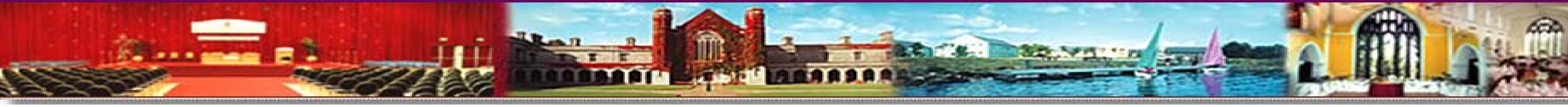
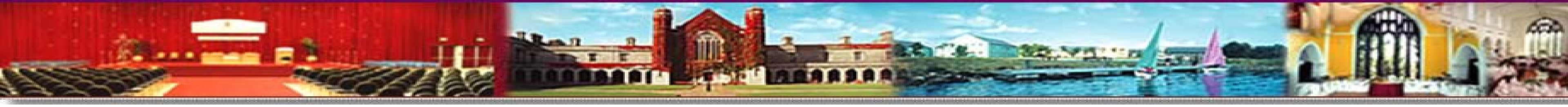


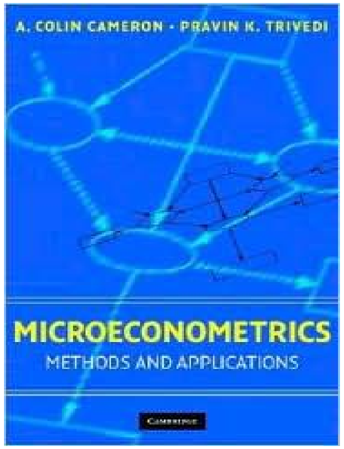
Topics in Microeconometrics

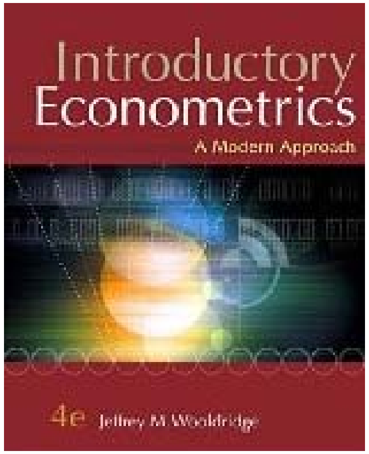
William Greene

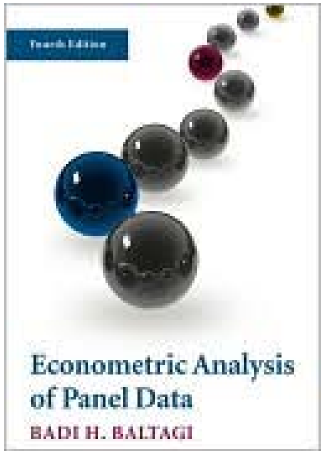
Department of Economics

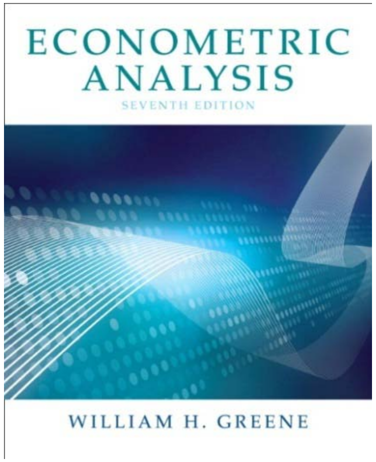
Stern School of Business



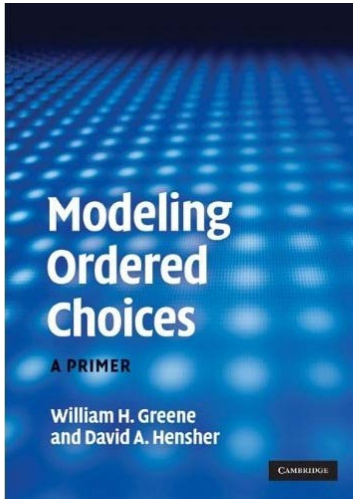


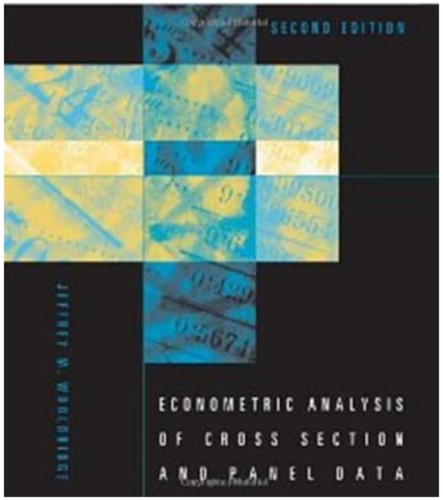


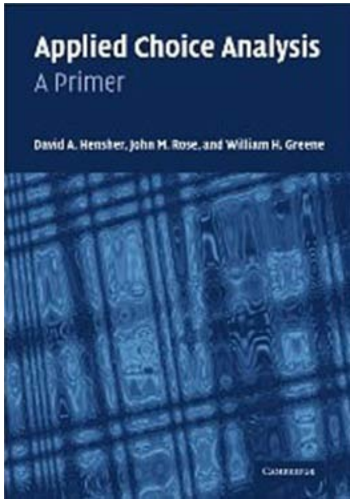


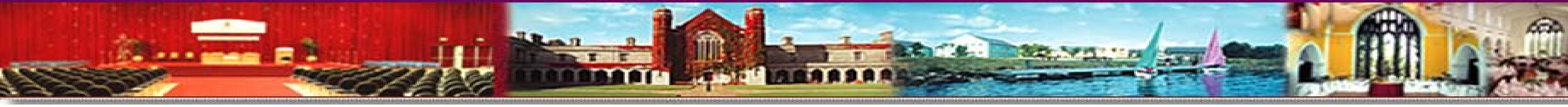


Part 2: Endogenous Variables in Linear Regression



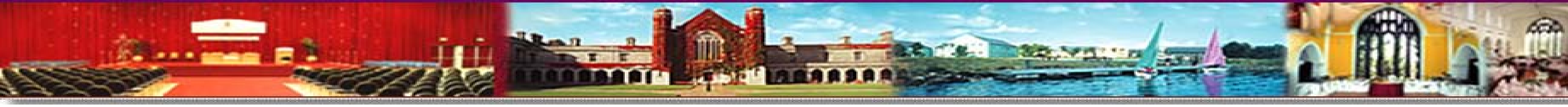






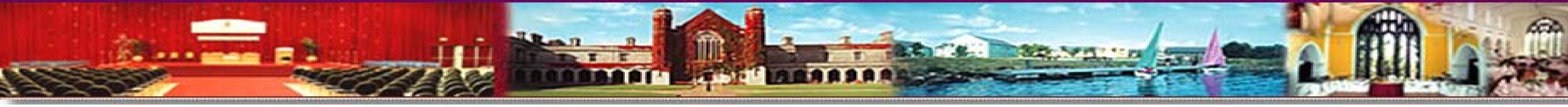
Endogeneity

- $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$,
- Definition: $E[\varepsilon|\mathbf{x}] \neq 0$
- Why not?
  - Omitted variables
  - Unobserved heterogeneity (equivalent to omitted variables)
  - Measurement error on the RHS (equivalent to omitted variables)
  - Structural aspects of the model
  - Endogenous sampling and attrition
  - Simultaneity (?)



Instrumental Variable Estimation

- One “problem” variable – the “last” one
- $y_{it} = \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_K x_{Kit} + \varepsilon_{it}$
- $E[\varepsilon_{it}|x_{Kit}] \neq 0$. (0 for all others)
- There exists a variable z_{it} such that
  - $E[x_{Kit}|\ x_{1it}, x_{2it},\ldots, x_{K-1,it}, z_{it}] = g(x_{1it}, x_{2it},\ldots, x_{K-1,it}, z_{it})$
    In the presence of the other variables, z_{it} “explains” x_{it}
  - $E[\varepsilon_{it}|\ x_{1it}, x_{2it},\ldots, x_{K-1,it}, z_{it}] = 0$
    In the presence of the other variables, z_{it} and ε_{it} are uncorrelated.
- A projection interpretation: In the projection
  $X_{Kt} = \theta_1 x_{1it}, + \theta_2 x_{2it} + \ldots + \theta_{k-1} x_{K-1,it} + \theta_K z_{it},$
  $\theta_K \neq 0$.

