PPOL 503-03, PPOL503-04, Fall 2016
Course Notes #10: Instrumental Variables & Two Stage Least Squares

I. Overview

- Suppose that we have *observational* data and we wish to estimate the model $y = \beta_0 + \beta_1 x + u$. Recall that one of our main threats to internal validity is *omitted variable bias*. This leads to biased coefficient estimates because there is a correlation between the error term (u) and the independent variable of interest (x).

- If the omitted variables do not vary by time, then we can use panel data (if available) to difference them out. However, omitted variable bias persists if there are time-varying, unobservable determinants of the dependent variable that are correlated with the variable of interest.

- **General idea behind IV:** Find a variable, z, that is highly correlated with the independent variable, x, but that has no direct effect on the dependent variable, y.

    - We want to estimate the effect of x on y: $y = \beta_0 + \beta_1 x + u$.

    - We suspect that $cov(x,u) \neq 0$. This violates Guass-Markov assumption #4, a key condition needed for $\hat{\beta}_1$ to be unbiased and consistent

    - An instrument is a variable, z, that satisfies two conditions:

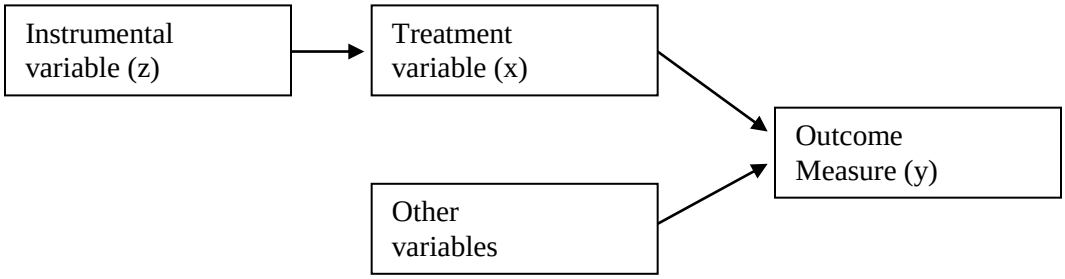
        a. ***Instrument Exogeneity:*** $Cov(z,u) = 0$ (i.e., the instrument is uncorrelated with the error term). This condition says that z is exogenous in our estimation equation (it has no effect on y other than through its effect on x). This is sometimes called the *exclusion restriction*.

        b. ***Instrument Relevance:*** $Cov(z,x) \neq 0$ (i.e., the instrument is correlated with the independent variable of interest). This condition, known as *the first-stage condition*, says that z must be related (positively or negatively) to x.

- **Note:** We can test the instrument relevance by regressing x against z: $x = \pi_0 + \pi_1 z + v$. Then $\pi_1 = Cov(z,x)/Var(z)$, and assumption (*a)* holds only if $\pi_1 \neq 0$. Thus, in order for our instrument to be valid (under condition *a*), we should be able to reject the null hypothesis H_0: $\pi_1 = 0$.

- **Note:** It is much harder to test for instrument exogeneity. This is frequently a matter of judgment.

- **Simplified graphical representation:**



- Key assumption: No arrow going from the instrumental variable to the outcome except through the treatment.
- **Example:**
  - Suppose we want to estimate the effect of education on wages. That is, we want to estimate the following wage equation:

    $\log(wage) = \beta_0 + \beta_1 educ + u$
  - What is the threat to internal validity?
  - How to assess the validity of an instrumental variable?
    - Conceptually: Need to think about whether it's reasonable to assume that a given z is uncorrelated with log(wage) except through variables that are included in the equation explaining log(wage).

- Empirically: Need to assess statistical significance of the instrument. Look at p-value of significance of instrument (t-test if one instrument; F-test if more than one instrument) in the first-stage regression.

- Are any of the following valid instruments for education?
  - The last digit of someone's social security number?
  - A measure of IQ, which is a proxy for ability?
  - Family background, such as mother's education?
  - The number of siblings?
  - Proximity to college (in the form of a dummy variable for whether someone grew up near a four-year college)?

