

Tutorial Business Analytics

Homework 3 - Solution

Exercise 3.4

Install the “AER” (Applied Econometrics with R) package and open the “CPS1988” data set.

a) Briefly describe the data set:

- i. Name the dependent variable and the independent variables.
- ii. Which scales of measurement do the variables belong to (e.g. nominal, ordinal, interval or ratio)?
- iii. Does the data set consist of cross-sectional, time-series or panel data?

b) Plot the dependent variable against each independent variable and transform the variables if necessary. Which transformations would you carry out and why?

Estimate the following model (mr_1):

$$\ln(wage_i) = \beta_0 + \beta_1 \cdot education_i + \beta_2 \cdot ethnicity_i + \beta_3 \cdot experience_i + \beta_4 \cdot experience_i^2$$

c) Interpret the above model (mr_1):

- i. Which variables are statistically significant?
- ii. Is the entire model statistically significant?
- iii. What is the explanatory power of the model and why?
- iv. Interpret each regression coefficient.

Now, estimate the following model (mr_2):

$$\ln(wage_i) = \beta_0 + \beta_1 \cdot education_i + \beta_2 \cdot ethnicity_i + \beta_3 \cdot education_i * ethnicity_i + \beta_4 \cdot experience_i + \beta_5 \cdot experience_i^2$$

d) What is the difference between both models from above (mr_1 and mr_2)?

e) Repeat c) with model mr_2.

Note: Use R to solve this exercise (exercise 3.4_R-template.R). Be aware that the natural logarithm “ln” corresponds to “log” in R.

Solution

a)

- i. dependent variable: wage
independent variables: education, experience, ethnicity, smsa, region, parttime
- ii. ratio: wage, education, experience
nominal: ethnicity, smsa, region, parttime
- iii. cross-sectional: 28155 different men in 1988

b)

```
plot(experience,wage,pch="+")
```

Problem: too many observations of wage close to the origin and only a few very far away

Solution: weigh the observations with the logarithmic function

```
plot(experience,log(wage),pch="+")
```

Problem: quadratic relationship observable

Solution: include square experience in the model

c) Interpret model (mr_1):

- i. All variables, including the intercept, are statistically significant (t value) at a significance level of 0.1 percent ($\Pr[>|t|]$).
- ii. The entire model is statistically significant (F-statistic) at a significance level of 0.1 percent (p-value).
- iii. Adjusted R-squared: 0.3346 (rather low explanatory power)
Reason: too many important variables missing (e.g. ability)
- iv. S. below

$$\beta_0 = 4.321$$

$$wage_i = e^{4.321} = 75.26$$

Wage per week for Caucasian-American worker with no education and no experience.

$$\beta_1 = 0.08567$$

$$\ln(wage_i'/wage_i) = 0.086$$

Wage increases by a factor of $e^{0.086} \approx 1.09$, or 9% for each additional year or education. (C)

$$\beta_2 = -0.2434$$

$$\ln(wage_i'/wage_i) = -0.2434$$

Wage decreases by a factor of $e^{-0.2434} \approx 0.79$, or 21% if worker is African-American (ethnicity_i= 1) instead of Caucasian-American (ethnicity_i=0). (C)

$$\beta_3 = 0.07747 \text{ and } \beta_4 = -0.001316$$

Be careful here and define the following two equations:

$$(I) \quad \ln(wage_i) = \mu + \beta_3 \cdot experience_i + \beta_4 \cdot experience_i^2$$

$$(II) \quad \ln(wage_i') = \mu + \beta_3 \cdot (experience_i + 1) + \beta_4 \cdot (experience_i + 1)^2$$

Here μ contains all other variables and their coefficients (cancels out in next step).

$$(II) - (I): \ln(wage_i'/wage_i) = 0.07747 - 2 \cdot 0.001316 \cdot experience_i - 0.001316$$

Note that because experience enters the linear regression as a linear and a quadratic term, the effect of an increase in experience on wage depends on the level of experience.

Suppose: $experience_i = 40$

$$\ln(wage_i'/wage_i) = -0.029$$

Wage decreases by a factor of $e^{-0.029} \approx 0.97$, or 3% if worker incurs one additional year of experience, if s/he has already been on the job market for 40 years. (C)

Suppose: $experience_i = 10$

$$\ln(wage_i' / wage_i) = 0.05$$

Wage increases by a factor of $e^{0.05} \approx 1.05$, or 5% if worker incurs one additional year of experience, if s/he has already been on the job market for 10 years. (C)

To find the number of years of experience at which further experience decreases the wage:

$$\ln(wage_i' / wage_i) = 0.07747 - 2 \cdot 0.001316 \cdot experience_i - 0.001316 = 0$$

$$\Rightarrow experience_i = 28.934$$

- d) Model mr_2 contains an interaction term in addition to model mr_1. The interaction term between education and ethnicity allows us to distinguish between the marginal effect of education on the wage of an African-American worker and on the wage of a Caucasian-American worker.

Model mr_1:

$$\ln(wage_i) = \beta_0 + \beta_1 \cdot education_i + \beta_2 \cdot ethnicity_i + \beta_3 \cdot experience_i + \beta_4 \cdot experience_i^2$$

Model mr_2:

$$\ln(wage_i) = \beta_0 + \beta_1 \cdot education_i + \beta_2 \cdot ethnicity_i + \beta_3 \cdot education_i * ethnicity_i + \beta_4 \cdot experience_i + \beta_5 \cdot experience_i^2$$

The interaction term in model mr_2 is captured in $education_i * ethnicity_i$ with coefficient β_3 .

e) Interpret model (mr_2):

- i. Ethnicity and the interaction effect are statistically significant (t value) at a significance level of 5 percent ($Pr[>|t|]$). All other variables, including the intercept, are statistically significant (t value) at a significance level of 0.1 percent ($Pr[>|t|]$).

