Tutorial Business Analytics

Homework 3 - Solution

Exercise 3.3

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.
   i. Which transformations would you carry out and why?
   ii. Estimate the following model (mr_1):

$$ln(\widehat{wage_\iota}) = \widehat{\beta_0} + \widehat{\beta_1} \cdot education_i + \widehat{\beta_2} \cdot ethnicity_i + \widehat{\beta_3} \cdot experience_i + \widehat{\beta_4} \cdot {experience_i}^2$$

c) Interpret the model above (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.

d) Estimate the following model (mr_2):

$$ln(\widehat{wage_\iota}) = \widehat{\beta_0} + \widehat{\beta_1} \cdot education_i + \widehat{\beta_2} \cdot ethnicity_i + \widehat{\beta_3} \cdot education_i * ethnicity_i + \widehat{\beta_4} \cdot experience_i + \widehat{\beta_5} \cdot {experience_i}^2$$

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

e) Repeat c) with model mr_2.

<u>Note:</u> Use R to solve this exercise (exercise 3.3_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) i.

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

Problem: too many observations of wage close to the origin and only a few very far away
Solution: transform the dependent variable, wage, with the logarithmic function

ii.

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

Problem: quadratic relationship observable
Solution: include square of experience in the model

c) Interpret model (mr_1):

i. All variables, including the intercept, are statistically significant at level *α* = 0.01 (look at Pr[>|t|]).

ii. The entire model is statistically significant (F-statistic) at level *α* = 0.001

iii. Adjusted R-squared: 0.3346 (rather low explanatory power)
Reason: too many important variables missing (e.g. ability)

iv. $\widehat{\beta_0} = 4.321,$
$\Rightarrow ln(\widehat{wage_i}) = 4.321, \quad \widehat{wage_i} = e^{4.321} = 75.26$
Wage per week for Caucasian-American worker is $75.26 with no education and no experience.

$\widehat{\beta_1} = 0.08567,$

(I) education at x:

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

(II) education at x + 1:

$$ln\left(\widehat{wage_i{}'}\right) = \widehat{\beta_0} + \widehat{\beta_1} \cdot (x+1) + \widehat{\beta_2} \cdot ethnicity_i + \widehat{\beta_3} \cdot education_i * ethnicity_i + \widehat{\beta_4} \cdot experience_i + \widehat{\beta_5} \cdot {experience_i}^2$$

(II) – (I): $ln\left(\widehat{wage_i{}'}/\widehat{wage_i}\right) = \widehat{\beta_1}$

where $ln\left(\widehat{wage_\iota}'/\widehat{wage_\iota}\right) = ln\left(1 + \frac{(\widehat{wage_\iota}' - \widehat{wage_\iota})}{\widehat{wage_\iota}}\right) \approx \frac{(\widehat{wage_\iota}' - \widehat{wage_\iota})}{\widehat{wage_\iota}}$

$$\therefore \widehat{\beta_1} \approx \frac{(\widehat{wage_\iota}' - \widehat{wage_\iota})}{\widehat{wage_\iota}} = 0.08567$$

Wage in dollars increases by 8.6 percent for each additional year of education, keeping ethnicity and experience constant.

$\widehat{\beta_2} = -0.2434,$

Wage is 24.34 percent lesser for African-America worker ($ethnicity_i = 1$) as compared to Caucasian-American ($ethnicity_i = 0$), keeping education and experience constant.

For $\widehat{\beta_3} = 0.07747\ and\ \widehat{\beta_4} = -0.001316,$

(I) $ln(\widehat{wage_\iota}) = \mu + \widehat{\beta_3} \cdot experience_i + \widehat{\beta_4} \cdot {experience_i}^2$

(II) $ln\left(\widehat{wage_\iota}'\right) = \mu + \widehat{\beta_3} \cdot (experience_i + 1) + \widehat{\beta_4} \cdot (experience_i + 1)^2$

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

(II) – (I): $ln\left(\widehat{wage_\iota}'/\widehat{wage_\iota}\right) = 0.07747 - 2 \cdot 0.001316 \cdot experience_i - 0.001316$

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: $\boldsymbol{experience_i = 40}$

$ln\left(\widehat{wage_\iota}'/\widehat{wage_\iota}\right) = -0.029$

Wage decreases by 2.9 percent when worker with at least 40 years of experience accumulate one additional year of experience, keeping other independent variables constant.

