Assessment guidance, exam SØK3001 Advanced Econometrics, spring 2023

Question 1

Briefly explain the following terms

- *Panel data*

Panel data refers to data with information in two dimensions. Most often across cross-section units (individuals, firms, etc) and time.

- *Average treatment effect (ATE)*

Treatment effects refer to estimates of an intervention. The intervention can either be

- random, for example that the treated and control units are drawn from a lottery
- a natural experiment (or quasi-experiment) where the intervention is not random but where an econometric approach is used such that the estimate can be interpreted as a treatment (for example, the Difference-in-Differences approach).

The effect is an average effect when it is estimated on a sample. Individual effects cannot be estimated.

- *Stationarity*

A (stochastic) time series process is said to be weakly (or covariance) stationary if the expected value, the variance and the covariances are all constant over time. Example of a stationary process is an first order autoregressive, AR(1), process with autoregressive parameter between 0 and 1. Can compare with a random walk where the autorefressive parameter = 1. Can also illustrate what happens after a shock in the stationary case and in the non-stationary case (random walk). In the first case the variable will return to it's equilibrium value (mean reversion). In the non-stationary case, the effect of a shock will never die out.

- *Proxy variable*

A proxy variable is an observed variable related to the unobserved variable of interest. The proxy variable and the variable of interest are correlated but not identical. It might be an advantage to say something about challenges with measurement error, but it is not asked for such a discussion.

Question 2

Politicians are concerned about low degree of completion of high school education because dropout from high school is correlated with crime, low attachment to the labor market and the use of public benefits. To improve the situation, the politicians need knowledge of factors predicting dropout. We have access to a random sample of students at the time they finish lower secondary education. The data includes a dummy variable that is equal to unity if completion of high school within five years and zero otherwise (Comp), the average grade from lower secondary education (GPA), a dummy variable equal to unity if male and zero otherwise (Male), and a dummy variable equal to unity if at least one of the parents has higher education and zero otherwise (PHE). We are interested in variations of the following model:

$$Comp_i = \beta_0 + \beta_1 GPA_i + \beta_2 Male_i + \beta_3 PHE_i + u_i$$

where subscript i denotes individual and u is the error term.

a) What are the necessary assumptions in order to estimate unbiased coefficients by the Ordinary Least Square (OLS) method?

These assumptions are (i) linearity in parameters, (ii) random sampling, (iii) no perfect collinearity, and (iv) zero conditional mean. In reality, the last assumption is the most challenging in economic analyses. It is expected that some explanations are provided for the assumptions, in particular (ii) and (iv).

b) Explain the term heteroskedasticity. Consider whether it is likely that the error term in the model is heteroskedastic. Describe how one can test for heteroskedasticity.

Heteroskedasticity is when the variance of the error term is not constant. The equation above is a linear probability model because the dependent variable is a dummy variable. In this case, there must be heteroskedasticity in the model. It is not important to present the intuition for this result, but it follows from the fact that the variance of the dependent variable depends on its mean value and thus also of the independent variables.

A test for heteroskedasticity is a test of whether the squared error term is constant or related to something. The Breusch-Pagan test for heteroskedasticity has the following form. Estimate the equation and calculate the square of the residuals ($\hat{u}_i^2$). Regress $\hat{u}_i^2$ on all the independent variables. Test whether the coefficients in this regression are jointly significant. It will be useful to write down the auxiliary regression, and to explain the test (either an F-test or a LM-test).

Results for different variants of the model are presented in Table 1. The table presents the estimated coefficients and heteroskedastic-robust standard errors in parentheses.

c) Explain the statistic R-squared (coefficient of determination).

R-squared is the ratio of the explained variation to the total variation. It is expected that the definitions of the explained variation and the total variation are provided.

d) Interpret the estimated coefficients in columns (1) and (2).

The mean value of the dependent variable is equal to the share of the students that complete high school. Column (1) distinguishes between males and females. The constant term is then the average completion rate for females (0.739) and the coefficient for Male is the difference between males and females. Thus, the average completion rate for males is 0.739 – 0.087 = 0.652. Column (2) takes parental education into account, such that the average completion rates for females with parents without higher education is 0.652, females with parents with higher education is 0.652 + 0.217 = 0.869, males with parents without higher education is 0.652 – 0.089 = 0.563, and males with parents with higher education is 0.652 + 0.217 – 0.089 =0.780.

e) The estimated effects of Male and PHE changes when GPA is included in the model in column (3). Explain why.