II. IV ESTIMATION IN UNIVARIATE MODEL WITH ONE INSTRUMENTAL VARIABLE (2SLS Framework)

- Suppose that you would like to estimate β_1 from $y = \beta_0 + \beta_1 x + u$, using z as an instrument for x. (We call this the *structural equation*.)
- First stage: Run the regression: $x = \pi_0 + \pi_1 z + v$, and get predicted values for x (call these predicted values $\hat{x}$). We call this the *reduced form equation* (i.e., an equation that writes the endogenous variables in terms of the exogenous variables).
- Second stage: Run the following regression: $y = \alpha_0 + \alpha_1 \hat{x} + e$.
- β_1 represents the causal effect of *x* on *y*. Under conditions (a) and (b) above, $\hat{\alpha}_1$ is a **consistent** estimator of β_1. To repeat, these conditions are the following: (a) *z* is uncorrelated with *u*; (ii) The partial effect of *z* on *x*, is different from zero, i.e. $\pi_1 \neq 0$.
- In practice, packages like *Stata* implement IV with one command. If you do it using the two stages explicitly (as described above), you need to correct standard errors in the second stage regression.

III. BOOK EXAMPLES

- Example 15.1 (Estimating the Return to Education for Married Women)

  - OLS estimates:

$$\log(wage) = -.185 + .109 educ$$
$$\quad\quad\quad\quad (.185)\ (.014)$$
$$n = 428, R^2 = .118$$

  - Is father's education a good instrument? First-stage estimate:

$$\hat{educ} = 10.24 + .269\, fatheduc$$
$$\quad\quad\quad (.28)\ (.029)$$
$$n = 428, R^2 = .173$$

  - IV estimates:

$$\log(wage) = -.441 + .059 educ$$
$$\quad\quad\quad\quad (.446)\ (.093)$$
$$n = 428, R^2 = .118$$

  - The IV estimate of the return to education is approximately 5.9%, while the OLS estimate is approximately 11%. This suggests the OLS estimate might be too high, which is what one would expect due to omitted variable bias.

  - Note, however, that we cannot know for certain which of the two estimates is the true return to education. In fact, the standard errors of the IV estimation in this example are large, so the confidence interval contains the OLS estimate. That is, we cannot reject the null that the coefficient estimates are the same.

- Example 15.2 (Estimating the Return to Education for Men)

```
. regress lwage educ

      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  1,   933) =  100.70
       Model |  16.1377042     1  16.1377042           Prob > F      =  0.0000
    Residual |  149.518579   933  .160255712           R-squared     =  0.0974
-------------+------------------------------           Adj R-squared =  0.0964
       Total |  165.656283   934  .177362188           Root MSE      =  .40032

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0598392   .0059631    10.03   0.000     .0481366    .0715418
       _cons |   5.973063   .0813737    73.40   0.000     5.813366    6.132759
------------------------------------------------------------------------------
```

The OLS results indicate the return to education is 5.9%.

```
. regress educ sibs

      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  1,   933) =   56.67
       Model |  258.055048     1  258.055048           Prob > F      =  0.0000
    Residual |   4248.7642   933  4.55387374           R-squared     =  0.0573


-------------+------------------------------           Adj R-squared =  0.0562
       Total |  4506.81925   934  4.82528828           Root MSE      =   2.134

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sibs |  -.2279164   .0302768    -7.53   0.000     -.287335   -.1684979
       _cons |   14.13879   .1131382   124.97   0.000     13.91676    14.36083
------------------------------------------------------------------------------
```

This reduced form equation implies that every sibling is associated with, on average, about .23 less a year of schooling.

```
. ivreg lwage (educ=sibs)

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  1,   933) =   21.59
       Model | -1.51973315     1 -1.51973315           Prob > F      =  0.0000
    Residual |  167.176016   933  .179181154           R-squared     =       .
-------------+------------------------------           Adj R-squared =       .
       Total |  165.656283   934  .177362188           Root MSE      =   .4233

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1224326   .0263506     4.65   0.000     .0707194    .1741459
       _cons |   5.130026   .3551712    14.44   0.000     4.432999    5.827053
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   sibs
------------------------------------------------------------------------------
```

The IV estimates indicate the return to education is 12.2%.

