РЕГРЕССИОННЫЙ И КОРРЕЛЯЦИОННЫЙ АНАЛИЗЫ

Практическое занятие 4

к.т.н., доцент кафедры, Томин Н.В.

Содержание

1. Проверка статистических гипотез
2. Отсев грубых нарушений
3. Доверительные интервалы

Корреляция

- Корреляция отражает степень связи между двумя переменными
- Коэффициент корреляции выражает эту степень количественно
- $-1 \leq r \leq +1$

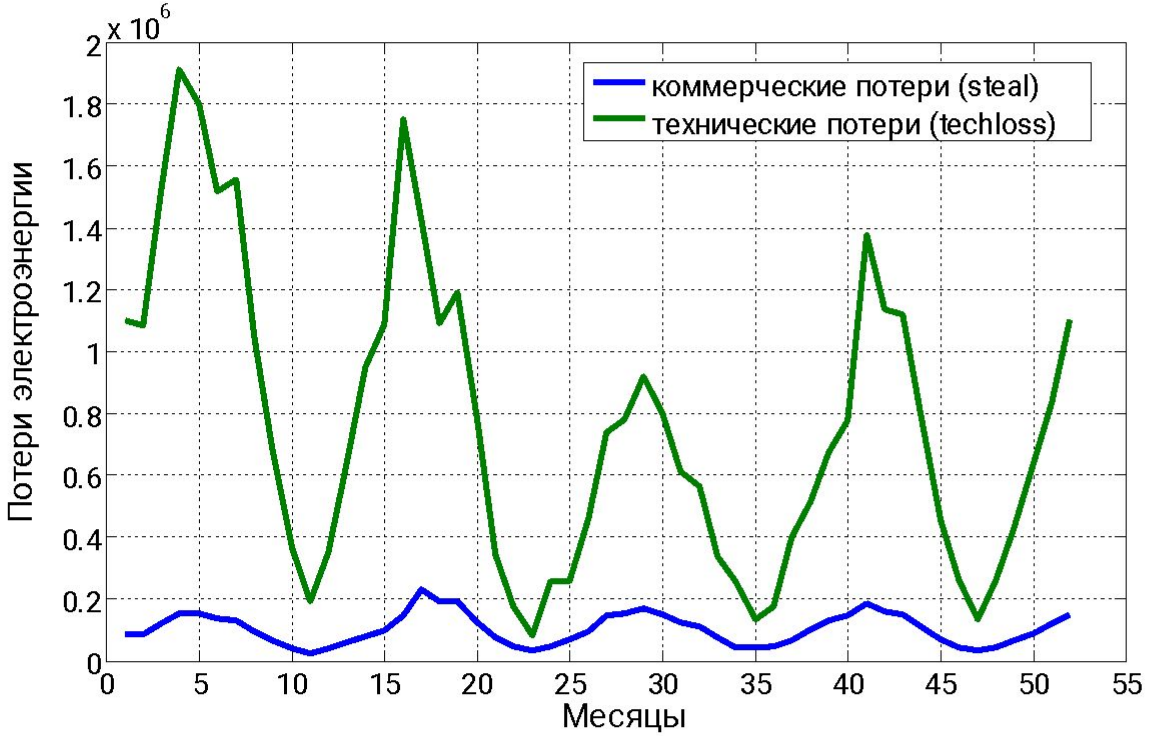
Коэффициент корреляции Пирсона

- Предполагает, что:
  - обе переменные распределены нормально
  - связь линейна
- Коэффициент корреляции Пирсона основан на расчете ковариации между двумя перемен-ными:

