Økonometri I

Efterår 2021

Obligatorisk Opgave 3

Deadline: 2. december, 23:59

Lavet af:

Emil Møller Rasmussen **mgn531**

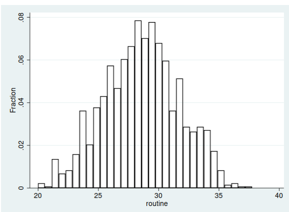
1 Problem 1

1. For this group the lowest student number ends with 1, so throughout this assignment we will work with **groupdata1.dta**. In the stated variables there are 1324 observations spread between two commuting zones in the United States denoted as *czone*. Additionally data is divided into two separate time periods, 1990-2000 and 2000-2007 which is indicated by $t2 = 0 \vee t2 = 1$. In this assignment we will be working with the following variables ΔIPW_{ct}^{USCH},ΔIPW_{ct}^{OTCH}, ΔIPW_{ct}^{UKCH},ΔIPW_{ct}^{USMX} also we will work with a combined variable of $foreignborn, college, routine$ denoted X_{ct}. As such we will only include descriptive statistics about these. The relevant statistics have been included in Table 1

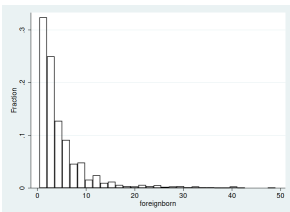
Variables	*mean*	*sd*	*min*	*max*	*count*	*datatype*	$datatype_{compress}$
$t2$	0.5	0.5	0	1	1324	double	byte
ΔIPW_{ct}^{USCH}	1.920	2.599	-0.629	43.085	1324	double	double
ΔIPW_{ct}^{USMX}	0.902	1.482	-6.337	12.596	1324	double	double
ΔIPW_{ct}^{OTCH}	1.773	2.087	-0.723	28.655	1324	double	double
ΔIPW_{ct}^{UKCH}	1.209	1.516	-0.964	22.502	1324	double	double
$college$	45.117	9.151	19.943	70.555	1324	double	double
$foreignborn$	5.041	6.027	.385	48.908	1324	double	double
$routine$	28.596	3.123	19.992	37.748	1324	double	double

Table 1: Variables, descriptive statistics and data types of **groupdata1.dta**

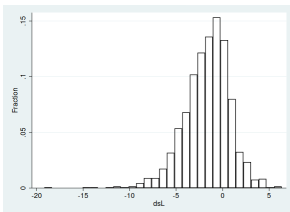
To give a further overview of the relevant variables, histograms have been calculated and included in Figure 1. From these it can be seen that most of the data is right-skewed, but with an overall uniformity to the standard deviations.



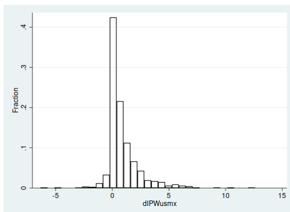
(a) $routine$



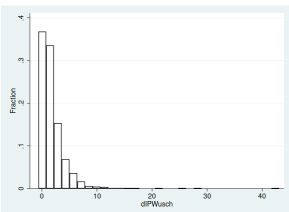
(b) $foreignborn$



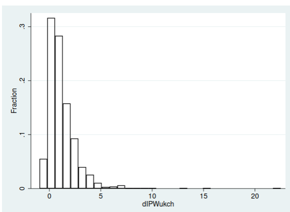
(c) dsl



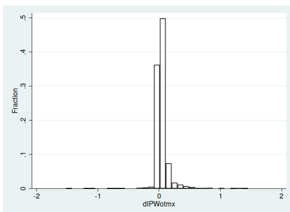
(d) $dIPWusmx$



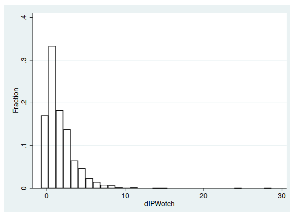
(e) $dIPWusch$



(f) $dIPWukch$



(g) $dIPWotmx$



(h) $dIPWotmx$

Figure 1: Histograms of variables

2. We are provided with the equation seen in (1.1). Here the β_1 coefficient reflects how the percentage change in US imports from China per worker (ΔIPW_{ct}^{USCH})

changes the manufacturing employment share in the same commuting zone c and time t in percent. I.e. a one unit, or percentage point, change in US Import from China changes the percentage of workers employed in manufacturing by β_1.

