Problem Set 2: IV & SEM

Truncated Poisson

Exercise 1 (Optional)**.** For this exercise you will need the dataset zerotrunc.dta and the problems MUST be implemented in STATA where indicated. For this you will need to provide your STATA program and regression output. This data set has a dependent variable called *stay* which indicates the number of days of hospital stay. Length of hospital stay is recorded as a minimum of at least one day. There are three predictor variables: *age, hmo and die.* We will treat the variables age as continuous. The variable hmo and die as binary. hmo indicates wether the patient was insured or not. die indicates whether the patient died during his stay in hospital. Create a histogram of the dependent variable and explain the truncation problem. Conduct a tpoisson and ols regression with robust standard errors and interpret the coefficients. Compare the results from the 2 regressions and explain which model you prefer. Justify your answers.

Solution 1 (Truncated Poisson)**.**
The following Stata code 1 produces the OLS estimates.

```
reg stay age i.hmo i.died
```

Listing 1: OLS for Truncated Poisson.

The value of the coefficient for age, -.14, suggests that the log count of stay decreases by .14 for each year increase in age. This coefficient is not statistically significant. The coefficient for hmo, -1.26, is significant and indicates that the log count of stay for HMO patient is 1.26 less than for non-HMO patients. The log count of stay for patients who died while in the hospital was -1.90 less than those patients who did not die. Finally, the value of the constant, 11.32 is log count of the stay when all of the predictors equal zero.

The following Stata code 2 produces the Truncated Poisson estimates.

```
tpoisson stay age i.hmo i.died, ll(0) vce(robust)
```

Listing 2: Truncated Poisson.

The value of the coefficient for age, -.014442, suggests that the log count of stay decreases by .014442 for each year increase in age. This coefficient is not statistically significant. The coefficient for hmo, -.1359, is significant and indicates that the log count of stay for HMO patient is .1359 less than for non-HMO patients. The log count of stay for patients who died while in the hospital was .20377 less than those patients who did not die. Finally, the value of the constant, 2.4358 is log count of the stay when all of the predictors equal zero.

Omitted Variables and IV

Exercise 2 (25 Marks)**.** For this exercise you will need the dataset bwght.dta and the problems MUST be implemented in STATA where indicated. For this you will need to provide your STATA program and regression output. We are interested in the effect of cigarette smoking (packs) on child birth weight (bwght). Consider the model:

$$\log(\text{bwght}_i) = \beta_0 + \beta_1 \text{packs}_i + u_i$$

We might worry that packs is correlated with other health factors or the availability of good prenatal care, so that packs and u might be correlated. A possible instrumental variable for packs is the average price of cigarettes in the state of residence, cigprice. Conduct an OLS and 2SLS regression and interpret the coefficients. Compare the results from the two regressions and explain which model you prefer. Justify your answers.

Solution 2 (Omitted Variables & IV)**.**
Stata code 3 produces OLS estimates while Stata code 4 produces IV estimates.

```
reg lbwght packs
```

Listing 3: Omitted Variables: OLS.

```
ivreg lbwght (packs = cigprice ), first
```

Listing 4: Omitted Variables: IV.

We might worry that packs is correlated with other health factors or the availability of good prenatal care, so that packs and u might be correlated. A possible instrumental variable for packs is the average price of cigarettes in the state of residence, cigprice. We will assume that cigprice and u are uncorre- lated (even though state support for health care could be correlated with cigarette taxes). If cigarettes are a typical consumption good, basic economic theory suggests that packs and cigprice are negatively correlated, so that cigprice can be used as an IV for packs. To check this, we regress packs on cigprice, using the data in BWGHT: This indicates no relationship between smoking during pregnancy and cigarette prices, which is perhaps not too surprising given the addictive nature of cigarette smoking. Because packs and cigprice are not correlated, we should not use cigprice as an IV for packs in (15.21). But what happens if we do? The coefficient on packs is huge and of an unexpectedly positive. The standard error is also very large, so packs is not significant. But the estimates are meaningless because cigprice fails the one requirement of an IV that we can always test: assumption (15.5).