This is because GPA is correlated with Male and PHE. It is probably easiest to explain this by considering column (2) as a model with an omitted variable. An omitted variable gives biased estimates because the included variables pick up some of the effects of the omitted variable. It might be useful to show this by using formal notational expressions, but it is sufficient to do this for the simplest possible case. Consider the true model

$$(1) \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

and we estimate the simple model excluding x_2. The estimate from the simple model (denoted with a ~) is

$$(2) \quad \tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1$$

where ^ denotes estimates from model (1) and δ_1 follows from the regression

$$(3) \quad x_2 = \delta_0 + \delta_1 x_1 + v$$

where v is an error term.

It follows from equation (2) that a condition for bias is that $\beta_2 \neq 0$. We know from column (3) that $\beta_2 > 0$ in our case.

δ_1 is basically the correlation between x_1 and x_2. It follows from equation (2) that the higher this correlation, the larger is the bias. It also follows from equation (2) that the estimate is negatively biased if $\delta_1 < 0$ and positively biased when $\delta_1 > 0$ when $\beta_2 > 0$. Because the effect of Male decreases when GPA is excluded, the correlation between Male and GPA has to be negative in the data (it is a negative bias in column (2)). Because the effect of PHE increases when GPA is excluded, the correlation between PHE and GPA is positive in the data.

f) The dependent variable in the model is a dummy variable. Discuss challenges by using OLS in this case.

All four assumptions presented in a) might be fulfilled. However, this linear probability model does not have a correct functional form. A probability, here *Comp*, must be between 0 and 1. The model estimated implies that the predicted probability might be below 0 and above 1. The predictions might be obviously wrong. Related to this is that the model has linear effects, and thus the partial effect is

the same in percentage points (change in predicted *Comp*) independent of the value of *Comp*. This seems unlikely to be the case, in particular when *Comp* approach 0 or 1.

The country consists of several regions, which are responsible for high school education. In one region the local politicians considered lack of competition as one factor that could explain low completion rate. Consequently, they changed the admission policy. Initially, the students were enrolled at the closest high school. After the reform, enrollment was based on the average grade from lower secondary education (GPA). We get access to a random sample of students at two points in time – prior to the reform and after the reform – with the same variables as above.

g) *Suggest an econometric approach to estimate the causal effect of the reform.*

The natural approach to use is the Difference-in-Differences (DiD) model. This approach "difference out" of the model a lot of unobserved factors that might be correlated with the reform. The region implementing the reform might differ from other regions in many ways, but this is taken into account in the DiD approach.

It is expected that the model is formalized. It might be easiest to start with the model given in the question.

$$Comp_i = \beta_0 + \beta_1 GPA_i + \beta_2 Male_i + \beta_3 PHE_i + u_i$$

Due to the fact that we have panel data, we expand the model with two variables and the interaction between these two new variables: One variable for the last time period *D* (using the first period as the reference category) and another variable for the reform region *T*. The variable *T* represents the treatment region, and the reference category (all other regions) is the control regions. We add the notation for time, *t*, because the model has a time dimension. The model becomes

$$\begin{aligned} Comp_i &= \beta_0 + \beta_1 GPA_i + \beta_2 Male_i + \beta_3 PHE_i + \delta_0 D_t + \delta_1 T_i + \delta_2 D_t T_i + u_i \\ &= \delta_0 D_t + \delta_1 T_i + \delta_2 D_t T_i + X_i \mathrm{B} + u_i \end{aligned}$$

where $X_i \mathrm{B} = \beta_0 + \beta_1 GPA_i + \beta_2 Male_i + \beta_3 PHE_i$

The symbols used in this guideline is, of course, arbitrary. Whether the presentation of the approach includes the other variables (*GPA, Male, PHE*) is also without substance, likewise for whether these variables have the time notation *t* or not.

This model includes everything specific for the reform region (the variable *T*) and the time periods (the variable *D*), in addition to everything specific for the period after the reform in the reform region (the interaction term *D*T*). The effect of the latter (δ_2) is therefore the estimate of the reform. To show this, it might be useful to make a Table that illustrates the DiD approach. The table presents the predicted values of *Comp* in the four different cases and illustrates the DiD estimate.

