Econometric Methods: Solutions to Empirical Exercise 8.2 Chapter 8: Non-Linear Regression Functions Stock & Watson, 3^{rd} Edition

Zaeen de Souza [*] Deepti Goel[†]

Azim Premji University

09 February 2022

[*]

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

Contents

Background: Empirical Exercise 8.2

These are the solutions to **E8.2** from **Chapter 8** 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
  - stargazer
  - dplyr
- The datasets called CPS92_12, CPS12.
- The data description pdf to understand the variables being used.
- For this problem set, we will start presenting regression output as formatted tables, instead of raw R output.

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.

```r
# Loading excel files
library(readxl)
# Making graphs
library(ggplot2)

# Setting working directory - this is unique to your computer
setwd("~Zaeen de Souza/Chapter 8 Nonlinear Regression Fnts")

# Loading the data as 'pset_data1' and 'pset_data2'
pset_data1 <- read_excel("CPS12.xlsx")
pset_data2 <- read_excel("CPS92_12.xlsx")
```

E8.2 Problem Context

On the text website www.pearsonglobaleditions.com/Stock_Watson you will find a data file CPS12, which contains data for full-time, full-year workers, ages 25–34, with a high school diploma or B.A./B.S. as their highest degree. A detailed description is given in CPS12_Description, also available on the website. (These are the same data as in CPS92_12, used in Empirical Exercise 3.1, but are limited to the year 2012.) In this exercise, you will investigate the relationship between a worker's age and earnings. (Generally, older workers have more job experience, leading to higher productivity and higher earnings.)

Exercise E8.2

a. Run a regression of average hourly earnings (AHE) on age (Age), gender (Female), and education (Bachelor). If Age increases from 25 to 26, how are earnings expected to change? If Age increases from 33 to 34, how are earnings expected to change?

```
model_a <- lm(ahe ~ age + female + bachelor, data = pset_data1)
```

Table 1: Exercise E8.2 a

	Dependent variable:
	ahe
age	0.510***
	(0.040)
female	−3.810***
	(0.230)
bachelor	8.319***
	(0.227)
Constant	1.866
	(1.188)
Observations	7,440
R^2	0.180
Adjusted R^2	0.180
Residual Std. Error	9.678
F Statistic	544.495***
Note:	*p<0.1; **p<0.05; ***p<0.01

Since we are modelling earnings as a *linear* function of age and other variables, the effect of a one unit (year) increase in age on earnings is the same at every level of age. In this example, if age were to increase from 25 to 26, earnings are predicted to increase by 0.51 USD per hour[1]. This is the same if age increases from 33 to 34. You can verify this in the following way:

$$\Delta \widehat{earnings} = 0.51 \cdot (26 - 25) = 0.51$$
$$\Delta \widehat{earnings} = 0.51 \cdot (34 - 33) = 0.51$$

[1] Please note that wages and earnings/income are all flow (not stock) concepts. You should therefore always remember to include the time dimension whenever you talk about them. Sadly, many people have become sloppy and do not follow this practice.

b. Run a regression of the logarithm of average hourly earnings, ln(AHE), on Age, Female, and Bachelor. If Age increases from 25 to 26, how are earnings expected to change? If Age increases from 33 to 34, how are earnings expected to change?

```
model_b <- lm(log(ahe) ~ age + female + bachelor, data = pset_data1)
```

Table 2: Exercise E8.2 b

	Dependent variable:
	log(ahe)
age	0.026^{***}
	(0.002)
female	-0.192^{***}
	(0.011)
bachelor	0.438^{***}
	(0.011)
Constant	1.941^{***}
	(0.059)
Observations	7,440
R^2	0.196
Adjusted R^2	0.196
Residual Std. Error	0.478
F Statistic	605.726^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01

