Problem Set 3
ECON 60303
(Due: Friday, March 1, 2013)

Bill Evans
Spring 2013

1. Consider a regression of y_i on a dummy variable (x_i). The regression is of the form $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$ and we know that OLS estimate for β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Show that because x_i is a dummy variable that the OLS estimate for β_1 equal to $\hat{\beta}_1 = \bar{y}_1 - \bar{y}_0$ where $\bar{y}_1 = (\bar{y} \mid x_i = 1)$ *and* $\bar{y}_0 = (\bar{y} \mid x_i = 0)$.

Some notation. There are n observations, and
$\sum_{i=1}^{n} x_i = n_1$ *and* $\sum_{i=1}^{n}(1 - x_i) = n_0$ *and* $n_1 + n_0 = n$.
Note also that $\bar{y}_1 = \sum_{i=1}^{n} y_i x_i / \sum_{i=1}^{n} x_i$ *and* $\bar{y}_1 = \sum_{i=1}^{n} y_i(1 - x_i) / \sum_{i=1}^{n}(1 - x_i)$

2. Suppose $y = x\beta + \varepsilon$ where x (n x k) is and assume $E[x'\varepsilon] \neq 0$ but there is a valid matrix of instruments z (n x q where q≥k) where $E[z'\varepsilon] = 0$. Define the first stage as $x = z\pi + v$ and define $\hat{v} = x - z\hat{\pi}$ where $\hat{\pi} = (z'z)^{-1}z'x$. Show that the OLS estimate for β from a regression of $y = x\beta + \hat{v}\lambda + \varepsilon^*$ will produce an estimate for $\hat{\beta}$ that is identical to the 2SLS estimate $\hat{\beta} = (x'P_z x)^{-1}(x'P_z y)$.

3. Consider the bivariate model $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$ where we suspect $\text{cov}(x_i, \varepsilon_i) \neq 0$ so we consider a 2SLS model with a single instrument z. Show that $\text{var}(\hat{\beta}_1^{2SLS}) = \text{var}(\hat{\beta}_1^{ols}) / R_x^2$ where R_x^2 is the R-squared from the first stage regression $x_i = \pi_0 + z_i\pi_1 + v_i$.

4. The STATA data set *data_for_ps3.dta* has 10 observations each from 400 groups. The data set has the following variables:

Variable	Definition
Groupid	Identifies groups (1-400)
Personid	Identifies people within a group (1-10)
y1	Continuous outcome variable
y2	Dummy (0-1) outcome variable

In this problem, you are to perform a simulation exercise similar to the one in Bertrand et al. Specifically, I want you to draw a random number from a uniform distribution over the 0-1 interval for each group. Call this variable temp. For those with temp<0.5, assign a random dummy variable that equals 1 for the group. This procedure will on average assign 50% of the groups a dummy variable that equals 1. Call this variable *dummylaw*. For each iteration, regress y1 (the continuous outcome) on *dummylaw* and save the t-statistic assuming there are no errors across observations within the group. Next, estimate the model but cluster the standard errors at the *groupid* level and again save the t-statistic. Finally, estimate a random effects model with the random effect at the group level and again, save the t-statistic. Do this 1000 times. I want the fraction of the times in all three cases you can reject the null that the coefficient on the *dummylaw* is zero (that is, where the absolute value of the t-statistic is in excess of 1.96.

Next, you are to do this for only 100 groups (groupid<=100), for 75 groups (groupid<=75), for 50, 25 and 10 groups. Then, you are to re-do the full set of results use y2 (a dummy variable) as the dependent variable. You essentially want to fill out the table below.

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	0.206					
100 groups						
75 groups						
50 groups						
25 groups						
10 groups						

To generate a dummy that turns on the a group, you will need to run the following text

```
qui by groupid: gen temp = uniform() ;
qui by groupid: gen dummylaw = (temp[1] < .5) ;
```

Because *dummylaw* is a random variable with no meaning, it should be the case we have a 5% Type I error rate. But because of within group correlation in errors, this number will be well in excess of 5%. In the table above, I get a Type I error rate of 20.6%. we also know that when we cluster, so long as there are large numbers of clusters, this Type I error rate should be close to 5%.

Using Y1 as the outcome, how well does the random effects model do compared to the clustered standard error in reducing Type I error rates.

The Random Effect model has been discounted as a way to deal with within group correlation in errors in the case when the outcome is a dummy variable because the model assume equal correlation across any two observations and by construction, the model cannot have equal correlation when outcomes arwhen y1 =1 for 1 person and y1=0 for another, the correlation will be different