Linguistic Data: Quantitative Analysis and Visualisation

Correlation coefficients and simple linear regression

Ilya Schurov, Olga Lyashevskaya, George Moroz, Alla Tambovtseva

Part 1: Pearson's coefficient vs Spearman's coefficient

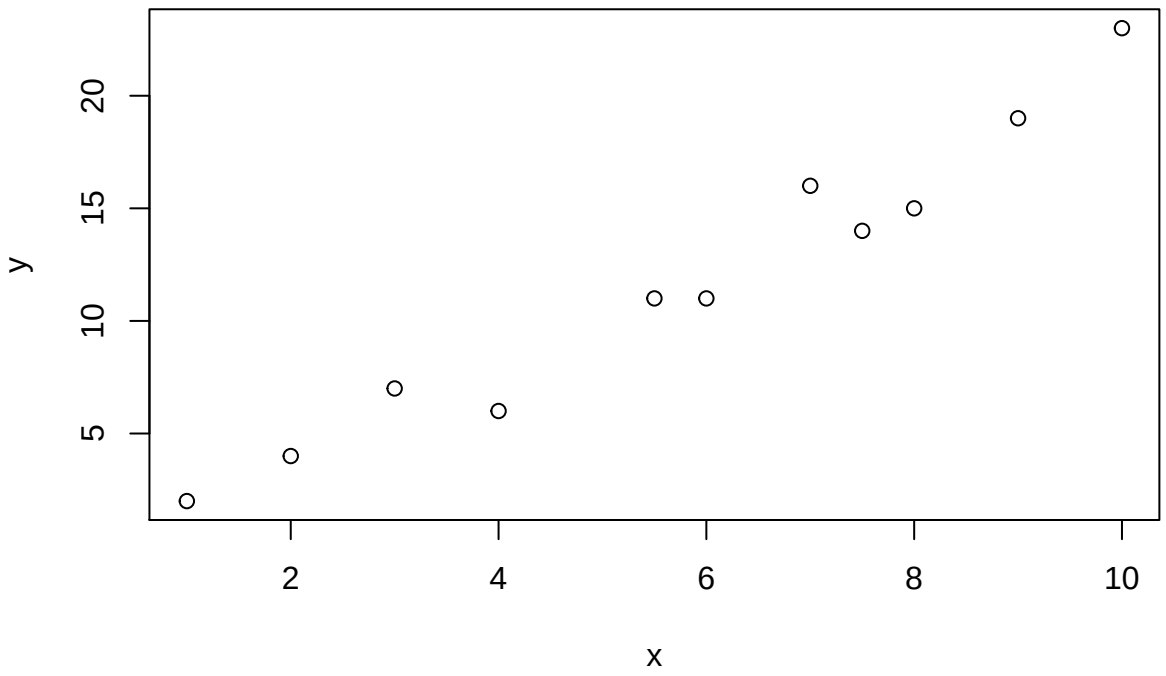
As we discussed, there are two widely used correlation coefficients, a Pearson's one and a Spearman's one. Since the latter is a measure of the rank correlation, it is usually used for variables in an ordinal scale. However, it can be helpful for quantitative variables as well because it is robust (not sensitive to outliers).

Consider two variables: `x` and `y`.

```
x <- c(1, 2, 6, 8, 9, 7, 7.5, 10, 3, 4, 5.5)
y <- c(2, 4, 11, 15, 19, 16, 14, 23, 7, 6, 11)
```

Let's plot a simple scatterplot first:

```
plot(x, y)
```



As we can see, although there are only few points, variables `x` and `y` seem to be positively associated (as `x` increases, `y` increases). We can even say that this association is pretty strong. Let's calculate two correlation coefficients and test their statistical significance.

```
# Pearson's coefficient
cor.test(x, y)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
```

```
## t = 13.862, df = 9, p-value = 2.234e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9124984 0.9942928
## sample estimates:
##       cor
## 0.9773737
```

What can we see in the output? The correlation coefficient itself is `cor` and here it is 0.977. So, we can conclude that the association between `x` and `y` is positive and very strong (the coefficient is approximately 1). Is it statistically significant at the 5% level of significance? Let us see.