This is a log linear specification of the model. If age increases from 25 to 26, earnings are predicted to increase by 2.6%[2]. If age increases from 33 to 34, earnings are expected to change by 2.6%. This is because the regression is linear in age. You can verify this in the following way:

$$\Delta ln(\widehat{earnings}) \approx \frac{\Delta Earnings}{Earnings} = 0.0255 \cdot (26 - 25) = 0.0255$$

$$\Delta ln(\widehat{earnings}) \approx \frac{\Delta Earnings}{Earnings} = 0.0255 \cdot (34 - 33) = 0.0255$$

[2] Recall that in the log linear model: $ln(Y)_i = \alpha + \beta_1 X_i + \varepsilon_i$, β_1 is to be interpreted as follows: A unit change in X, is associated with a $(\beta_1 \cdot 100)$ % change in Y.

c. Run a regression of the logarithm of average hourly earnings, ln(AHE), on ln(Age), Female, and Bachelor. If Age increases from 25 to 26, how are earnings expected to change? If Age increases from 33 to 34, how are earnings expected to change?

```r
model_c <- lm(log(ahe) ~ log(age) + female + bachelor, data = pset_data1)
```

Table 3: Exercise E8.2 c

	Dependent variable:
	log(ahe)
log(age)	0.753^{***} (0.057)
female	-0.192^{***} (0.011)
bachelor	0.438^{***} (0.011)
Constant	0.150 (0.194)
Observations	7,440
R^2	0.197
Adjusted R^2	0.196
Residual Std. Error	0.478
F Statistic	606.413^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01

This is a log-log specification. The coefficient on $ln(age)$ has the following interpretation: If age goes up by 1%, earnings are predicted to go up by 0.75%. An increase from 25 to 26 constitutes a 4% increase in age. So the associated increase in earnings is 3% $(4 \cdot 0.75)$. An increase from 33 to 34 constitutes a 2.9% increase in age. So the associated increase in earnings is 2.2% $(2.9 \cdot 0.75)$.

$$\Delta ln(\widehat{earnings}) = 0.75 \cdot (ln(26)) - ln(25)) = 0.029$$

$$\Delta ln(\widehat{earnings}) = 0.75 \cdot (ln(34)) - ln(33)) = 0.022$$

d. Run a regression of the logarithm of average hourly earnings, ln(AHE), on Age, Age2, Female, and Bachelor. If Age increases from 25 to 26, how are earnings expected to change? If Age increases from 33 to 34, how are earnings expected to change?

```
model_d <- lm(log(ahe) ~ age + I(age^2) + female + bachelor, data = pset_data1)
```

Table 4: Exercise E8.2 d

	Dependent variable:
	log(ahe)
age	0.104^{**} (0.046)
I(age^2)	-0.001^{*} (0.001)
female	-0.192^{***} (0.011)
bachelor	0.437^{***} (0.011)
Constant	0.792 (0.670)
Observations	7,440
R^2	0.197
Adjusted R^2	0.196
Residual Std. Error	0.478
F Statistic	455.156^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01

This is a quadratic specification for log average hourly earnings.

- **Case 1:** When age changes from 25 to 26: The change in log average hourly earnings is given by:

  $$0.104 \cdot 26 - (0.001 \cdot (26^2)) - (0.104 \cdot 25 - (0.001 \cdot (25)^2)) = 0.053$$

  Note: 0.053 is in units of log AHE. Earnings are therefore expected to go up by 5.3% (approx).

- **Case 2:** When age changes from 33 to 34: The change in log average hourly earnings is given by:

  $$0.104 \cdot 34 - (0.001 \cdot (34^2)) - (0.104 \cdot 33 - (0.001 \cdot (33^2)) = 0.037$$

  Note: 0.03 is in units of log AHE. Earnings are therefore expected to go up by 3.7% (approx)

e. Do you prefer the regression in (c) to the regression in (b)? Explain.

Model (c) is a log-log specification and (b) is a log-linear specification. There are two ways to think about the answer to part (e). First is, how do you think about the relationship between earnings and age? Do you think that a unit change in X leads to a constant percentage change in Y or do you think that a one percent change in X leads to a certain fixed percent change in Y.

