Chapter Thirteen: Correlation and Linear Regression

Simona Simona

6/25/2020

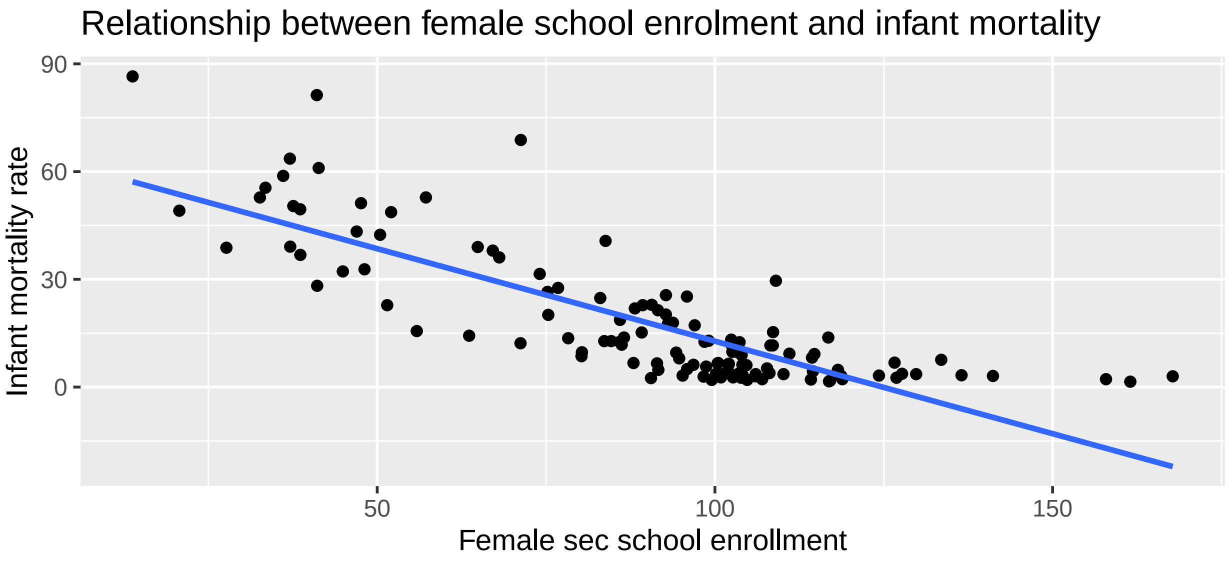
In this chapter we shall look at:
- What is correlation
- What is regression

In the previous chapter we have looked at bivariate relationships involving categorical variables. In this chapter, are considering relationships between variables measured at the interval and ratio levels. Again, as is the case in all bivariate analysises, we ask the 3 questions. Is there a relationship between the variables concerned?, What is the strength of the relationship? and What is the direction of the relationship?

In nominal data, the answers to these questions were found by using contigency tables and chisquare while in ordinal data, we used contigency tables and gamma. These methods would be inappropriate to use with interval and ratio data. This is because these variables tend to have many data points which may make contigency tables displaying the values as many rows or columns as the number of values in the data. As such, interval and ratio bivariate associations are visualised by scatter plots, which gives an idea whether a relationship extists between the two variables. We use correlation coefficients to measure the strengths and direction of the relationship. Regression adds another layer to this by showing how much an independent variable affects the dependent variable. This is what we will consider in this chapter

Scatter plot

We have already studied scatter plots in details in our data visualisation chapter. We will use it here just to show the joint distributions of our two variables we wish to analyse in this chapter which are again taken from the World Indicators data of the World Bank: Infant mortality rate and female secondary school enrollement:



It doesn't matter much how the variables are arranged on the x,y coordinates but you should always put your indepdent variable on the x-axis and the dependent variable on the y-axis. Each point on the scatter plot represents an individual country and the combination of its measure of female secondary enrolment and infant mortality rate. This is real data of the 2016 measurements from the World Bank data bank. Always remember to put the independent variable on the x-axis and the dependent variable on the y-axis. In this case the independent variable is female secondary school enrolment and the dependent variable is infant mortality rate. You can see that there's a pattern being created by this data. The clarity of the pattern can be enhanced by drawing the **line of best fit** or **regression line** to the pattern of the data points, which we have alread fitted in the scatter plot.

