Estimating the Population Correlation

James M. Murray, Ph.D.
University of Wisconsin - La Crosse

Updated: September 24, 2017

PDF file location: http://www.murraylax.org/rtutorials/correlation.pdf

HTML file location: http://www.murraylax.org/rtutorials/correlation.html

---

Note on required packages: The following code requires the packages in the `tidyverse`. The `tidyverse` contains many packages that allow you to organize, summarize, and plot data. If you have not already done so, download and install the libraries (needed only once per computer), and load the libraries (need to do every time you start R) with the following code:

```
install.packages("tidyverse") # This only needs to be executed once for your machine
library("tidyverse") # This needs to be executed every time you load R
```

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

1. Download the Data

Stock and Watson's *Introduction to Econometrics* textbook includes a data set with economic growth and education data for 65 countries from 1960-1995. A subset of that data set is available to download from `http://murraylax.org/datasets/growth.RData`

The code below downloads the data file assigns the data set to an object we create and call `growthdata`.

```
load(url("http://murraylax.org/datasets/growth.RData"))
```

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 scatter plot 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 `ggplot()` function as follows:

```
growthplot <- ggplot(growthdata, aes(x=yearsschool, y=growth)) +
  geom_point() +
  xlab("Years Schooling") +
  ylab("Growth Rate %")
```

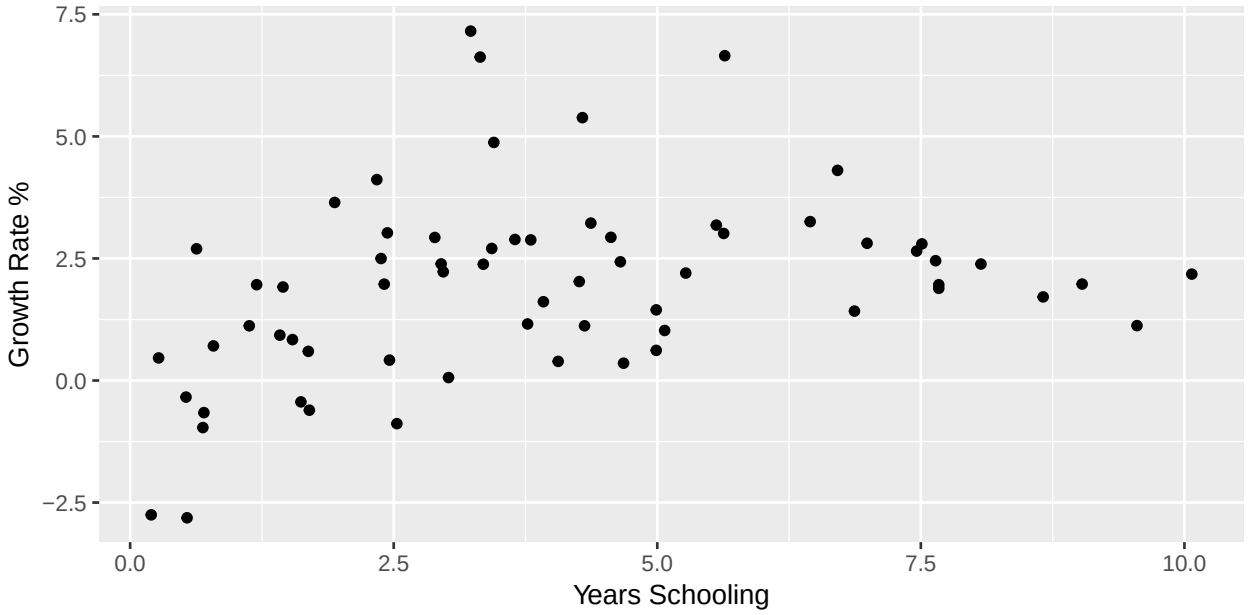
The first line of the `ggplot()` function call sets the data and aesthetic layers of the plot. The first parameter `growthdata` tells ggplot where to find the data. The second parameter, `aes(x=yearsschool, y=growth)` maps the variable `yearsschool` to the x-axis and the variable `growth` to the y-axis.

The next line adds the geometry layer. In this case, `geom_point()` creates a point for each pair of observations.

The last two lines set the labels for the x-axis and y-axis.

