Econometric Methods:
Solutions to Empirical Exercise 11.1
Chapter 11: Binary Dependent Variables
Stock & Watson, 3^{rd} Edition

Zaeen de Souza [*] Deepti Goel[†]

Azim Premji University

22 April 2022

[*]

[†]Solution key prepared jointly by Zaeen and Deepti. R code and presentation in Rmarkdown by Zaeen.

Contents

Background: Empirical Exercise 11.1

These are the solutions to **E11.1** from **Chapter 11** of *Introduction to Econometrics (Updated Third edition)* by Stock & Watson. You should have the following on your computer in order to check answers/run the code and follow the questions in this assignment:

- An updated version of `R` and `Rstudio`.
- The following packages installed:
  - `ggplot2`
  - `readxl`
  - `fixest`
- The datasets called `employment_06_07.xlsx`, `employment_08_09.xlsx`
- The data description pdf to understand the variables being used.

Reading guide

All the code needed to complete the assignments is within this document. R code will be in a grey box and will look like this:

```
summary(iris)
```

And all R output i.e what R shows you once you run some code, will have # signs next to it, and will look like this:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.300   5.100   5.800   5.843   6.400   7.900
```

As far as possible these guides will show the **exact output** that comes from running code in `R`, and at times will use formatted tables made in latex. The results themselves, will be identical. Some things to note, that might make output look different accross different computers:

- R reports things like p-values using scientific notation, but some computers report the numbers with many trailing zeros.
- If you have an old version of `R` or `Rstudio` it is highly recommended that you update it using the following code:

```
# Use this to update R from within RStudio
install.packages("installr")
library(installr)
# This last command, will open up a download prompt; choose yes/no accordingly.
updateR()
```

For updating `Rstudio`, un-install your version of `RStudio`, and download a fresh version from the `RStudio` website.

Loading the data and libraries

The following code sets the working directory to the folder where you have downloaded the data, loads the libraries needed fo the assignment and loads the excel dataset.

```
# Loading excel files
library(fixest)
library(readxl)
library(ggplot2)
library(dplyr)

# Setting working directory - this is unique to your computer
setwd("~Zaeen de Souza/Chapter 11 Binary Dependent Variables")

# Loading the data as 'pset_data'
pset_data1 <- read_excel("employment_06_07.xlsx")
pset_data2 <- read_excel("employment_08_09.xlsx")
```

E11.1 Problem Context

In April 2008 the unemployment rate in the United States stood at 5.0%. By April 2009 it had increased to 9.0%, and it had increased further, to 10.0%, by October 2009. Were some groups of workers more likely to lose their jobs than others during the Great Recession? For example, were young workers more likely to lose their jobs than middle-aged workers? What about workers with a college degree versus those without a degree, or women versus men? On the textbook website, www.pearsonglobaleditions.com/Stock_Watson, you will find the data file Employment_08_09, which contains **a random sample of 5440 workers who were surveyed in April 2008 and reported that they were employed full time**. A detailed descriptionis given in Employment_08_09_Description, available on the website. These workers were surveyed one year later, in April 2009, and asked about their employment status (employed, unemployed, or out of the labor force). The data set also includes various demographic measures for each individual. Use these data to answer the following questions.

Exercise E11.1

a. What fraction of workers in the sample were employed in April 2009? Use your answer to compute a 95% confidence interval for the probability that a worker was employed in April 2009, conditional on being employed in April 2008.

We will answer this by first writing our own function to estimate this fraction and construct the confidence interval and then we will use regression to do the same.

```r
my_estimator <- function(X = x, Z = z) {
  sample_mean <- mean(X, na.rm = T)
  sample_n <- length(X)
  sample_sd <- sd(X, na.rm = T)
  sample_se <- sample_sd / sqrt(sample_n)
  degrees_freedom = sample_n - 1
# the qt() function is an inverse cumulative density function for a t-distribution.
# it takes as its arguments/inputs a quantile (Z/2=0.05/2) and dof (sample size -1)
# and returns the corresponding t score value.
  t_stat = qt(p = Z / 2, df = degrees_freedom)
  margin_error <- t_stat * sample_se
  lower_bound <- sample_mean - margin_error
  upper_bound <- sample_mean + margin_error
  cat("The probability is",
      round(sample_mean, 3),
      "\nThe 95% confidence interval is [",
      round(upper_bound, 3),
      round(lower_bound, 3),
      "]")
}
my_estimator(X = pset_data2$employed, Z = 0.05)
```

```
## The probability is 0.875
## The 95% confidence interval is [ 0.867 0.884 ]
```

An easier way to do this is to regress *Employed* on a constant, and then use the `confint()` function to get the 95% confidence interval.