Now we can try to answer our three questions. Is there a relatioship between the two variables? A relationship between two variables is determined by observing whether thw two variables **vary** together. In other word, if a change to the values one variables enlists similar changes in the values of another. In this case, does the value of infant mortality (the dots above each score on the x-axis) change when the value of secondary school enrolment changes? The answer to this question is yes. You can see that the values that are furthest on the x-axis are located around zero on the y-axis and the data points rise as you approach zero on the x-axis.

The line of best fit also gives a good clue that there is a relationship between two variabes. The line of best fit would be exactly parallel to the horizontal axis (x-axis) if there was not relationship between the two variables. In our case, the line lies at an angle to the x-axis.

How about the strength of the relatioship? From the scatter plot the strength of the relationship is determined by how the data points are spread around the line of best fit. The more the data points cluster along the line of best the stronger the relationship. In fact, if the two variables have a perfect relatioship, the data points will form a straight line. Of course you will not find any two variables obtaining perfect correlation. Perfect correlation is only attainable when a variable is correlated with itself.

Direction? On a scatter plot, direction is determined by looking at the angle of the line of best fit. The direction of a relationship denotes the extent to which the two variables 'move' together and it is either positive or negative. Positive direction entailes both variables moving in the same direction: one one variable incereases, the other also increases or one variable decreases, the other decreases too. Negative direction is the case when one variable decreases the other increases or the other way around. In this regard, our graph shows a negative relationship, when secondary school enrolment is high, infant mortality goes down.This means that those countries who have high female secondary school enrollment levels tend to have lower rates of infant mortality. This makes sense. That's is the beauty of quantitative methods, it makes sense. If you find that your outputs are against commonsense, you might want to have a look at what you are doing, there could be something wrong. Of course this is not to say that all results should align with commonsense understanding. Sometimes it is not the case.

Put a graph of correlation types here!!!!

Pearson correlation coefficient

Sometimes we can not be absolutely sure about the strength and direction of the relationship between two variables by just looking at the scater plot and the line of best fit. The graph only gives us an impression of the relationship but to be sure that the nature of the relationship, we have to statistically measure it. For that, we will use **Person's correlation coefficient (r)**, which gives us a value indicating the extent of a **linear** relationship between two interval/ratio variables. **Pearson's r** as it is often called ranges from -1 to 1[1], with a coefficient of 1 indicating perfect positive correlation, meaning as one variable increases, the other increases by the same amount. Conversely, a coefficient of -1 means perfect negative correlation so that as one variable increases, the other decreases by proportional amount. A coefficient of zero says there is no relationship between the two variables: when one variable changes, the other remains the same.

The formular for calculating Pearson's correlation coefficient is:

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{[\sum(X - \bar{X})^2][\sum(Y - \bar{Y})^2]}}$$

This formular might look scary but what it actually does is calculate the extent to which the two variables correlate. We will not go in details to make manual computations as we are lucky enough to rely on software to do that for us. The final result you will get is unlikely to be perfect correlation but you will get a value between -1 and 1. A relationship between two variables exists if the correlation coefficient is not zero. The direction of the relationship determined by the sign associated with the value of the correlation coefficient. A negative sign is the case when the relationship is negative and a positive sign when the relationship is positive. The strength of the relationship is determined by how close the absolute value[2] of the correlation coefficient is to 1. An absolute value closest to 1 shows a stronger relationship while a value closer to zero indicates a weaker relationship.

We often use descriptive words when interpreting pearson's correlation coefficient values. Although there are disciplinary differences, it generally accepted that the pearson's r values between the coefficient of 0.00 and 0.3 would indicate a weak relationship, the values between 0.3 and 0.6 would be a moderate relationship, between 0.6 ad 0.8, a moderately strong relationship and above 0.8 a strong relationship. Bearing in mind the direction of the relationship of course, so that if you have an 0.9 value you would say strong positive relationship.

[1] This means if you find you have calculated the correlation cofficient of less than -1 or more than 1, just know that you have made a grave mistake!

[2] Regardless of the sign associated with the value

Coefficient of determination (r^2) is a much less ambigous and more direction interpretation of pearson's r that we can use. It indicates the amount of variation in the dependent variable that is explained by the independent variable. This is why it is impoartant when you are doing correlation to determine from the word go what would be your dependent and independent variables. This statistic is calculated by just multiplying the correlation coefficient value by itself. We will discuss this more when we implement correlation in R.