We save the output to an object we call `growthplot`. We can view the plot by entering the name of this object at the R console:

```
growthplot
```



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(growth ~ yearsschool, data=growthdata)
```

The first parameter we passed to the function `lm()` is a *formula* of the form `y ~ 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
```

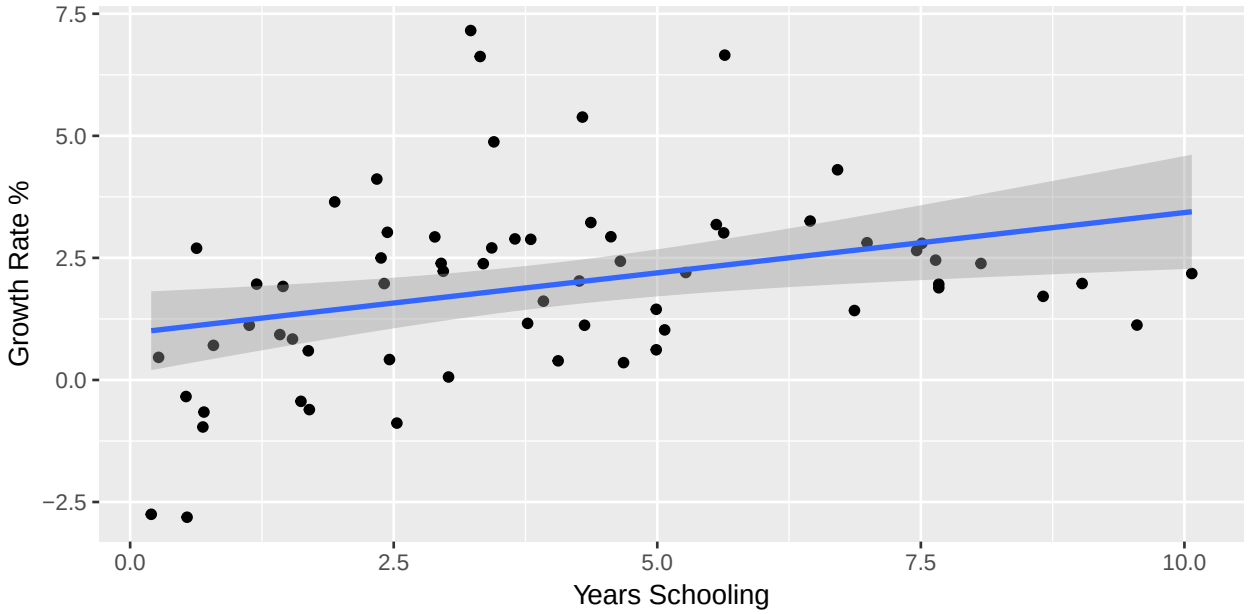
```
## (Intercept) yearsschool
##   0.9582918   0.2470275
```

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.

Below, we add another geometry layer to our existing `growthplot` object. The call to function `geom_smooth(method="lm")` computes the same linear relationship that we computed above and plots a line and gives shaded areas representing the margin of error for a 95% confidence.

```
growthplot + geom_smooth(method="lm")
```



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 co-movement*.

3. Estimating the Pearson correlation coefficient

The **Pearson correlation coefficient** is a measure of the strength of a **linear** co-movement between two interval or ratio variables. **Linear co-movement** 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, method='pearson')
```

```
## [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 for the Pearson correlation coefficient

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 the Greek letter 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 result in the null hypothesis 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, method='pearson')
```

```
##
##  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.