	Before the reform	After the reform	After – Before
Control regions	$\hat{\beta}_o + X_i \hat{\mathrm{B}}$	$\hat{\beta}_o + \hat{\delta}_0 + X_i \hat{\mathrm{B}}$	$\hat{\delta}_0$
Treatment region	$\hat{\beta}_o + \hat{\delta}_1 + X_i \hat{\mathrm{B}}$	$\hat{\beta}_o + \hat{\delta}_0 + \hat{\delta}_1 + \hat{\delta}_2 + X_i \hat{\mathrm{B}}$	$\hat{\delta}_0 + \hat{\delta}_2$
Treatment – Control	$\hat{\delta}_1$	$\hat{\delta}_1 + \hat{\delta}_2$	$\hat{\delta}_2$

There are two differences in the model: The difference between the treatment region and the control regions and the difference between the two periods. Subtracting one difference from the other, in

either way (either horizontally in the table or vertically in the table) shows that the DiD parameter is δ_2. This is the effect of the reform.

h) Explain the identifying assumption in your suggestion in g). Can the assumption be tested?

The model assumes that the difference between the regions (δ_1) is constant and do not change over time, except for the reform. The only reason for a change between the regions from the pre-reform period to the post-reform period is the reform. That is, one assumes that the development in the reform region would have been as in the other regions in the absence of the reform. This is called the parallel trend assumption. The trend in *Comp* would have been the same as in the control regions without the reform.

i) Discuss the external validity of the findings from your suggested approach.

External validity is a question of whether the effect estimated by the model is of relevance in other cases/contexts. One might argue that regions in a country are reasonably similar such that the estimated effect of a reform in one region must be expected to give a good prediction of the effect of similar reforms in other regions in the country. There is high external validity in this case. One might also argue that the educational institutions/context vary to a large extent across countries, such that the findings from the model cannot be expected in another country. There is low external validity in this case. This is probably a difficult question because validity is not much mentioned in the textbook, but it is covered in the last lecture on panel data.

Question 3

A student conducts an empirical analysis of supply and demand for locally produced strawberries. The candidate assumes that demand depends on the price of locally produced strawberries, p, and on the price of imported strawberries, p. It is assumed that the supply depends on the price of locally produced strawberries. Furthermore, the candidate takes into account that supply depends on the weather during the season by including a dummy variable, uw, equal to 1, for "unfavorable weather conditions", 0 otherwise.*

a) Let x be the quantity of strawberries sold and formulate the market model given the information above. Thereafter discuss whether the supply and demand equations are identified.

Given the information above, the market model can be formulated by the equations

$$(1)\ x_t = \beta_0 + \beta_1 p_t + \beta_2 p_t^* + u_{1t}$$

$$(2)\ x_t = \alpha_0 + \alpha_1 p_t + \alpha_2 uw_t + u_{2t}$$

where (1) is the demand equation and (2) the supply equation. In addition to the variables defined in the text, the demand equation contains an error term, u_{1t}, and the supply equation an error term, u_{2t}. (These error terms may alternatively be introduced under b)).

Identification: Briefly argue that p* and uw are both exogenous. The equilibrium price, p, is obviously endogenous, so both equations contain one endogenous explanatory variable. The supply equation contains one exogenous variable, uw, with nonzero coefficient, which is excluded from the demand equation. Therefore, the demand equation is exacrly identified. Further, the demand equation contains one exogenous variable, p*, with nonzero coefficient, which is excluded from the supply equation. So supply is also exactly identified.

Intuitively, uw shifts the supply equation which make the demand equation identified, and p* shifts the demand equation which make the supply equation identified.

b) Explain why ordinary least squares (OLS) applied to the demand equation gives biased estimators.

Use (1) and (2) to find the solution (reduced form) for the market price given by

$$(3)\ p_t = \frac{\beta_0 - \alpha_0}{(\alpha_1 - \beta_1)} + \frac{\beta_2}{(\alpha_1 - \beta_1)} p_t^* - \frac{\alpha_2}{(\alpha_1 - \beta_1)} uw_t + \frac{u_{1t} - u_{2t}}{(\alpha_1 - \beta_1)}$$

Now start by stating that OLS gives unbiased and consistent estimators if the all explanatory variables in the equation to be estimated are uncorrelated with the error term.

For a brief answer, refer to (3) where the market price depends on the error term in the demand equation so p must be therefore be correlated with u_{1t} and conclude that OLS applied to (1) will produce biased estimators.