$$\Delta sL_{ct}^{man} = \beta_0 + \delta_0 t2 + \beta_1 \Delta IPW_{ct}^{USCH} + \delta X_{ct} + u_{ct} \tag{1.1}$$

3. Our intuition is that that as the imports from China to the US rises, manufacturing in the United States will decrease, and therefore the amount of workers employed in this sector. This understanding implies a negative sign, as this reflects a negative relationship between these two variables. The magnitude of the variable indicates how large of a percentage of the workforce will leave manufacturing as the independent variable changes by one percent.

4. We perform a regression analysis of which the results can be found in Table 2. The

Variables	*coeff*	*se*	*t*	*P* >*t*
δ_0*	-1.009	0.141	-7.17	0.000
β_1*	-0.312	0.0255	-12.25	0.015
$\delta*$	0.0097	0.005	1.83	0.067
β_0*	-1.33	0.410	-3.24	0.001
$R^2 = 0.1775$	$\overline{R}^2 = 0.1756$			

Table 2: Estimated parameters of equation (1.1) with standard error and R^2. * indicates statistical significance.

results of the regression show that all the coefficients are statistically significant to a large degree. The coefficient of the variable X, δ, is the least statistically significant level, at a $\approx$ 10% level. The statistical significance for this coefficient is also almost irrelevant, as a change here has an almost nonexistant effect on the dependent variable. As expected the β_1 coefficient is negative, indicating that our intuition was correct - an increase in imported goods from China decreases employed workers in the manufacturing sector in the zone that imports them. We also see that variable $t2$ indicates that there is an additional negative effect on manufacturing in the United States in the second period.

2 Problem 2

1. The validity of an instrument can be divded into two aspects:

   (a) To be a valid instrument a variable must be uncorrelated with the error term, u, and it must be correlated with the endogenous variable. Finally there must not be any perfect multicollinerarity between the used instruments. Using STATA it is found ΔIPW_{ct}^{OTCH} is correlated with the used variables to some degree, and does not suffer from perfect multicollinerarity.

(b) Additionally the condition $Cov(u, z) = 0$ is not possible to test mathematically. Therefore, there must be a causal interpretation of using the specific variable as an instrument. If the given variable is to be a valid instrument it should affect ΔsL_{ct}^{man} not directly, but only through it's effect own on ΔIPW_{ct}^{USCH}. The very definition of the variable seems to confirm the second part of the previous statement, as the US is not included in the variable. Given the global market economy it seems dubious that increased export from China to other countries would not have some effect on employment in manufacturing in the United States, and this effect will most likely be measured through Chinas exports to the United States - as the use of the instrument suggests. On the other hand it cannot be excluded that the change in import/export between other countries and China has a direct effect on the composition of the labor force within the US.

Even though it is infeasible to adequately argument for the validity of the suggested instrument within this paper, based on the above assumptions it seems that ΔIPW_{ct}^{OTCH} fulfills the conditions needed being a valid instrument.

2. We will now estimate (1.1) using IV. The results of this regression can be found alongside values from Table 2 in Table 3.

Variables	*coeff*	*se*	*t*	$P > t$
δ_0*	-1.009	0.141	-7.17	0.000
$\delta_0^{IV}*$	-0.614	0.040	-15.17	0.000
β_1*	-0.312	0.025	-12.25	0.015
$\beta_1^{IV}*$	-0.485	0.157	-3.09	0.002
$\delta*$	0.01	0.005	1.83	0.067
δ^{IV}	0.0005	0.006	0.09	0.930
β_0*	-1.33	0.410	-3.24	0.001
β_0^{IV}	-0.285	0.444	-0.64	0.520

Table 3: Estimated parameters of equation (1.1) with standard error and R^2. * indicates statistical significance. IV indicates coefficients from regression with instrument variable.

