

Estimating Correlation

Note on required packages: The following code required the package `readxl` to read in Excel files. If you have not already done so, download, install, and load the library with the following code:

```
install.packages("readxl")  
library("readxl")
```

A **correlation** exists between two variables when one is related to the other such that there is **comovement**. **Positive comovement** means as one variable increases, the other variable also increases. **Negative comovement** means as one variable increases, the other variable decreases.

Example: The website for Stock and Watson's *Introduction to Econometrics* textbook includes a dataset with economic growth data for 65 countries from 1960-1995, along with variables that may be related to growth.

1. Download the dataset

The code below downloads the Excel file from the textbook's website and assigns the dataset to a variable we create and call `growthdata`.

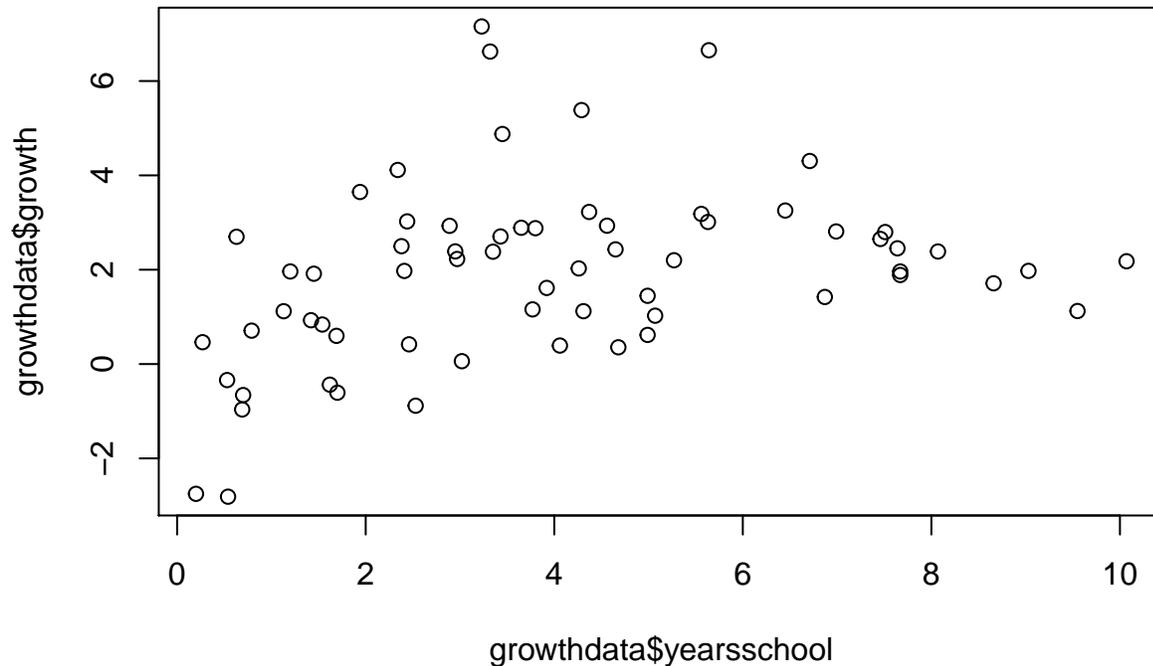
```
download.file(  
  url="http://wps.aw.com/wps/media/objects/11422/11696965/data3eu/Growth.xlsx",  
  dest="Growth.xlsx")  
growthdata <- read_excel("Growth.xlsx");
```

The dataset includes several variables. For this tutorial, we will focus on the average annual growth rate of real GDP from 1960-1995 (labeled `growth`) and the average number of years of schooling for adult residents in the country in 1960 (labeled `yearsschool`).

