Homework 2
Intermediate Econometrics

October 20, 2023

Exercise 1

See R code.

Exercise 2

Question 1

We estimate the following model and obtain the following results using OLS:

$$logsulfdm_i = \beta_0 + \beta_1 inc_i + \beta_2 pwtopen_i + \beta_3 polity_i + \beta_4 lareapc_i + \varepsilon_i$$

	Dependent variable:
	logsulfdm
inc	-0.361^{*} (0.189)
pwtopen	-0.005^{**} (0.002)
polity	-0.090^{***} (0.031)
lareapc	-0.162^{*} (0.084)
Constant	7.685^{***} (1.480)
Observations	41
R^2	0.537
Adjusted R^2	0.486
Residual Std. Error	0.679 (df = 36)
F Statistic	10.456^{***} (df = 4; 36)
Note:	*p<0.1; **p<0.05; ***p<0.01

The coefficients β_2 (associated with *pwtopen*) and β_3 (associated with *polity*) are significant at the 5% level because their associated p-values are smaller than 0.05.

Let's interpret the coefficients on *inc* and *polity*:

- When the real per capita GDP increases by 1%, the mean SO^2 concentration in micrograms per cubic meter decreases by 0.361% (ceteris paribus).
- When the democracy index increases by one unit, the mean SO^2 concentration in micrograms per cubic meter decreases by 9% (ceteris paribus).

Question 2

We add an interaction term of the logarithm of income (*inc* is already in log!) and OECD country to the model, together with the OECD dummy variable. We thus estimate the following model and obtain the following results using OLS:

$$\begin{aligned} logsulfdm_i = \beta_0 + \beta_1 inc_i + \beta_2 pwtopen_i + \beta_3 polity_i + \beta_4 lareapc_i \\ + \beta_5 oecd_i + \beta_6 (inc_i \times oecd_i) + error_i \end{aligned}$$

	Dependent variable:
	logsulfdm
inc	0.583 (0.346)
pwtopen	-0.009^{***} (0.003)
polity	-0.088^{***} (0.029)
lareapc	-0.195^{**} (0.079)
oecd	11.768^{***} (3.655)
inc × oecd	-1.397^{***} (0.423)
Constant	0.438 (2.632)
Observations	41
R^2	0.653
Adjusted R^2	0.592
Residual Std. Error	0.605 (df = 34)
F Statistic	10.658^{***} (df = 6; 34)
Note:	*p<0.1; **p<0.05; ***p<0.01

The average marginal effect of income on concentration of sulphur is:

$$\frac{\partial E(logsulfdm \mid inc, pwtopen, polity, lareapc, oecd, inc \times oecd)}{\partial inc} = \beta_1 + \beta_6 oecd.$$

We can see that if $oecd = 1$ then the marginal effect is $\beta_1 + \beta_6$, and if $oecd = 0$ then the marginal effect is β_1.

We now want to test if there is a significant difference between the marginal effect of income for OECD countries (i.e. when $oecd = 1$) v.s. non OECD country (i.e. when $oecd = 0$), i.e. if there is a significant difference between $\beta_1 + \beta_6$ and β_1. Thus we do the following t-test:

- The hypotheses are:

$$H_0 : \beta_6 = 0 \text{ against } H_1 : \beta_6 \neq 0.$$

- We use the following test statistic:

$$\hat{t} = \frac{\hat{\beta}_6}{s.e.(\hat{\beta}_6)}.$$

  Under H_0, $\hat{t} \xrightarrow{d} \mathcal{N}(0, 1)$.

- The rejection rule for a 5% level test is to reject H_0 if:

$$\left|\hat{t}\right| > z_{1-\frac{0.05}{2} \,=\, 1-0.025 \,=\, 0.975} = 1.96.$$

- The value of the test statistic is:

$$\hat{t} = \frac{-1.397}{0.423} = -3.3.$$

- Since 3,3 is larger than 1.96, we reject the null in favor of the alternative at 5%. We can thus conclude that there is a significant difference between the marginal effect of income for OECD countries v.s. non OECD country.