Simultaneous Equations Models

Exercise 3 (25 Marks)**.** For this exercise you will need the dataset iv2sls.dta and the problems MUST be implemented in STATA where indicated. For this you will need to provide your STATA program and regression output. Consider the 2 equation model:

$$\text{hours}_i = a_1 * \log(\text{wage}_i) + b_{10} + b_{11}\text{educ}_i + b_{12}\text{age}_i + b_{13}\text{kidslt6}_i + b_{14}\text{nwifeinc}_i + u_i$$

$$\log(\text{wage}_i) = a_2\text{hours}_i + b_{20} + b_{21}\text{educ}_i + b_{22}\text{exper}_i + b_{23}\text{exper}_i^2 + u_i.$$

where *age* is the womans age, in years; *kidslt*6 is the number of children less than six years old. $nwifeinc$ is the womans non-wage income (which includes husbands earnings), and *educ* and *exper* are years of education and prior experience, respectively. Conduct two 2SLS regressions with the dependent variables being i) hours_i ii) $\log(\text{wage}_i)$. Compare the 2SLS and OLS results and explain which model you prefer. Justify your answers.

Solution 3 (SEM)**.**
Stata code 5 produces OLS estimates for the SEM while Stata code 6 produces 2SLS estimates for the SEM.

```
reg hours lwage educ age kidslt6 nwifeinc
```

Listing 5: SEM: OLS.

```
ivreg hours (lwage = exper expersq ) educ age kidslt6 nwifeinc
```

Listing 6: SEM: 2SLS.

We use the data on working, married women in MROZ.RAW to estimate the labor supply equation (16.19) by 2SLS. The full set of instruments includes educ, age, kidslt6, nwifeinc, exper, and exper2. The estimated labor supply curve slopes upward. The estimated coefficient on log(wage) has the following interpretation: holding other factors fixed, hours = 16.4(%△ wage). We can calculate labor supply elasticities by multiplying both sides of this last equation by 100/hours: which implies that the labor supply elasticity (with respect to wage) is simply 1,640/hours. [The elasticity is not constant in this model because hours, not log(hours), is the dependent variable in (16.24).] At the average hours worked, 1,303, the estimated elasticity is 1,640/1,303 = 1.26, which implies a greater than 1% increase in hours worked given a 1% increase in wage. This is a large estimated elasticity. At higher hours, the elasticity will be smaller; at lower hours, such as hours = 800, the elasticity is over two. For comparison, when (16.19) is estimated by OLS, the coefficient on log(wage) is -2.05 (se = 54.88), which implies no wage effect on hours worked. To confirm that log(wage) is in fact endogenous in (16.19), we can carry out the test from Section 15.5. When we add the reduced form residuals v_2 to the equation and estimate by OLS, the t statistic on v2 is -6.61, which is very significant, and so log(wage) appears to be endogenous. This differs from previous wage equations in that hours is included as an explanatory variable and 2SLS is used to account for endogeneity of hours (and we assume that educ and exper are exogenous). The coefficient on hours is statistically insignificant, which means that there is no evidence that the wage offer increases with hours worked. The other coefficients are similar to what we get by dropping hours and estimating the equation by OLS.

Exercise 4 (30 Marks)**.**

1. Consider the simple macro model

$$C = a_0 + a_1 Y + u \tag{1}$$
$$I = b_0 + b_1 Y_{-} + b_2 r + v \tag{2}$$
$$Y = C + I + G \tag{3}$$

   where a_1 is the marginal propensity to consume, Y_{-} is lagged income (last year's income may be a proxy for the profitability of firm's, i.e. invest on basis of last year's performance) and introduces dynamics into the system and r is the interest rate. Label each equation in terms of its type, e.g. behavioural, technological, identity or equilibrium condition. List the variables according to their category: predetermined, exogenous and endogenous. Is this system complete? Are Y and u correlated? Explain.

2. Now consider the perfect competition example from class, except here we look at variables in deviation from the mean form:

$$\text{Demand: } q_D = \alpha_1 p + \alpha_2 y + u \tag{4}$$
$$\text{Supply: } q_S = \beta_1 p + v \tag{5}$$
$$\text{Equilibrium: } q_D = q_S \tag{6}$$

   Derive the simultaneous equation bias for the OLS estimator of β_1.

