# MIS771 Descriptive Analytics and Visualisation

# Topic 5 Tutorial – Multiple Regression, Model Building and Advanced Topics

**Introduction**

In this tutorial you will cover multiple regression, model building and advanced topics in regression analysis.

Specifically, the aims of this tutorial are to:

- To generate correlation matrices to help identify linear relationships.
- To generate scatter plots to help identify relationships.
- To generate regression models to illustrate the key features of multiple regression modelling.
- Define and incorporate interaction terms to model interaction between two independent variables.
- Create interaction plots to illustrate and explain the nature of interaction effects.
- Define quadratic terms to model non-linear relationships.
- Use regression models to make a prediction.

**Scenario**

We continue with the analysis of the BLITZ employee case study. Management is concerned about the wide variation in productivity between employees. Last week we identified *unpaid overtime hours worked* as an important factor in explaining variation in productivity. Management now wishes to take this a step further and has asked you to identify other factors that might affect productivity. The aim is to develop a regression-based prediction model that could help explain variation in productivity.

**Open the data file and install the Data Analysis Tool Pak**

a) Download the file ***BLITS_Dataset_Tut5.xls*** from

"Content → Learning Resources → Topic 5 Folder" in Cloud Deakin. **Save it** to your hard drive.

b) Open the file in Excel.

c) Install the data Analysis Tool Pak.

> **Instruction:**
>
> From the top of *Excel* (*Microsoft Office Ribbon*), click on **File** tab (Figure 1a), select **Options** (Figure 1b), choose **Add-ins** (Figure 1c), and then press **Go…** button to *manage excel add-ins* (Figure 1c). Finally, select **Analysis Tool Pack** and press **OK** (Figure 1d).

**Q1. Scatter diagrams and Correlation analysis**

Before constructing scatterplots and correlation matrix, we need to dummy code *Gender* variable so that it could be incorporated in the regression model.

> **Instruction:**
>
> To dummy code a categorical variable in excel we can us **"IF"** statement. Insert new columns to the right of variable ***Gender*** by *right-clicking* **column C** and selecting **Insert**. This creates a new blank column next to the *Gender* column.
>
> Label new columns as: ***Gender_Dummy.***
>
> To dummy code *Gender*:
>
> - Go to the first blank cell under ***Gender_Dummy*** **column** and type: *=IF(B2="Female",0,1)*. This statement recodes Females to be 0 and others (i.e., Males) as 1.
> - We will use this newly created variable in building the regression model.

a) Three scatter diagrams have already been given in the *Working* worksheet. Construct three more scatter diagrams between *Gender* vs. *Productivity*, *Job Satisfaction* vs. *Productivity*, and *Job Security* vs. *Productivity*. Discuss key features of these scatterplots. In particular, is there any unusual (hidden) feature you could observe in any of the scatterplots?

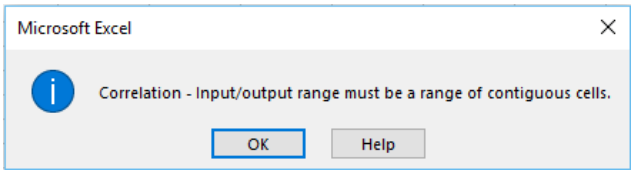
See Tutorial 1 instructions for how to create a scatter diagram.

b) Create a correlation table (matrix) for the following variables: *Productivity*, *Weekly Salary, Age*, *Days Absent*, *Gender*, *Job Satisfaction*, and *Job Security*.

**Instruction:**

From the *Data* tab, select *Data Analysis* and then choose ***Correlation***. The pop-up window in Figure 1 will appear. Enter the ***Input Range*** (i.e., all variables mentioned above together with their corresponding values), select the ***Labels*** checkbox and specify the ***Output Range*** (i.e., where you want to place the correlation table) as shown in Figure 1. Press *OK*.

Dealing with the following *error*:




When creating correlation tables or building regression models, you need to ensure all study variables are place right next to each other (no gaps between columns). In other words, **input/output variables must be a range of contiguous cells**. To overcome this issue, **rearrange columns so all independent variables are together**. It is always *advisable* (but not necessary) to place the *dependent* variable in the last column.