$H_0 : corr(x, y) = 0$ (no linear association between x and y)

This null hypothesis should be rejected at the 5% significance level since p-value < 0.05. So, variables `x` and `y` are associated.

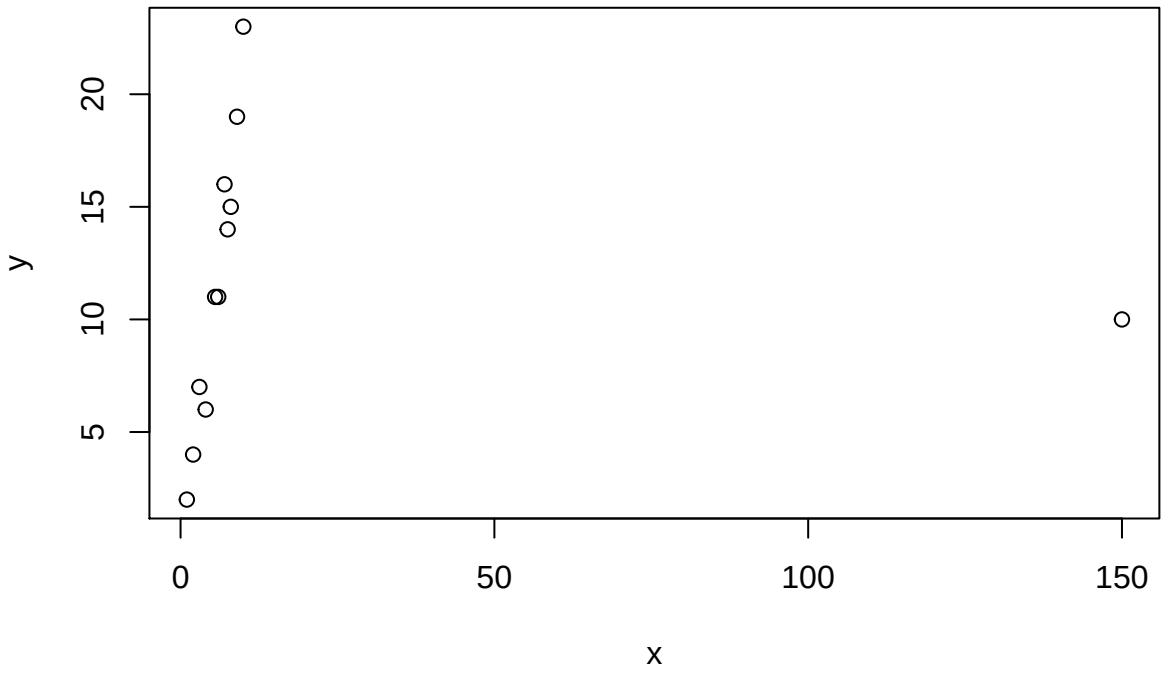
```
# Spearman's coefficient
cor.test(x, y, method = 'spearman')
```

```
## Warning in cor.test.default(x, y, method = "spearman"): Cannot compute
## exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  x and y
## S = 8.5188, p-value = 2.449e-06
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9612781
```

Here we also get a very high positive coefficient (0.96). Now let us add an outlier, a non-typical observation to our data, a point (150, 10).

```
x <- c(1, 2, 6, 8, 9, 7, 7.5, 10, 3, 4, 5.5, 150)
y <- c(2, 4, 11, 15, 19, 16, 14, 23, 7, 6, 11, 10)
plot(x, y)
```



It seems that this point can spoil everything! We can calculate correlation coefficient for updated variables:

```
cor.test(x, y)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = -0.033164, df = 10, p-value = 0.9742
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5808924  0.5668262
## sample estimates:
##         cor
## -0.01048683
```