Question 3

We want to test for the presence of heteroskedasticity in the original model of question 1. We suppose that potential heteroskedasticity might be coming from *inc*, *pwtopen*, *polity* and *lareapc* and their squared terms. In other words, we suspect the variance of the error term to be in the following way:

$$\begin{aligned} Var(\varepsilon_i \mid inc_i, pwtopen_i, polity_i, lareapc_i) = \alpha_0 &+ \alpha_1 inc_i + \alpha_2 pwtopen_i \\ &+ \alpha_3 polity_i + \alpha_4 lareapc_i \\ &+ \alpha_5 inc_i^2 + \alpha_6 pwtopen_i^2 \\ &+ \alpha_7 polity_i^2 + \alpha_8 lareapc_i^2 \end{aligned}$$

The hypotheses to be tested are the following:

$$H_0 : Var(\varepsilon_i \mid inc_i, pwtopen_i, polity_i, lareapc_i) = \alpha_0$$

$$\text{v.s. } H_1 : Var(\varepsilon_i \mid inc_i, pwtopen_i, polity_i, lareapc_i) = \alpha_0 + \alpha_1 inc_i + \alpha_2 pwtopen_i + \alpha_3 polity_i + \alpha_4 lareapc_i + \alpha_5 inc_i^2 + \alpha_6 pwtopen_i^2 + \alpha_7 polity_i^2 + \alpha_8 lareapc_i^2$$

From the estimation of the original model, we collect the residuals $\hat{\varepsilon}_i$ and we estimate the following auxiliary regression:

$$\hat{\varepsilon_i}^2 = \alpha_0 + \alpha_1 inc_i + \alpha_2 pwtopen_i + \alpha_3 polity_i + \alpha_4 lareapc_i + \alpha_5 inc_i^2 + \alpha_6 pwtopen_i^2 + \alpha_7 polity_i^2 + \alpha_8 lareapc_i^2 + \eta_i \text{ (for all i)}$$

We can stop here, as the linearHypothesis() function in R, which tests the usefulness of this model i.e. with the same H_0 and H_1 as below, gives a p-value $= 0.7775 > 0.05$. So we do not reject H_0 below at 5%, we do not reject homoskedasticity. We can also do a Wald test, by using the $\hat{F} = 0.5914$ that we obtained when testing the usefulness of this model.

We can conduct a Wald test (as stated in the assignment, even though the sample size of the dataset is small, we can apply asymptotic results):

- The hypotheses are $H_0 : \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \\ \alpha_8 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ against $H_1 : \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \\ \alpha_8 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$.
- We use the following test statistic:

  $$\hat{W} = 8 \times \hat{F}.$$

  Under H_0, $8 \times \hat{F} \xrightarrow{d} \mathcal{X}_8^2$.
- The rejection rule for a 5% level test is to reject H_0 if:

  $$8 \times \hat{F} > \mathcal{X}^2_{8,1-0.05} = \mathcal{X}^2_{8,0.95} = 15.507$$
- The value of the test statistic is:

  $$\hat{W} = 8 \times \hat{F} = 8 \times 0.5914 = 4.7312$$
- Since 4.7312 is smaller than 15.507, we cannot reject the null hypothesis at 5%. So we cannot reject homoskedasticity.

Question 4

(a) Let's reestimate the model of question 1 when we assume that we have no information on $Var(\varepsilon_i \mid x_i)$.

In this case, we just use the Heteroskedasticity-Consistent Covariance Matrix Estimator (HCCME) to estimate the model.

We get the following results:

	Dependent variable:
	logsulfdm
inc	-0.361^{*} (0.212)
pwtopen	-0.005^{**} (0.002)
polity	-0.090^{**} (0.038)
lareapc	-0.162^{**} (0.079)
Constant	7.685^{***} (1.749)
Note:	*p<0.1; **p<0.05; ***p<0.01

The estimated coefficients are the same as in question 1, but the standard errors are different.