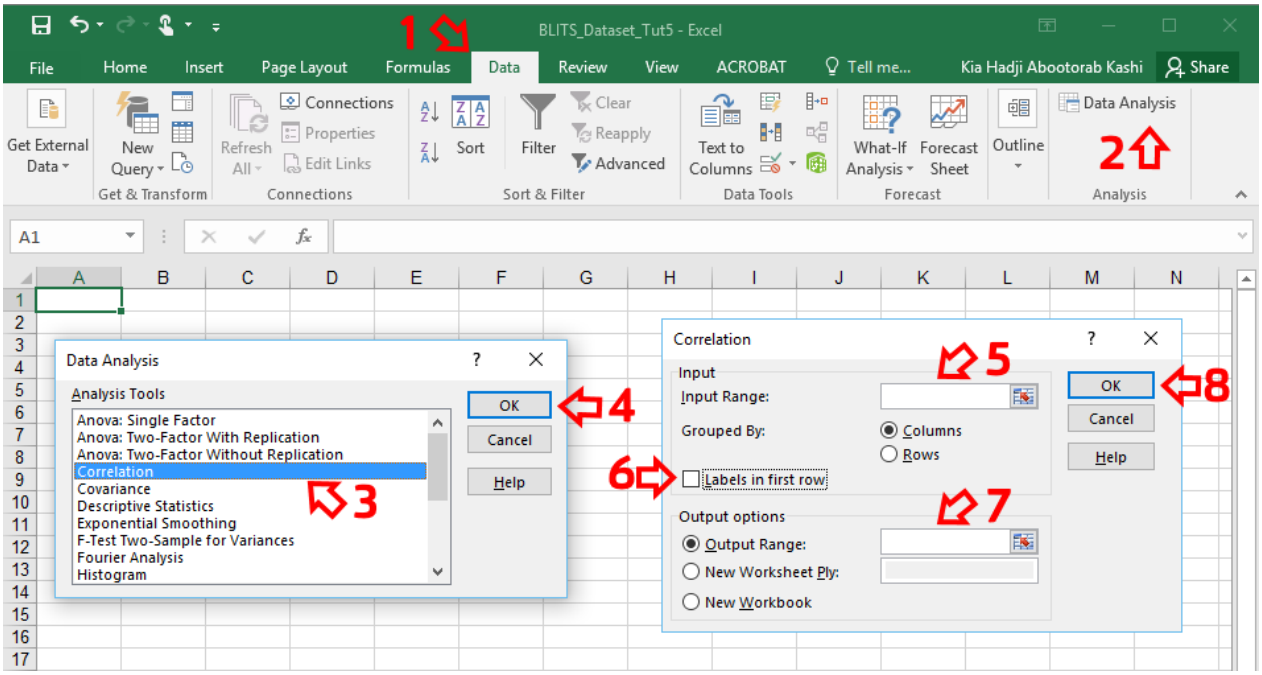



**Figure 1.**

c) Based on your answers from (a) and (b) above answer the following questions:

i. Which of the independent variables would be useful for a regression model because they seem to have a linear relationship with Productivity?
ii. Of the potential independent variables you have chosen, could some pairs cause multi-collinearity?
iii. Decide on your list of potential independent variables for a useful model with Productivity as the dependent variable.

**Q2. Building a Regression Model**

a) Build a **regression model** between Productivity and the potential independent variables you have identified in Q1 (c) (iii) above.

See **Tutorial 4** for instructions for how to create a regression model. Note that since we are conducting a ***Multiple Regression Analysis***, your **Input X Range** must include ***all*** independent variables to be included in the model.

b) Based on your answers from (a):

i. Evaluate your model overall (use the *F*-test).
ii. Are all of the independent variables, individually, significant?
iii. Comment on the explanatory power of the model (use $R^2$).

c) If necessary, adjust the variables in your model in (b) above. Perform residual analysis.