A Pearson's correlation coefficient has broken down! Now it is negative, very small by absolute value and, what is more, insignificant! This coefficient is very sensitive to outliers, so here it "reacts" on a non-typical point in a very dramatic way. Now let's look at a Spearman's coefficient:

```
cor.test(x, y, method = 'spearman')
```

```
## Warning in cor.test.default(x, y, method = "spearman"): Cannot compute
## exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  x and y
## S = 64.613, p-value = 0.003127
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.7740817
```

Magic! This coefficient has not undergone serious changes, it is still positive and high. Besides, it is significant at the 5% significance level. So, with the help of this illustration we made sure that a Spearman's correlation coefficient is more robust than Pearson's one.

```
cor.test(x, y, method = 'kendall')
```

```
## Warning in cor.test.default(x, y, method = "kendall"): Cannot compute exact
## p-value with ties
```

```
##
##  Kendall's rank correlation tau
##
## data:  x and y
## z = 3.093, p-value = 0.001981
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.6870429
```

Part 2: real data

Data description

Two hundred observations were randomly sampled from the *High School and Beyond* survey, a survey conducted on high school seniors by the National Center of Education Statistics. Source: UCLA Academic Technology Services.

Variables

- `id`: Student ID.
- `gender`: Student's gender, with levels `female` and `male`.
- `race`: Student's race, with levels `african american`, `asian`, `hispanic`, and `white`.
- `ses`: Socio economic status of student's family, with levels `low`, `middle`, and `high`.
- `schtyp`: Type of school, with levels `public` and `private`.
- `prog`: Type of program, with levels `general`, `academic`, and `vocational`.
- `read`: Standardized reading score.
- `write`: Standardized writing score.
- `math`: Standardized math score.
- `science`: Standardized science score.
- `socst`: Standardized social studies score.

Let's load data first:

```
educ <- read.csv("education.csv")
```

And load `tidyverse` package and install `GGally` package that is a useful extension of `ggplot2`:

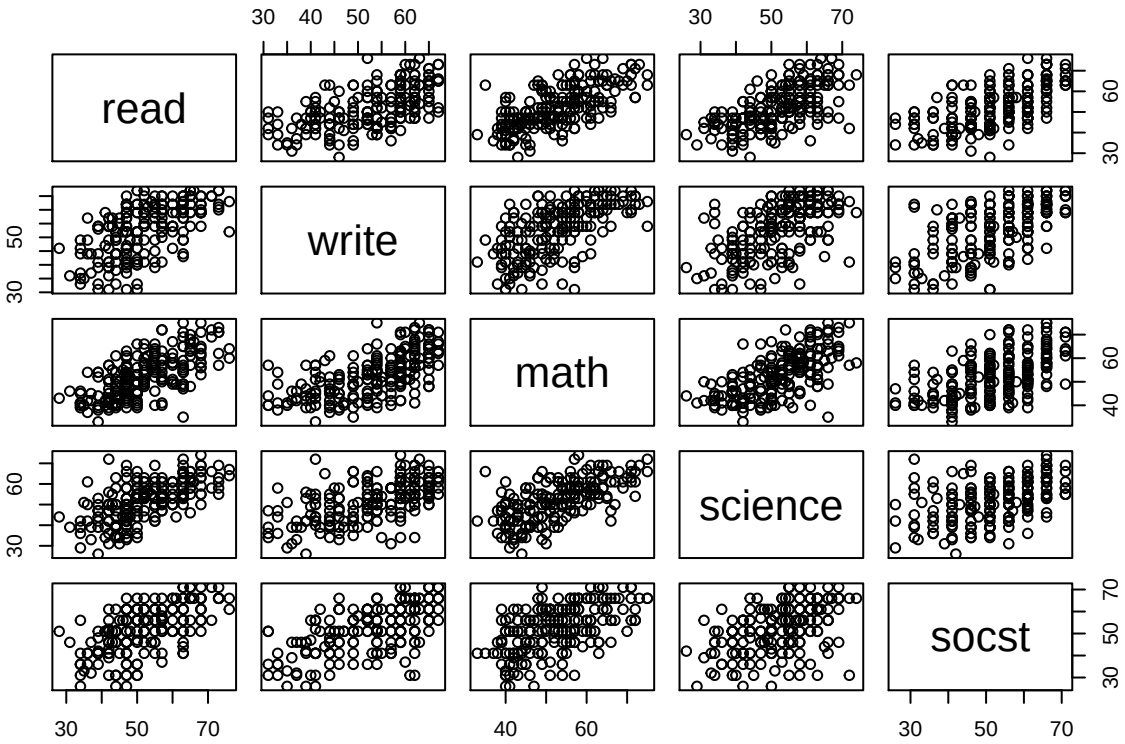
```
library(tidyverse)
library(GGally)
```