(b) Now let's assume that:

$$Var(\varepsilon_i \mid x_i) = e^{\theta_0 + \theta_1 inc_i + \theta_2 pwtopen_i + \theta_3 polity_i + \theta_4 lareapc_i}$$

(θ not known)

(FGLS)

We first regress the log of the square of the OLS residuals on an intercept and all the explanatory variables from the question 1 model.

From this auxiliary regression, we extract the fitted values

$$\hat{\theta}_0 + \hat{\theta}_1 inc_i + \hat{\theta}_2 pwtopen_i + \hat{\theta}_3 polity_i + \hat{\theta}_4 lareapc_i.$$

Then we compute an estimator of the variance

$$\hat{\sigma}_i^2 = e^{\hat{\theta}_0 + \hat{\theta}_1 inc_i + \hat{\theta}_2 pwtopen_i + \hat{\theta}_3 polity_i + \hat{\theta}_4 lareapc_i}.$$

Then estimate the initial model by WLS using $\frac{1}{\sqrt{\hat{\sigma}_i^2}}$ as individual weights.

We get the following results:

	Dependent variable:
	logsulfdm
inc (weighted)	-0.490^{**} (0.204)
pwtopen (weighted)	-0.004^{**} (0.002)
polity (weighted)	-0.065^{**} (0.030)
lareapc (weighted)	-0.148^{*} (0.082)
Associated with β_0	8.551^{***} (1.653)
Observations	41
R^2	0.509
Adjusted R^2	0.455
Residual Std. Error	1.771 (df = 36)
F Statistic	9.338^{***} (df = 4; 36)
Note:	*p<0.1; **p<0.05; ***p<0.01

(c) Now let's assume the heteroskedasticity is of the form:

$$Var(\varepsilon_i \mid x_i) = \sigma^2 inc_i$$

(σ^2 is known)

(GLS/WLS)

We transform the initial model to obtain a model in which the homoskedasticity assumption holds.

In the initial model, we have $Var(\varepsilon_i \mid x_i) = \sigma^2 inc_i$.

We need to divide every term by $\sqrt{inc_i}$. We get:

$$\frac{logsulfdm_i}{\sqrt{inc_i}} = \beta_0 \frac{1}{\sqrt{inc_i}} + \beta_1 \frac{inc_i}{\sqrt{inc_i}} + \beta_2 \frac{pwtopen_i}{\sqrt{inc_i}} + \beta_3 \frac{polity_i}{\sqrt{inc_i}} + \beta_4 \frac{lareapc_i}{\sqrt{inc_i}} + \frac{\varepsilon_i}{\sqrt{inc_i}}$$

$$\frac{logsulfdm_i}{\sqrt{inc_i}} = \beta_0 \frac{1}{\sqrt{inc_i}} + \beta_1 \sqrt{inc_i} + \beta_2 \frac{pwtopen_i}{\sqrt{inc_i}} + \beta_3 \frac{polity_i}{\sqrt{inc_i}} + \beta_4 \frac{lareapc_i}{\sqrt{inc_i}} + u_i$$

$$(\text{where } u_i = \frac{\varepsilon_i}{\sqrt{inc_i}})$$

So we get the following results:

	Dependent variable:
	logsulfdm
inc (weighted)	-0.332^{*} (0.185)
pwtopen (weighted)	-0.005^{*} (0.002)
polity (weighted)	-0.091^{***} (0.029)
lareapc (weighted)	-0.152^{*} (0.086)
Associated with β_0	7.403^{***} (1.436)
Observations	41
R^2	0.537
Adjusted R^2	0.485
Residual Std. Error	0.230 (df = 36)
F Statistic	10.430^{***} (df = 4; 36)
Note:	*p<0.1; **p<0.05; ***p<0.01

