Econometrics - Homework Assignment

Maïté Lamothe - Florentine Oliveira - Tom Verrier

10/11/2019

Contents

1] Description of the sample

Question 1

- interprération écart type EDUC
- comparaison sal et salbegin : gros écart d'écart type, différence éducation, salaire d'efficience (mean)
- interprétations sur gender et minotity: sample composed of more women ou jsp + regarder interprétation dummies gender et tout : much diversity
- ajouter tableau stats jobcat + commenter

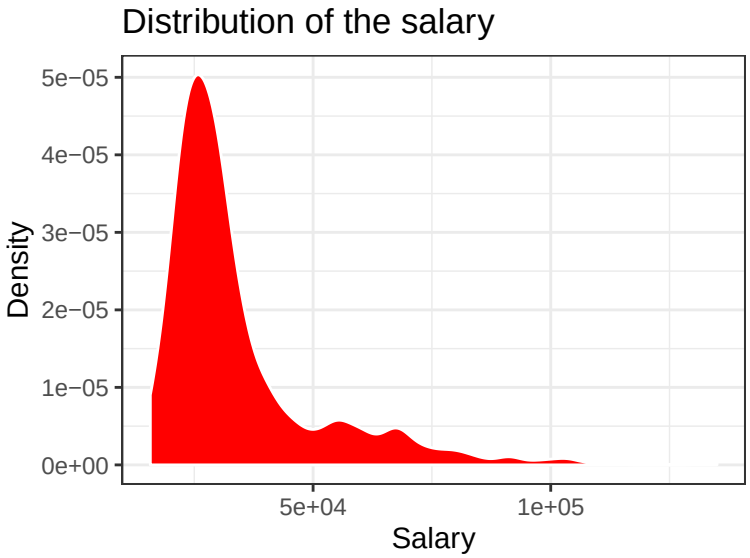
Table 1: Some statistics

variable	Mean	Sd	Min	Max
EDUC	13.49156	2.884846	8	21
SALARY	34419.56751	17075.661465	15750	135000
SALBEGIN	17016.08650	7870.638155	9000	79980

Table 2: Some other statistics

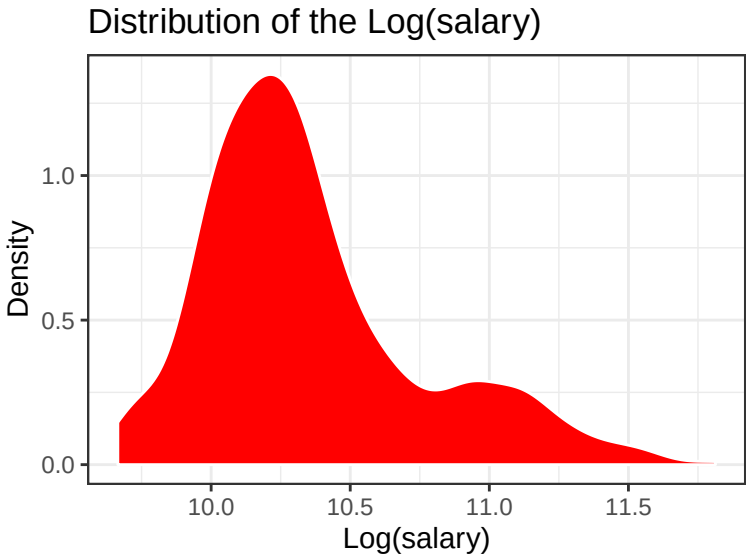
variable	Mean
GENDER	0.5443038
MINORITY	0.2194093

Question 2



It could be a good idea to use the logarithm of the variable Salary because the logarithm linearizes and smoothes the variable. In facts it decreases the expanse of the values that the variable takes (max-min is lower). Moreover, the interest to use the logarithm of this variable is that we can easily interpret the coefficient as an elasticity in a log-log model.

Question 3



We see that the variable LogSal is much more readable. The distribution is less extensive.