A second way to think about this is, since the LHS is the same in both models (namely, log(AHE)), we pick the one with a higher R-squared. In this case, the adjusted R-squared is the same for both, I would put more weight to what is the right way to think about the relationship.

f. Do you prefer the regression in (d) to the regression in (b)? Explain.

The regression in (d) is preferred over (b). (d) allows for a non-linear relationship between earnings and age, and given that the coefficient on Age^2 is significant (although, only at the 10% level), we could conclude that there is indeed, a non-linear relationship between these two variables.

g. Do you prefer the regression in (d) to the regression in (c)? Explain.

Here we are comparing log log with quadratic specification for logY. Pick the one with higher R squared.

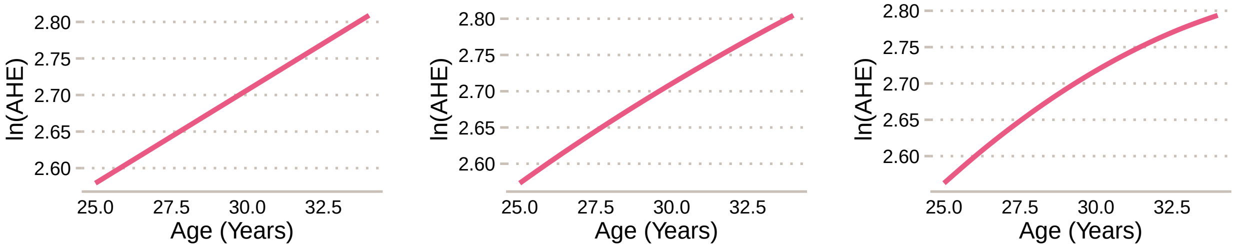
h. Plot the regression relation between Age and ln(AHE) from (b), (c), and (d) for males with a high school diploma. Describe the similarities and differences between the estimated regression functions. Would your answer change if you plotted the regression function for females with college degrees?

In order to do this, we will use the `predict` function to estimate the $\widehat{ln(ahe)}$ from regressions (b), (c) and (d). In all three panels, we are plotting ln(ahe) against age. In (b), which is a log-linear model we therefore see that the line of best fit has same slope. In (c), which is a log-log model, the line of best fit has a curve to it (note if we had plotted log AHE against log age, this line would have a constant slope). In (d), as in (c), the line of best fit, has a curve to it and the curvature is more exaggerated.

```
# this generates the yhat from each of the regressions
pset_data1$yhat_b <- predict(model_b)
pset_data1$yhat_c <- predict(model_c)
pset_data1$yhat_d <- predict(model_d)

# dplyr makes it easy to use IF conditions. Note the %>% or "pipe" operator.
library(dplyr)
pset_data1 %>%
  filter(female == 0 & bachelor == 0) %>%
  ggplot() + geom_smooth(aes(age, yhat_b))

pset_data1 %>%
  filter(female == 0 & bachelor == 0) %>%
  ggplot() + geom_smooth(aes(age, yhat_c))

pset_data1 %>%
  filter(female == 0 & bachelor == 0) %>%
  ggplot() + geom_smooth(aes(age, yhat_d))
```

```
pset_data1 %>%
  filter(female == 1 & bachelor == 1) %>%
  ggplot() + geom_smooth(aes(age, yhat_b))

pset_data1 %>%
  filter(female == 1 & bachelor == 1) %>%
  ggplot() + geom_smooth(aes(age, yhat_c))

pset_data1 %>%
  filter(female == 1 & bachelor == 1) %>%
  ggplot() + geom_smooth(aes(age, yhat_d))
```

