Extra Exercise to Understand Dummy Variables

Background:

Understanding Dummy Variables is critical to understanding Linear/Logistic Regression and then other more advanced models. A fictitious dataset was created on 14 Sep 2020 for you to explore and understand the concept more clearly with numbers. As R, Python and other Analytics software (but not Excel) automates the creation of dummy variables, to further boost your understanding, we use Excel to force students to create dummy variables manually so that students will know what is happening “behind the scene”. At the end, we will use R to get the same results as Excel, but it will be much faster as dummy variables’ creation are automated in R.

Files:

1. CSV: Occupation.csv
2. Excel: Occupation.xlsx

Link to learn how to use Excel Analysis Toolpak for Linear Regression:

https://www.excel-easy.com/examples/regression.html

Questions:

1. In the Excel file, using Analysis Toolpak, conduct Linear Regression to estimate Salary with Occ Code. [Recall what is the meaning of the linear regression line?].
    a. What is the linear regression equation?
    b. What is the estimated salary for Clerk?
    c. What is the estimated salary for Manager?
    d. What is the estimated salary for Director?
    e. What is the estimated salary for CEO?

2. In the Excel file, create a copy of the worksheet and then create dummy variables for occupation code using Clerk as the baseline reference level. Conduct Linear Regression to estimate Salary with the dummy variables only.
    a. How many dummy variables are needed?
    b. What is the linear regression equation?
    c. What is the estimated salary for Clerk?
    d. What is the estimated salary for Manager?
    e. What is the estimated salary for Director?
    f. What is the estimated salary for CEO?

3. Compare your answers in (1) vs (2), what are your conclusion and insights?

4. In the Excel file, create a copy of the worksheet and then create dummy variables for occupation code using Manager as the baseline reference level. Conduct Linear Regression to estimate Salary with the dummy variables only.
   a. How many dummy variables are needed?
   b. What is the linear regression equation?
   c. What is the estimated salary for Clerk?
   d. What is the estimated salary for Manager?
   e. What is the estimated salary for Director?
   f. What is the estimated salary for CEO?

5. Compare your answers in (2) vs (4), what are your conclusion and insights?

6. In R, import the CSV dataset and set Manager as the baseline reference level [Hint: Use levels() function to check the default baseline reference level and relevel() function to change the baseline reference if necessary]. Conduct Linear Regression to estimate Salary with Occupation. R will auto-create all the required dummy variables for you. Observe how R reports on the model coefficients for dummy variables.
   a. What is the linear regression equation?
   b. What is the estimated salary for Clerk?
   c. What is the estimated salary for Manager?
   d. What is the estimated salary for Director?
   e. What is the estimated salary for CEO?

Solutions: Q1 – Q5 in Excel file and Rscript for Q6.