Chapter Fourteen: Regression Analysis

Simona Simona

7/14/2020

In this chapter,we shall look at the following
- The line of best fit
- Regression analysis
- Eye ball method
- Least squares method

In the previous chapter, we looked at how can test the relationships between two variables. We used Pearson and Spearman's correlations for this. We noted the caveats of these methods though and major among them was the fact that they only provide us with the extent of correlation between two interval/ratio variables and nothing else. Well, the world is far more complicated and just knowing that the two continuous variables are positively or negatively correlated is not enough to even begin to solve problems that we face everyday.

Regression analysis is a much more versetile tool than correlation. With regression analysis, we are able to calculate the effects of one variable on another. We can also predict the value of our outcome variable based on the predictor variable and yes we can add many more different types of variables to our analysis. Regression can even get more complicated than that. In fact, most of what is called quantitative social science is regression analysis in different forms and shapes. That's why we spent sometime in this chapter to develop intuitions about regression analysis and then move from bivariate linear regression to multiple regression. It is important that you master the basics of regression analysis because it forms the foundation of most of quantitative analyses in the social sciences.

Basics

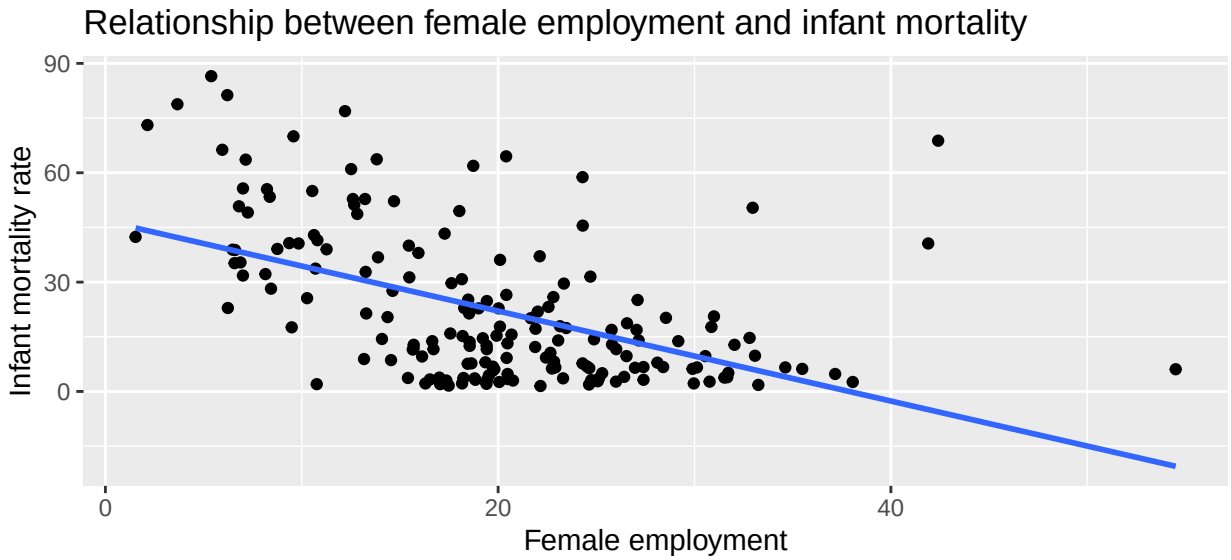
Regression builds from the correlation analysis that we looked at in the previous chapter but this time around, we will be able to do more with our two continuous variables. In the example of two variables we used in the previous chapter (female secondary enrolment and infant mortality), we were not able to predict the amount of reduction in infant mortality based on the value of female secondary school enrolment. We wouldn't say for example that 10% increase in female secondary school enrolment induces a 15% reduction in infant mortality rate. I am sure you can see how much of a difference being able to make this determination can bring to our understanding of the two variable. Even better, imagine the potential of being able to extend this tool to include different combinations of variables. We could solve a few social, health and policy problems, not so? That is the essence of regression analysis. Being able to predict the value of our outcome variable from one or more predictor variables. But the question is what is the logical foundation of prediction?