Good candidates are expected to derive the covariance between p and u_{1t}. Given that p* and uw are true exogenous variables, uncorrelated with both error terms, and further assume that the error terms in (1) and (2) are uncorrelated, we find:

$$(4)\ \operatorname{cov}(p_t, u_{1t}) = \frac{\operatorname{var}(u_{1t})}{(\alpha_1 - \beta_1)}$$

Assuming that the supply curve is upward sloping and the demand curve is downward sloping, we conclude that the denominator in (4) is positive. Further, since the variance of the error term must be positive $\operatorname{cov}(p_t, u_{1t}) > 0.$ Intuitively, positive covariance between p and the error term in (1) implies that OLS give an estimator of the price effect on demand which is positively biased (the estimated price effect will be too small).

c) Explain how the parameters in the demand equation can be estimated using the instrumental variable (IV) method (or 2-stage least squares). Also explain what assumptions are needed to ensure that this method gives consistent estimators.

Here it suffice to explain 2SLS more or less mechanically: Write the reduced form price equation (3) more compactly as

$$(5)\ p_t = \pi_0 + \pi_1 p_t^* + \pi_2 uw_t + e_t$$

First stage: Estimate the parameters in (5) and compute the predicted values of p.

Second stage: Replace actual values of p in the demand equation with the predicted values obtained after estimating the first stage and apply OLS to this equation. Alternatively, use predicted values to instrument actual values of p. The two alternatives will give exactly the same estimates, but the second alternative will give the correct standard errors.

This procedure will give consistent estimators if

(i) $\text{cov}\left(p_t^*, u_{jt}\right) = \text{cov}\left(uw_t, u_{jt}\right) = 0,\ j = 1,2$ (instrumental validity)

(ii) $\text{cov}\left(uw_t, p_t\right) \neq 0.$ (instrumental relevance)

Comment 1: Given that (i) holds, the variation in predicted price is driven by variation in the exogenous variables p* and uw so the predicted price is considered exogenous.
Comment 2: (ii) requires that the exogenous variable, uw, which makes the demand equation identified is in fact correlated with the endogenous variable to be instrumented. From (5) we see that this holds if $\pi_2 \neq 0 \Leftrightarrow \alpha_2 \neq 0.$

The candidate uses weakly observations for one region over ten consecutive seasons. Table 2, columns (1) – (4), reports estimated coefficients with standard errors in parentheses. The table also shows, for each regression, the dependent variable, the explanatory variables, and the estimation method, using data for ln x, ln p, ln p, and the dummy variable uw.*

d) Explain how we can test the strength (or relevance) of the instrumental variable used to estimate the demand equation and perform the test using the information provided in the table.

The question here is related to comment comment 2 above. The question is now whether uw is sufficiently correlated with p. To test this empirically we estimate (5) and test the null hypothesis that $\pi_2 = 0.$ In this case, with only one instrument, we can test the null hypothesis using a t-test. The null hypothesis should be rejected with good margin.

Column (2) in Table 2 reports results for the reduced form equation. The estimated parameter of uw is 0.6 with an estimated standard error = 0.06 and t=10 so we reject the null hypothesis with good margin. Could also notice that the F-value is given by the square of t so F = 100 and far above the rule of thum that the F-value should exceed 10. So the overall conclusion is that uw is a really strong or relevant instrumental variable.

e) Compare the estimates of the own price elasticity based on OLS and the IV method. Explain why the observed difference between the OLS- and IV-estimates is as expected. Also explain why the estimated standard error based on IV-estimation is higher than the estimated standard error using OLS.

Using OLS the candidate obtain an estimated own price elasticity equal to -1.25 whereas the IV estimate is -1.75. That is what we should expect since the OLS-estimator is biased upwards, cf answer to b) above. Intuitively one can ask: What are we estimating when we run an OLS-regression between lx and lp (pluss exogenous variables)? Are we estimating the demand equation, or the supply equation, or a mixture of supply and demand?
Concerning the estimated standard errors, we notice that this is higher for 2SLS compared to OLS. Here candidates can simply refer to the conclusion that the variance of the IV-estimator will always

be highet than the variance of the OLS-estimator (or write down and comment the formulas for these variances, cf Wooldridge equations 15.12 and 15.13).

It should also be commented that the difference between the reported standard errors (0.5 versus 0.6) is rather small which is consistent with the conclusion in d).

f) Use results provided in Table 2 to test the hypothesis that the absolute value of the own price elasticity is equal to the elasticity with respect to the import price.