Figure 1: Males with Highschool Diploma

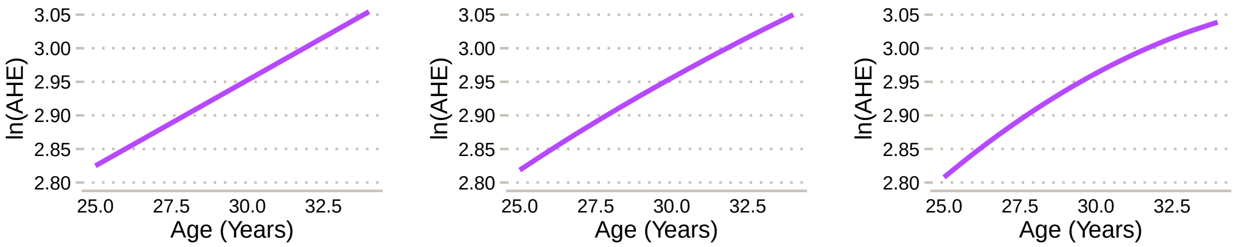


(a) ln(ahe), linear in age

(b) ln(ahe), ln(age)

(c) ln(ahe), quadratic age

Figure 2: Females with College Degree



(a) ln(ahe), linear in age

(b) ln(ahe), ln(age)

(c) ln(ahe), quadratic age

h.i. Run a regression of ln(AHE) on Age, Age^2, Female, Bachelor, and the interaction term Female $\times$ Bachelor. What does the coefficient on the interaction term measure? Alexis is a 30-year-old female with a bachelor's degree. What does the regression predict for her value of ln(AHE)? Jane is a 30-year-old female with a high school degree. What does the regression predict for her value of ln(AHE)? What is the predicted difference between Alexis's and Jane's earnings? Bob is a 30-year-old male with a bachelor's degree. What does the regression predict for his value of ln(AHE)? Jim is a 30-year-old male with a high school degree. What does the regression predict for his value of ln(AHE)? What is the predicted difference between Bob's and Jim's earnings?

```
model_1 <- lm(log(ahe) ~ age +
                       I(age^2) + female + bachelor +
                       I(female * bachelor), data = pset_data1)
```

Table 5: Exercise E8.2 h.i

	Dependent variable:
	log(ahe)
age	0.104**
	(0.046)
I(age^2)	$-$0.001*
	(0.001)
female	$-$0.242***
	(0.017)
bachelor	0.400***
	(0.015)
I(female * bachelor)	0.090***
	(0.023)
Constant	0.804
	(0.669)
Observations	7,440
R^2	0.198
Adjusted R^2	0.198
Residual Std. Error	0.478
F Statistic	367.940***
Note:	*p<0.1; **p<0.05; ***p<0.01

The coefficient on the interaction term measures the CHANGE in earnings PREMIUM for having a Bachelor's degree (relative to being only a high school graduate) that females experience compared to the bachelor's premium that males experience. In other words, it shows how the earnings value of having a Bachelor's degree differs for males and females.

- Alexis: $0.804 - 0.242 \cdot 1 + (0.104 \cdot 30) - (0.001 \cdot (30^2)) + 0.400 \cdot 1 + 0.09 \cdot 1 = 3.272$
- Jane: $0.804 - 0.242 \cdot 1 + (0.104 \cdot 30) - (0.001 \cdot (30^2)) = 2.782$

The difference is: $3.272 - 2.782 = 0.490$. The units are in ln(ahe). In other words, the difference in earnings between Alexis (bachelor's graduate) and Jane (high school graduate), is approximately, 49%.

- Bob: $0.804 + 0.104 \cdot 30 - (0.001 \cdot (30^2)) + 0.400 \cdot 1 = 3.424$
- Jim: $0.804 + 0.104 \cdot 30 - (0.001 \cdot (30^2)) = 3.024$

The difference is $3.424 - 3.024 = 0.400$. The units are in ln(AHE). In other words, the difference in earnings between Bob (Bachelor's degree) and Jim (high school graduate), is approximately, 40%.

Thus the Bachelors' premium for women is **9 PERCENTAGE POINTS** higher, compared to men.

j. Is the effect of Age on earnings different for men than for women? Specify and estimate a regression that you can use to answer this question.