We find that the IV estimation has increased the statistical significance of β_1, but greatly decreased the significances for β_0 and δ - making them statistically insignificant. Especially δ has decreased. Through these results we find that by using the instrument models explanatory power increases - as can be seen especially by the increase in the level of statistical significance of β_1.

3. By including ΔIPW_{ct}^{OTCH} and ΔIPW_{ct}^{UKCH} we are effectively including imports to the UK as a factor twice into the regression. We will now test to see if one of the instruments is exogenous. We start out by estimating the regression (1.1), after which we regress the residuals on all the exogenous variables as such:

$$\hat{u} = \beta_0 + \beta_1 X + \beta_2 t2 + \beta_3 \Delta IPW_{ct}^{OTCH} + \beta_4 \Delta IPW_{ct}^{UKCH}. \tag{2.1}$$

With H_0 : *All IV's are uncorrelated with u*, we test $nR_1^2 \overset{a}{\sim} \chi_q^2 \implies 1324 \cdot 0.0044 \overset{a}{\sim} \chi_{4-2}^2 = 5.8256$, the variables therefore do not pass the overidentification test at 1% critical level, but do at a 5% level. This means at a high level of statistical significance our model is not correctly specified, but can be accepted as such at a 5% level. The final model with IV estimated coefficients is as follows:

$$\Delta sL_{ct}^{man} = -0.522 - 0.604 t2_t - 0.546 \Delta IPW_{ct}^{USCH} + 0.003 \delta X_{ct} + u_{ct} \tag{2.2}$$

4. We calculate the parameters of (2.3), giving us the results found in (2.4).

$$\begin{gathered} \Delta sL_{ct}^{man} = \beta_0 + \delta_0 t2_t + \beta_1 \Delta IPW_{ct}^{USCH} + \delta X_{ct} + \\ \beta_3 \Delta IPW_{ct}^{USMX} + u_{ct} \end{gathered} \tag{2.3}$$

$$\begin{gathered} \Delta sL_{ct}^{man} = -0.258 - 0.728^* t2_t - 0.531^* \Delta IPW_{ct}^{USCH} + 0.001^* \delta X_{ct} \\ -0.123^* \Delta IPW_{ct}^{USMX} + u_{ct} \end{gathered} \tag{2.4}$$

We find that all the coefficients except for X and β_0 are statistically significant (* indicates statistical significance). Comparing the coefficients between (2.2) and (2.4) there is a large difference between the effect on the labor force composition the United States when importing products from China vs Mexico. First off, ΔIPW_{ct}^{USMX} is statistically significant at a 1% level, so was most likely an omitted variable in the model. Additionally, it is seen that whilst both variables have a negative effect on manufacturing employment in the united states, the effect is substantially larger for every unit increase of imported goods from China. Thus when importing more from Mexico, compared to China, fewer workers lose their jobs in the US manufacturing sector.

3 Problem 3

1. We perform a 2SLS estimation on (3.4) manually, using ΔIPW_{ct}^{OTCH} and $(t2_t \times \Delta IPW_{ct}^{OTCH})$ as instruments. We estimate 2SLS in two stages. First, we create the

reduced formula of all the exogenous variables for each relevant variable, as seen in (3.1) and (3.2)

$$\Delta \widehat{IPW_{ct}^{USCH}} = \pi_0 + \pi_1 t2_t + \pi_3 X_{ct} + \pi_4 \Delta IPW_{ct}^{OTCH} + \pi_5 (t2_t \times \Delta IPW_{ct}^{OTCH}) + u_{ct} \tag{3.1}$$

$$(t2_t \times \widehat{\Delta IPW_{ct}^{USCH}}) = \pi_0 + \pi_1 t2_t + \pi_3 X_{ct} + \pi_4 \Delta IPW_{ct}^{OTCH} + \pi_5 (t2_t \times \Delta IPW_{ct}^{OTCH}) + u_{ct} \tag{3.2}$$

The estimated values of these equations are then included as variables as seen in (3.3).

$$\Delta sL_{ct}^{man} = \beta_0 + \delta_0 t2_t + \beta_1 \Delta \widehat{IPW_{ct}^{USCH}} + \beta_2 (t2_t \times \widehat{\Delta IPW_{ct}^{USCH}}) + \delta X_{ct} + u_{ct} \tag{3.3}$$

Estimating these values concludes the second step of the 2SLS method, and the results are found in Table 4. The results show that we sucessfully can generate a model with the two instruments.