To test the hypothesis in the text we can re-write the demand equation as:

(6) $x_t = \beta_0 + \beta_1\left(p_t - p_t^*\right) + \left(\beta_1 + \beta_2\right)p_t^* + u_{1t} = \beta_0 + \beta_1\left(p_t - p_t^*\right) + \theta p_t^* + u_{1t}$

Under the null hypothesis $\theta = \beta_1 + \beta_2 = 0$.

Column (4) in Table 2 reports results for the transformed demand equation. We see that the estimate of $\theta = -0.25$ with an estimated standard error = 0.12. The relevant t-value is -2.08. Since the absolute value of t exeeds 2 we can reject the null hypothesis (the main point here is how to go on testing the null hypothesis).

g) The candidate discusses the results with a fellow student who thinks the analysis is problematic since there is probably little variation in the import price. What is your response to this comment?

In general, the fellow student has a point: Small variation in an explanatory variable will give large estimated standard errors and impresise estimated. But if we use the results in column (3) we see that the estimate of p* = 1.50, the estimated standard error is 0.6 so the t-value is 2.5 which implies significant effect. The conclusion is that “probably little variation” is not a serious problem in our case.

The candidate’s supervisor suggests that the candidate can go a step further and define two dummy variables reflecting weather conditions assumed to affect the supply of local strawberries: One dummy variabel, dw, for dry weather and another dummy variable, crw, for "cold and rainy" weather. These new dummy variables can be used instead of uw.

h) Explain how the candidate can estimate the demand equation using the two new dummy variables as instruments. Compare the results in columns (3) and (5) in Table 2 and discuss briefly why they are different.

Here we can start with re-writing the first stage regression (5) as:

(7) $p_t = \pi_0 + \pi_1 p_t^* + \pi_2 dw_t + \pi_3 crw_t + e_t$

We now implement 2SLS by estimating the parameters in (7) by OLS – the first stage – compute predicted values of p, replace actual values of p in (1) and estimate this equation by OLS.

Comparing the results for the own price elasticity in (3) and (5) we see that using two instruments give a bith higher own price elasticity (in absolute value) compared to the case where only one instrument is used. Also notice that the estimated standard error decreases. So we can say a marginal

gain is obtained by using two instead of only one instrument. Also, reporting results in (3) and (5) serves as a robustness check.

i) Explain how the strength of two new instrumental variables can be tested.

The null hypothesis is now: $\pi_2 = \pi_3$ which means that neither dw nor crw affect p (the instruments give no information that can be utilized in the IV / 2SLS estimation). To test the null hypothesis, we use an F-test. If the null hypothesis is rejected with good margin F > 10), we conclude that the two instruments are sufficiently strong or relevant.

j) Explain how the candidate can test the validity of the instruments (overidentifying restrictions). Use the results in column (6) to perform the test, where v is the residual from the model in column (5). Critical values can be found in Table 3.

Overidentifying restrictions (Wooldridge ch 15-5b):
The null hypothesis is that the variables assumed to be exogenous are in fact uncorrelated with the error term in the demand equation. Since we do not observe the error term, this can not be tested directly. What we can do is to find the residuals based on the IV-regression. (The residuals are the observable or empirical counterparts to the error terms). Then investigate whether the variables in question are empirically uncorrelated with the error term.

A formal test can be performed by running a regression with the IV-residual, v in the table, as the left hand side variable, and all exogenous variables on the right hand side. This is what's done in column (6) in the Table.

To performe a formal test of the null hypothesis that the parameters of lp*, dw and crw are jointly equal to zero, we can use the R-square from (6) and multiply ny the number of observations to obtain a test statistic with a Chi.square distribution with one degree of freedom. Degrees of freedom is equal to the number of instrumental variables minus the number of endogenous variables to be instrumentet, in our case 2 – 1 = 1 (we have one overidentifying restriction).

Using the information for the auxailary regression (5) we find the value of the test statistic = 80 x 0.05 = 4 which is approximately equal to the 5% critical value (3.84). Exactly how the candidate colcludes is not the most important here, but may be we could question the validity of the instruments.

Could notice that in the case with only one instrument, we are not able to test the validity of the instrumental variable – the equation estimated by IV / 2SLS must be overidentified to perform the test (for overidentification).