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

Расчёт коэффициента Пирсона в R

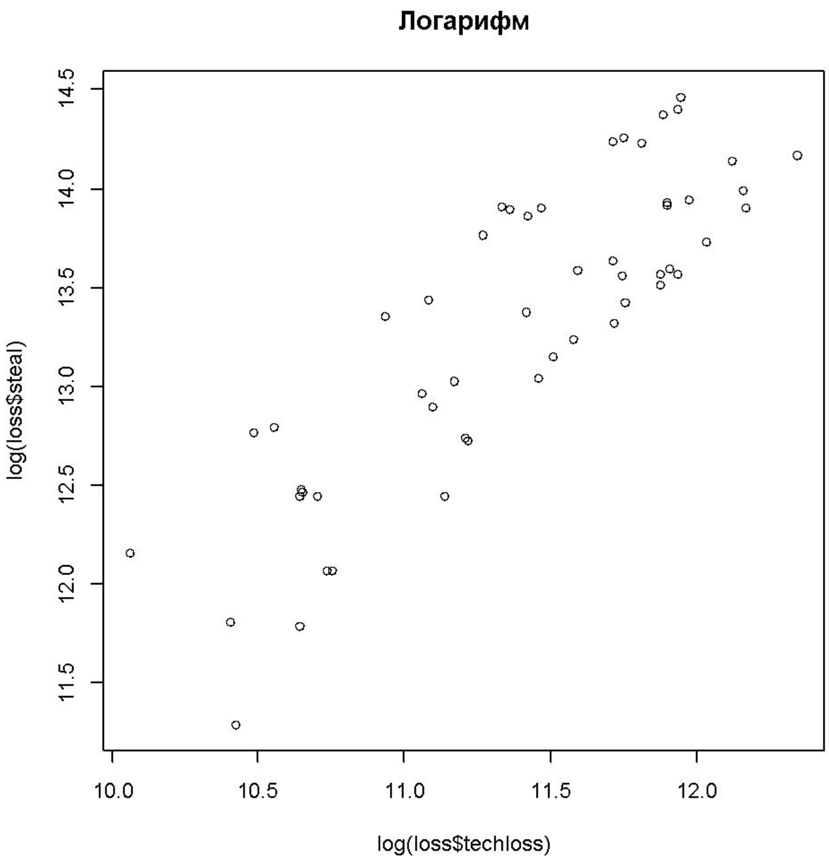
Пример. Даны выборки данных по техническим и коммерческим потерям электроэнергии в электрических сетях г. Братска за 2 года. Необходимо найти коэффициент корреляции между этими параметрами и проверить его статис