**Instruction:**

Repeat the steps in *Q2(a)* above (i.e., building a regression model). Enter the **X Range** to match your new *independent* variables. This time set select the ***Residuals***, ***Standardized Residuals*** and ***Residual Plots*** checkboxes (see Figure 2). Press **OK**.

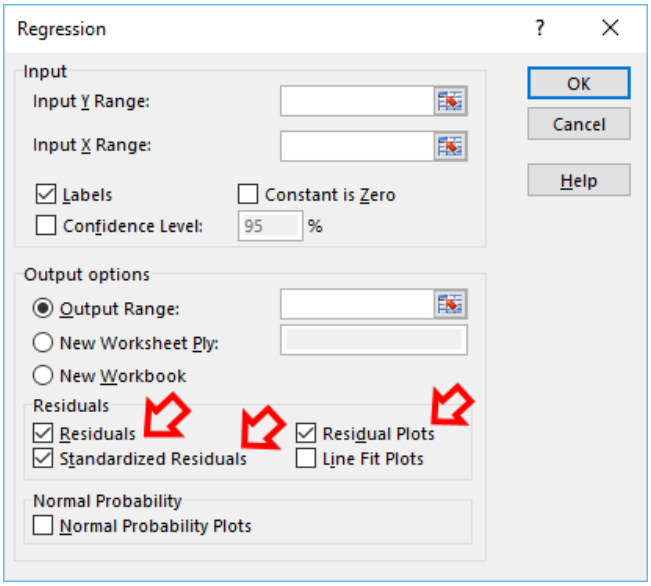



**Figure 2.**

i. Interpret the residual plots.
ii. Interpret the independent variables' coefficients from your model.

**Q3. Working with Non-linear relationship**

A review of scatterplots in **Q1a** indicated that the relationship between *Days Absent* and *Productivity* may be ***non-linear***. BLITZ management is interested in investigating this relationship further. Specifically, you are instructed to model this non-linear relationship in the final regression model developed in **Q2**.

a) Build a new **regression model** by adding *Days Absent* as a second-order polynomial independent variable.

**Instruction:**

To define a second-order polynomial term to your model, create a new independent variable by *squaring* the *Days Absent* measure variable.

- Use Excel equation to create a new variable in Column D (i.e., to create the first polynomial data value in column D, use **D247 = C247^2.** Then copy down the formula to the last row (see Figure 3).
- Rearrange the columns so all independent variables (i.e., *Job Satisfaction*, *Job Security*, *Days Absent*, and *Second-order polynomial term* are together.
- Run a regression model using regression function from *Analysis Tool Pack*.

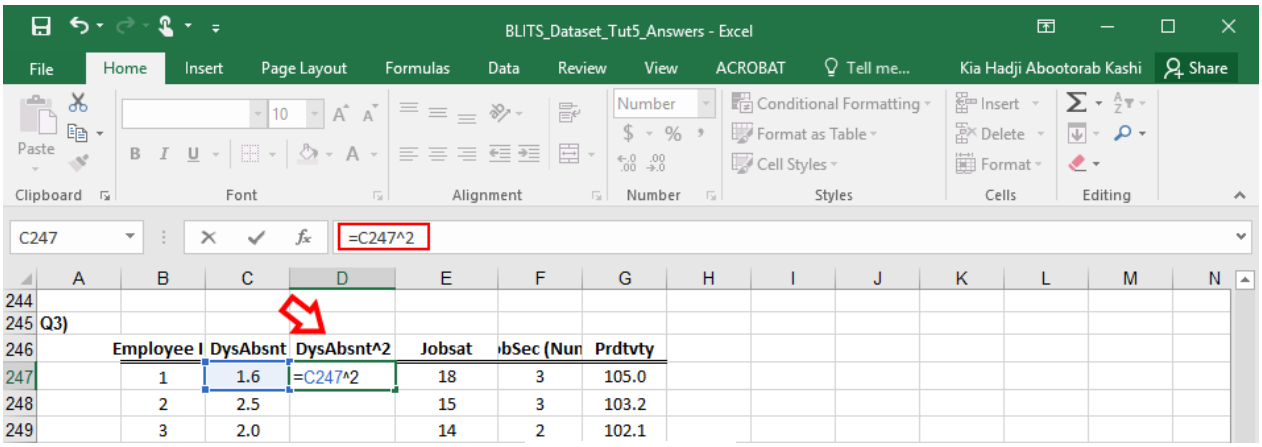



**Figure 4.**

b) Based on your answers from (a):

