HW1

Liliya

2023-01-22

## Загркзка библиотек

## Задача 1

*Размер выборки*: s = 100 *Показатели:* - креатинин k - мочевина m

*Выборочные характеристики:* Средние значения: - median(k) = 88.5 - median(m) = 5.4 Стандартные отклонения: - sd(k) = 13.25 - sd(m) = 1.45 Коэффициент корреляции: cor(k,m) = 0.6

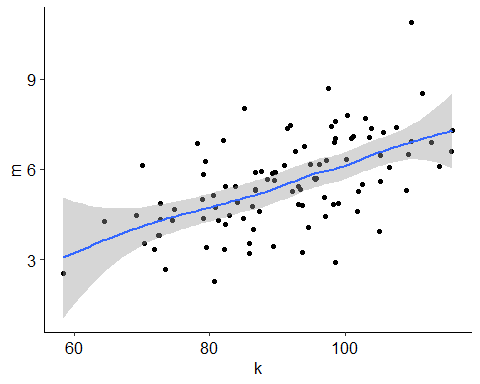
### Моделируем выборку

### Оценим получившуюся модель

## k m   
## Min. : 58.42 Min. : 2.265   
## 1st Qu.: 82.29 1st Qu.: 4.468   
## Median : 92.12 Median : 5.473   
## Mean : 91.12 Mean : 5.531   
## 3rd Qu.: 99.29 3rd Qu.: 6.596   
## Max. :115.83 Max. :10.870

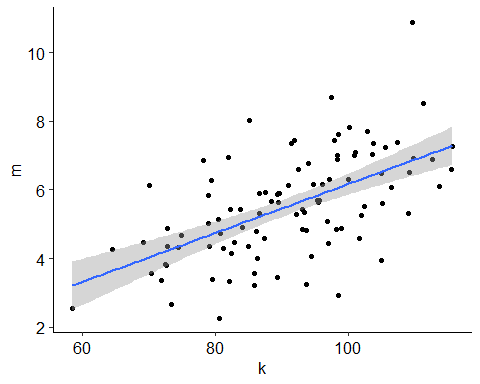
## k m  
## k 1.0000000 0.5776499  
## m 0.5776499 1.0000000

### Визуализация исходных данных



### Построим модель линейной регрессии креатинина k на мочевину m

##   
## Call:  
## lm(formula = m ~ k, data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.1427 -0.7703 -0.0182 0.6604 4.0122   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.95166 0.93360 -1.019 0.311   
## k 0.07114 0.01015 7.005 3.13e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.24 on 98 degrees of freedom  
## Multiple R-squared: 0.3337, Adjusted R-squared: 0.3269   
## F-statistic: 49.08 on 1 and 98 DF, p-value: 3.131e-10

 #### Как полученные оценки связаны с выборочными характеристиками (средним, дисперсией, корреляцией) выборки S?

1. Коэффициент детерминации (R-квадрат) — это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными. В частности, для модели парной линейной регрессии коэффициент детерминации равен квадрату обычного коэффициента корреляции между y и x. В нашем случае R^2 = 0.3016, а cor(k,m) = 0.6.
2. Наилучшая прямая линейной регрессии проходит через точку (; ) и имеет угловой коэффициент и свободный член, соответственной:

r\_{xy}$ \* ) / , где \* — коэффициент корреляции

$= - \* $

### Проверка гипотезы о нормальном распределении остатков

**Гипотеза** **H0:** Распределение нормальное.

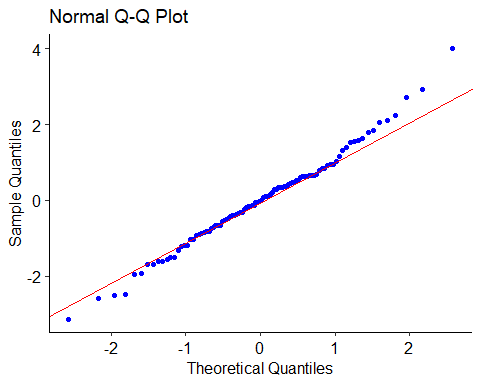
The type of test to use depends on the number of observations. In general, if you have less than 50 observations, you should use the Shapiro-Wilk test. Otherwise, the Kolmogorov-Smirnov test is better.

## -----------------------------------------------  
## Test Statistic pvalue   
## -----------------------------------------------  
## Shapiro-Wilk 0.9902 0.6839   
## Kolmogorov-Smirnov 0.0701 0.7088   
## Cramer-von Mises 6.4692 0.0000   
## Anderson-Darling 0.2918 0.6001   
## -----------------------------------------------

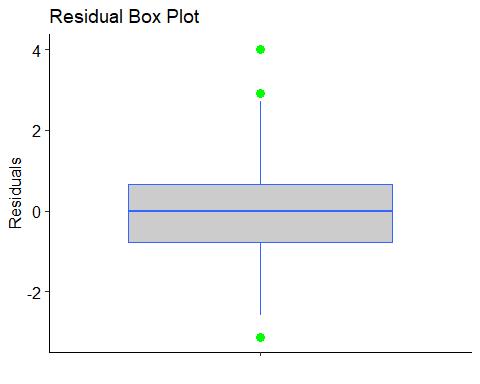
*Результат:* The criteria to reject the null hypothesis, and therefore conclude that the data is not normally distributed, depends on the p-value. Typically, if the p-value is <0.05 we reject the null hypothesis.

Так как значения p-value > 0.05, H0 принимается.

#### График Q-Q Plot

 *Результат:* Looking at the Q-Q plot above, we see that most of the points are located on the diagonal line. Though the extremes deviate from the line and you might observe an “S”-shape. Therefore, this Q-Q plot is not conclusive regarding the normality of the residuals.

#### Box-plot остатков

 *Результат:* Looking at the boxplot above, we see that residuals are located around zero and the whiskers are almost of the same size. We might assume that the residuals are normally distributed.

Источники: [1] <https://handbook.mathpsy.com/?page_id=238> [2] <https://www.codingprof.com/5-ways-to-check-the-normality-of-residuals-in-r-examples/> [3] <https://milnepublishing.geneseo.edu/natural-resources-biometrics/chapter/chapter-7-correlation-and-simple-linear-regression/>