2. Plot the data

Let us first create a graph that illustrates the relationship between average years of schooling of adult residents and the subsequent average growth rate over the next 35 years. We can create a scatter plot using the `plot()` function as follows:

```
plot(x=growthdata$yearsschool, y=growthdata$growth)
```



The first parameter, `x=growthdata$yearsschool`, tells the `plot()` function to put years of schooling on the horizontal axis (aka the ‘x-axis’). The second parameter, `y=growthdata$growth`, tells the function to put the real GDP growth rate on the vertical axis.

It appears that years of schooling and real GDP growth may have a positive relationship. We can compute the best fitting straight line that describes this relationship with the function `lm()` which stands for ‘linear model’. In the code below, we call the `lm()` function and assign its output to a variable we call `growthmodel`.

```
growthmodel <- lm(growthdata$growth ~ growthdata$yearsschool)
```

We passed to the function `lm()` a single parameter which was a *formula* of the form $y \sim x$. This notation means to fit a function that has the linear form $y = a + bx$. The output variable `growthmodel` includes a lot of objects and statistical tests that describe the linear relationship between the x and y variables.

We can find out what precisely what the equation of the line is by calling the `coefficients` variable in the `growthmodel` object as follows:

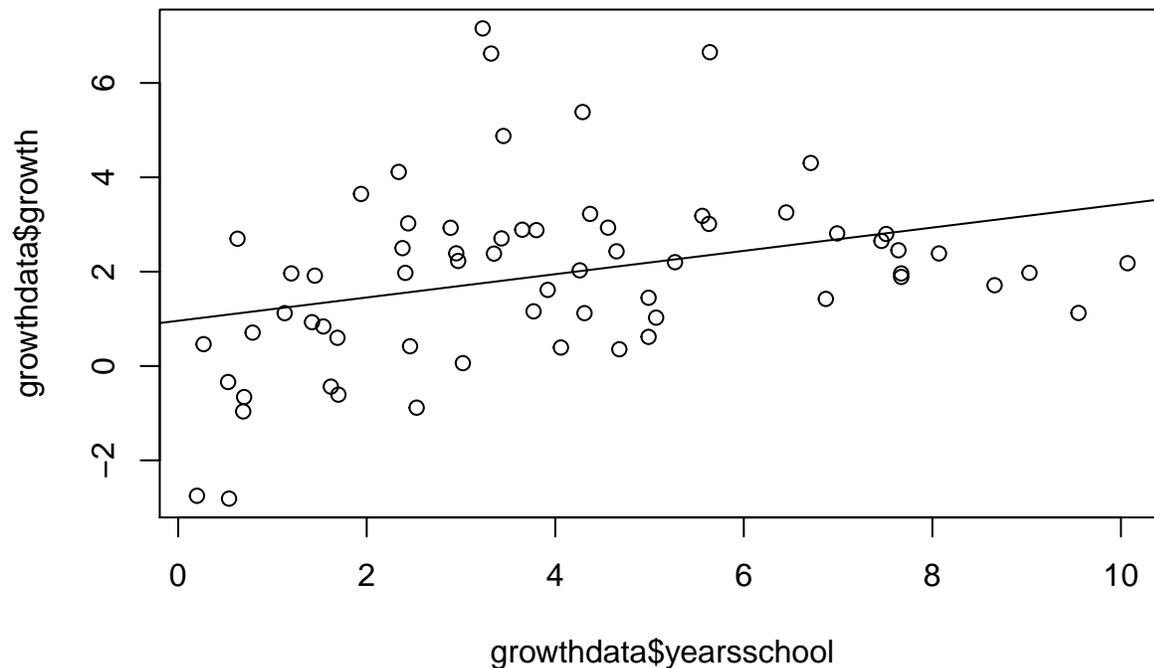
```
growthmodel$coefficients
```

```
##          (Intercept) growthdata$yearsschool
##          0.9582918          0.2470275
```

The output means when $Y =$ (growth rate of real GDP) and $X =$ (average years of schooling of adults in 1960), the equation of the line that best describes the linear relationship between these two variables is $Y = 0.958 + 0.247X$.