2] Linear Regression

Question 1

a)

We estimate the model: $LogSal = \alpha + \beta Education + \epsilon$ (**R1**)

Table 3:

	Dependent variable:
	LOGSAL
EDUC	0.096^{***}
	(0.005)
Constant	9.062^{***}
	(0.063)
Observations	474
R^2	0.485
Adjusted R^2	0.484
Residual Std. Error	0.285 (df = 472)
F Statistic	445.300^{***} (df = 1; 472)
Note:	*p<0.1; **p<0.05; ***p<0.01

We can see on the *Table 2* that the variable Education is statistically significant at the threshold of 1%, as well as the intercept.

b)

This is a log-level linear model. Then, an increase in one year of education lead to an increase of $100.\hat{\beta} = 100 * 0.096 = 9,6$ in the Salary.

c)

Question 2

We now estimate the model: $LogSal = \alpha + \beta_1 Education + \beta_2 LogSalBegin + \epsilon$ (**R2**). The results are shown in *Table 4*.

a)

The impact of education on LogSal is different from the first model **R1** because we have added an explanatory variable in the model.
Mathematically, the matrix X of the explanatory variable is now different. Therefore the vector of estimated coefficients, which is equal to $(X'X)^{-1}X'Y$, differs. There was an omitted variable bias.

b)

Theoretically: total effect is the effect shown with model R1 direct effect is the one of model R2 the indirect effect is the one captures by the regression : logsalbegin on a constant and Education We are supposed to

Table 4:

	Dependent variable:
	LOGSAL
EDUC	0.023***
	(0.004)
LOGSALBEGIN	0.869***
	(0.032)
Constant	1.647***
	(0.275)
Observations	474
R^2	0.801
Adjusted R^2	0.800
Residual Std. Error	0.178 (df = 471)
F Statistic	945.421*** (df = 2; 471)
Note:	*p<0.1; **p<0.05; ***p<0.01

find that R1=R2+

c)

We regress *LogSalBegin* on a constant and *Education*. Results are shown in *Table 5*.

Table 5:

	Dependent variable:
	LOGSALBEGIN
EDUC	0.084***
	(0.004)
Constant	8.538***
	(0.057)
Observations	474
R^2	0.470
Adjusted R^2	0.469
Residual Std. Error	0.257 (df = 472)
F Statistic	418.920*** (df = 1; 472)
Note:	*p<0.1; **p<0.05; ***p<0.01

d)

Question 3

Results of the regression of the model $\overline{LogSal} = \beta_1 \ \overline{education} + \beta_2 \ \overline{LogSalBegin} + \epsilon$ are shown in *Table 6*.

Results of the regresion of the model $DMLogSal = \beta_1 DMeducation + \beta_2 DMLogSalBegin + \epsilon$ are shown in *Table 7*.

Table 6:

	Dependent variable:
	resid_LOGSAL
resid_EDUC	0.023***
	(0.004)
resid_LOGSALBEGIN	0.869***
	(0.032)
Observations	474
R^2	0.801
Adjusted R^2	0.800
Residual Std. Error	0.178 (df = 472)
F Statistic	947.428*** (df = 2; 472)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 7:

	Dependent variable:
	DMLogSal
DMeducation	0.023***
	(0.004)
DMLogSalBegin	0.869***
	(0.032)
Observations	474
R^2	0.801
Adjusted R^2	0.800
Residual Std. Error	0.178 (df = 472)
F Statistic	947.428*** (df = 2; 472)
Note:	*p<0.1; **p<0.05; ***p<0.01

a)

As we can see in *Table 6* and *Table 7*, the estimated coefficients are the same. Proove that residuals is the same as demean.

b)

Proove that demin without constant ad demean with constant leads to the same estimates.

Question 4

We test the model: $LogSal = \beta_1 + \beta_2 education + \beta_3 LogSalBegin + \beta_4 gender + \beta_5 minority + \epsilon$ (**R3**). Results are shown in *Table 8*.

Table 8:

	Dependent variable:
	LOGSAL
EDUC	0.023^{***}
	(0.004)
LOGSALBEGIN	0.822^{***}
	(0.036)
GENDER	0.048^{**}
	(0.020)
MINORITY	-0.042^{**}
	(0.020)
Constant	2.080^{***}
	(0.315)
Observations	474
R^2	0.804
Adjusted R^2	0.802
Residual Std. Error	0.177 (df = 469)
F Statistic	481.321^{***} (df = 4; 469)
Note:	*p<0.1; **p<0.05; ***p<0.01

We can make a student test: $H_0 : \beta_5 = 0$ versus $H_1 : \beta_5 \neq 0$. The test statistic is: $t_{\hat{\beta_5}} = \frac{\hat{\beta_5}}{\hat{\sigma_5}} \sim t_{0,975}(474-4-1)$.