It starts from the line that we were able to fit across the scatter plot in the previous chapter. We

called it the line of best fit. Well, the other name of the line of best fit is the **regression line**. I didn't want to use that name then to avoid confusions because we had not yet encountered regression analysis. But hence forth, we will be using the regression line. Fitting a line to data to be used to predict values of outcome variables based on the values of predictor variables is also technically called **modelling**. What this entails is that the outcome we want to predict for a particular case or observation can be predicted by whatever model we fit to the data. As you can imagine, there are so many types of models out there that can be used to analyse different types of data but in this chapter we use **linear models**. This is because we are going to analyse variables that have a linear relationship. Remember what linear relationships are: **relationships that can be represented by a straight line on a scatter plot**. This gives a good intuition that you can use in more complex analyses.

Simple linear regression

To understand linear relationships (linear models), we have to start simple. What better place to start than **simple linear regression**. It is basically the type of regression analysis that involves two interval/ratio variable, meaning one dependent and one independent variable. In other words, we predict the value of one dependent variable based on the values of one independent variable. How do we do this? Look at the figure below:



We use the regression line to describe the best representation of relationship between the two variables. To make a prediction of the value of the dependent variable based on the independent variable, we observe the value of the independent variable on the x-axis until we hit the regression line and we then follow it until we find the corresponding value on the y-axis. For example, if we want to find the infant mortality rate induced by a 20% female employment levels in the population, we follow the 20% value on the x-axis until we hit the regression line and then find the corresponding value on the y-axis, which in this case is approximately 20. We can then say that countries with a 20% female employment levels in the population are expected to have 20 infant mortality per thousand of the population.

The method we used above to find the corresponding value on the y-axis is called the **eyeball method** and it is just exactly as the name suggests: using your eye balls to draw a line across the data points on a scatter plot and find the corresponding value of the dependent variable given the

value of the independent variable. I guess you already have doubts on the accuracy of this method. Yes! your eyes may look pretty good but they probably found themselves on your head for other purposes not to draw regression lines.

However, although the eyeball method is not very effective, we have managed to arrive at an approximate value that we believe represents the corresponding y value from the given x value. This was possible because the graph is two dimensional, owing to the fact that we are working with only two variables. However, things can get much worse when things get more complicated. Can you imagine appproximating the value of your dependent variable when you three or more independent variables. First of all we can not even represent such an analysis graphically. Good luck on drawing a five dimensional graph much less six or more.

Another thing that we seem to have simplified too much is the regression line itself. It is certainly not that simple. Do you know that you can draw many regression lines on the same scatter plot? The question is, how do you decide where the line starts and the angle it takes going forward? These are all very legitimate questions. Since this line is so important that if we messy it up it messes up our entire regression analysis, we need some reliable method that we can use to accurately draw it. Enter the **regression equation**

The regression equation

The regression equation offers a much more reliable way of calculating the the dependence of the outcome variable on the independnet variable. So much so that even if we have more relationships to analyse which we can not be represented on a two dimensional or three dimensional graph, we can just extend it to include such complexity. Also, the regression equation gives accurate starting points in drawing our linear regression line and the angle the line takes is given.

The equation line comes from the equation of a straight line, if you still remember your high school arithmetic. For you to draw a straight line, you requare two pieces of information. Either I give you two points, the first point is where you are supposed to start from and the second is where you should end. Or I give you one point, where the line begins from and the angle or slope it should take. We use the later. Below is the equation of a line

$$Y = mx + c$$

This equation is what has changed by statistician as follows

$$Y = a + bx$$

Where `y` is the variable on the y-axis, `x` is the x-axis variable, a is the *intercept* and `b` is the slope. The intercept is the point at which the regression line crosses the y-axis. It is usually defined as the value of y when x is zero. The slope is the angle that the regression line takes. Now we have the two pieces of information we need to make a regression line but we still a few issues we need to sort out. Before the analysis is done, we normally have both the dependent y and the independent x variable(s). Like in the example above, we are given infant mortality rate as our dependent variable and female employment as the independent variable but we don't normally know the intercet (a) and the slope (b) before the analysis is done. In statistics, the parameters we don't know before the analysis is done are usually represented by Greek letters. Therefore a and b change to α and β in our regression equation.