Variables	*coeff*
$\delta_0^{(3.1)}$	0.576
$\beta_1^{(3.1)}$	-0.0142
$\beta_2^{(3.1)}$	0.846
$\delta^{(3.1)}$	-0.005
$\beta_0^{(3.1)}$	0.365
$\delta_0^{(3.2)}$	0.396
$\beta_1^{(3.2)}$	0.984
$\beta_2^{(3.2)}$	-0.152
$\delta^{(3.2)}$	-0.004
$\beta_0^{(3.2)}$	0.489

(a) First stage results

Variables	*coeff*	*se*	*t*	$P > t$
δ_0*	-.661	.193	-3.43	0.001
β_1*	-.723	.0837	-8.64	0.000
β_2	.129	.0919	1.41	0.160
δ	-.0003	.005	-0.05	0.958
β_0*	-.0989	.423	-0.23	0.815

(b) Second stage results

Table 4: Results of manual implementation of 2SLS

2. Using the IV's from before, we estimate the following regression:

$$\Delta sL_{ct}^{man} = \beta_0 + \delta_0 t2_t + \beta_1 \Delta IPW_{ct}^{USCH} + \beta_2 (t2_t \times \Delta IPW_{ct}^{USCH}) + \delta X_{ct} + u_{ct} \tag{3.4}$$

The calculated parameters can be found in (3.5)

$$\begin{gathered} \Delta sL_{ct}^{man} = -0.0989 - 0.662 t2_t - 0.723 \Delta IPW_{ct}^{USCH} + \\ 0.129 (t2_t \times \Delta IPW_{ct}^{USCH}) - 0.0003 X_{ct} + u_{ct} \end{gathered} \tag{3.5}$$

After having estimated the parameters we will now test H_0 : *No difference in the impact of import competition.* We do this by doing an F-test on β_1 and β_2 across $t2 = 1$ and $t2 = 0$. The two equations we will be testing are as follows:

$$\begin{gathered} \Delta sL_{ct}^{man} = \beta_0 + \beta_1 \Delta IPW_{ct}^{USCH} + \delta X_{ct} + u_{ct} \\ \Delta sL_{ct}^{man} = \beta_0 + \delta_0 1 + \beta_1 \Delta IPW_{ct}^{USCH} + \beta_2 \Delta IPW_{ct}^{USCH} + \delta X_{ct} + u_{ct} \end{gathered} \tag{3.6}$$

Using STATA, with a p-value of 0.0000, we find that we reject our H_0 and thus there is an impact of import competition.

3. Testing exogeneity of ΔIPW^{USCH} and $t2 \times \Delta IPW^{USCH}$ is done by using the residuals ($\hat{\epsilon}$) of the following regression $\Delta IPW^{USCH} = \pi_1 \Delta IPW^{OTCH} \pi_2 \Delta IPW^{UKCH} + \pi_3 X_{ct} + \pi_4 t2_t + \epsilon$, in the following OLS regression $\Delta sL_{ct}^{man} = \gamma_1 t2_t + \gamma_2 X_{ct} + \gamma_3 \Delta IPW_{ct}^{USCH} + \gamma_4 \hat{\epsilon}_{ct}$. Examining the t-statistic of $\hat{\epsilon}$ gives us an indication of the statistical significance of the relevant variables. We find that the coefficient is statistically significant at a 1% level.

