Problem Set 4
ECON 60303
(Due: Friday, February 28, 2014)

Bill Evans
Spring 2014

1. Using the psid1.dta data set and the areg procedure, run a fixed-effect model of ln(wages) on tenure and union status absorbing the person id. Next, construct within-panel deviations in mean for tenure and union ONLY. Next, run a two-stage least square model of ln(wages) on tenure and union using the within-panel deviations in union and tenure as instruments. Fill in the table below

Parameter Estimates and Standard errors

Covariates	Fixed Effect	2SLS
Union		
Tenure		

Kinda neat hah? Why are the standard errors so much lower in the fixed-effect model?

2. Consider the simple linear model with panel data

$Y_{it} = X_{it}\,\beta + \eta_{it}$

where i=1,2,...N and t=1,2,...T, and X_{it} is a scaler. Suppose the error term has a two-part structure where $\eta_{it}=u_i+\varepsilon_{it}$ and ε_{it} is a zero mean random error with $var(\varepsilon_{it})=\sigma^2_{\varepsilon}$ and u is an individual-specific error that is potentially correlated with X.

a. Show that the variable $\tilde{X}_{it} = X_{it} - \bar{X}_{i}$ (where $\bar{X}_{I}$ is the within panel mean of X) is by construction uncorrelated with u_i. (Don't overthink this – this is not a question about expectations – this is a question about finite samples).

b. Consider an instrumental variables model where one were to use $\tilde{X}_{it}$ as an instrument for X_{it} to eliminate the potential covariance between X and u. Show that the IV estimate of β where $\tilde{X}_{it}$ is used as an instrument for X is identical to a fixed-effects estimate for β. (This is much easier to write in matrix form – use the matrix characterization of the problem).

3. Consider the indirect least squares model where the structural equation of interest is

$$y_i = \beta_0 + x_i\beta_1 + w_i\beta_2 + \varepsilon_i$$

Where x is a scaler and w_i is a g x 1 vector of other exogenous factors. The first-stage is estimated as

$$x_i = \pi_0 + z_i\pi_1 + w_i\pi_2 + u_i$$

Where z is a scaler. The reduced form is estimated by the equation

$$y_i = \theta_0 + z_i\theta_1 + w_i\theta_2 + v_{2i}$$

And the 2SIV estimate of $\hat{\beta}_1 = \hat{\theta}_1 / \hat{\pi}_1$.

Using the "delta method" and assuming $\text{cov}(\hat{\theta}_1, \hat{\pi}_1) = 0$, what is $\text{var}(\hat{\beta}_1)$?

Define $t(\hat{\beta}_1)$ to be the t-ratio for a parameter. With the results from above, show that

$$t(\hat{\beta}_1)^2 = \frac{1}{\frac{1}{t(\hat{\theta}_1)^2} + \frac{1}{t(\hat{\pi}_1)^2}}$$

4. In Angrist, Imbens and Rubin, why is the monotonicity assumption important in the definition of the LATE? Consider the results in the Angrist and Evans paper on children and female labor supply. What is the LATE? Suppose that among families with two or more children, there are two types of families – those that prefer a mix of children and those that prefer single sex pairings – but those that prefer a sex mix outnumber those that prefer same sex child groups – which is why there is a first stage.

5. Following the notation in class, consider the RDD model with a structural equation of interest

$$y_i = \beta_0 + x_i\beta_1 + w_i\beta_2 + h^1(z_i) + \varepsilon_i$$

Where x is the treatment variable and w is a vector of covariates. The first-stage regression (how does treatment change at the discontinuity) is of the form

$$x_i = \pi_0 + D_i\pi_1 + w_i\pi_2 + h^2(z_i) + v_i .$$

Where $D_i = I(z_i \geq z_0)$ and π_1 is the coefficient of interest. Assume also that h^m is defined as $h^m(z_i) = \sum_{j=1}^{\rho}[D_i\delta_j^{m+}(z_i - z_0)^j + (1 - D_i)\delta_j^{m-}(z_i - z_0)^j]$. Show that if we DO NOT subtract off z_0 in the term $z_i - z_0$ we cannot treat π_1 as the impact of crossing the threshold on treatment. To make your life easy, assume linear terms in $z_i - z_0$ only.

Empirical Portion of the Problem Set

Women with children work less than women without kids. In a model where labor supply is regressed on the number of children in a household, the coefficient on the number of children is negative, large in magnitude, and statistically significant. This does not mean that the drop in work is actually caused by the presence of children in the house. To obtain a consistent estimate of the impact of kids on labor supply, some authors have suggested using whether a mother had twins on their first birth as an instrument for the number of children in the household. Twins are in many respect random and by definition, the realization of a twin increases the number of children in the household by 1. Using data from the 1980 Public Use Micro Sample 5% Census data files, I constructed a sample of women aged 21-40 with at least one kid. The 1980 PUMS identifies a person's age at the time of then census and their quarter of birth. Because the census is taken on April 1st, we know a person's year and quarter of birth and we can infer that any two kids in the household with the same age and quarter of birth are twins. There are roughly 6,000 1st births to mothers that are twins. There are over 800,000 observations in the original data set so to make the problem manageable, I select a random sample of about 6,500 non-twin births for a total of about 12,500 observations. The STATA data file data is called twins1sta.dta.