The effect of being African-American on wage now splits up between the dummy and the interaction effect and therefore is weaker for each variable.

- ii. The entire model is statistically significant (F-statistic) at a significance level of 1 percent (p-value).
- iii. Adjusted R-squared: 0.3347 (rather low explanatory power) has increased slightly

Reason: still far too many important variables missing (e.g. ability)

- iv. To interpret the coefficients of model `mr_2`, let us define the following simplified version:

$$\ln(wage_i) = \beta_0 + \beta_1 \cdot education_i + \beta_2 \cdot ethnicity_i + \beta_3 \cdot education_i * ethnicity_i + \mu$$

Again, μ contains all other variables and their coefficients. Their interpretation does not differ from c).

$ethnicity_i = 0$ (Caucasian-American worker)

$$\Rightarrow \ln(wage_i) = \beta_0 + \beta_1 \cdot education_i + \mu$$

$$\beta_0 = 4.313$$

$$wage_i = e^{4.313} = 74.66$$

Wage per week for Caucasian-American worker with no education and no experience.

$$\beta_1 = 0.08631$$

$$\ln(wage_i' / wage_i) = 0.08631$$

Wage increases by a factor of $e^{0.08631} \approx 1.09$, or 9% for each additional year of education for Caucasian-American worker.

$ethnicity_i = 1$ (African-American worker)

$$\Rightarrow \ln(wage_i) = \beta_0 + \beta_1 \cdot education_i + \beta_2 + \beta_3 \cdot education_i + \mu$$

$$\Leftrightarrow \ln(wage_i) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot education_i + \mu$$

$$\beta_0 + \beta_2 = 4.313 - 0.124 = 4.189$$

$$wage_i = e^{4.189} = 65.96$$

Wage per week for African-American worker with no education and no experience.

$$\beta_2 = -0.124$$

$$\ln(wage_i' / wage_i) = -0.124$$

Wage decreases by a factor of $e^{-0.124} \approx 0.88$, or 12% if worker with no education and no experience is African-American instead of Caucasian-American: $74.66 \cdot 0.88 = 65.7$ (≈ 65.96).

$$\beta_1 + \beta_3 = 0.08631 - 0.00965 = 0.07666$$

$$\ln(wage_i' / wage_i) = 0.07666$$

Wage increases by a factor of $e^{0.07666} \approx 1.08$, or 8% for each additional year of education for African-American worker.

$$\beta_3 = -0.00965$$

Wage decreases by a factor of $e^{-0.00965} \approx 0.99$, or 1% less for African-American worker than for Caucasian-American worker if education rises by one year.

Exercise 3.5

Install the “AER” (Applied Econometrics with R) and the “plm” (Panel Data Econometrics in R) packages and open the “Grunfeld” data set. See the R-Script from the lecture Multiple Regression & Panel Data (3.Reggression.R) to solve this exercise.

- a) Briefly describe the data set:
 - i. Name the dependent variable and the independent variables.
 - ii. Which scales of measurement do the variables belong to (e.g. nominal, ordinal, interval or ratio)?
 - iii. Does the data set consist of cross-sectional, time-series or panel data?
- b) Plot the dependent variable against each independent variable and transform the variables if necessary. Which transformations would you carry out and why?

Consider the following model:

$$invest_{it} = \beta_0 + \beta_1 \cdot value_{it} + \beta_2 \cdot capital_{it}$$

- c) How can you test the presence of unobserved individual specific effects in the above model? Do so.
- d) Should you use a Random Effects Regression or a Fixed Effects Regression to take into account the unobserved individual specific effects?

Note: Use R to solve this exercise (exercise 3.5_R-template.R).

Solution

a)

i. dependent variable: invest
independent variables: value, capital, firm, year

ii. ratio: invest, value, capital, year
nominal: firm

iii. panel: 11 firms from 1935 to 1954

b) We simply transform the data frame by deleting 8 firms from the data set. This is not a necessary transformation, but carried out in the script. It is, however, important to give the data set a panel structure with the following command:

```
panel_grunfeld = plm.data(grunfeld, index = c("firm", "year"))
```

c) First carry out a pooled linear regression (a simple linear regression for panel data), that does not take into account the possibility of unobserved individual specific effects.

```
grunfeld_pool = plm(invest~value+capital, data=panel_grunfeld,  
model="pooling")
```

Then use the Lagrange Multiplier Test for Panel Models to test for unobserved individual specific effects.

```
plmtest(grunfeld_pool)
```

The zero hypothesis assumes no presence of unobserved individual specific effects. As the $p - value < 2.2 \cdot 10^{-16}$ is very low, we reject the zero hypothesis and conclude that significant unobserved individual specific effects are present.

d) In principle, if the Random Effects Regression is adequate to take into account unobserved individual specific effects, the Fixed Effects Regression is not needed anymore. Carry out a Random Effects Regression and a Fixed Effects Regression:


```
grunfeld_re = plm(invest ~ value + capital, data = panel_grunfeld, model =  
"random", random.method="walhus")
```

```
grunfeld_fe = plm(invest~value+capital, data=panel_grunfeld, model="within")
```

Conduct a Hausman test to check whether the unobserved individual specific effects are “problematic” in our panel data model and have to be taken into account by a Fixed Effects Regression.

```
phptest(grunfeld_re, grunfeld_fe)
```

The Hausman test assumes as zero hypothesis that the Random Effects Regression is adequate to take into account unobserved individual specific effects. As the $p - value = 0.98$ is very high, we cannot reject the zero hypothesis and conclude that a Random Effects Regression is adequate to take into account unobserved individual specific effects.