Exercise 3

Question 1

By OLS: $logwage_i = \beta_0 + \beta_1 education_i + \beta_2 experience_i + \beta_3 married_i + \beta_4 south_i + \varepsilon_i$

	Dependent variable:
	logwage
education	0.092^{***} (0.008)
experience	0.010^{***} (0.002)
married	0.096^{**} (0.044)
south	-0.124^{***} (0.045)
Constant	0.653^{***} (0.128)
R^2	0.229
Adjusted R^2	0.223
Residual Std. Error	0.465 (df = 529)
F Statistic	39.321^{***} (df = 4; 529)
Note:	*p<0.1; **p<0.05; ***p<0.01

Question 2

From the previous results, we can see that *education*, *experience* and *south* are significant at 1% because their associated p-values are smaller than 0.01 (which isn't the case for *married*).

Since we assume that the standard OLS assumptions hold, then the error terms are assumed to be homoskedastic (and no need at this point to compute confidence interval with robust Variance-Covariance Matrix). So we use the basic R function confint() and we get:

- For *education*: $[0.070125384; 0.113291150]$
- For *experience*: $[0.005702115; 0.015076878]$
- For *married*: $[-0.017920022; 0.210101515]$
- For *south*: $[-0.239413177; -0.007823336]$

(We can also compute with the formula, for example for *married*:)

$$[\hat{\beta}_3 - z_{0.995} s.e.(\hat{\beta}_3); \hat{\beta}_3 + z_{0.995} s.e.(\hat{\beta}_3)]$$

$$[0.096 - 2.5758 \cdot 0.044; 0.096 + 2.5758 \cdot 0.044]$$

$$[-0.017; 0.21]$$

For *married*, the parameter estimate is statistically insignificant (at 1%) because 0 belongs to the confidence interval (which is in line with what we found through the p-values).

Question 3

We want to check if the model from question 1 is the same for men and women. We first create subdatasets for females and for males in R. In other words, we now have the following models:

$$logwage_i = \beta_0^f + \beta_1^f education_i + \beta_2^f experience_i + \beta_3^f married_i + \beta_4^f south_i + error_i$$

$$logwage_i = \beta_0^m + \beta_1^m education_i + \beta_2^m experience_i + \beta_3^m married_i + \beta_4^m south_i + error_i$$

We now do a Chow (or Stability) test:

- The hypotheses are $H_0 : \begin{pmatrix} \beta_0^f \\ \beta_1^f \\ \beta_2^f \\ \beta_3^f \\ \beta_4^f \end{pmatrix} = \begin{pmatrix} \beta_0^m \\ \beta_1^m \\ \beta_2^m \\ \beta_3^m \\ \beta_4^m \end{pmatrix}$ against $H_1 : \begin{pmatrix} \beta_0^f \\ \beta_1^f \\ \beta_2^f \\ \beta_3^f \\ \beta_4^f \end{pmatrix} \neq \begin{pmatrix} \beta_0^m \\ \beta_1^m \\ \beta_2^m \\ \beta_3^m \\ \beta_4^m \end{pmatrix}$.

- We can directly see in R (with chow.test()) that the p-value of this test is 3.100914e-24 which is greatly smaller than 0.01. So we reject H_0 at 1%. We can also keep doing the test on paper as follow:

- We use the following test statistic, a Chow F statistic given by chow.test() in R:

$$\hat{F} = 55.5$$

  Under H_0, $\hat{F} \xrightarrow{d} \mathcal{F}_{k,n_f+n_m-2k} = \mathcal{F}_{5,524}$.

- The rejection rule for a 1% level test (we do a 1% level test because we were asked to work with such a level in question 2) is to reject H_0 if:

$$\hat{F} > \mathcal{F}_{5,524,1-0.01} = \mathcal{F}_{5,524,0.99} = 3.05227585$$

- Since 55.5 is much greater than 3.05227585, we reject the null hypothesis at 1%. So there are differences in the model from question 1 between men and women.