Suggested Answers Problem set 3
ECON 60303

Bill Evans
Spring 2014

1. In the simple bivariate regression $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$ we know the estimate for β_1 can be written as

$\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ but in this case x_i =1 or 0. There are n observations in the sample and $n_1 = \sum_{i=1}^{n} x_i$ observations for which x_i=1 and $n_0 = \sum_{i=1}^{n}(1 - x_i)$ for which x_i=0 and n_1+n_0=n. Recall also that

$$\bar{y}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i} \text{ and } \bar{y}_0 = \frac{\sum_{i=1}^{n} y_i(1 - x_i)}{\sum_{i=1}^{n}(1 - x_i)}$$

Work with the numerator for $\hat{\beta}_1$ first.

$$\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n}(y_i - \bar{y})x_i = \sum_{i=1}^{n} y_i x_i - \bar{y}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i x_i - \bar{y}n_1$$

Note that $\sum_{i=1}^{n} y_i x_i = n_1\bar{y}_1$ and $\bar{y}$, the sample mean of y, is simply a weighted average of $\bar{y}_1$ *and* $\bar{y}_0$ where $\bar{y} = \frac{n_1}{n}\bar{y}_1 + \frac{n_0}{n} y_0$. Therefore, the numerator can be written as

$$n_1\bar{y}_1 - n_1\left(\frac{n_1}{n}\bar{y}_1 + \frac{n_0}{n} y_0\right) = n_1\bar{y}_1 - \frac{n_1^2}{n}\bar{y}_1 - \frac{n_1 n_0}{n} y_0 = \frac{nn_1\bar{y}_1 - n_1^2\bar{y}_1 - n_1 n_0\bar{y}_0}{n} = \frac{n_1(n - n_1)\bar{y}_1 - n_1 n_0\bar{y}_0}{n}$$

and because $n = n_1 + n_0$ *then* $n_0 = n - n_1$ *and the numerator equals*

$$\frac{n_1 n_0}{n}\left(\bar{y}_1 - \bar{y}_0\right)$$

Now work with the denominator. Note that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i$

Remember that $n_1 = \sum_{i=1}^{n} x_i$ and since x_i = 1 or zero then $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i = n_i$ so

$$\sum_{i=1}^{n} x_i^2 - \bar{x}n_1 = n_1 - \frac{n_1}{n}(n_1) = n_1 - \frac{n_1^2}{n} = \frac{n_1 n - n_1^2}{n} = \frac{n_1(n - n_1)}{n} = \frac{n_1 n_0}{n}$$ and therefore

$$\hat{\beta}_1 = \frac{\frac{n_1 n_0}{n}(\bar{y}_1 - \bar{y}_0)}{\frac{n_1 n_0}{n}} = (\bar{y}_1 - \bar{y}_0)$$

2. The 2SLS estimate is $\hat{\beta} = (x' P_z x)^{-1}(x' P_z y)$ and using partitioned inverses, the estimate for $\hat{\beta}$ would be $\hat{\beta} = (x' M_v x)^{-1}(x' M_v y)$ where $M_v = I - \hat{v}(\hat{v}'\hat{v})^{-1}\hat{v}'$. The easiest way to prove that these two estimates are the same is to show that $M_v x = P_z x$.

   From the first, stage regression, by construction $z'\hat{v} = 0$ and $\hat{v}'z = 0$ so whenever we see these terms, they drop out.

   Note that $M_v x = (I - \hat{v}(\hat{v}'\hat{v})^{-1}\hat{v}') = x - \hat{v}(\hat{v}'\hat{v})^{-1}\hat{v}'x$. Note also that $\hat{v} = x - z\hat{\pi}$ so $x = \hat{v} - z\hat{\pi}$. Substituting this into the equation for $M_v x$

$$M_v x = x - \hat{v}(\hat{v}'\hat{v})^{-1}\hat{v}'x = x - \hat{v}(\hat{v}'\hat{v})^{-1}\hat{v}'(\hat{v} - z\hat{\pi}) = x - \hat{v} = z\hat{\pi} = z(z'z)^{-1}z'x = P_z x$$

$$\hat{\beta} = (x' M_v x)^{-1}(x' M_v y) = (x' M_v x)^{-1}(M_v x)' y = (x' P_z x)^{-1}(P_z x)' y = (x' P_z x)^{-1} x P_z' y$$

3. $Var(\hat{\beta}_1^{2SLS}) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (\hat{x}_i - \bar{\hat{x}})^2}$ and $Var(\hat{\beta}_1^{OLS}) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$ so

$$\frac{Var(\hat{\beta}_1^{OLS})}{Var(\hat{\beta}_1^{2SLS})} = \frac{\frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}{\frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (\hat{x}_i - \bar{\hat{x}})^2}} = \frac{\sum_{i=1}^n (\hat{x}_i - \bar{\hat{x}})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

   The properties of the OLS are such that $\bar{\hat{x}} = \bar{x}$ and therefore $\frac{Var(\hat{\beta}_1^{OLS})}{Var(\hat{\beta}_1^{2SLS})} = \frac{\sum_{i=1}^n (\hat{x}_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = R^2_{1st\ stage}$

Empirical portion

A sample program to construct the estimates below is in answers_ps3.do. This produces the estimates for y1. Just global search and replace reg y1 with reg y2 to produce the other set of estimates. My estimates seem to be slightly different from yours because you need to adjust the critical value of the t-statistic based on the appropriate degrees of freedom.

A couple of notes about the results. First, the clustered standard errors seems to nail the correct rejection rate when groups are in excess of 100, but certainly as the number of groups starts to fall below 75 then the type I error rates really takes off. Second, the random effects model is thought to be a poor alternative because it imposes equal

correlations across all pairs of observations, which cannot be correct, especially for dummy endogenous variables. That said, the model does reasonably well compared to the clustered model. Oh well.

Fraction of 1000 draws where the reject the H_0: $\beta_{dummylaw} = 0$

	Using Y1 as the outcome			Using Y2 as the outcome		
# groups	OLS	Clustered	Random Effect	OLS	Clustered	Random Effect
400 groups	0.208	0.041	0.041	0.137	0.047	0.047
100 groups	0.220	0.052	0.051	0.118	0.049	0.044
75 groups	0.214	0.060	0.057	0.106	0.062	0.063
50 groups	0.239	0.063	0.060	0.118	0.070	0.058
25 groups	0.265	0.085	0.074	0.164	0.084	0.077
10 groups	0.345	0.171	0.104	0.158	0.168	0.079