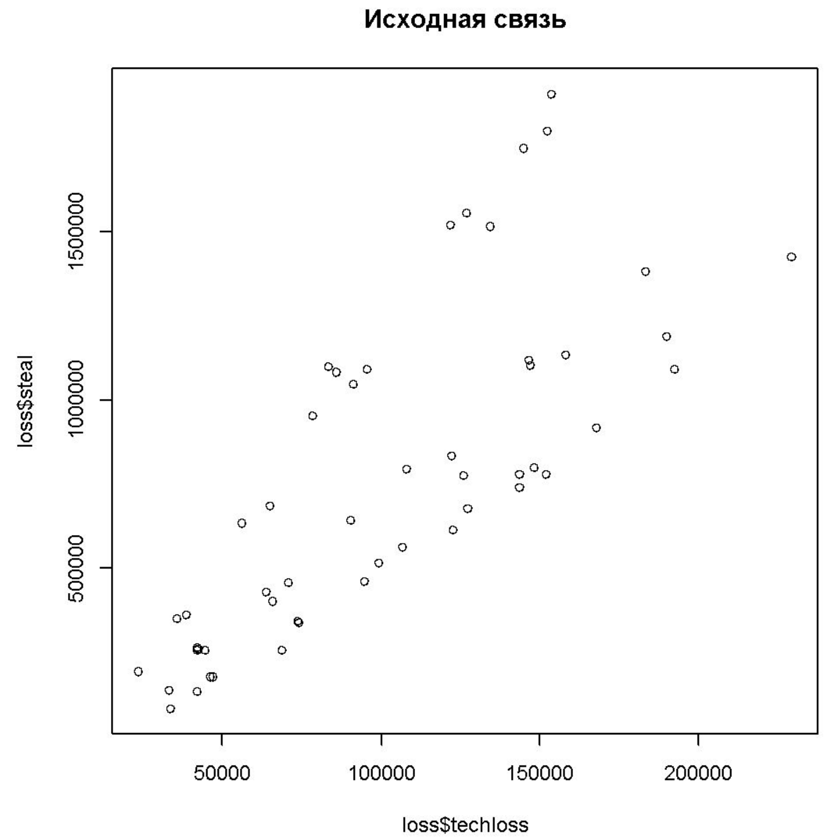


Расчёт коэффициента Пирсона в R

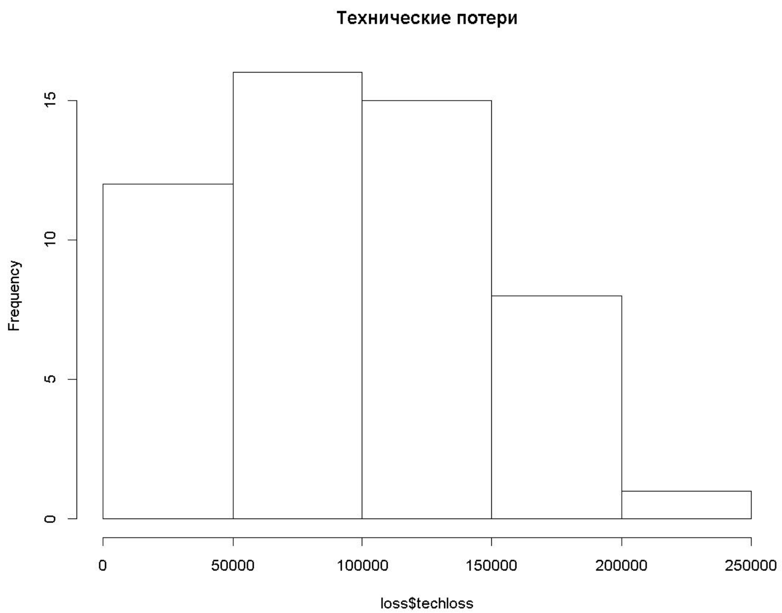
```
< loss <- read.csv ("loss.csv", sep = ";", header=TRUE)

#корреляционный анализ
< cor.test (loss$techloss, loss$steal)

     Pearson's product-moment correlation

data:  loss$techloss and loss$steal
t = 8.4983, df = 50, p-value = 2.848e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6274242 0.8609867
sample estimates:
      cor
0.7687038
```

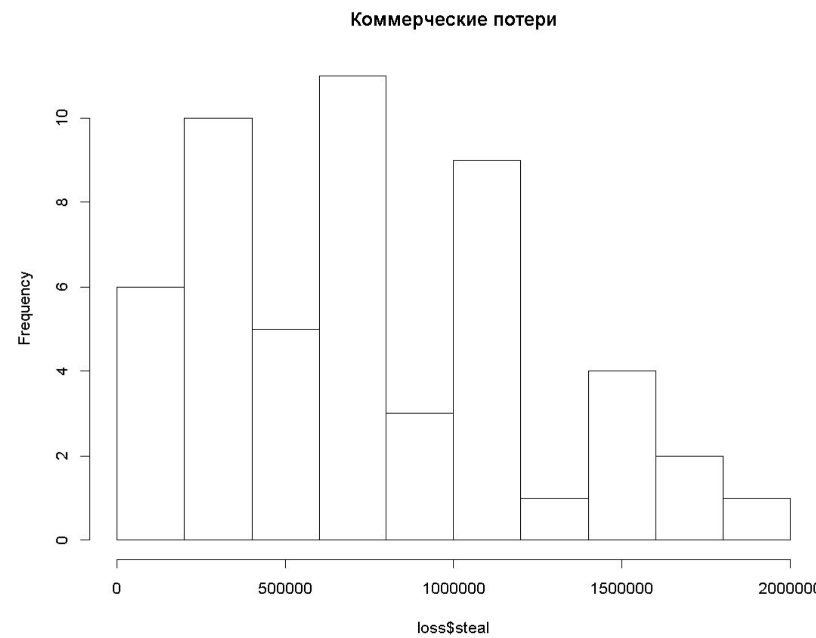
Связь между потерями нелинейна (на исходной шкале)





Ни одна из переменных не распределена нормально






```
Shapiro-Wilk normality test

data: loss$techloss
W = 0.95535, p-value = 0.04928
```

```
Shapiro-Wilk normality test

data:  loss$steal
W = 0.94266, p-value = 0.01438
```