You should have noticed that our regression line did not strike every data point in our scatter plot. In fact it is impossible for us to draw a line that strikes all the data points unless we are measuring a variable by itself, also known as **perfect correlation**. We will always have some gap between the data points and the regression line. That gap is called the **error term** or **residuals**. The error term is usually included in the equation with a Greek symbol ε. That's why you have often seen the equation line looking like:

$$y = \alpha + \beta x + \varepsilon$$

Now, it is important to note that the α and β values we get in the analysis represent predicted values of the population also called parameters. Predicted parameter values in statistics are often represented by a$\hat{}$, which in this we have $\hat{\alpha}$ and $\hat{\beta}$. As a matter of fact, we also predict the y based on the sample values of x. In our example above when we were trying to find the infant mortality rate induced by a 20% female employment levels in the population, we knew the value of x (20%) from the sample and it is on the basis of this value that we performed an eyeball analysis of following the value of x up to the regression line to get the corresponding value of y. In essence, the y should be $\hat{y}$. The regression equation should therefore look like:

$$\hat{y} = \hat{\alpha} + \hat{\beta} x + \hat{\varepsilon}$$

where $\hat{y}$ is the predicted outcome variable, $\hat{\alpha}$ is the predicted value of the intercept and $\hat{\varepsilon}$ is the error term.

Ordinary least squares

There are many regression lines that can possibly be drawn across the data points on a scatter plot to approximate a relationship between two variables but as we have already observed, every line that we are able to draw will not strike all the data points on the scatter plot resulting in the error term. For linear regression to be viable, we need a method that we can use to ensure that we maintain the error term to its minimum. In linear regression, that method is called **ordinary least squares**. Ordinary least squares is a method of approximating the regression line which results into the least amount of difference between the line and the data points. In other word, the method of ordinary least squares lets us pick the best line that represents the relationship between the independent variable and the dependent variable among all the possible lines.

To understand the logic that the OLS uses to approximate the regression line, you need to think back to **Chapter Four** where we discussed measures of central tendency and measures of dispersion from the mean. The regression line is like the mean, in that it is the model that we use to predict the value of the dependent variable (y) based on the value the independent variable (x). Data points that fall above and below it are deviations akin to the variance. When data points fall along the regression line, it means we have less deviation and thus our model accurately describes the data.The further away the data points are above the regression line, the more the model underestimates their value and in the same vein, the further the data points are below the regression line, the more the model underestimates their value. The accuracy of the model is determined by how much the regression line minimises the the distance between each of the data points and the regression line.

If we think of the regression line as the mean then the data points along it have minimum variation. Therefore for us to calculate the distance between each data point and the regression, we just substracting the value of each data point from the conditional mean which happens to be the regression line. We can denote it as $\varepsilon = y - \bar{y}$. Where ε is the error term, y represents a data point

and $\bar{y}$ is the conditional mean. However the problem is similar to what we found with the variance. Data points are on both sides of the regression line which means we are going to have both positive and negative values when calculating the error. We elimate this problem by squaring the error we find and after we do that we can add the squared differences. We then use this value to gauge how our regression line estimates the relationship between the two variables. If the difference is big, then our model is a shordy one and if the value of the squared differences is small then the model is good. The regression line is the one that gives us the lowest value of the squared differences.

What the method of OLS is used to find the value at which the sum of squared difference is at its minimum. OLS does this by locating the means of both variables and assign them as coordinates through which the line must pass. Then we have the slope which we have to calculate using the the following formular:

$$b = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

We don't really have to care much about it because this method is built in R and it will calculate the sum of squares for us and give us the correct results. Our main task is to know what has happened under the hood and interpret the out of results.

Simple regression in R

In R, simple regression is calculated by creating using the `lm()` function representing linear model. We start with the dependent variable (mri) followed by a tilde `~` followed our dependent variable (employ) then the dataset same, `data` in our case. The model is stored in the `model` object and we need to run the `summary()` function to see the output.

```
model1 <- lm(mri ~ employ, data = data )
summary(model1)
```

```
##
## Call:
## lm(formula = mri ~ employ, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.464 -12.351  -3.038   7.250  74.401
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.7513     3.3833  13.818  < 2e-16 ***
## employ       -1.2345     0.1578  -7.823 4.49e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.64 on 177 degrees of freedom
##   (38 observations deleted due to missingness)
## Multiple R-squared:  0.2569, Adjusted R-squared:  0.2527
## F-statistic:  61.2 on 1 and 177 DF,  p-value: 4.493e-13
```

Interpretation of output

The output shows the call, which is basically the contents in the `lm()` function that we provided. Residuals shows the max value, minimum value and the median. The coefficients section shows the intercept and the independent variable values. For the intercept the value of 46.75 is the intercept and it represents the value of y when x is zero. In countries were there is not employment opportunities for women, infan mortaliy rate is expected to be at 46 per 1000 of the population.