3. Derive the reduced form system for the very simple macro model with structural form

$$C = \alpha + \beta Y + u$$
$$Y = C + Z$$

Solution 4 (Nature, OLS Bias and Reduced Forms with SEMs)**.**

1. Equations (1) & (2) are *behavioural* equations as they describe the behaviour of economic agents (consumption and investment), while (3) is an identity since there is nothing to estimate, (3) is simply

the definition of national income.
There are 4 *predetermined* variables: Y_{-} (lagged endogenous) and 1, r and G (exogenous). Note that 1 is the constant / intercept across equations (1) & (2). Remember, predetermined variables include exogenous and lagged endogenous, i.e. endogenous variables determined in previous periods. There are 3 *exogenous* variables: 1, r and G. Exogenous variables are variables that are not determined by the operation of the system – they are determined outside the system/model. There are 3 *endogenous* variables: C, I and Y. Endogenous variables are variables that are jointly determined by the operation of the system – they are determined within the model.
As the number of equations (three) equals the number of endogenous variables (three), the system is *complete*.

Remark. The disadvantage of structural form is that a_1, which is the marginal propensity to consume does not pick up the effects of simultaneity: changes in Y lead to changes in C, which lead to changes in Y, which lead to changes in I in the next period since the lag of Y determines I, etc. So, the structural form does not allow us to measure the overall effect on the system.

Since Y is a function of C from (3) and C is a function of u from (1), Y is a function of u and so the explanatory variable Y and consumption disturbance u will be correlated, i.e. $Cov(Y, u) \neq 0$. Similarly, Y is a function of I from (3) and I is a function of v from (2), so Y will also be a function of v, the investment disturbance; thus, $Cov(Y, v) \neq 0$. As the explanatory variable is correlated with the disturbance, Classical assumptions are violated.

2. The existence of an endogenous variable on the right-hand side produces a stochastic link with disturbance terms, which leads to bias and more importantly this bias does not go to zero as sample size tends to infinity (inconsistency). The bias and inconsistency from using OLS is called *simultaneous equation bias*. Note that (4) will not be identified and (5) may be exactly identified since it omits one variable. Suppose we have a time series of observations. Let us derive the bias from using the OLS estimate $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum q_t p_t}{\sum p_t^2}$$

where in deviation from the mean form, $x_i = X_i - \bar{X}$. Substituting for q_t:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum (\beta_1 p_t + v_t) p_t}{\sum p_t^2} \\ &= \beta_1 \frac{\sum p_t^2}{\sum p_t^2} + \frac{\sum v_t p_t}{\sum p_t^2} \\ &= \beta_1 + \frac{\sum v_t p_t}{\sum p_t^2}\end{aligned}$$

We know from the weak law of large numbers that averages tend to converge to the true parameters, under certain conditions:

$$E(\hat{\beta}_1) = \beta_1 + E\left(\frac{\sum v_t p_t}{\sum p_t^2}\right)$$

Note that in general, the expectation of a ratio will not be the ratio of expectations, i.e. $E\left(\frac{A}{B}\right) \neq \frac{E(A)}{E(B)}$ in general. However, since $Cov(p_t, v_t) \neq 0$, the right-hand side expectation will be non-zero, so $\hat{\beta}_1$ will be biased. We can use *plim* to derive the exact simultaneous equation bias:

$$plim(\hat{\beta}_1) \overset{\text{Slutsky}}{=} plim\left(\frac{\frac{1}{N}\sum v_t p_t}{\frac{1}{N}\sum p_t^2}\right)$$

Remember that the *plim* is the point at which the sampling distribution collapses. Consistency is defined by the sampling distribution of $\hat{\beta}_1$ going to or collapsing to β_1 as the sample size tends to

infinity, i.e. $plim\hat{\beta}_1 = \beta_1$. So

$$\begin{aligned} plim\left(\frac{\sum v_t p_t}{\sum p_t^2}\right) &= plim\left(\frac{\frac{1}{N}\sum v_t p_t}{\frac{1}{N}\sum p_t^2}\right) \\ &\overset{\text{Slutsky}}{=} \frac{plim\left(\frac{1}{N}\sum v_t p_t\right)}{plim\left(\frac{1}{N}\sum p_t^2\right)} \\ &= \frac{Cov(v_t, p_t)}{Var(p_t)} \neq 0 \end{aligned}$$

A couple of points are worth mentioning here. The second equality is possible for the *plim* but not for expectations; also, while there are $N-1$ degrees of freedom, this is not really relevant here since N will be much like $N-1$ as $N \longrightarrow \infty$. The ratio in the last line is a measure of the asymptotic bias, which is what we were asked to derive. So, $\hat{\beta}_1$ is inconsistent as $plim(\hat{\beta}_1) \neq \beta_1$.