i. Evaluate your model overall (use the *F*-test).
ii. Are all of the independent variables, individually, significant?
iii. Compare the explanatory power of this model (using $R^2$) with that of the regression model developed in **Q2**.
iv. Write the regression formula and interpret the intercept and all regression coefficients.
v. Use the regression model to predict the productivity of an employee who has 3 days of absence, with a job satisfaction rating of 15, and job security score of 2.

**Q3. Analysing Interaction Effects**

The Human Resource Manager at BLITZ has recently come across an article entitled "*Men are from Mars, Women are from Venus: The role of Gender in the relationship between Job Satisfaction and Productivity*". The article concludes that there is a significant difference between male and female workers in the relationship between Job Satisfaction and Productivity.

Now the HR manager wonders if this is the case of employees at BLITZ and has asked you to test this interaction effect.

a) Define the interaction term (*Gender × Job Satisfaction*)

**Instruction:**

Use excel equation to create a new variable that represents the product of Gender and Job Satisfaction (see Figure 4).

You simply need to multiply values of Gender by Job Satisfaction to create the interaction term. Label the interaction term as ***"Gen*JobSat"***

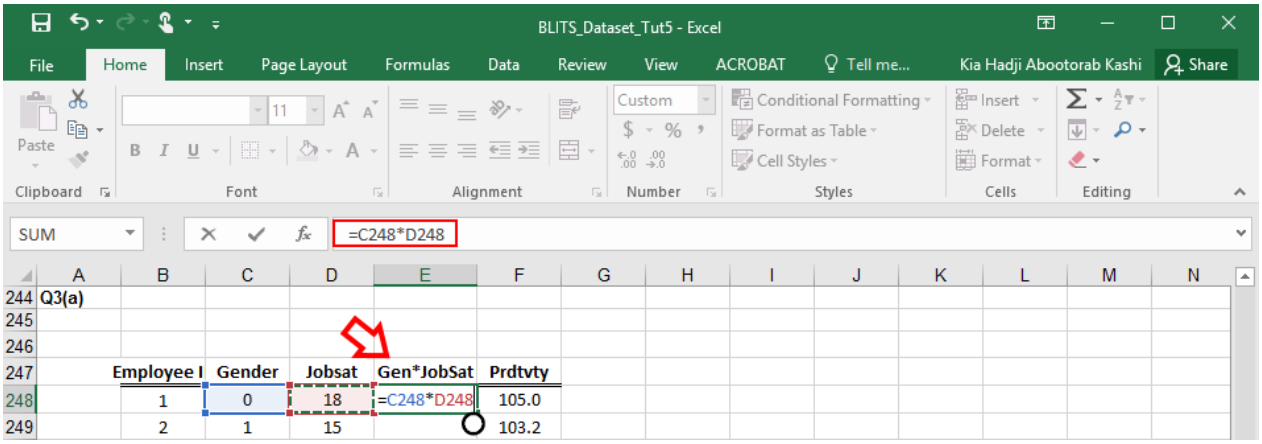



**Figure 4.**

b) Build a new regression model by including the newly created interaction term as an independent variable.

**Instruction:**

- Make sure that columns are rearranged such that all independent variables are together.
- Select **Data → Data Analysis → Regression**.
- Define **Y Variable Range** (*Productivity*) and the **X Variable Range** (*Gender*, *Job Satisfaction* and Interaction term – *Gen*JobSat*).
- Click **Labels**.
- Specify output location and click **OK**.

c) Based on the output from (b), evaluate the overall model, significance of independent variables and specifically, the interaction term.

d) If interaction term is found to be significant, plot the interaction effect using "***Interaction – Binary Var***" template.

e) Interpret the interaction plot in plain language and with respect to the HR Managers' speculation.