Коэффициент Спирмена

- Не предполагает, что данные распределены каким-то особым образом
- Вместо исходных значений использует их ранги
- (!) Интерпретация не настолько проста, как в случае с коэффициентом Пирсона (т.к. связь необязательно линейна)

$$\rho = 1 - \frac{6}{n(n-1)(n+1)} \sum_{i=1}^{n} (R_i - S_i)^2$$

Расчёт коэффициента Спирмена в R

```
#корреляционный анализ по Спирмену
< cor.test (loss$techloss, loss$steal, method = "spearman")

    Spearman's rank correlation rho

data:  loss$techloss and loss$steal
S = 3968, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.8306156
```

Оценка значимости корреляции

Для проверки гипотезы о значимости коэффициента корреляции используется критерий Стьюдента в виде:

$$t_{\text{набл}} = \frac{r_B \sqrt{N-2}}{\sqrt{1-r_B^{\;2}}}$$

В этом случае, распределение Стьюдента имеет степень свободы равную.

Проверяемый коэффициент корреляции считается значимым, если значение tнабл по модулю будет больше, чем величина tкр, определенная по таблицам t-распределения

Расчётный пример

Пример. В испытательной лаборатории изучалось влияние переменного магнитного поля на микропроцессорные реле. Был сформирован двумерный массив данных, содержащий значения напряжённости магнитного поля, H и времени срабатывания реле t. По выборке объёмом N=122, извлечённой из двумерного массива, найден коэффициент корреляции $r_в$=0.4. Необходимо, при уровне значимости 0.05, проверить гипотезу о значимости выборочного коэффициента корреляции необходимо. Другими словами, узнать действительно ли напряжённость магнитного поля влияет на эффективность работы исследуемых реле.

Данные по скорости движения галактик

Freedman et al. (2001) опубликовали данные по расстоянию до 24 галактик, а также по скорости удаления этих галактик, полученные при помощи космического телескопа "Хаббл". Данные были собраны в рамках проекта (т.н. Key Project – "ключевой проект"), целью которого являлось уточнение значения постоянной Хаббла.

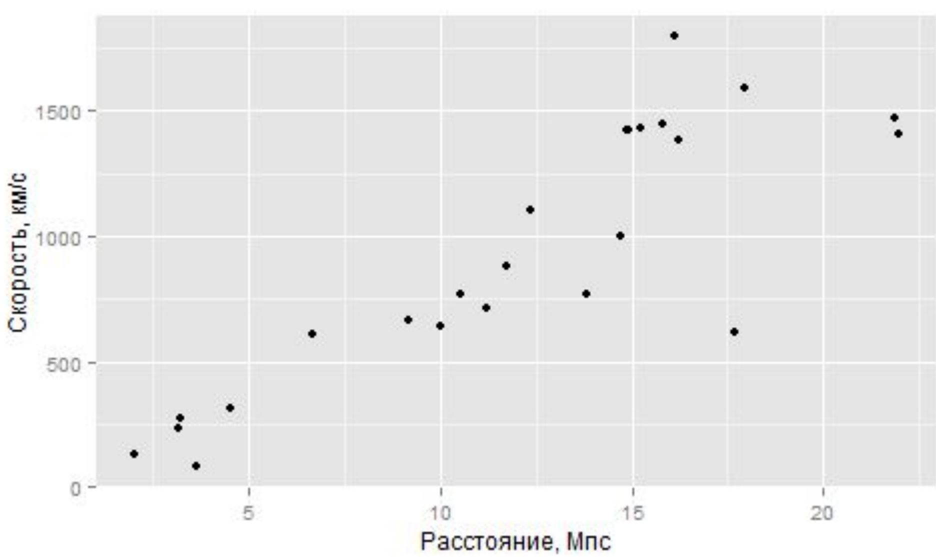
Эта постоянная представляет собой коэффициент в уравнении закона Хаббла, который описывает связь между расстоянием до внегалактического объекта (например, галактики, квазара) и скоростью его удаления, обусловленного расширением Вселенной после Большого взрыва.