IV. IV ESTIMATION IN MULTIVARIATE MODEL WITH ONE OR MORE INSTRUMENTAL VARIABLES (2SLS Framework)

- The general IV regression model has **four types of variables**: the *dependent variable*, Y; the *problematic endogenous explanatory variables* (which we label X) that are potentially correlated with the disturbance term; the *included exogenous explanatory variables* (which we label W), which are additional regressors that are not correlated with the disturbance term; and the *instrumental variables* (Z).

- In general, there can be multiple endogenous explanatory variables (X's), multiple included exogenous explanatory variables (W's), and multiple instrumental variables (Z's).

- The general IV regression model (structural equation) is as follows:

  - $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \beta_{k+1} W_1 + \ldots + \beta_{k+r} W_r + u$

- For IV regression to be possible, there must be at least as many instrumental variables (Z's) as endogenous explanatory variables (X's). The IV regression coefficients is said to be *exactly identified* if the number of instruments equals the number of endogenous explanatory variables. The coefficients are *over-identified* if the number of instruments exceeds the number of endogenous explanatory variables. The coefficients are *under-identified* if the number of instruments is less than the number of endogenous explanatory variables. In order to conduct an IV regression, the coefficients must be either exactly identified or over-identified.

- If there are multiple endogenous explanatory variables, each one requires a first-stage regression. That is, you need to separately regress each X against all the instruments (Z's) and all the included exogenous variables (W's). These first-stage regressions produce predicted values for each of the endogenous explanatory variables, which you then use to estimate the structural equation.

- More formally, the 2SLS estimator in the general IV regression model is computed as follows:

- **First-stage (reduced form) regression(s):** Regress X_1 on the instrumental variables (Z_1, … Z_m) and the included exogenous variables (W_1, …, W_r) using OLS. Compute the predicted values from this regression; call these $\hat{X}_1$. Repeat this for all the endogenous explanatory variables X_2, …, X_k, thereby computing the predicted values $\hat{X}_2, \text{K}, \hat{X}_k$.

- **Second-stage regression:** Regress Y on the predicted values of the endogenous variables ($\hat{X}_1, \text{K}, \hat{X}_k$) and the included exogenous variables (W_1, …, W_r) using OLS. The 2SLS estimators are the estimators from this second-stage regression. Given valid instruments, these estimators are biased but consistent.

- **Note:** The standard errors reported by OLS estimation of the second-stage regression are incorrect because they do not recognize that it is the second stage of a two-stage process. That is, the second-stage OLS standard errors fail to adjust for the fact that the second-stage regression uses the predicted values of the endogenous explanatory variables. Formulas for standard errors that make the necessary adjustment are incorporated into (and automatically used by) 2SLS regression commands in econometric software such as Stata.

- Conditions for Valid Instruments

  - **Order Condition:** We need at least as many instruments as endogenous explanatory variables.

  - **Instrument Relevance:** When there is one endogenous explanatory variable but multiple instruments, the condition is that at least one instrument is useful for predicting X, conditional on W. That is, if you regress X on the instruments (Z's) and on the exogenous explanatory variables (W's), then the coefficient estimates for the instruments must be jointly significant. When there are multiple endogenous variables, this

condition is more complicated because we must rule out perfect multicollinearity in the regression. That is the instruments must provide enough information about the exogenous movements in the endogenous explanatory variables to sort out their separate effects on Y.

- **Note:** For the instrument relevance condition, keep in mind that a more relevant instrument produces a more accurate IV estimator. It is important that the instruments should not just be relevant, but be highly relevant. The rule of thumb is that you want your F-statistic for the joint significance of you instruments to be at least 10.

- **Instrument Exogeneity:** The instruments must be uncorrelated with the disturbance term. That is $\text{cov}(Z_1, u) = 0$, … , $\text{cov}(Z_m, u) = 0$. If the instruments are not exogenous, 2SLS yields inconsistent coefficient estimates.