We have: $t_{\hat{\beta_5}} = \frac{\hat{\beta_5}}{\hat{\sigma_5}} = \frac{-0,042}{0,020} \simeq -2.1$.

Hence, $\mid t_{\hat{\beta_5}} \mid = 2.1 > t(469) \in [1,960; 1,984]$. We rejct H_0, which means that the variable *minority* is significant and relevant to explain wages.

Question 5

We can make a Fisher test and test the hypothesis: $H_0 : \beta_4 = \beta_5 = 0$. We can rewrite the hypothesis: H_0:

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} = \quad \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

We can test whether this hypothesis is accepted or reject by making a Fisher test. The statistic is:

$F = \frac{RSS_c - RSS_{nc}}{RSS_{nc}} \cdot \frac{df_c}{df_c - df_{nc}} = \frac{R^2_{nc} - R^2_c}{1 - R^2_{nc}} \cdot \frac{df_c}{df_c - df_{nc}}$

We have that SSR of the constrained model is equal to 18.532 (with 471 df). The SSR of the unconstrained model is 16,900 (with 469 df).

Hence, $\hat{F} = \frac{18,532-14.627}{14.627} . \frac{469}{471-469} = 59,97 > F(2,469) \in [3,01; 3,03]$. We then reject H_0, which means that the variables *minority* and *gender* are jointly statistically significant.

Question 6

We can make a Chow test to test whether the effect of one more year of education is the same for both groups (a group with at most 16 years of education and the other with at least 17 years of education).

Let us call:

- S the residual sum of squares of the global model **R3**;
- $S_{\leq 16}$ the residual sum of squares of the same model but with observations of individuals whom education is below 16 years: $LogSal_{\leq 16} = \gamma_1 + \gamma_2 education_{\leq 16} + \gamma_3 LogSalBegin_{\leq 16} + \gamma_4 gender_{\leq 16} + \gamma_5 minority_{\leq 16} + \epsilon_{\leq 16}$;
- $S_{\geq 17}$ the resiual sum of squares of the same model but with observations of individuals whom education is above 17 years: $LogSal_{\geq 17} = \eta_1 + \eta_2 education_{\geq 17} + \eta_3 LogSalBegin_{\geq 17} + \eta_4 gender_{\geq 17} + \eta_5 minority_{\geq 17} + \epsilon_{\geq 17}$.

We test whether the hypothesis $H_0 : \beta_i = \gamma_i = \eta_i, \forall i \in [\![1; 4]\!]$ is satisfied or not.

The statistic of the Fisher test is:

$$\hat{F} = \frac{S-(S_{\leq 16}+S_{\geq 17})}{(S_{\leq 16}+S_{\geq 17})} . \frac{N-2(K+1)}{K}$$

We then have: $\hat{F} = \frac{16.900-(12.4866+3.3569)}{(12.4866+3.3569)} . \frac{474-2(4+1)}{4} = 7.735286 > F(5,464) \in [2.23; 2,24]$. Hence, we reject H_0, which means that $\exists i \in [\![1; 4]\!]$ such that $\beta_i \neq \gamma_i \neq \eta_i$.

If we focus only on the regression of LogSal on education, we obtain that: $\hat{F} = \frac{38.424-(23.523+6.4651)}{(23.523+6.4651)} . \frac{474-2(1+1)}{2} = 66.1074 > F(2,470) \in [3.01; 3.03]$. Hence, we reject H_0, which means that the effect of *education* in the first group is not the same as the effect in the second groupe.

