Regression Analysis

For this last workshop, we will learn about regression analysis, a powerful statistical method that allows you to examine the influence of one or many independent variables on a dependent variable. In our introductory unit, we focus on simple linear regression and multiple linear regression. We will also learn how to use dummy variables to incorporate categorical variables. We continue use the Crime Rates Datasets of NSW Local Government Areas (NSW Crime).

1. **Simple linear regression**

Following from the last week example, we will explore how unemployment rates in in NSW Local Government Areas (LGAs) affect robbery rates. In regression model, there is always one dependent variable (in our case, the robbery rate). And if the model uses only one independent variable to predict the dependent variable, it is called a simple linear regression model. In our case, the model can be expressed as following:

Robbery = α+ βUnemployment + ε

where α is an intercept, β is the regression coefficient, and ε is the error. The regression model is to estimate α and β.

To conduct a linear regression analysis, go to Analyze > Regression > Linear (see <Figure 1>) . In the popped-up box, 1) move *robbery* to Dependent, 2) move unemploy to Independent(s) and 3) click OK (see <Figure 2>.



<Figure 1>



<Figure 2>

<Table 1> is the output. Focus on two tables named Model Summary and Coefficients. First, R-square is a measure of model fit, indicating how much your regression model explains the variation in the data. R square of .068 means that our model explains 6.8% of the variance.



<Table 1>

Next, the Coefficients table provides information on the constant (intercept) and coefficients (slope) for robbery rate (note we look at the unstandardized coefficients here). It also provides information on whether the unemployment rate reaches statistical significance (note it does this through a t-Test). Conventionally, three significance levels are set at .001, .05, and .01. In our case, the unemployment rate is significant at .05 level because .014 is greater than .001 and smaller than .05 (we don’t care p-value of intercept). Therefore, the regression model is estimated as following:

Robbery = 1.504 + 4.519Unemployment

The key findings from the regression analysis are that:

* Our model explains 6.8% of the overall variance in the data (R2=.068)
* There is a significant positive relationship between unemployment rate and robbery rate with an increase of the robbery rate by 4.519 for each increase of the unemployment rate by 1, and
* We can predict the robbery rate in the NSW by using the prediction equation $\hat{Y}$ = 1.504+4.519×unemployment rate
1. **Multiple Linear Regression**

The procedure for inputting multiple regression models into SPSS is similar to those for simple linear regression. You simply add more independent variables into the box of Independent(s) (see <Figure 3>). In addition to unemployment, we now add income inequality (giniinc), rental dwelling rate (pctrent), and median age of residents (medage).



<Figure 3>

Again, we focus on Model Summary and Coefficients. Let’s go over what it all means again (see <Table 2>)

* R square of .304 means that our new model including several independent variables can explain 30.4% of the overall variance in the data. This is a quite big improvement in the model fit from our previous model.
* After controlling for other factors, unemployment rate still has a significant positive effect on robbery rate. It is now significant at .001 level and its coefficient becomes larger as one unit increase in unemployment rate leads to an increase of the robbery rate by 7.661
* Income inequality has a significant positive effect on the robbery rate (p<.05). One unit increase in the Gini coefficient of total income rises the robbery rate by 118.235.
* Median age of residents has a significant negative influence on the robbery rate (p <.05). LGAs with older residents have less robbery rates.
* Percentage of residents who rent does not have any significant impact on robbery rate.



<Table 2>

1. **Dummy Variables**

So far we have used only continuous variables. How can we include the effect of a categorical variable? The typical example is to test the effect of gender on wage. If the variable of gender is coded as male=1 and female =2, we know that these values assigned are arbitrary and do not have any numerical value. In other word, being female (=2) isn’t greater than being male (=1). In order to include such categorical variables, we need to transform them into dummy variables to represent subgroups of the categorical variable. Dummy variable is a numerical variable that has either 0 or 1 as its value. 1 indicates belonging to the subgroup and 0 indicates non-belonging to the subgroup. Going back to the example of gender, if you create a dummy variable of female, all the female respondents are given 1 and others are given 0.

This dummy transformation is useful because it enables us to include a category of subgroup in a regression equation. This is done by adjusting the intercept by subgroups. Using the example about the effect of gender on wage, you can write a single equation as following:

Wage = α + β1Female + β2X

Where α is the constant (this is the intercept of male respondents, see below), β1 is the regression coefficient of female dummy variable, and β2 is the regression coefficient of another independent variable (for example, education). β1 is the estimated effect of gender on wage and it adjust the intercept of female subgroup. To figure out how this works, consider the two possible cases below.

For male respondents whose value of the female dummy variable is 0, this is equivalent to

Wage = α + β1(0) + β2X = α + β2X

For female respondents whose value of the female dummy variable is 1, this is equivalent to

Wage = α + β1(1) + β2X = (α + β1) + β2X

And you see that for female respondents, (α + β1) is the intercept. It may be more intuitive if you depict this graphically (see <Figure 4>). If β1 is estimated to have a negative value (meaning females earn less than males), the figure shows the lower intercept of the female regression line than that of the male regression line. In sum, dummy variable enables us to estimate the effect of categories in regression analysis by capturing the group difference via different intercepts between subgroups.



Women’s intercept

Men’s intercept

Source: Wikipedia [https://en.wikipedia.org/wiki/Dummy\_variable\_(statistics)](https://en.wikipedia.org/wiki/Dummy_variable_%28statistics%29)

<Figure 4>

The procedure of including a dummy variable into SPSS is remarkably simple. You just need to recode a variable into a dummy variable and treat it as another independent variable. Going back to our NSW crime data, suppose we want to examine the effect of metropolitan area on robbery rate. The variable *region2* lists regions of LGAs (see <figure 5>). We will make a dummy variable of *metro* by recoding this variable so that greater metropolitan Sydney is coded as 1 and all the other regions as 0 (see <figure 6)



<Figure 5>





<Figure 6>

As you finished the creation of metro dummy variable, you can simply add the dummy into your multiple regression model as below.



<Figure 7>

Based on the result (see <table 3>), do you think being within the metropolital area has a significant effect on robbery rate?



<Table 3>

**Workshop Activities**

Q1. Conduct a simple linear regression analysis between sexual offences and unemployment rate. Interpret the results (the R square, regression coefficient and its significance)

Q2. Conduct a multiple linear regression analysis about sexual offences using all of the independent variables we used with robbery rate. That is unemploy, giniinc, pctrent, medage, and metro. Again, interpret the results like you did with Q1, but when you do so, make comparison about what they showed and did not show.

Q3. Suppose we want to include a dummy variable of rural area into your regression model about sexual offences. To do so, first make a dummy variable by recoding the *urban* variable and then include *rural* into the regression model you used for Q2. Interpret your finding.