Correlation is not Causation

This is very important to stress here that the correlation coefficient and the coefficient of determination do not both give us causality. Meaning that they do not tell us whether one variable causes the other. In our example, we can say that female school enrolment is associated with infant mortality but we can not say school enrolment causes mortality. That's why even if we swapped the variables and put school enrolment on the y-axis, the shape will remain and the results of computation will not change.

The reason why we need to apply this caution is twofold: **the third variable problem** and **direction of causality**. The third variable problem simply means the two variables we have are not enough to determine causality because there could be other variables that responsible for the variation in the outcome variable. For example we can not attribute all variations in infant mortality to secondary school enrolment alone. That would be too simplistic. There are obviously many other factors that could affect child mortality. For us to determine causality, we need to account for all the factors that can affect our outcome variable. There are methods outside the scope of this book that can do that.

Direction of causality: Correlation coefficients only tell us about the relationship between two variables and nothing about which variable variable cuases the other. Which means even if the third variable wasn't a problem we would still have problems diciphering which variable is the cause and which one is the effect. Therefore, it may make sense to assume that hight female secondary school enrolment causes less mortality, there is no statistically reason to suggest that the opposite may be also true: that lower child mortality may cause female secondary school enrolment to be higher.

Implementing correlation in R

We will use the variables we are working with: infant mortarity rate and female secondary school enrolment from the world bank dataset. In R, we calculate correlation using the *cor()* function which takes on the two variables. The function will produce NA if there are missing values in any of the variables. We use the argument *complete.obs* so that the correlation coefficient can be calculated only on complete cases.

```r
cor(data$mri, data$secnrolf, use = "complete.obs")
```

```
## [1] -0.8173838
```

The correlation coefficient is -0.8173838 which signifies a strong negative correlation between female secondary school enrolment and infant mortality rate. We can interpret this relationship as *it is expected that as the female secondary school enrolment increases in a country, infant mortality will decrease.* Notice that we do not use the actual correlation coefficient value in our interpretation. This is because of what noted earlier that the correlation coefficient cannot interpreted directly. We can only do that if we convert it to the coefficient of determination. We do that by multiplying the correlation coefficient by itself.

```
-0.8173838^2
```

```
## [1] -0.6681163
```

We can use a percentage of the coefficient of determination and say that female secondary school enrolment explains approximately 65% of variations in infant mortality. This makes a better direct interpretation because it is much more intuitive.In other words we can attribute 65% of variations in infant mortality to female secondary school enrolment. However, another caveat with both the correlation coefficient and the coefficient of determination is that they don't tell us whether the relationship between the two variables is statisically significant. For that, we have to conduct a hypothesis testing of correlation.

Linearity and non-linearity

The underlying assumption of the pearson correlation is that the relationship should be linear. What is lenearity? Linearity means the data points on a scatter plot forms a pattern that can be approximated by a straight line. If this assumption is broken, you may not us pearson's correlation. All the more reason that you should always run scatter plots before doing any correlation analysis.

Spearman's correlation (ρ) is an alternative to pearson's correlation which can be used when the linearity between the two variables is not necessarilly achieved. Spearman's correlation is not an option for **non-linearity** but for monotonic relationships. Monotonic relationships occur in both negative and positive direction. A positive direction monotonic relationship occurs when one variable increases the other increases or when one variable increases the other decreases. A negative monotonic relationship happens when one variable increases the other decreases. An important component in monotonic relaionships is that they don't necessarily have to be linear. This removes the restrictions that Pearson's r has.

The procedure of correlation should start with a scatter plot and observe the pattern produced by the data points. If a linear relationship is observed, you may carry out Pearson's correlation but you produce a pattern that doesn't neccesarily form linearity, you may want to use Spearman's correlation instead. In R, the implemation of Spearman's correlation is achived by supplying an additional argument of *method = "spearman"*. The interpretation of the output is not different from Pearson's r.

```
cor(data$mri, data$secnrolf,method = "spearman", use = "complete.obs")
```

```
## [1] -0.7981627
```

You can see that the results here are different from the Pearson's correlation. This is because the restrictions of linearity is removed here. Therefore there is a slight difference in the formular for calculating the (rho) as opossed to the (r) but we will conveniently avoid those details here. If you want a good treatment on these, you may want to look at *Discovering Statistics Using SPSS* by Andy Field or *Statistics* by Joseph F. Healey

Hypothesis testing for correlation significance

To conduct hypothesis testing, we follow all the steps we have met in **Chapter Five**. We need to note before doing this though, that we are using random variables. That is to say we are working with the sample that was collected using random sampling methods (**See chapter seven**) and we wish to test if the correlations we are observation is true in our population of interest. Accordingly, we need to state the null and alternative hypothesis then calculate the Pearson's correlation coefficient. The test statistic is calculated after this and compared to the values in the t-distribution to decide whether to reject or fail to reject the null hypothesis.