3. Derivation of reduced form for very simple macro model where the structural form provides an economic description of the system: this is a two equation macroeconomic model acknowledging simultaneity and endogeneity. We have 2 exogenous variables, 1 (dummy variable for the intercept, α) and Z. Solving for endogenous variables, we could use matrix methods or Cramer's rule. Elementary substitution shows

$$\begin{aligned} C &= \alpha + \beta C + \beta Z + u \\ Y &= \alpha + \beta Y + Z + u \end{aligned}$$

So rewriting with endogenous variables on the left hand side and exogenous variables on the right hand side:

$$\begin{aligned} (1-\beta)C &= \alpha + \beta Z + u \\ (1-\beta)Y &= \alpha + Z + u \end{aligned}$$

So, the reduced form is given by

$$\begin{aligned} C &= \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta}Z + \frac{u}{1-\beta} \\ Y &= \frac{\alpha}{1-\beta} + \frac{1}{1-\beta}Z + \frac{u}{1-\beta} \end{aligned}$$

Or

$$\begin{aligned} C &= \pi_{11} + \pi_{12}Z + w_1 \\ Y &= \pi_{21} + \pi_{22}Z + w_2 \end{aligned}$$

where $\pi_{22} = \frac{1}{MPS} = \frac{1}{1-MPC}$ where MPC stands for marginal propensity to consume and MPS stands for marginal propensity to save.

Remark. Note that macro models for the Irish economy in the central bank typically have over 100 equations. Reduced form parameters π's are complicated function of structural parameters (α, β, ...) and reduced form disturbances are complicated functions of these and structural disturbances (u, v, ...). We can sometimes interpret reduced form parameters as multipliers but sometimes there is no interpretation. We can estimate reduced form parameters via OLS since classical assumptions hold (assuming no heteroscedasticity, autocorrelation, etc.). We can use reduced form equations for forecasting since exogenous variables are usually under government control.
Note that final form of simultaneous equation models for dynamic models have endogenous variables as functions of purely exogenous variables through a process of continuous substitution. Reduced form can include lagged endogenous variables to pick up some of the dynamics, so while there is no difference in static models between reduced and final form, there tends to be a difference in dynamic models between both forms, usually in terms of the multipliers and parameters.

Exercise 5 (20 Marks)**.** Consider the following macroeconomic model in structural form:

$$\begin{aligned}\text{Consumption: } C_t &= a_0 + a_1 Y_t - a_2 T_t + u_t \\ \text{Investment: } I_t &= b_0 + b_1 Y_{t-1} + v_t \\ \text{Tax: } T_t &= c_0 + c_1 Y_t + w_t \\ \text{GNP identity: } Y_t &= C_t + I_t + G_t\end{aligned}$$

Note that the last equation is an identity since there is nothing to estimate – all coefficients are one.

What are the endogenous variables? What are the predetermined variables? Check the identifiability of the consumption equation.

Solution 5 (Identification of SEMs)**.**
Recall from lectures the notation: M is te number of endogenous variables / equations in the system; m is the number of endogenous variables included in the equatoin of interest; K is the number of predetermined (exogenous plus lagged endogenous) variables in the system; k is the number of predetermined variables included in the equation of interest. The order condition (necessary):

$$\begin{aligned}K - k &\geq m - 1 \\ M + K - (m + k) &\geq M - 1\end{aligned}$$

where the first says that the numbe of excluded predetermined variables from the equation you are focusing on must be no less than the number of endogenous variales on the right hand side of the structural equation you are focusing on; and the second says that the total number of excluded variables (endogenous and predetermined) from the equation you are focusing on must be no less than the number of equations *minus one*. The rank condition (necessary & sufficient) stated that

$$\rho(\Lambda) = M - 1$$

where $\rho(\Lambda)$ is the rank of the matrix Λ and Λ is the matrix formed by the coefficients of variables not included in the equation of interest; we saw this more clearly in class and will see it more clearly once more in answering this question.

