\* =left

Если зависимости нет, то

Логика та же, что и с t-критерием

\*\*\* =right

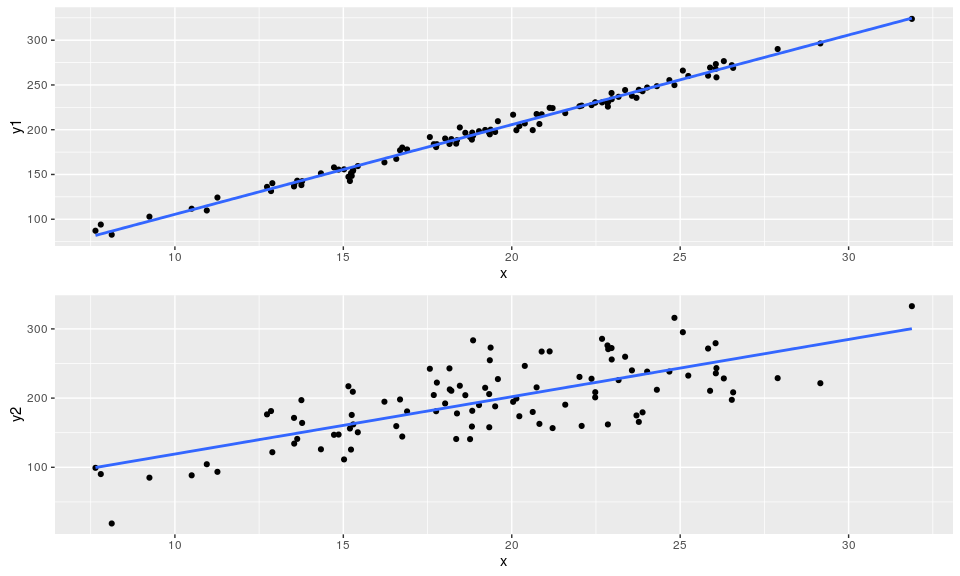


Форма F-распределения зависит от двух параметров

и

## 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

## 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

- 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 мделью вида IQ = 1.74 + 0.00012 MRINACount ( = 6.686, p = 0.0136, = 0.149)

## Summary

* Модель простой линейной регрессии
* 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.
* Состоятельность модели можно проверить при помощи t- или F-теста. )
* Качество подгонки модели можно оценить при помощи коэффициента детерминации

## Что почитать

* Гланц, 1999, стр. 221-244
* [Open Intro to Statistics](https://docs.google.com/viewer?docex=1&url=http://www.openintro.org/stat/down/OpenIntroStatSecond.pdf): [Chapter 7. Introduction to linear regression](https://docs.google.com/viewer?docex=1&url=http://www.openintro.org/stat/down/oiStat2_07.pdf), pp. 315-353.
* Quinn, Keough, 2002, pp. 78-110