We will answer this question using the following regression.

$$ln(Y_i) = \alpha + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Female_i + \beta_4 (Age \times Female)_i + \beta_5 (Age^2 \times Female)_i + \varepsilon_i$$

The results are reported in table 6, column (1). The coefficients on the interaction terms, i.e β_4, β_5 are both significant, indicating that there association between earning and age is, indeed different for men and women. The test of joint significance also confirms this.

```
model_2 <- lm(log(ahe) ~ age +
                         I(age^2) +
                         female +
                         I(age * female) +
                         I(age^2 * female),
                         data = pset_data1)
```

Test of joint significance for $H_0 : \beta_4 = \beta_5 = 0, H_1 : \beta_4 \neq 0$ and/or $\beta_5 \neq 0$:

```
library(car)
linearHypothesis(model_2, c("I(age * female)=0",
                            "I(age^2 * female) = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## I(age * female) = 0
## I(age^2 * female) = 0
##
## Model 1: restricted model
## Model 2: log(ahe) ~ age + I(age^2) + female + I(age * female) + I(age^2 *
##     female)
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1   7436 2046.4
## 2   7434 2044.1  2    2.3611 4.2935 0.01369 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

k. Is the effect of Age on earnings different for high school graduates than for college graduates? Specify and estimate a regression that you can use to answer this question.

We will answer this question using the following regression.

$$ln(Y_i) = \alpha + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Bachelor_i + \beta_4 (Age \times Bachelor)_i + \beta_5 (Age^2 \times Bachelor)_i + \varepsilon_i$$

The results are in table 6, column (2). Here, the coefficients on the interaction terms, i.e β_4, β_5 are both insignificant, indicating there there are no differences in the association between earnings and age, between workers who are college graduates, and those who are not. The test of joint significance as in (j) also shows that these there is indeed no difference.

```
model_3 <- lm(log(ahe) ~ age +
                         I(age^2) +
                         bachelor +
                         I(age * bachelor) +
                         I(age^2 * bachelor),
                         data = pset_data1)
```

```
linearHypothesis(model_3, c("I(age * bachelor)=0",
                            "I(age^2 * bachelor) = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## I(age * bachelor) = 0
## I(age^2 * bachelor) = 0
##
## Model 1: restricted model
## Model 2: log(ahe) ~ age + I(age^2) + bachelor + I(age * bachelor) + I(age^2 *
##     bachelor)
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   7436 1765.7
## 2   7434 1765.0  2   0.68432 1.4411 0.2367
```

Table 6: Exercise E8.2 k and j

	Dependent variable:	
	log(ahe)	
	(1)	(2)
age	0.055 (0.067)	0.034 (0.067)
I(age^2)	-0.0004 (0.001)	-0.0002 (0.001)
female	-2.773^{*} (1.482)	
I(age * female)	0.190^{*} (0.101)	
I(age^2 * female)	-0.003^{**} (0.002)	
bachelor		-1.575 (1.367)
I(age * bachelor)		0.131 (0.093)
I(age^2 * bachelor)		-0.002 (0.002)
Constant	1.665^{*} (0.976)	1.781^{*} (0.986)
Observations	7,440	7,440
R^2	0.034	0.166
Adjusted R^2	0.033	0.165
Residual Std. Error	0.524	0.487
F Statistic	52.474^{***}	295.821^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

l. After running all these regressions (and any others that you want to run), summarize the effect of age on earnings for young workers.

Based on the regressions we ran, we can come to the following conclusion. There is strong evidence that the association between log earnings and age is increasing and concave[3]. There is some evidence (based on the p-values of the interaction terms in table 6, column 1) that this association is more pronounced and is higher in magnitude, for female workers. We cannot conclusively say that this association is different for workers with and without a college degree.

[3] For example, $Earnings = f(Age)$, with $f'(Age) > 0$, $f''(Age) < 0$