3. Heteroscedasticity

Question 1

I think it is likely that the variation in wages differs among these categories. It might be the case that the variation in lower in custodial jobs and in administratives jobs than in management jobs, because: - make plots - compute variances

Question 2

We estimate the model: $LogSal = \beta_1 + \beta_2 education + \beta_3 gender + \beta_4 minority + \beta_5 JobCat_2 + \beta_6 JobCat_3 + \varepsilon$ (**R4**)

We get (*Table 9*):

Table 9:

	Dependent variable:
	LOGSAL
EDUC	0.044^{***}
	(0.004)
GENDER	0.178^{***}
	(0.021)
MINORITY	-0.075^{***}
	(0.022)
JOBCAT2	0.170^{***}
	(0.043)
JOBCAT3	0.539^{***}
	(0.030)
Constant	9.575^{***}
	(0.054)
Observations	474
R^2	0.761
Adjusted R^2	0.758
Residual Std. Error	0.195 (df = 468)
F Statistic	297.663^{***} (df = 5; 468)
Note:	*p<0.1; **p<0.05; ***p<0.01

a)

We do not include a dummy for the first job category because If we do so, we would have that the first column of the X matrix, which is a vector composed of 1, is equal to the sum of the variables $JobCat_1$, $JobCat_2$ and $JobCat_3$. This leads to a problem for estimation because the matrix $X'X$ won't be invertible anymore because it is not a full rank matrix.

The parameters β_5 and β_6 are interepreted as the additionnal effect on the log of salary of having a job corresponding to, respectively, category 2 and 3.

b)

We can conclude from these plots that there is an increasing function between the log of the salary and the residuals: the more the log salary, the more we over estimate it.

Question 3

We estimate the model: $LogSal = \beta_1 + \beta_2 education + \beta_3 gender + \beta_4 minority + \varepsilon$ (**R5**).

For subsample n_1 that includes employees of administratives jobs:

Table 10:

	Dependent variable:
	LOGSAL
EDUC	0.046***
	(0.004)
GENDER	0.169***
	(0.021)
MINORITY	−0.099***
	(0.023)
Constant	9.556***
	(0.057)
Observations	363
R^2	0.419
Adjusted R^2	0.414
Residual Std. Error	0.188 (df = 359)
F Statistic	86.292*** (df = 3; 359)
Note:	*p<0.1; **p<0.05; ***p<0.01

For subsample n_2 that includes employees of management jobs:

Table 11:

	Dependent variable:
	LOGSAL
EDUC	0.067^{***}
	(0.017)
GENDER	0.211^{**}
	(0.081)
MINORITY	0.261^{**}
	(0.120)
Constant	9.676^{***}
	(0.274)
Observations	84
R^2	0.309
Adjusted R^2	0.283
Residual Std. Error	0.227 (df = 80)
F Statistic	11.922^{***} (df = 3; 80)
Note:	*p<0.1; **p<0.05; ***p<0.01

The Goldfeld-Quandt test is the most general test in which we assume there are 2 types of individuals (the ones who have a management job and the others who have an administrative job).

We find that: $\hat{\sigma_1}^2 = \frac{12.7142}{363-3-1} = 0.0354156$ and $\hat{\sigma_2}^2 = \frac{4.1396}{84-3-1} = 0.051745$.

We now perform a unilateral test:

$H_0 : \sigma_1^2 = \sigma_2^2 \setminus H_1 : \sigma_1^2 \neq \sigma_2^2 \setminus$

The test statistic is: $\hat{F} = \frac{\hat{\sigma_2}^2}{\hat{\sigma_1}^2} \underset{H_0}{\hookrightarrow} F(n_2 - K, n1 - k)$.

We have: $\hat{F} = \frac{0.051745}{0.0354156} = 1.461079 > F(80, 359) \in [1.30; 1.32]$.

Hence, we reject H_0 which means that $\sigma_1^2 \neq \sigma_2^2$. We consider that there is heteroscedasticity in perturbations.

Question 4

The effects of the presence of heteroskedasticity on the asymptotic properties of the OLS estimator are :

- the estimator is still convergent and unbiased
- it is still asymptotically normal

Question 5

Table 12:

	Dependent variable:
EDUC	0.044^{***}
	(0.004)
GENDER	0.178^{***}
	(0.020)
MINORITY	-0.075^{***}
	(0.021)
JOBCAT2	0.170^{***}
	(0.033)
JOBCAT3	0.539^{***}
	(0.036)
Constant	9.575^{***}
	(0.054)
Note:	*p<0.1; **p<0.05; ***p<0.01