Now let us choose variables that correspond to abilities (`read` and `write`) and scores for subjects (`math`, `science`, 'socst).

```
scores <- educ %>% select(read, write, math, science, socst)
```

Let's create a basic scatterplot matrix, a graph that includes several scatterplots, one for each pair of variables.

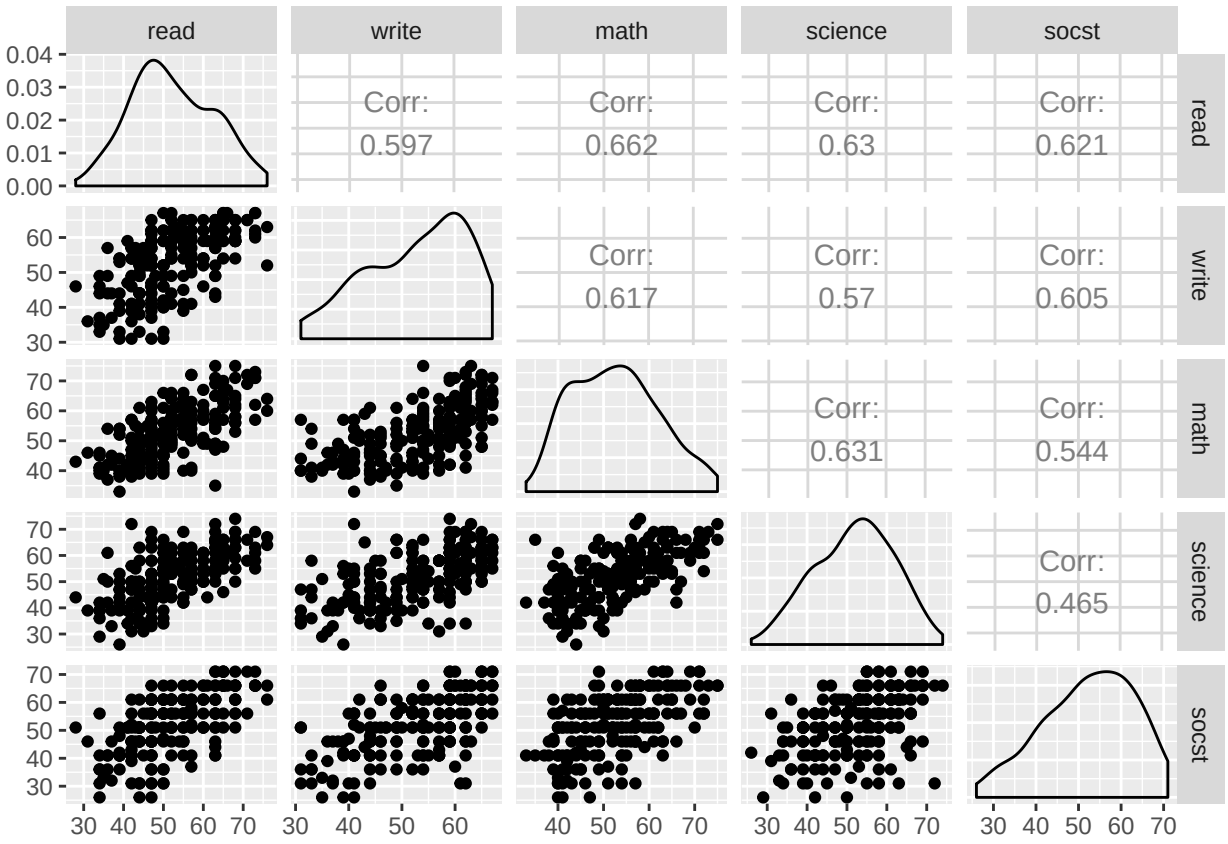
```
pairs(scores)
