Pset 3 - Empirical

Romain Fernex

2025-02-24

Contents

1 Load Data

```r
root_dir =  "/Users/rfernex/Documents/Education/SciencesPo/Courses"
Sub_dir = "/S2/Econometrics/TD/Psets/Pset1/Data/ee2002ext.dta"
fullpath = paste0(root_dir,Sub_dir)
employment_data = read_dta(fullpath)
```

2 Question 2 & 3

```r
# Creates the new variables and filters out excluded values
employment_data <- employment_data %>% na.omit() %>%
  select(salfr, ddipl1, s, agd) %>%
  mutate(log_salfr = log(salfr),
         agd_sqd = agd^2) %>%
  filter(log_salfr > quantile(log_salfr, probs = 0.005),
         log_salfr < quantile(log_salfr, probs = 0.995),
         ddipl1!= 7)
```

3 Question 4

```r
# turns the education and sex variables into categorical variables
employment_data$ddipl1 <- as.factor(employment_data$ddipl1)
employment_data$s <- as.factor(employment_data$s)

# runs model
model <- lm(log_salfr ~ s + agd + agd_sqd + ddipl1, data = employment_data)

stargazer(model, type = "text", title = "Regression Results",
          dep.var.labels = "Log(salfr)",
          omit.stat = c("f", "ser"))
```

```
Regression Results
========================================
                 Dependent variable:
             ---------------------------
                     Log(salfr)
----------------------------------------
s2                    -0.334***
                       (0.007)

agd                   0.054***
                       (0.002)

agd_sqd               -0.001***
                      (0.00003)

ddipl12               0.270***
                       (0.014)

ddipl13               0.231***
                       (0.009)

ddipl14               0.415***
                       (0.011)

ddipl15               0.602***
                       (0.012)
```

```
ddipl16                0.823***
                        (0.012)

Constant               7.565***
                        (0.048)

----------------------------------------
Observations            21,573
R2                       0.287
Adjusted R2              0.287
========================================
Note:        *p<0.1; **p<0.05; ***p<0.01
```

```r
original_coef <- coef(model)
original_coef_no_intercept <- original_coef[names(original_coef) != "(Intercept)"]

transformed_coef <- (exp(original_coef_no_intercept) - 1) * 100

coef_matrix <- rbind(Original = original_coef_no_intercept,
                     `(Exp(Coef)-1)100` = transformed_coef)
coef_matrix_rounded <- round(coef_matrix, 4)

kable(coef_matrix_rounded, caption = "Table with transformed coefficients (Intercept
 ↪  excluded)")
```

Table 1: Table with transformed coefficients (Intercept excluded)

	s2	agd	agd_sqd	ddipl12	ddipl13	ddipl14	ddipl15	ddipl16
Original	-0.334	0.0536	-0.0005	0.2700	0.2314	0.4149	0.6023	0.8225
(Exp(Coef)-1)100	-28.396	5.5080	-0.0515	30.9913	26.0371	51.4249	82.6347	127.6208

4 Question 5

General observations

- Based on p-values, all independent variables are significant at the 1
- For instance, having the BEPC (ddipl1 = 2) is associated with a 30.99% increase in wage.
- The R-squared (including the adjusted one) appears to be quite low with only 28.72% of the variations in monthly log wages explained by this model. This signifies that the model might potentially benefit from the inclusion of additional variables with a good explanatory power.

On the age and age-squared variables

- We notice that age squared is negatively related to wages (-0.052% decrease) when age is positively related (5.508 % increase). Since we include a squared term we are in fact modeling a quadratic relationship and this indicates that, while wage increases with age, it does so at a decreasing rate over time. In fact, it eventually reaches a maximum before declining.
- This makes sense since, passed a certain age, salaries can't be expected to increase further.

5 Question 6

```
SSR <- sum(residuals(model)^2)
N <- length(employment_data$log_salfr)
K <- length(coef(model))

cat("Sum of squared residuals : ", sprintf("%.2f", SSR))
```

```
Sum of squared residuals :  5129.30
```

```
cat("Number of independent variables (+intercept) : ", K)
```

```
Number of independent variables (+intercept) :  9
```

```
cat("Number of observations : ", N)
```

```
Number of observations :  21573
```

6 Question 7

```
restricted_model <- lm(log_salfr ~ agd + agd_sqd, data = employment_data)
anova_results <- anova(restricted_model, model)
kable(anova_results, digits = 3, caption = "ANOVA Comparison Results") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
              full_width = FALSE)
```

Table 2: ANOVA Comparison Results

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
21570	7003.105	NA	NA	NA	NA
21564	5129.303	6	1873.802	1312.936	0

After computing the F test statistic and the corresponding p-value, we observe the latter is low enough to reject the null hypothesis (≈ 0). This supports the fact that the level of education is indeed significantly related to wages.

