Pre-Live-Session Coding Exercise 1

Dr. Jeffrey Yau

Contents

DUE: 11:59pm Pacific Time on Monday, August 30.

Question 1: Linear Probability Model

- **What are the advantages of the linear probability model?**
- **What are the drawbacks of the linear probability model?**

Question 2: Examining the dataset befor conducting EDA

Insert the function to *tidy up* the code when they are printed out

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

Practical Tips for Implementing Binary Logistic Regression

When solving data science problems, always begin with the understanding of the underlying (business, policy, scientific, etc) question; our first step is typically **NOT** to jump right into the data.

In this example, suppose the question is *"Do females who have higher family income (excluding wife's income) have lower labor force participation rate?" If so, what is the magnitude of the effect?* Note that this was not objective in *Mroz (1987)'s* paper. For the sake of learning to use logistic regression in answering a specific question, we stick with this question in this example.

Understanding the sample data: Remember that this sample comes from *1976 Panel Data of Income Dynamics (PSID)*. PSID is one of the most popular datasets used by labor economists.

First, load the `car` library in order to use the Mroz dataset and understand the structure dataset. Once the `car` library is loaded, the `Mroz` library can be called simply by calling its name `Mroz`.

Typical questions you should always ask when examining a dataset include the following.

Exercise: Write your own codes to answer the following questions:

- What is the number of variables and the number of observations in the Mroz dataset?
- Are these variables sufficient for you to answer you questions?
- If not, what other variables would you like to have? What impact (qualitatively) might not having these variables have on your models?
- Are there any missing values (in each of the variables)? If so, how many missing values in each of the variables?
- Are there any abnormal values in each of the variables in the raw data?

```
# Import libraries
library(car)
```

```
## Loading required package: carData
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
# Load the first 3 rows of the Mroz dataset
head(Mroz,3)
```

```
##   lfp k5 k618 age wc hc       lwg   inc
## 1 yes  1    0  32 no no 1.2101647 10.91
## 2 yes  0    2  30 no no 0.3285041 19.50
## 3 yes  1    3  35 no no 1.5141279 12.04
```

```
## YOUR CODE TO BE HERE
```

Question 3: Descriptive statistical analysis of the data

Exercise: Conduct EDA on the Mroz Dataset

```
# YOUR CODE TO BE HERE
```

As a best practice, we will need to incorporate insights generated from EDA on model specification. In what follows, we employ a very simple specification that uses all the variables "as-is", but the focus in this exericse is on how to interpret the coefficients.

Question 4: Estimate a Binary Logistic Regression

Exercises: * Estimate a Binary Logistic Regression * Print the summary of the model results * Intrepret the model results

- Dependent variable: lfp ** Explanatoryvariables: k5, k618, age, wc, hc, lwy, inc

```
# YOUR CODE TO BE HERE
```