```r
round(confint(lm(employed ~ 1, data = pset_data2)),3)
```

```
##             2.5 % 97.5 %
## (Intercept) 0.867  0.884
```

This works because when you regress Y on only a constant, the intercept is equal to $E[Y]$ if you run the regression in the population, and $\bar{Y}$ (i.e. sample mean of Y), if you run it in a sample. This is always the case irrespective of whether Y is binary or not.

b. Regress Employed on Age and Age^2, using a linear probability model.

We will estimate the following LPM (Linear Probability Model) using OLS,

$$Y_i = \alpha + \beta_1 Age_i + \beta_2 Age_i^2 + \varepsilon_i$$

Note,

$$Y_i = \begin{cases} 1, & if \text{ individual } i \text{ was employed in 2009} \\ 0, & if \text{ individual } i \text{ was unemployed in 2009} \end{cases}$$

```r
model_1 <- feols(employed ~ age + age_sq, data = pset_data2, vcov = "HC1")
```

Table 1: E.11.1 b

	Employed
	(1)
Age	0.028***
	(0.003)
Age^2	-0.0003***
	(3.88×10^{-5})
(Intercept)	0.307***
	(0.067)
R^2	0.020
Observations	5,412

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

b i. Based on this regression, was age a statistically significant determinant of employment in April 2009?

Age is a significant predictor of employment status in April 2009 as the coefficients on both Age and Age squared are statistically significant at the 1 percent level. Because Age squared is also significant, the relationship between Employment and Age is not linear but non-linear. In fact since age squared is negative, it is concave (if age squared would have been positive, it would be convex). We can calculate the age at which the relationship between employment and age changes from being positive to negative as follows:

Set the partial derivative equal to 0,

$$0.028 - 0.0003 \cdot 2 \cdot Age = 0$$

Solving,

$$-0.0006 Age = -0.028$$

$$Age = \frac{0.028}{0.0006} = 44.7$$

The probability of being employed is increasing in Age until 44.7 years, and then starts to decline.

b ii. Is there evidence of a nonlinear effect of age on the probability of being employed?

The coefficient on Age^2, which is negative *and* significant, indicates that there is some evidence that the relationship between Age and employment status is indeed non-linear; specifically, it is increasing (as the first derivative of $\widehat{Y}$ w.r.t. age is positive) and concave (as second derivative is negative).

b iii. Compute the predicted probability of employment for a 20-yearold worker, a 40-year-old worker, and a 60-year-old worker.

We need to calculate:

1. $P(Y = 1|Age = 20)$
2. $P(Y = 1|Age = 40)$
3. $P(Y = 1|Age = 60)$

Using the coefficients and intercept from table 1, we have:

```r
b1 <- model_1$coefficients[2]
b2 <- model_1$coefficients[3]
employed_20_lpm <- 0.307 + (b1 * 20) + (b2 * (20^2))
employed_40_lpm <- 0.307 + (b1 * 40) + (b2 * (40^2))
employed_60_lpm <- 0.307 + (b1 * 60) + (b2 * (60^2))
# 20 year old
round(employed_20_lpm,3)
```

```
##   age
## 0.742
```

```r
# 40 year old
round(employed_40_lpm,3)
```

```
##   age
## 0.915
```

```r
# 60 year old
round(employed_60_lpm,3)
```

```
##   age
## 0.827
```

c. Repeat (b) using a probit regression.

We will estimate

$$P(Y = 1|Age, Age^2) = \Phi(\alpha + \beta_1 Age_i + \beta_2 Age_i^2)$$

Where Φ is a cumulative standard normal distribution function. The probit results are below, in table 2.

```
model_2 <- feglm(employed ~ age + age_sq, data = pset_data2, family= binomial("probit"))
```

Table 2: E.11.1 c: Probit

	Employed
	(1)
Age	0.122***
	(0.013)
Age^2	-0.001***
	(0.0002)
(Intercept)	-1.26***
	(0.247)
Pseudo R^2	0.023
Observations	5,412

IID standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

The estimated coefficients have the same *direction* as the LPM. After transforming them, we can see that they are *qualitatively*, similar in size to the LPM coefficients.

```
b1 <- model_2$coefficients[2]
b2 <- model_2$coefficients[3]
#pnorm is CDF of the normal distribution
employed_20_pro <- pnorm(-1.26 + b1*20 + b2*20^2)
employed_40_pro <- pnorm(-1.26 + b1*40 + b2*40^2)
employed_60_pro <- pnorm(-1.26 + b1*60 + b2*60^2)
round(employed_20_pro,3);round(employed_40_pro,3);round(employed_60_pro,3)
```

```
##   age
## 0.729
```

```
##   age
## 0.911
```

```
##   age
## 0.831
```

d. Repeat (b) using a logit regression.

We will estimate

$$P(Y = 1|Age, Age^2) = \frac{1}{1 + e^{-(\alpha + \beta_1 Age_i + \beta_2 Age_i^2)}}$$

```
model_3 <- feglm(employed ~ age + age_sq, data = pset_data2, family= binomial("logit"))
```

Table 3: E.11.1 d: Logit

	Employed (1)
Age	0.225^{***}
	(0.023)
Age^2	-0.003^{***}
	(0.0003)
(Intercept)	-2.49^{***}
	(0.437)
Pseudo R^2	0.023
Observations	5,412

IID standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

As can be seen, the estimated coefficients have the same direction as the LPM *and* the Probit regressions. The estimates themselves are also similar.