Variable name	Description
agem	Mother's current age in years
agefst	Mom's age when she first gave birth
race	1=white, 2=black, 3=other race
educm	Mother's years of education
married	Dummy variable for current marital statue, 1= married, 0=not
kids	Number of children ever born to the mother
boy1st	Dummy variable, =1 if first kid is a boy, =0 otherwise.
twin1st	Dummy variable, =1 if the first pregnancy ended in a twin birth
weeks	Weeks worked in previous year (from 0-52)
worked	Dummy variable, = 1 if the Mom worked at all in the previous year
lincome	Labor income earned in the previous year
mysteryz	Mystery instrument – will be used in part 8

1. What fraction of women work? What is average weeks worked among women that work? What is median labor earnings for women who worked?

2. Construct an indicator that equals 1 for women than have a second child. Call this variable SECOND. What fraction of women had a second child? Consider a simple bivariate regression where WEEKS (Y) is regressed on SECOND (X) such as $Y = \beta_0 + \beta_1 X_i + \varepsilon_i$. What is the coefficient for β_1 in this regression? Because of the concern that X and ε are correlated, use twins on 1st birth (Z) as an instrument for X in an instrumental variables model. What is the first-stage and reduced-form estimates for this model? Interpret these coefficients, that is, what do these coefficients measure? Consider the regression of X on Z. Why is the coefficient on Z not 1 - e..g, don't twins increase the number of kids in the house by 1? What is the Wald estimate for β_1 and compare the coefficient to the OLS estimate you produced above? Repeat this exercise using whether a mom worked at all as the outcome of interest. What is the R2 from the 1st stage regression of SECOND on TWIN1ST? What is the ratio of $Var(\hat{\beta}_1^{OLS}) / Var(\hat{\beta}_1^{2SLS})$?

3. A number of authors have used twins as an instrument for fertility in a number of different papers. The argument is that twins are "random" but the question is whether twins convey information about the mother. Construct three indicators for the mother's race. Run a series of regressions with 6 different outcomes (EDUC, AGEFST, AGEM, and whether the mother is

white, black, or some other race) on a single indicator: TWIN1ST. What coefficients are statistically significant? Are these differences economically meaningful, that is, are the coefficients large in magnitude? What do these results suggest about the "randomness" of twins on first birth?

4. Now that we know twins are correlated with some observed characteristics, run two structural labor supply models with weeks worked and whether a mom worked as outcomes and control for agem, age1st, educm, black, other race, and whether the mom had a second child. What is the impact of a second child on labor supply and weeks worked? Now, use twin1st as an instrument for SECOND in these models. Compare these estimates to the Wald estimates in b). What has happened to the labor supply impacts of having a second child?

5. The results in 3) suggest that twins might signal something about the mother that is correlated with labor supply, and as a result, the Wald estimates in 2) and the estimates 2SLS estimates in 4) may be more inconsistent than OLS estimates. Calculate the correlation coefficient between Z and X. Given this value, is this a concern?

6. Construct three dummy variables that indicate whether the mother's first birth was before age 20, between ages 20 and 24, or after age 24. Call these agegrp1, agegrp2 and agegrp3. Next, interact twin first with these three variables to construct three instruments. Call these twin1st1, twin1st2 and twin1st3. Estimate the 1st stage regression, add agefst1 and agefst2 to the model, take out twin1st, then add the three instruments agefst1 agefts2 agefts3. Using an F test, test two different hypotheses. The first is that the instruments are all the same value and the second being that the instruments are all equal to zero. Can you reject or not reject the null hypotheses in these cases? Is finite sample bias an issue in this case?

7. Using weeks worked and whether the mother worked as outcomes and the same covariates as in (4), the new covariates added in (6) as well as the instruments from (6) in a 2SLS model where SECOND is considered an endogenous variable. What has happened to the coefficient on SECOND in the WEEKS and WORKED equations? Do tests of over-identifying restrictions for these two models. What is the degrees of freedom on this test statistic? Can you accept or reject the null hypothesis that the model is correctly specified? Provide an intuitive explanation for this result -- you may have to estimate some extra models to answer this question.

8. a. (More difficult) Run an OLS using "weeks" as the dependent variable with the same covariates as in part 4) (second agem agefst educm black otherrace) and then 2SLS using mysterz as the instrument.

```
reg weeks second agem agefst educm black otherrace
reg weeks second agem agefst educm black otherrace (mystery agem
agefst educm black otherrace)
```

   b. Next, what you need to do is write a program that draws 5 instruments at random, each from a standard normal distribution, run the first stage (a regression of second on the exogenous variables, mysteryz and the 5 instruments), get the first-stage f-test that mysteryz and the 2SLS model using mysteryz and the five random instruments. Save the first-stage F and the 2SLS coefficients on second. Do this 1000 times.

   c. Redo part b) drawing 10 random, standard normal instruments.

   d. Redo part b) drawing 30 random, standard normal instruments.

   e. Redo part b) drawing 10 random, standard normal instruments but dropping mysterz as the

instrument.

Fill in the table below

Instruments	1st stage F (actual or average)	2SLS (actual or average)
Mysteryz		
1000 replications, mysterz and 5 random instruments		
1000 replications, mysterz and 10 random instruments		
1000 replications, mysterz and 30 random instruments		
1000 replications, 10 random instruments		