  - **Note:** If the coefficients are exactly identified, then it is impossible to develop a statistical test of the hypothesis that the instruments are in fact exogenous. In this case, the only way to assess whether the instruments are exogenous is to draw on expert opinion and your personal knowledge of the empirical problem at hand. Later we will examine an exogeneity test when the coefficients are over-identified.

- Testing for Endogeneity

  - If the explanatory variables are all exogenous, using 2SLS is less efficient than using OLS. Therefore, it is useful to test for endogeneity of an explanatory variable in order to determine whether 2SLS is even necessary.

  - Suppose we want to estimate the return to education for working women given by the following structural equation: $\log(\text{wage}) = \beta_0 + \beta_1\text{Educ} + \beta_2\text{Exper} + \beta_3\text{Exper}^2 + u$, where we suspect that Education might be endogenous. Our instruments

are mother's education and father's education. We can test whether education is endogenous as follows:

- Estimate the reduced form equation: $Educ = \pi_0 + \pi_1 Exper + \pi_2 Exper^2 + \pi_3 Motheduc + \pi_4 Fatheduc + v$. Obtain the residuals from this regression, called v_hat.

- Estimate the following equation: $\log(wage) = \beta_0 + \beta_1 Educ + \beta_2 Exper + \beta_3 Exper^2 + \delta v\_hat + u$. If the coefficient estimate δ is statistically different from zero, we conclude that education is indeed endogenous.

- The Over-identifying Restrictions Test

  - In cases where there are more instruments than endogenous explanatory variables, the over-identifying test can be helpful in testing whether the instruments are exogenous.

  - Suppose we have the following structural equation: $Y = \beta_0 + \beta_1 X + u$ and that we have two instrumental variables (Z_1 and Z_2) for the endogenous explanatory variable (X).

  - Because we have two instruments for one endogenous explanatory variable, we could compute two different 2SLS estimators, one using the first instrument, the other using the second instrument. For each of these, we can compute the residuals. If we assume Z_1 is exogenous, then the residuals from 2SLS using Z_1 should not be correlated with Z_2. If we find that Z_2 is correlated with these residuals, then this casts doubt on the exogeneity of Z_2. Similarly, if we assume that Z_2 is exogenous, then the residuals from 2SLS using Z_2 should not be correlated with Z_1. If we find that Z_1 is correlated with these residuals, then this casts doubt on the exogeneity of Z_1.

  - There's an easier way to conduct this test:

- Estimate the structural equation by 2SLS (using all instruments) and obtain the 2SLS residuals.
- Regress these residuals on *all* exogenous variables (including instruments). Obtain the R-squared from this regression.
- Under the null hypothesis that all IVs are uncorrelated with the disturbance term, $nR_1^2 \overset{a}{\sim} \chi_q^2$, where q is the number of instrumental variables from outside the model minus the total number of endogenous explanatory variables. If nR_1^2 exceeds (say) the 5% critical value in the χ_q^2 distribution, we reject the null hypothesis and conclude that at least some of the IVs are not exogenous.