4 Problem 4

1. We will now provide the *plim* for the following equation:

$$y_i = \beta_0 + \beta_1 x_i^* + u_i \tag{4.1}$$

We are given the following assumptions. $x_i = x_i^* + \epsilon_i$, $cov(x^*, \epsilon) = 0$. Instrument $z_i = z_i^* + \eta_i$. $cov(z^*, x^*) = \theta\sigma_x^2$. $\sigma_x^2 = Var(x^*)$. $cov(\epsilon, \eta) = \rho\sigma_\epsilon^2$. $Cov(u, z) = 0$ and $cov(z^*, \epsilon) = cov(x^*, \eta) = 0$. $z_i* = z_i + \eta_i$

We know that $x_i^* = x_i - e_i$. We plug this into (4.1):

$$y_i = \beta_0 + \beta_1(x_i - e_i) + u_i \iff y_i = \beta_0 + \beta_1 x_i + u_i - \beta_1 e_i \tag{4.2}$$

This is an example of the classical errors-in-variables assumptions, that $Cov(x_1^*, e_1) = 0$. From Wooldridge p. 311 we know that in this instance there is an attenuation bias. To get this bias we write the formula from (4.2) and rewrite it for β_1.

$$\begin{gathered} cov(y, z) = Cov(\beta_0, z) + Cov(\beta_1 x, z) + Cov(u, z) - Cov(\beta_1 e, z) \\ cov(y, z) = 0 + \beta_1 Cov(x, z) + 0 - \beta_1 Cov(e, z) \\ cov(y, z) = \beta_1(Cov(x, z) - Cov(e, z)) \\ \beta_1 = \frac{Cov(y, z)}{Cov(x, z) - Cov(e, z)} \end{gathered} \tag{4.3}$$

We now insert our value for y.

$$\begin{aligned}\beta_1 &= \frac{Cov(\beta_0, z) + Cov(\beta_1 x^*, z) + Cov(u, z)}{Cov(x, z) - Cov(e, z)} \\ \beta_1 &= \frac{0 + Cov(\beta_1 x^*, z) + 0}{Cov(x, z) - Cov(e, z)}\end{aligned} \tag{4.4}$$

To get the estimator for our IV we insert $z_i = z_i^* + \eta_i$.

$$\begin{aligned}\beta_1^{IV} &= \beta_1 \frac{Cov(x^*, z^* + \eta)}{Cov(x^* + \epsilon, z^* + \eta) - Cov(\epsilon, z^* + \eta)} \\ \beta_1^{IV} &= \beta_1 \frac{Cov(x^*, z^*) + Cov(x^*, \eta)}{Cov(x^*, z^*) + Cov(x^*, \eta) + Cov(\epsilon, \eta) + Cov(\epsilon, z^*) - Cov(\epsilon, z^*) + Cov(\epsilon, +\eta)}\end{aligned} \tag{4.5}$$

Giving us our probability limit as:

$$\hat{\beta}_1^{IV} = \beta_1 \cdot \frac{\theta\sigma_{x^*}^2}{\theta\sigma_{x^*}^2 + 2 \cdot \rho\sigma_\epsilon^2} \tag{4.6}$$

The IV estimator is consistent when $\beta_1^{IV} = \beta_1$, this is evidently so when $\frac{Cov(x^*, z^*)}{Cov(x^*, z^*) + Cov(\epsilon + \eta)} = 1$. So we solve the equation as seen in (4.7)

$$\begin{aligned}\frac{Cov(x^*, z^*)}{Cov(x^*, z^*) + 2 \cdot Cov(\epsilon + \eta)} &= 1 \\ \frac{\theta\sigma_{x^*}^2}{\theta\sigma_{x^*}^2 + 2 \cdot \rho\sigma_\epsilon^2} &= 1\end{aligned} \tag{4.7}$$

It is evident that by evaluating the expression in (4.7), it will only be true for values of $\rho = 0$. Therefore the equation is only consistent for this value.

2. The asymptotic bias in OLS is defined as $plim\hat{\beta}_1 = \beta_1 + \frac{Cov(x,u)}{var(x)}$. We therefore try to solve the equation in (4.8).

$$plim\hat{\beta}_1^{OLS} = plim\hat{\beta}_1^{IV} \tag{4.8}$$

I have not been able to solve this part of the assignment, but if the previous answer holds it would seem intuitive that they would be asymptotically equal in situations where $\rho = 0$ and θ is any number.