Endogenous variables: C, I, T, Y; hence, $M = 4$.

Predetermined variables: I, G, Y_{t-1}; hence, $K = 3$.

The structural parameters (arranged) are

$$\begin{array}{cccccccc} C & I & T & Y & 1 & G & Y_{t-1} \\ 1 & 0 & a_2 & -a_1 & -a_0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -b_0 & 0 & -b_1 \\ 0 & 0 & 1 & -c_1 & -c_0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & -1 & 0 \end{array}$$

Focusing on the consumption function, the order condition is checked by:

$$\begin{aligned}K - k &= 2 \\ m - 1 &= 2 \\ \therefore\ K - k &\geq m - 1\end{aligned}$$

Alternatively

$$\begin{aligned}M + K - (m + k) &= 3 \\ M - 1 &= 3 \\ \therefore\ M + K - (m + k) &\geq M - 1\end{aligned}$$

In both cases, the order condition is satistfied as an equality, so the consumption function *may* be *just* identified. We say *may* be since the order condition is not the sufficient condition – we will know with certainty once we have checked the rank condition, which is both necessary and sufficient as a check for identifiability of an equation in a simultaneous equation model. Checking the rank condition:

$$\Lambda_C = \begin{bmatrix} 1 & 0 & -b_1 \\ 0 & 0 & 0 \\ -1 & -1 & 0 \end{bmatrix}$$
$$\rho(\Lambda_C) = 2 \neq M - 1 = 3$$

Therefore, the consumption function is *under* identified – it cannot be meaningfully estimated.
To see that the rank of Λ_c is less than full rank, we can proceed in two ways. Firstly, we can show that the determinant is equal to zero, which is obvious from multiplying diagonally to the right from top to bottom and adding each diagonal product and then subtracting each diagonal product (to the left from top to bottom):

$$det(\Lambda_C) = 0$$

The second way to see that the columns/rows are linearly dependent is by checking to see whether there exists a set α_1, α_2 and α_3, at least one non-zero such that a linear combination of rows is equal to zero:

$$\alpha_1(1 \;\; 0 \;\; -b_1) + \alpha_2(0 \;\; 0 \;\; 0) + \alpha_3(-1 \;\; -1 \;\; 0) = 0$$

which is true if and only if

$$\begin{aligned} \alpha_1 - \alpha_3 &= 0 \\ \alpha_3 &= 0 \\ -\alpha_1 b_1 &= 0 \end{aligned}$$

which is true if and only if

$$\alpha_1 = \alpha_3 = 0$$

but α_2 can be anything, i.e. it does not necessarily have to be zero! So, there exists $\alpha_1 = \alpha_3 = 0$ and $\alpha_2 \neq 0$ (e.g. $\alpha_2 = 1$) such that the linear combination of rows of the matrix Λ_C sum to zero. Therefore, the rows of Λ_C are linearly dependent, so Λ_C has less than full rank, i.e. $\rho(\Lambda_C) < 3$ but $M - 1 = 3$, so the consumption function is under-identified (it is not identified); therefore, no meaningful estimation of the coefficients in the consumption function is possible.
This is as far as you would need to go to get full marks, but note that you can tell that $\rho(\Lambda_C) = 2$ because the first row and the last row are linearly independent: the determinant is $-1 + b_1 - b_1 = -1 \neq 0$ and

$$\alpha_1(1 \;\; 0 \;\; -b_1) + \alpha_2(-1 \;\; -1 \;\; 0) = 0$$

if and only if

$$\begin{aligned} \alpha_1 - \alpha_2 &= 0 \\ -\alpha_2 &= 0 \\ -\alpha_1 b_1 &= 0 \end{aligned}$$

if and only if

$$\alpha_1 = \alpha_2 = 0$$

and so the first and last row of Λ_C are linearly independent, so we have two linearly independent rows in Λ_C. One definition of the rank (there are many equivalent definitions) is that the rank is equal to the number of *linearly independent* rows. So $\rho(\Lambda_C) = 2$.