The null hypothesis for Pearson's r is that there is no linear relationship between the independent and the dependent variables (x and y) in our population of interest. In other words, Pearson's r has the value of zero. This can symbolically be represented as

$$H_0 : r = 0$$

The alternative hypothesis is that there is a linear relationship between the indepedent and dependent variables in our population of interest. This also means Pearson's r is not equal to zero. Symbolically, we represent the alternative hypothesis as

$$H_1 : r = 0$$

Instead of calculating the correlation coefficient and associated t-statistic mannually, we use R and it will do all the hard work for us. All we need to do is understand and interpret the output. The `cor.test()` function performs the analysis on the supplied variables. Notice that here, we use the argument `na.action = na.rm` to ignore missing values in the calculation. We use the same variables female secondary school enrolment and infant mortality rate using the world data.

```r
cor.test(data$mri, data$secnrolf, na.action = na.rm)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$mri and data$secnrolf
## t = -15.736, df = 123, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8683495 -0.7493349
## sample estimates:
##        cor
## -0.8173838
```

The output reports Pearson's product moment coefficient. The *data* is the two variables that we have supplied.The t-statistic is reasonably bigger (-15.736) with a degree of freedom of 123 and a very small resultant p-value. As usual our main concerned here is the p-value. Is is less that .05? In this case it is, so we make the decision of rejecting the null hypothesis of no relationship between the two variables. We can say with confidence that the relationship between female secondary school enrolment and infant mortality rate is statistically significant.

We can see from the output that R has also kindly given us the alternative hypothesis as well as calculated Pearson's r and it's confidence intervals. The 95% confidence intervals are between -0.8683495 on the lower bound and -0.7493349 on the upper bound. We interpret confidence intervals by saying that we are 95% confidence that the true correlation coefficient in the population falls between -0.87 and -0.75.

Implementing Spearman's correlation significant test is R is not different from Pearson's significant test. All we need to do is supply the argument `method = "spearmen"` as we did before. The interpretations is the same as the Pearson's correlation significant test.

```
cor.test(data$mri, data$secnrolf, method = "spearman", na.action = na.rm)
```

```
## Warning in cor.test.default(data$mri, data$secnrolf, method = "spearman", :
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$mri and data$secnrolf
## S = 585302, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.7981627
```

Assumption of Pearson's r

Assumptions are basically things that needs to be in place for the test to work properly. The Pearson's correlation coefficient significant test has some assumptions that needs to be met for it to ptoduce reliable results. The first and probably most important assumptions is that the two variables have to be measured at the interval/ratio level. This means if any of the variables is measured at the ordinal or nominal level, the test will be invalid.

The second assumption is one we have met already. That the relationship between the two variables should be linear. This is important because the correlation coefficient measures a linear relationship between the two variables. In cases where the two variables are associated in a none-linear way, the line of best fit and the significant test may not appropriately describe the relationship.

The third assumption is that the two variables should be normally distributed. It is important to check that the two variables have a bell shape before carrying out the analysis. It is also important to check if there are some outliers which might skew the test statistic especially in small samples. Both of these procedures can easily be implemented in R and we will demonstrate how to do this later in **Chapter Fithteen**

The fourth assumption is called **homoskedasticity**, meaning the relative dispersion of scores in the two variables is about the same. That is to say that the Y scores are evenly spread above and below the line of best fit throughout the length of the line. The opposite of homoskedasticity is heteroskedasticity and it occurs when the variance or Y scores are not uniform for all the X values. Tests for homoskedasticity are also discussed in **Chapter Fifteen** ahead. We are not discussing these diagnostic tests in this chapter because most serious studies do not solely use correlation analysis. Correlation analysis in real studies is often done in conjuction with regression analysis. Which is why we will discuss the diagnostic tests after discussion regression analysis in the nest chapter.

Correlation matrix