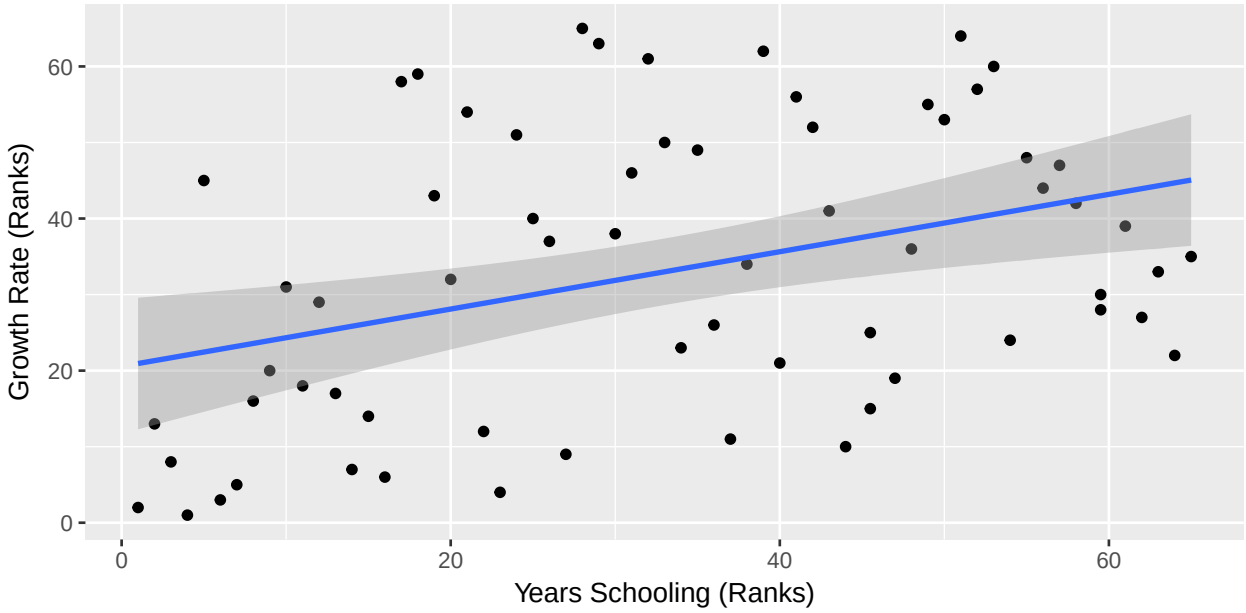
5. Spearman Correlation Coefficient

The **Spearman correlation coefficient** is a non-parametric measure of the relationship between two variables, which measures the strength of the relationship between the **ranks** of observations in the two variables. Because the calculation relies only on ranks, the method is appropriate for ordinal data as well as interval or ratio data.

Because the Spearman correlation measures the strength of a **linear relationship on ranks** between two variables, and ***not*** a **linear relationship on the actual data** of two variables, it can be used to determine positive or negative relationships that are not best described by straight lines.

We can create a scatter plot that illustrates the relationship **in the ranks** of average years of schooling of adult residents and the subsequent average growth rate over the next 35 years. We call the `ggplot()` function like before, but instead of passing in the raw data for years of schooling and real GDP growth, we pass in the ranks, with inner calls to the function, `rank()`

```
ggplot(growthdata, aes(x=rank(yearsschool), y=rank(growth))) +
  geom_point() +
  xlab("Years Schooling (Ranks)") +
  ylab("Growth Rate (Ranks)") +
  geom_smooth(method="lm")
```



The Spearman correlation coefficient estimates the strength of the linear relationship between the ranks. It is exactly the same as the Pearson correlation coefficient, expect applied to the ranks of the data. We can compute the estimate with either of the following methods:

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

```
## [1] 0.376975
```

```
cor(x=rank(growthdata$yearsschool), y=rank(growthdata$growth), method='pearson')
```

```
## [1] 0.376975
```

Hypothesis Testing

We can conduct hypothesis tests on the Spearman correlation coefficient in the same manner as the Pearson correlation coefficient. If we want to test for evidence for a relationship between schooling and economic growth, we would consider the following null and alternative hypotheses:

Null hypothesis: $\rho = 0$
Alternative hypothesis: $\rho \neq 0$

Again we call the function `cor.test()` to conduct the hypothesis test, this time specifying the Spearman method with the parameter, `method='spearman'`.

```
cor.test(x=growthdata$yearsschool, y=growthdata$growth,
         alternative="two.sided", method='spearman')
```

```
## Warning in cor.test.default(x = growthdata$yearsschool, y =
## growthdata$growth, : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
```

```
## data:  growthdata$yearsschool and growthdata$growth
## S = 28510, p-value = 0.001966
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.376975
```

Equivalently, we can conduct hypothesis tests and confidence intervals on the Spearman correlation coefficient with a call `cor.test()` using the *Pearson* method, but submit the ranks of the data instead of the raw data.

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

```
## Warning in cor.test.default(x = rank(growthdata$yearsschool), y =
## rank(growthdata$growth), : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  rank(growthdata$yearsschool) and rank(growthdata$growth)
## S = 28510, p-value = 0.001966
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.376975
```

Using either method, we find a p-value equal to `0.001966`. Since this is below common significance levels, we reject the null hypothesis and conclude there is sufficient statistical evidence that there is a correlation between schooling and economic growth.