```



Question: Judging by this graph, can you say which scores have the strongest association? Try to guess the values of correlation coefficient for each pair of variables.

Now let's create a more beautiful graph via `GGally` library and check whether your guesses were true:

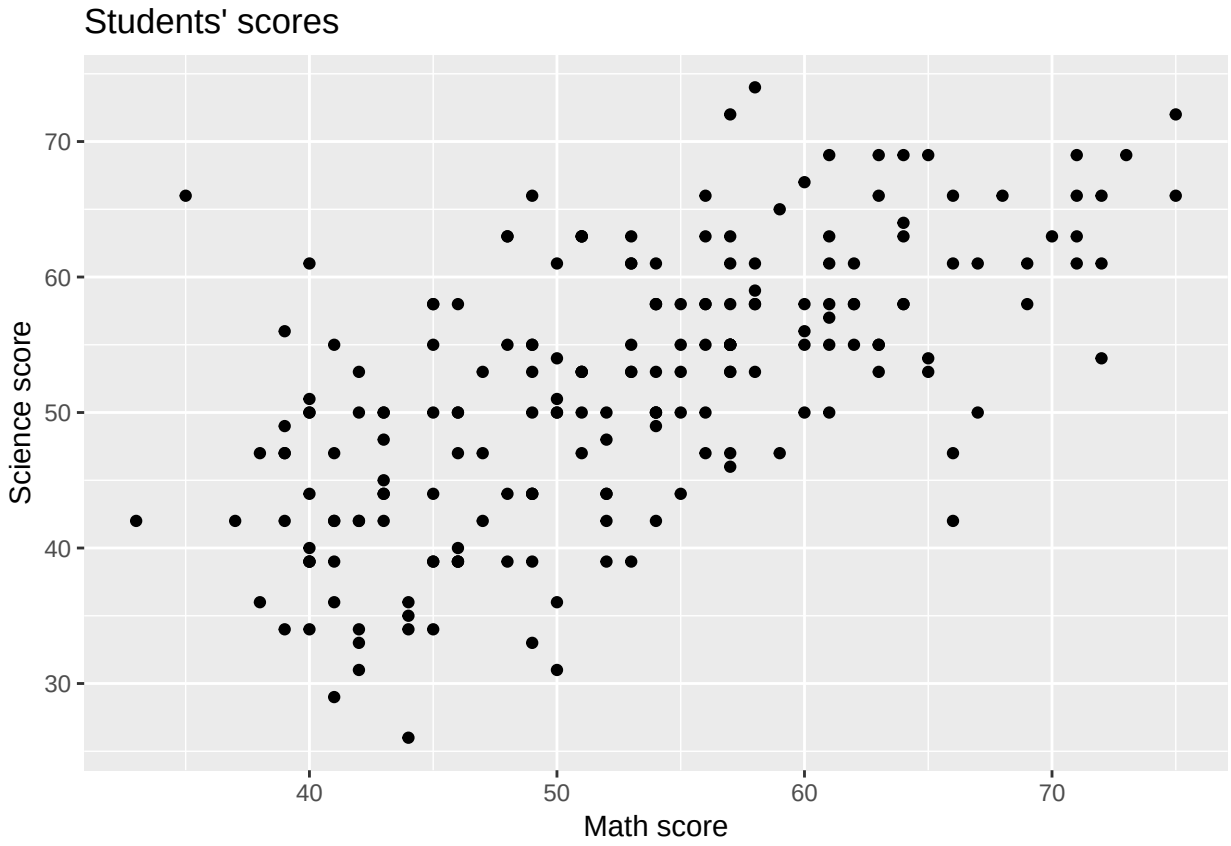
```
ggpairs(scores)  # dataset inside
```