- Example: We want to estimate the return to education for working women given by the following structural equation: log(wage) = β_0 + β_1Educ + β_2Exper + β_3Exper^2 + u, where we suspect that Education is endogenous. Suppose that we use father's education and mother's education as instruments:

```
. regress educ motheduc fatheduc exper expersq

      Source |       SS       df       MS              Number of obs =     753
-------------+------------------------------           F(  4,   748) =   66.52
       Model |  1025.94324     4   256.48581           Prob > F      =  0.0000
    Residual |   2884.0966   748  3.85574412           R-squared     =  0.2624
-------------+------------------------------           Adj R-squared =  0.2584
       Total |  3910.03984   752  5.19952106           Root MSE      =  1.9636

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    motheduc |   .1856173   .0259869     7.14   0.000     .1346014    .2366331
    fatheduc |   .1845745   .0244979     7.53   0.000     .1364817    .2326674
       exper |    .085378   .0255485     3.34   0.001     .0352228    .1355333
     expersq |  -.0018564   .0008276    -2.24   0.025    -.0034812   -.0002317
       _cons |   8.366716   .2667111    31.37   0.000     7.843125    8.890307
------------------------------------------------------------------------------

. test motheduc fatheduc

 ( 1)  motheduc = 0.0
 ( 2)  fatheduc = 0.0

       F(  2,   748) =  124.76
            Prob > F =    0.0000

Motheduc, fathedu and are strong predictor of a married woman's education

. predict v_hat, residuals
v_hat already defined
```

```
. ivreg lwage exper expersq (educ=motheduc fatheduc), robust

IV (2SLS) regression with robust standard errors       Number of obs =     428
                                                       F(  3,   424) =    6.15
                                                       Prob > F      =  0.0004
                                                       R-squared     =  0.1357
                                                       Root MSE      =  .67471

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0613966   .0333386     1.84   0.066    -.0041329    .1269261
       exper |   .0441704   .0155464     2.84   0.005     .0136128     .074728
     expersq |   -.000899   .0004301    -2.09   0.037    -.0017443   -.0000536
       _cons |   .0481003   .4297977     0.11   0.911    -.7966992    .8928998
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper expersq motheduc fatheduc
------------------------------------------------------------------------------

. predict u_hat, residual
(325 missing values generated)

. regress u_hat exper expersq motheduc fatheduc

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  4,   423) =    0.09
       Model |  .170502982     4  .042625746           Prob > F      =  0.9845
    Residual |  192.849518   423  .455909026           R-squared     =  0.0009
-------------+------------------------------           Adj R-squared = -0.0086
       Total |  193.020021   427   .45203752           Root MSE      =  .67521

------------------------------------------------------------------------------
       u_hat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |  -.0000183   .0133291    -0.00   0.999    -.0262179    .0261813
     expersq |   7.34e-07   .0003985     0.00   0.999    -.0007825     .000784
    motheduc |  -.0066065   .0118864    -0.56   0.579    -.0299704    .0167573
    fatheduc |   .0057823   .0111786     0.52   0.605    -.0161902    .0277547
       _cons |   .0109641   .1412571     0.08   0.938    -.2666892    .2886173
------------------------------------------------------------------------------
```

Now let's see what happens if we add husband's education to the IV list.

```
. regress educ motheduc fatheduc huseduc exper expersq

      Source |       SS       df       MS              Number of obs =     753
-------------+------------------------------           F(  5,   747) =  130.16
       Model |  1820.49038     5  364.098077           Prob > F      =  0.0000
    Residual |  2089.54946   747  2.79725496           R-squared     =  0.4656
-------------+------------------------------           Adj R-squared =  0.4620
       Total |  3910.03984   752  5.19952106           Root MSE      =  1.6725

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    motheduc |    .130004   .0223789     5.81   0.000      .086071    .1739371
    fatheduc |   .1013613   .0214423     4.73   0.000      .059267    .1434556
     huseduc |   .3715645   .0220465    16.85   0.000     .3282839     .414845
       exper |   .0532406   .0218443     2.44   0.015     .0103571    .0961241
     expersq |  -.0007403    .000708    -1.05   0.296    -.0021303    .0006497
       _cons |   5.115778    .298017    17.17   0.000     4.530727    5.700828
```

```
------------------------------------------------------------------------------

. test motheduc fatheduc huseduc

 ( 1)  motheduc = 0.0
 ( 2)  fatheduc = 0.0
 ( 3)  huseduc = 0.0

       F(  3,   747) =  209.33
            Prob > F =    0.0000
```