Данные по скорости движения галактик

- Этот закон выражается простой линейной регрессией, которая может быть записана следующим образом:

$$y = \beta x + 0$$

где y - относительная скорость движения любых двух галактик, разграниченных в данный момент времени расстоянием y. Постоянная Хаббла, обозначенная здесь как β, выражается в км/с на мегапарсек. же, не могли удаляться друг от друга):



<u>Обратите внимание</u>: сво-бодный член уравнения регрессии здесь приравнен нулю, поскольку в момент, когда Вселенная находи-лась в состоянии сингуляр-ности, галактик не сущест-вовало и они, конечно

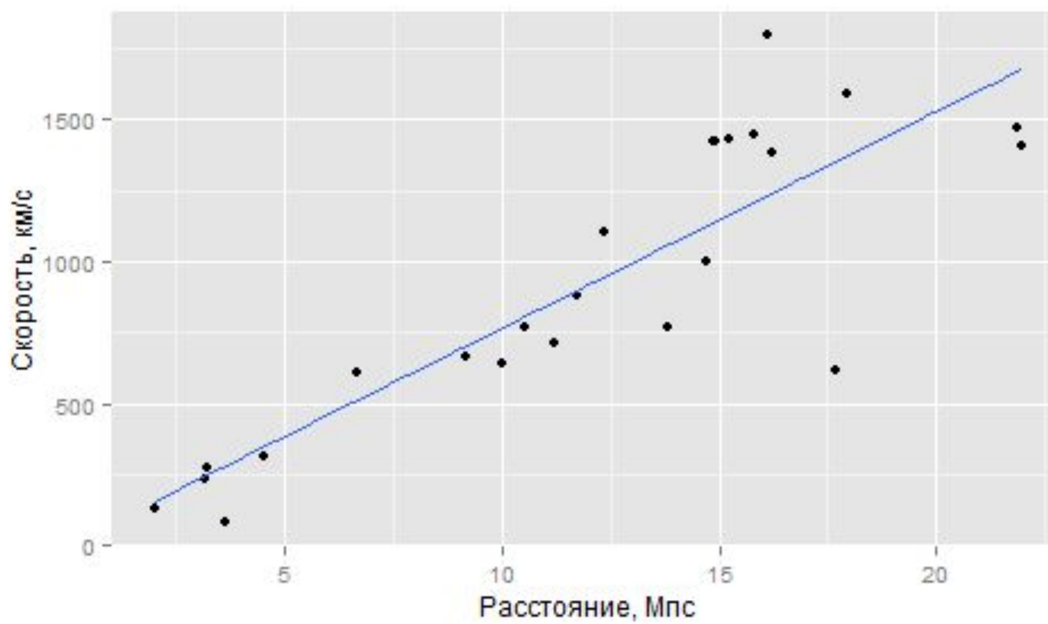
Данные по скорости движения галактик

```
> install.packages("gamair")
> library(gamair)
> data(hubble)

M <- lm(y ~ x - 1, data = hubble)
# -1 нужно для исключения свободного члена регрессионной модели
summary(M)
Call: lm(formula = y ~ x - 1, data = hubble)

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x   76.581      3.965   19.32 1.03e-15 ***
```

Данные по скорости движения галактик

Как видим, оцененное значение постоянной Хаббла составило 76.581 км/с на мегапарсек. Это значение существенно отличается от нуля (P-значение соответствующего t-теста в столбце Pr(>|t|)). На рисунке ниже приведена линия регрессии, описываемая полученным нами ура

$$y=76.581x$$



Данные по скорости движения галактик

- Расчет возраста Вселенной теперь не представляет труда. Один мегапарсек – это 3.09×10^9 км. Разделим полученную выше постоянную Хаббла на это значение, чтобы выразить ее в секундах:

```
hub.const <- 76.581/3.09e19
 [1] 2.47835e-18
```

Тогда возраст Вселенной, выраженный в секундах, составит:

```
age <- 1/hub.const
 [1] 4.034943e+17
```

Выполнив простое преобразование, получим возраст, выраженный в годах:

```
age/(60^2*24*365)
[1] 12794721567
```

Домашняя задача

Задача. Данные почасовые значения электрической нагрузки и температуры наружного воздуха для (по данным ОДУ Урала). Необходимо:

1) найти коэффициент корреляции между этими параметрами и оценить его статистическую значимость
2) найти в явном виде уравнение регрессии, связывающее эти параметры

Расчёты выполнить в R.