Suppose: $\boldsymbol{experience_i = 10}$

$ln\left(\widehat{wage_\iota}'/\widehat{wage_\iota}\right) = 0.05$

Wage increases by 5 percent when worker with at least 10 years of experience accumulate one additional year of experience, keeping other independent variables constant.

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

$ln\left(\widehat{wage_\iota}'/\widehat{wage_\iota}\right) = 0.07747 - 2 \cdot 0.001316 \cdot experience_i - 0.001316 = 0$

$\Rightarrow\ experience_i = 28.934$

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:

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

Model mr_2:

$$ln(\widehat{wage_\iota}) = \widehat{\beta_0} + \widehat{\beta_1} \cdot education_i + \widehat{\beta_2} \cdot ethnicity_i + \widehat{\beta_3} \cdot education_i * ethnicity_i + \widehat{\beta_4} \cdot experience_i + \widehat{\beta_5} \cdot {experience_i}^2$$

The interaction term in model mr_2 is captured in $education_i * ethnicity_i$ with coefficient $\widehat{\beta_3}$.

e) Interpret model (mr_2):

i. All variables, including the intercept, are statistically significant at a level α = 0.05
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 level α = 0.001

iii. Adjusted R-squared: 0.3347 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:

$$ln(\widehat{wage_\iota}) = \widehat{\beta_0} + \widehat{\beta_1} \cdot education_i + \widehat{\beta_2} \cdot ethnicity_i + \widehat{\beta_3} \cdot education_i * ethnicity_i + \mu$$

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

For $\boldsymbol{ethnicity_i} = \mathbf{0}$ (Caucasian-American worker)

$\Rightarrow \quad ln(\widehat{wage_\iota}) = \widehat{\beta_0} + \widehat{\beta_1} \cdot education_i + \mu$

$\widehat{\beta_0} = 4.313, \;\; \widehat{wage_\iota} = e^{4.313} = 74.66$
Wage per week for Caucasian-American worker with no education and no experience is $74.66.

$\widehat{\beta_1} = 0.08631,$
Wage increases by 8.6 percent for each additional year of education for Caucasian-American worker, keeping experience constant.

For $\boldsymbol{ethnicity_i} = \mathbf{1}$ (African-American worker)

$\Rightarrow \quad ln(\widehat{wage_\iota}) = \widehat{\beta_0} + \widehat{\beta_1} \cdot education_i + \widehat{\beta_2} + \widehat{\beta_3} \cdot education_i + \mu$

$\Leftrightarrow \quad ln(\widehat{wage_\iota}) = (\widehat{\beta_0} + \widehat{\beta_2}) + (\widehat{\beta_1} + \widehat{\beta_3}) \cdot education_i + \mu$

$\widehat{\beta_0} + \widehat{\beta_2} = 4.313 - 0.124 = 4.189, \quad \widehat{wage_\iota} = e^{4.189} = 65.96$

Wage per week for African-American worker with no education and no experience is $65.96.

$\widehat{\beta_2} = -0.124$

Wage is 12.4 percent lesser if worker with no education and no experience is African-American instead of Caucasian-American:

$74.66 \cdot (1 - 0.124) = 65.4$ (≈ 65.96).

$\widehat{\beta_1} + \widehat{\beta_3} = 0.08631 - 0.00965 = 0.07666$

Wage increases by 7.6 percent for each additional year of education for African-American worker, keeping experience constant.

$\widehat{\beta_3} = -0.00965$

Wage is 0.965 percentage lesser for African-American worker than for Caucasian-American worker for each additional year of education, keeping experience constant.

Exercise 3.4

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.Regression.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 model: $\widehat{invest}_{it} = \widehat{\beta_0} + \widehat{\beta_1} \cdot value_{it} + \widehat{\beta_2} \cdot capital_{it}$

c) How can you test the presence of unobserved individual specific effects in the above model?

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 3.4_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 form 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 null 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.

```
phtest(grunfeld_re, grunfeld_fe)
```

The Hausman test assumes as null 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 null hypothesis and conclude that a Random Effects Regression is adequate to take into account unobserved individual specific effects.