```
b1 <- model_3$coefficients[2]
b2 <- model_3$coefficients[3]
# plogis is the CDF of the logistic distribtion
employed_20_lo <- plogis(-2.49 + b1*20 + b2*20^2)
employed_40_lo <- plogis(-2.49 + b1*40 + b2*40^2)
employed_60_lo <- plogis(-2.49 + b1*60 + b2*60^2)
round(employed_20_lo,3)
```

```
##   age
## 0.725
```

```
round(employed_40_lo,3)
```

```
##   age
## 0.911
```

```
round(employed_60_lo,3)
```

```
##   age
## 0.831
```

e. Are there important differences in your answers to (b)–(d)? Explain.

For 40 and 60 year olds the predictions from all three models are very close, for 20 year olds also they are close but not as close. Thus, all three models give similar predictions for these ages. However, the limitation of the LPM is that for certain ages, it could give predictions less than 0 or more than 1.

f. The data set includes variables measuring the workers' educational attainment,sex, race, marital status,region of the country, and weekly earnings in April 2008.

f i. Construct a table like Table 11.2 to investigate whether the conclusions on the effect of age on employment from (b)–(d) are affected by omitted variable bias.

In table 4 we compare the results of the regression with controls.

```r
# making a dummy for race==black
pset_data2$Black <- ifelse(pset_data2$race==2, 1,0)

model_4 <- feols(employed ~ age + age_sq +
                   Black + female +
                   educ_lths + educ_hs +
                   educ_somecol+ educ_aa+
                   educ_bac + married + log(earnwke) |
                   ne_states + so_states + ce_states,
                   data = pset_data2, vcov = "HC1")
```

```
## NOTE: 645 observations removed because of NA and infinite values (RHS: 645).
```

We see that the associated between age and employment status is more or less, unchanged (the regression with no controls is in table 1). The squared term is significant as well, which suggests that the non-linear relationship between age and employment exists, even after adding additional covariates. The probability of being employed is increasing until 38 years (done by setting the partial derivative to 0 and solving the same way we did in **b.i**) and is decreasing after.

f ii. Use the regressions in your table to discuss the characteristics of workers who were hurt most by the Great Recession.

Younger workers, black workers, workers with below high school education and workers with lower earnings were less likely to be employed. In terms of magnitudes, workers with education less than high school were 7 percentage points less likely to be employed, compared to those with a advanced degree. Black workers were 3.7 percentage points less likely to be employed, compared to other workers. A one percent increase in earnings was associated with 0.0004 increase in probability of being employed or a 0.04 percentage point increase in the probability of being employed.

Table 4: E.11.1 f: LPM With Controls

	Employed (1)
Age	0.023***
	(0.004)
Age^2	-0.0003***
	(4.24×10^{-5})
Black	-0.037**
	(0.019)
Female	0.0005
	(0.010)
Below High School	-0.070***
	(0.027)
High School	-0.016
	(0.016)
Some College	0.006
	(0.017)
Assoc. Degree	0.009
	(0.018)
BA/BS Degree	-0.012
	(0.015)
Married	-0.004
	(0.010)
log(earnwke)	0.040***
	(0.009)
North-East State Fixed Effects	✓
Southern State Fixed Effects	✓
Central State Fixed Effects	✓
R^2	0.036
Observations	4,767

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

g. The results in (a)–(f) were based on the probability of employment. Workers who are not employed can either be (i) unemployed or (ii) out the labor force. Do the conclusions you reached in (a)–(f) also hold for workers who became unemployed? (*Hint: Use the binary variable Unemployed instead of Employed.*)

Table 5 presents the linear probability model, where we have regressed unemployment (0/1 binary variable) on covariates. We interpret the second column of results here. The relationship between unemployment and age is convex (U shaped), with young workers and old workers more likely to be unemployed.

Females are less likely to be unemployed compared to males (remember, out of labour force is also coded as 0, along with employed); females are 2.1 percentage points less likely to be unemployed compared to males.

Workers with education below highschool and only highschool are more likely to be unemployed compared to workers with an advanced degree; Below Highschool are 6.8 percentage points more likely to be unemployed compared to workers with an advanced degree, while those with a highschool degree are 3.3. percentage points more likely to be unemployed compared to those with advanced degrees.