Motheduc, fathedu and huseduc are strong predictor of a married woman's education

```
. ivreg lwage exper expersq (educ=motheduc fatheduc huseduc), robust

IV (2SLS) regression with robust standard errors       Number of obs =     428
                                                       F(  3,   424) =    9.19
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1495
                                                       Root MSE      =   .6693

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0803918   .0217033     3.70   0.000     .0377323    .1230512
       exper |   .0430973   .0153064     2.82   0.005     .0130114    .0731832
     expersq |  -.0008628   .0004217    -2.05   0.041    -.0016916    -.000034
       _cons |  -.1868574   .3012625    -0.62   0.535    -.7790113    .4052966
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper expersq motheduc fatheduc huseduc
------------------------------------------------------------------------------

. predict u_hat2, residuals
(325 missing values generated)


. regress u_hat2 exper expersq motheduc fatheduc huseduc

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  5,   422) =    0.22
       Model |  .494825844     5  .098965169           Prob > F      =  0.9537
    Residual |  189.439884   422  .448909678           R-squared     =  0.0026
-------------+------------------------------           Adj R-squared = -0.0092
       Total |   189.93471   427  .444811967           Root MSE      =  .67001

------------------------------------------------------------------------------
      u_hat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |    .000056   .0132285     0.00   0.997     -.025946     .026058
     expersq |  -8.88e-06   .0003956    -0.02   0.982    -.0007865    .0007687
    motheduc |  -.0103852   .0118688    -0.87   0.382    -.0337145    .0129442
    fatheduc |   .0006734   .0113798     0.06   0.953    -.0216948    .0230417
     huseduc |   .0067811   .0114259     0.59   0.553    -.0156776    .0292398
       _cons |   .0086063   .1772724     0.05   0.961    -.3398405    .3570531
------------------------------------------------------------------------------
```

Regressing the residuals from the log wage equation on the instruments and other exogenous variables shows that none are significant predictors.

V. PROPERTIES OF IV WITH A POOR INSTRUMENTAL VARIABLE

- IV estimates are consistent when a) *z* and *u* are uncorrelated and b) *z* and *x* are correlated. However, the estimates will have large standard errors if the first-stage correlation is weak.

- Even worse is that if the first-state correlation is weak, the IV estimator can have a large asymptotic bias (even when the sample is large) even if there is only a weak correlation between the instrument and the error term in the original equation. (Bound, Jaeger, and Baker 1995 JASA paper.)

- The probability limit (plim) of the IV estimator when *z* and *u* are positively correlated is as follows:

  - $\text{plim}\hat{\beta}_1 = \beta_1 + (\frac{Corr(z,u)}{Corr(z,x)} \times \frac{\sigma_u^2}{\sigma_x^2})$, where σ_u and σ_x are the standard deviations of *u* and *x* in the population.

- Thus, even if Corr(z,u) is small, the inconsistency of the IV estimator may be large if Corr(z,x) is also small.

- The probability limit (plim) of the OLS estimator is as follows:

  - $\text{plim}\tilde{\beta}_1 = \beta_1 + (Corr(x,u) \times \frac{\sigma_u^2}{\sigma_x^2})$

- Therefore, OLS is preferred to IV if Corr(z,u)/Corr(z,x)>Corr(x,u).

- As a rule of thumb, the instrumental variables in the first-stage regression should be declared weak if the F-statistic is less than ten. If the first-stage relationship is weak, it is better to use OLS rather than IV.

VI. EXAMPLES OF STUDIES THAT USE INSTRUMENTAL VARIABLES

- To estimate the effect of years of schooling on wages, Angrist and Krueger used birth quarter as an instrument (as earlier birth quarter allows you to drop out of school earlier).
- To estimate the effect of Catholic schools, researchers have used distance to Catholic schools as an instrument.
- To estimate the effect of family income on savings in a developing country, Paxson used rainfall as an instrument.
- To estimate the effect of class size on test scores, Hoxby used the timings of births in a school district (deviation of potential enrollment from long-term trends).
- To estimate the effect of cardiac catheterization on length of patient survival, McClellan, McNeil, and Newhouse used the difference between a patient's distance to a cardiac catheterization hospital and the distance to the nearest hospital of any sort.
- **Note:** The trick to finding a good instrumental variable is to look for some exogenous source of variation in x arising from what is, in effect, a random phenomenon that induces shifts x.