7 Question 8

```
## Method 1 : using the "car" package
lh_test <- linearHypothesis(model, "ddipl12 = ddipl13")

stats_df <- data.frame(
  Statistic = c("F statistic", "p-value"),
  Value = c(
    round(lh_test$F[2], 4),
    format.pval(lh_test$`Pr(>F)`[2], digits = 4)
  )
)

kable(stats_df, align = c("l", "r"),
      caption = "Test: ddipl12 = ddipl13")
```

Table 3: Test: ddipl12 = ddipl13

Statistic	Value
F statistic	7.9427
p-value	0.004832

```r
## Method 2 : manual computation
beta <- coef(model)
vcov_mat <- vcov(model)

# Estimate
estimate <- beta["ddipl12"] - beta["ddipl13"]

# Standard error of the parameter estimated :
SE <- sqrt(vcov_mat["ddipl12", "ddipl12"] + vcov_mat["ddipl13", "ddipl13"] -
                  2 * vcov_mat["ddipl12", "ddipl13"])

# T-statistic
t_stat <- estimate / SE

# Degrees of freedom
df <- df.residual(model)

# Calculate the two-sided p-value
p_value <- 2 * pt(-abs(t_stat), df)

results_df <- data.frame(
  Parameter = c("Estimate", "Standard Error", "t-statistic", "p-value"),
  Value = c(estimate, SE, t_stat, p_value)
)

# Print a nicely formatted table
kable(results_df, digits = 4, caption = "Parameter Estimates") %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = FALSE)
```

Table 4: Parameter Estimates

Parameter	Value
Estimate	0.0386
Standard Error	0.0137
t-statistic	2.8183
p-value	0.0048

Independently of the method used, we find the p-value to be low enough (<0.05) for the null hypothesis to be safely rejected at the 5% significance level. This implies that having only the BEPC or having the CAP, BEP,.. does not have the same impact on the dependent variable. This makes sense as we would expect a higher level of education to have a more significant impact on the independent variable.

8 Question 9

```r
# Reduces the dimension of the Education variable by merging level 2 and 3
employment_data_reduced <- employment_data %>%
  mutate(
    ddipl1_reduced = case_when(
      ddipl1 %in% c(2, 3) ~ "2_3",
      ddipl1 %in% c(4, 5, 6) ~ as.character(ddipl1),
      TRUE ~ "1"  # Reference level
    ),
    ddipl1_reduced = factor(ddipl1_reduced),
    agd_sqd = agd^2
  )

# Run the model with the reduced education variable
model2 <- lm(log_salfr~ s + agd + agd_sqd + ddipl1_reduced,
                      data = employment_data_reduced)

stargazer(model2, type = "text", title = "Regression Results",
          dep.var.labels = "Log(salfr)",
          omit.stat = c("f", "ser"))
```

```
Regression Results
=============================================
                      Dependent variable:
                  ---------------------------
                          Log(salfr)
---------------------------------------------
s2                         -0.333***
                            (0.007)

agd                         0.053***
                            (0.002)

agd_sqd                    -0.001***
                           (0.00003)

ddipl1_reduced2_3           0.240***
                            (0.009)

ddipl1_reduced4             0.415***
                            (0.011)

ddipl1_reduced5             0.602***
                            (0.012)

ddipl1_reduced6             0.823***
                            (0.012)

Constant                    7.566***
                            (0.048)

---------------------------------------------
```

```
Observations                21,573
R2                           0.287
Adjusted R2                  0.287
=============================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```

We notice grouping both categories does not change significantly the explanatory power of the model. Indeed the adjusted R-squared remains the same as the one in the model with full specifications for education. This seems counterintuitive given that, in question 8, we rejected the hypothesis that the effect of (ddiple1==2) and (ddipl1==3) is the same.

9 Question 10

```
model2_restricted <- lm(log_salfr ~ s + agd + agd_sqd, data = employment_data)

stargazer(model2_restricted, type = "text", title = "Regression Results",
          dep.var.labels = "Log(salfr)",
          omit.stat = c("f", "ser"))
```

```

Regression Results
========================================
                 Dependent variable:
             ---------------------------
                     Log(salfr)
----------------------------------------
s2                    -0.282***
                       (0.008)

agd                   0.054***
                       (0.003)

agd_sqd               -0.001***
                      (0.00003)

Constant              7.973***
                       (0.053)

----------------------------------------
Observations           21,573
R2                      0.086
Adjusted R2             0.086
========================================
Note:        *p<0.1; **p<0.05; ***p<0.01
```

We observe that all independent variables are significant at the 1% significance leveL.

10 Question 11

```
SSR2 <- sum(residuals(model2_restricted)^2)
cat("Sum of squared residuals : ", sprintf("%.2f", SSR2))
```

```
Sum of squared residuals :  6573.73
```

11 Question 12

To compute the F-statistic I use the new restricted model and the model we established in Question 9 (with the new education variable).