In a later tutorial, we will discuss the precise equation and hypothesis testing on that equation at length. For now, let us add a graph of this line to our scatter plot, so that we can see the data and the best fitting line together on one graph. The function `abline()` allows us to add a line graph to our scatter plot. We will add the best fitting line by passing `growthmodel` as a parameter to `abline()`. In the code below, we again produce our scatter plot using the same call to the `plot()` function as above. Then on the very next line of code, we call the `abline` function to produce the straight line.

```
plot(x=growthdata$yearsschool, y=growthdata$growth)
abline(growthmodel)
```



We can see from this graph that an upward sloping line describes well the relationship between years of schooling of adults in 1960 and the subsequent 35 year average growth rate of real GDP. That is, our variables seem to display a *positive, linear comovement*.

3. Estimating the Pearson correlation coefficient

The **Pearson correlation coefficient** is a measure of the strength of a **linear** co-movement between two variables. **Linear comovement** implies that either an upward sloping or downward sloping **straight line** best describes the relationship.

The Pearson correlation coefficient takes values only between -1.0 and +1.0. The stronger is the relationship, the closer the points on the scatter plot will be to the best fitting line. For a positive relationship, the stronger it is, the closer the correlation coefficient will be to +1.0. For a negative relationship, the stronger it is, the closer the correlation coefficient will be to -1.0. If the relationship is weaker, the observations will be farther from the best fitting line, and the correlation coefficient will be closer to 0.0.

The function `cor` can be used to compute the Pearson correlation coefficient for two variables as follows:

```
cor(x=growthdata$yearsschool, y=growthdata$growth)
```

```
## [1] 0.3309986
```

We see from our result that the sample estimate for the Pearson correlation coefficient is 0.33. Since this number is positive, the two variables are positively correlated.

4. Hypothesis testing and confidence intervals

Our sample estimate for the correlation coefficient is positive, but is this enough evidence that there is a relationship between years of schooling and real GDP growth in the population? To answer this, let us conduct a hypothesis test with the following null and alternative hypotheses:

Null hypothesis: $\rho = 0$

Alternative hypothesis: $\rho \neq 0$

Following common statistical notation, we use ρ to denote the *population* Pearson correlation coefficient. The null hypothesis says that the two variables are not correlated, i.e. that there is not a linear relationship. Like all null hypotheses, it states that a population parameter is *equal to* some specified value (zero in this case). The alternative hypothesis says that the two variables are correlated, that there is *some* linear relationship, either positive or negative. The not-equal sign in the alternative hypothesis implies that this is a *two-tailed* test, so either positive or negative Pearson correlation coefficients significantly far away from zero will allow the null to be rejected.

The function `cor.test` can be called to conduct this hypothesis test as follows:

```
cor.test(x=growthdata$yearsschool, y=growthdata$growth, alternative="two.sided", conf.level=0.95)

##
## Pearson's product-moment correlation
##
## data: growthdata$yearsschool and growthdata$growth
## t = 2.7842, df = 63, p-value = 0.007077
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.09474858 0.53195301
## sample estimates:
##      cor
## 0.3309986
```

The first two parameters tell the function which variables to estimate a Pearson correlation coefficient for. The parameter `alternative="two.sided"` tells the function to conduct a two-tailed hypothesis test. Finally the parameter `conf.level=0.95` is used to conduct a 95% confidence interval for the population Pearson correlation coefficient.

The p-value for the hypothesis test is 0.007, which is far below a common significance level of 0.05. With a high degree of confidence we can state we have found sufficient statistical evidence that the average years of schooling is correlated subsequent real GDP growth.

Confidence Interval

The 95% confidence interval is also included in the output to `cor.test`. The results reveal an interval estimate for the population Pearson correlation coefficient between 0.095 and 0.53. With 95% confidence, this interval contains the true population Pearson correlation coefficient. This range includes all positive numbers, but ranges from somewhat weak but positive correlation to strong positive correlation.