5 Problem 5

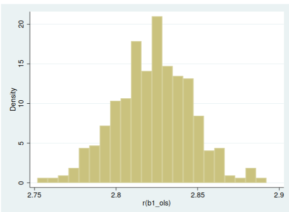
1. For this DPG we will be using the following seed number 1337. The simulation is run and we get the histograms seen in Figure 2. Evaluating the figures it becomes apparent that all IV's are inconsistent for all values except $\rho = 0$. This is more clearly demonstrated in Table 5, as only the mean for $IV_{\rho=0} = 3$. These observations make sense as our instrument is defined as (rewritten): $z_i = \theta x_i^* + \eta_i \implies z_i = \theta x_i + \rho\epsilon_i + \mu_i$, so when $\rho = 0$ the instrument is: $z_i = \theta x_i + 0 + \mu_i \iff z_i = \theta x_i + \mu_i$.

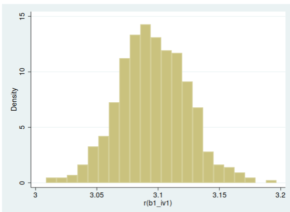
   We know that when the explanatory variable is measured with error, and this error is not correlated with the true but unobserved variable then the OLS estimate of β_1 will exhibit a so-called attenuation bias (as shown in Section 4). The data generation proces specifies $\sigma^2_{x^*} = 4$ and $\sigma^2_\epsilon = 1$, and with a value of $\beta_1 = 3$ and we would thus expect that the OLS estimator on average yields a consistent value when $plim(\hat{\beta}_1) = 3 \implies \beta_1 \frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + 2 \cdot \rho\sigma^2_\epsilon} = 3 \iff \frac{4}{4+(2 \cdot 1 \cdot \rho)} = 1 \iff \frac{4}{4+2\rho} = 1 \implies \rho = 0$. This is what we expected from our derived *plim* in Section 4. It seems that the IV estimator does suffer from attenuation bias in the prescence of measurement error.

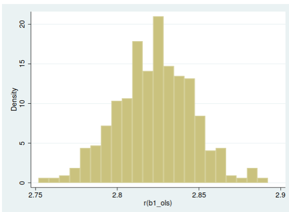
Variables	$OLS_{\rho=-0.5}$	$IV_{\rho=-0.5}$	$OLS_{\rho=0}$	$IV_{\rho=0}$	$OLS_{\rho=0.5}$	$IV_{\rho=0.5}$	$OLS_{\rho=1}$	$IV_{\rho=1}$
mean	2.823	3.095	2.823	2.999	2.822	2.908	2.823	2.823

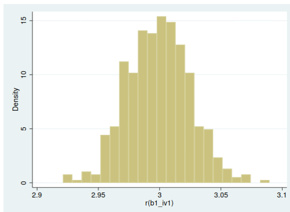
Table 5: Table of results of simulation experiment with different values for ρ

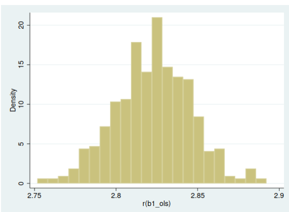
Attenuation bias is when the measurement error is random.

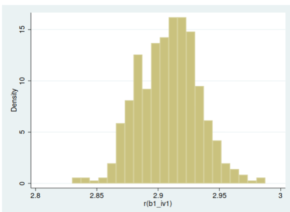

(a) $OLS\rho = -0.5$

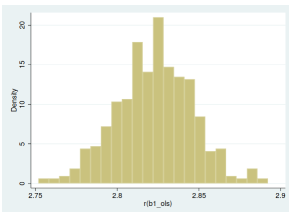

(b) $IV\rho = -0.5$


(c) $OLS\rho = 0$


(d) $IV\rho = 0$


(e) $OLS\rho = 0.5$


(f) $IV\rho = 0.5$


(g) $OLS\rho = 1$

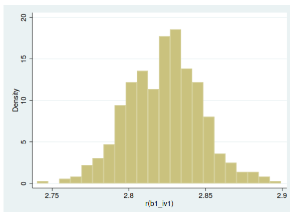

(h) $IV\rho = 1$

Figure 2: Histograms of results of simulation experiment with $\rho = [-0.5, 0, 0.5, 1]$