```
# method 1 : retrieves the F-statistic from the ANOVA table
test_result <- anova(model2_restricted, model2)
F_statistic <- test_result$F[2]

# method 2 : manually computes the F-statistic
F_statistic_manual <- ((SSR2-SSR)/4)/(SSR/(N-5))

cat("F-statistic for the education test (manual) :", sprintf("%.2f", F_statistic))
```

```
F-statistic for the education test (manual) : 1515.65
```

```
cat("F-statistic for the education test (manual) :", sprintf("%.2f", F_statistic_manual))
```

```
F-statistic for the education test (manual) : 1518.40
```

We observe there is a slight difference between the F-statistic obtained using the "anova" function and the one obtained through manual computations. Nonetheless, this does not impact the result of question 13 as both are widely superior to the critical value.

12 Question 13

```
df1 <- test_result$Df[2]
df2 <- length(employment_data$log_salfr)
critical_value <- qf(0.95, df1, df2)
cat("95th percentile of the Fisher Distribution :", sprintf("%.2f", critical_value))
```

```
95th percentile of the Fisher Distribution : 2.37
```

The critical value (2.37) is vastly under the value of the Fisher statistic we computed earlier (1515.65), thus we can safely reject the null hypothesis. This is in accordance with what we found earlier and seems to indicate that adding education greatly improves the model.

13 Question 14 :

13.1 a)

Set-up of the test :

$$\begin{aligned} H_0 &: \beta_5 + \beta_5^{Women} - \beta_3 = 0 \\ H_1 &: \beta_5 + \beta_5^{Women} - \beta_3 \neq 0 \end{aligned} \tag{1}$$

We note the t-statistic W, it is equal to the following :

$$W = \frac{\hat{T}}{SE} = \frac{\hat{\beta}_5 + \hat{\beta}_5^{Women} - \hat{\beta}_3}{SE} \tag{2}$$

```
model_interact <- lm(log_salfr ~ agd + agd_sqd + ddipl1 + s + ddipl1:s, data =
↪  employment_data)

# computes the estimate for \hat{T}
beta <- coef(model_interact)
```

```
estimate <- beta["ddipl15:s2"] + beta["ddipl15"] - beta["ddipl13"]
cat("we get the following estimate for the parameter : ", estimate, "\n")
```

```
we get the following estimate for the parameter :  0.586883
```

13.2 b)

Let us now find the standard error of the parameter we are trying to estimate ($\hat{T}$). We set : $\theta = (\beta_5, \beta_5^{Women}, \beta_3)$ and $f(\beta_) = \beta_5 + \beta_{5,Women} - \beta_3 \setminus$ According to the delta method, we have :

$$\sqrt(n)(f(\hat{\theta}) - f(\theta)) \xrightarrow{P} N(0, \sigma) \text{ with } \Sigma \text{ the covariance matrix for } \theta \tag{3}$$

So we can write :

$$Var(\hat{T}) = Var(g(\theta))' \approx \nabla g(\theta)^T Var(\theta) \nabla g(\theta) \text{with } Var(g(\theta)) = \begin{pmatrix} \frac{\delta g}{\delta \beta_5} = 1 \\ \frac{\delta g}{\delta \beta_5} = 1 \end{pmatrix} \tag{4}$$

If we rewrite it this gives us :

$$Var(\hat{T}) = \begin{pmatrix} 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} \text{Var}(\beta_5) & \text{Cov}(\beta_5, \beta_5^{\text{femme}}) & \text{Cov}(\beta_5, \beta_3) \\ \text{Cov}(\beta_5^{\text{femme}}, \beta_5) & \text{Var}(\beta_5^{\text{femme}}) & \text{Cov}(\beta_5^{\text{femme}}, \beta_3) \\ \text{Cov}(\beta_3, \beta_5) & \text{Cov}(\beta_3, \beta_5^{\text{femme}}) & \text{Var}(\beta_3) \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \tag{5}$$

Finally we know that : $SE = \sqrt{Var(\hat{T})}$

Now that we have the formula for it, we can compute the standard error using R :

```
# Get the variance-covariance matrix
vcov_matrix <- vcov(model_interact)
var_cov_subset <- vcov_matrix[c("ddipl15", "ddipl15:s2", "ddipl13"),
                              c("ddipl15", "ddipl15:s2", "ddipl13")]

# Define the gradient vector
gradient <- c(1, 1, -1)

# Computes Var(\hat{})
var_delta <- t(gradient) %*% var_cov_subset %*% gradient

# Compute the standard error
se_delta <- sqrt(var_delta)
cat("We find the following standard error :", sprintf("%.2f", se_delta),"\n")
```

```
We find the following standard error : 0.02
```

13.3 c)

Finally we compute the test statistic :

```
# Computes the test statistic
test_statistic <- estimate/se_delta

# Computes the p-value for the test
df_resid <- df.residual(model_interact)
p_value <- 2 * (1 - pt(abs(test_statistic), df = df_resid))
```

```r
# Compute the two-sided p-value
cat("test statistic : ", test_statistic, "\n")
```

```
test statistic :  29.65444
```

```r
cat("p-value:", sprintf("%.2f", p_value), "\n")
```

```
p-value: 0.00
```

The resulting t-statistic is $t = 29.65$ and the accompanying p-value p is ≈ 0. Given this extremely low p-value, we reject the null hypothesis that the effects of being a woman . Thus, at the 1% level of significance, we observe a significant difference between the effect on wage associated with having a CAP (or equivalent) while being a woman, and that associated with having the Baccalauréat while being a man.