Let's choose a pair of variables and proceed to formal testing. We will check whether students' score for Math and Science are associated.

First, create a simple scatterplot for these variables:

```
ggplot(data = scores, aes(x = math, y = science)) +
  geom_point() +
  labs(x = "Math score",
       y = "Science score",
       title = "Students' scores")
```



Again, as we saw, these variables seem to be positively associated.

Substantial hypothesis:

Math score and Science score should be associated. Explanation: most fields of Science require some mathematical knowledge, so it is logical to expect that people with higher Math score succeed in Sciences and vice versa.

Statistical hypotheses:

H_0 : there is no linear association between Math score and Science score, the true correlation coefficient R is 0.

H_1 : there is linear association between Math score and Science score, the true correlation coefficient R is not 0.

```
cor.test(scores$math, scores$science)
```

```
##
##  Pearson's product-moment correlation
##
## data:  scores$math and scores$science
## t = 11.437, df = 198, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5391745 0.7075569
## sample estimates:
##       cor
## 0.6307332
```

P-value here is approximately 0, so at the 5% significance level we reject H_0 about the absence of linear

association. Thus, we can conclude that Math score and Sciences score are associated. The Pearson's correlation coefficient here is 0.63, so we can say that the direction of this association is positive (the more is the Math score, the more the Science score is) and its strength is moderate.

Part 3: simple linear regression

Now suppose we are interested in the following thing: how does Science score change (on average) if Math score increases by one point? To answer this question we have to build a linear regression model. In our case it will look like this:

