ITM 6285 Data Mining Assignment 3 - Regression Analysis

Regression with Logarithm Transformation, Fixed Effect Modeling, and Logistic Regression (expected time - 1.5 hours)

Submit your **codes**, **the outputs**, and your **short answers** in a MS-word document because screenshot is needed in this assignment.

**Learning objective:**

1. Carry out linear regression analysis in R

2. Calculated Variables

3. Interpreting Logarithm transformed linear model

4. Fixed effect modeling

5. Logistic modeling

**Task 1: Import the Dataset and Background Information**

Import the data set “DairyFarm.xlsx” into RStudio. Beware this is a .xlsx file.

We would like to understand what factors can affect the milk production at Spanish dairy farms.

The data in “DairyFarm.xlsx” records the milk production and the amount of cows, land, labor, and feed for each of the 247 Spanish dairy farms for the six years from Year 1993 to Year 1998. “milk” records the annual milk production; “cows” records the average number of cows; “land” records the land size of the farm; “labor” records the average number of people working at the farm; “feed” records the amount of feed consumed. We have created dummy variables “year93”, “year94”, “year95”, “year96”, “year97”, “year98”; the log transformed variables “logmilk”, “logcows”, “logland”, “loglabor”, “logfeed”.

**Task 2: Understand the Logarithm**

Do a histogram on the variable milk with a break of 50. Copy the histogram to the answer sheet. Is it skewed?

What is the benefit to do a logarithm transformation?

Do a scatter plot for (1) milk ~ cows, and (2) logmilk ~ logcows, what are the difference between the two plots?

**Task 3: Understand Panel Data**

Short answer what is the cross section in this dataset?

If we regress “logmilk” onto “logcows” “logland”, “loglabor”, “logfeed” with no controls for the cross section fixed effect, are we going to get accurate insights on the problem? Why?

**Task 4:**

Please estimate a panel data model with “two-way fixed group effects” (both with farm fixed effects and year fixed effects), with “logmilk” as the dependent variable, and “logcows” “logland”, “loglabor”, “logfeed” as independent variables.

What is the R-squared value? What does it mean?

Please interpret the coefficients on logcows (hint: 1% change in the number of cows leads to ……).