```r
model_5 <- feols(unemployed ~ age + age_sq,
                 data = pset_data2, vcov = "HC1")

model_6 <- feols(unemployed ~ age + age_sq +
                            Black + female +
                            educ_lths + educ_hs +
                            educ_somecol+ educ_aa+
                            educ_bac + married + log(earnwke) |
                            ne_states + so_states + ce_states,
                 data = pset_data2, vcov = "HC1")
```

Table 5: E.11.1 g: Unemployment - with and without controls (LPM)

	Unemployed	
	(1)	(2)
Age	-0.007***	-0.005**
	(0.002)	(0.002)
Age^2	$7.77 \times 10^{-5***}$	$5.92 \times 10^{-5**}$
	(2.36×10^{-5})	(2.78×10^{-5})
Black		0.019
		(0.013)
Female		-0.021***
		(0.007)
Below High School		0.068***
		(0.019)
High School		0.033***
		(0.010)
Some College		0.010
		(0.010)
Assoc. Degree		0.009
		(0.011)
BA/BS Degree		0.012
		(0.009)
Married		-0.014**
		(0.007)
log(earnwke)		-0.004
		(0.005)
North-East State Fixed Effects		✓
Southern State Fixed Effects		✓
Central State Fixed Effects		✓
R^2	0.004	0.016
Observations	5,412	4,767

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

h. These results have covered employment transitions during the Great Recession, but what about transitions during normal times? On the textbook website, you will find the data file Employment_06_07, which measures the same variables but for the years 2006–2007. Analyze these data and comment on the differences in employment transitions during recessions and normal times.

During normal times (columns 1-2 in table 6), that the effect of age on employment status is similar as during recession (columns 3-4). The probability of employment is increasing until age 40, and is decreasing after (column 2, table 6), whereas during recession, it is increasing till age 38, and decreasing after. Workers with higher salaries are also more likely to be employed in both periods, with the coefficients similar in size—during normal times, a 1 percent increase in earnings is associated with a 0.052 percentage point increase in the probability of being employed, while it is associated with a 0.04 percentage point increase in the probability of being employed during a recession. During recessions workers with lower education are also less likely to be employed—workers with less than HS education are 7 percentage points less likely to be employed as compared with those with an advanced degree. This is not significant during normal times.

Note: The change in sample sized between columns in the table is explained by observations with MISSING or ZERO earnings. Missing observations are dropped because they are missing, but zero earnigns observations are dropped because $ln(0)$ is undefined.

```r
model_7 <- feols(employed ~ age + age_sq |
                 ne_states + so_states + ce_states,
                 data = pset_data1, vcov = "HC1")

model_8 <- feols(employed ~ age + age_sq +
                            Black + female +
                            educ_lths + educ_hs +
                            educ_somecol+ educ_aa+
                            educ_bac + married + log(earnwke) |
                            ne_states + so_states + ce_states,
                 data = pset_data1, vcov = "HC1")

model_9 <- feols(employed ~ age + age_sq |
                 ne_states + so_states + ce_states,
                 data = pset_data2, vcov = "HC1")

model_10 <- feols(employed ~ age + age_sq +
                            Black + female +
                            educ_lths + educ_hs +
                            educ_somecol+ educ_aa+
                            educ_bac + married + log(earnwke) |
                            ne_states + so_states + ce_states,
                 data = pset_data2, vcov = "HC1")
```

Table 6: E.11.1 h: Employment Transitions: 2006 vs 2009 (LPM)

	Employed			
	2006		2009	
	(1)	(2)	(3)	(4)
Age	0.024***	0.016***	0.028***	0.023***
	(0.003)	(0.003)	(0.003)	(0.004)
Age^2	-0.0003***	-0.0002***	-0.0003***	-0.0003***
	(3.66×10^{-5})	(4.06×10^{-5})	(3.87×10^{-5})	(4.24×10^{-5})
Black		-0.015		-0.037**
		(0.018)		(0.019)
Female		-0.024**		0.0005
		(0.009)		(0.010)
Below High School		-0.024		-0.070***
		(0.024)		(0.027)
High School		0.007		-0.016
		(0.015)		(0.016)
Some College		0.014		0.006
		(0.016)		(0.017)
Assoc. Degree		0.022		0.009
		(0.017)		(0.018)
BA/BS Degree		0.007		-0.012
		(0.014)		(0.015)
Married		0.020**		-0.004
		(0.010)		(0.010)
log(earnwke)		0.052***		0.040***
		(0.008)		(0.009)
North-East State Fixed Effects	✓	✓	✓	✓
Southern State Fixed Effects	✓	✓	✓	✓
Central State Fixed Effects	✓	✓	✓	✓
R^2	0.018	0.045	0.022	0.036
Observations	5,220	4,562	5,412	4,767

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*