$$Science = \beta_0 + \beta_1 \times Math$$

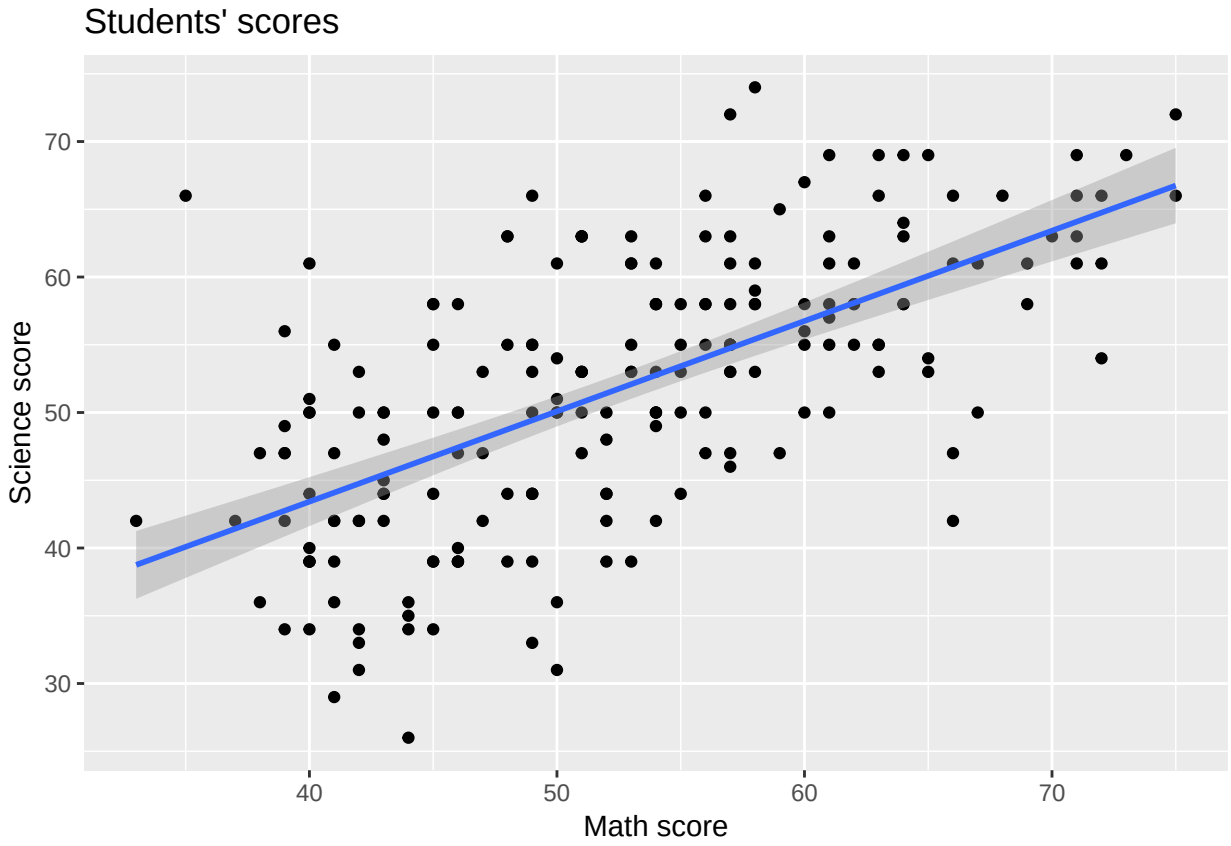
```
model1 <- lm(data = scores, science ~ math)
summary(model1)
```

```
##
## Call:
## lm(formula = science ~ math, data = scores)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.0874  -4.7524  -0.0859   4.9123  25.9118
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.75789    3.11623   5.378 2.11e-07 ***
## math         0.66658    0.05828  11.437  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.702 on 198 degrees of freedom
## Multiple R-squared:  0.3978, Adjusted R-squared:  0.3948
## F-statistic: 130.8 on 1 and 198 DF,  p-value: < 2.2e-16
```

How to interpret such an output?

1. `Intercept` is our β_0 and `math` is our β_1, the coefficient before the independent variable *Math score*. So, we can write a regression equation (try it).

```
ggplot(data = scores, aes(x = math, y = science)) +
  geom_point() +
  labs(x = "Math score",
       y = "Science score",
       title = "Students' scores") +
  geom_smooth(method=lm)
```



2. The coefficient β_1 shows how Science scores changes on average when Math scores increases by one unit. Now test its significance.

H_0 : the true correlation coefficient equals to 0 (Math score does not affect Science score).

H_1: the true correlation coefficient is not 0.

Should we reject our null hypothesis at the 5% significance level? Make conclusions.

3. `Multiple R-squared` is R^2, a coefficient of determination that shows what share of the reality our model explains. A more formal way to interpret it: it shows a share of variance of a dependent variable that is explained by an independent one.

Part 4: try yourselves

Here you are suggested to work with a dataset on Chekhov's stories (`chekhov.csv`).

Variables

- `n_words`: number of words in a
- `n_unique`: number of unique words in a

1. How do you feel: is there a linear relationship between the number of words and the number of unique words?
2. Plot a scatterplot for these variables and check whether your intuition was true. Interpret the scatterplot obtained.
3. Check using a proper statistical test, whether `n_words` and `n_unique` are associated: formulate a null hypothesis, test it and make conclusions.

4. Create a simple linear regression for the variables `n_words` and `n_unique`. Decide which one should be a dependent variable and which one should be independent. Perform regression analysis in R. Provide your conclusions.