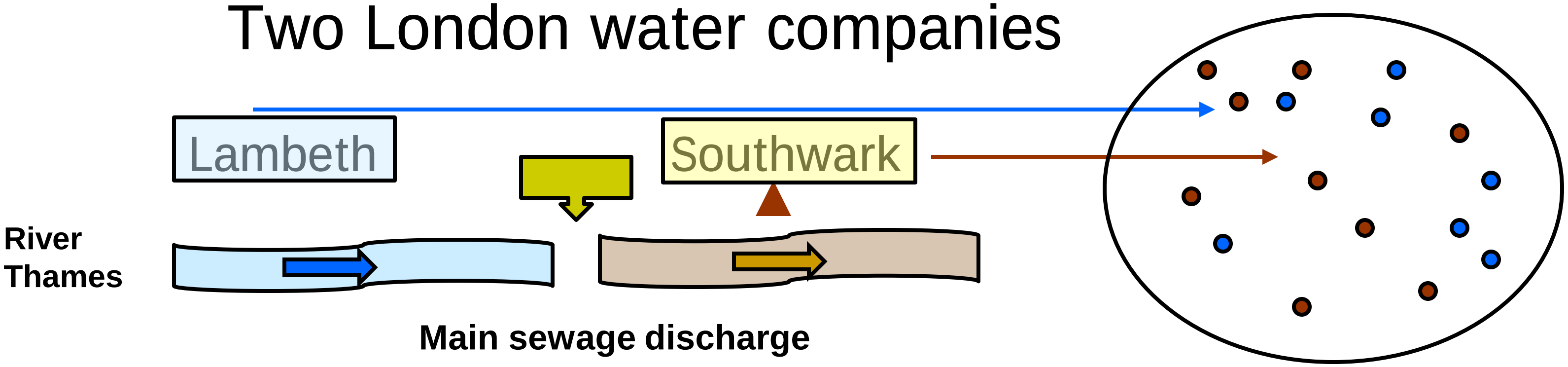


The First IV Study: Natural Experiment

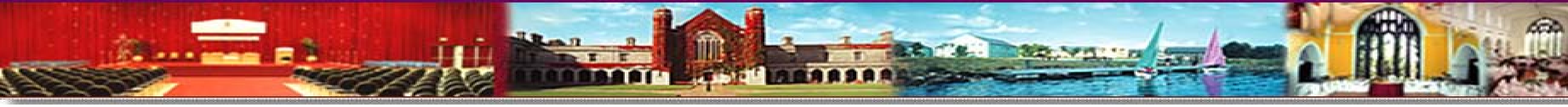
(Snow, J., On the Mode of Communication of Cholera, 1855) http://www.ph.ucla.edu/epi/snow/snowbook3.html

- London Cholera epidemic, ca 1853-4
- Cholera = f(Water Purity,u)+ε.
  - ‘Causal’ effect of water purity on cholera?
  - Purity=f(cholera prone environment (poor, garbage in streets, rodents, etc.). Regression does not work.

Two London water companies

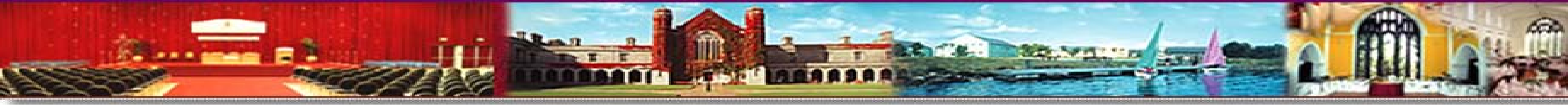


Paul Grootendorst: A Review of Instrumental Variables Estimation of Treatment Effects…
http://individual.utoronto.ca/grootendorst/pdf/IV_Paper_Sept6_2007.pdf



IV Estimation

- $\text{Cholera}=f(\text{Purity},u)+\varepsilon$
- Z = water company
- $\text{Cov}(\text{Cholera},Z)=\delta\text{Cov}(\text{Purity},Z)$
- Z is randomly mixed in the population (two full sets of pipes) and uncorrelated with behavioral unobservables, u)
- $\text{Cholera}=\alpha+\delta\text{Purity}+u+\varepsilon$
  - Purity = Mean+random variation+λu
  - $\text{Cov}(\text{Cholera},Z)= \delta\text{Cov}(\text{Purity},Z)$

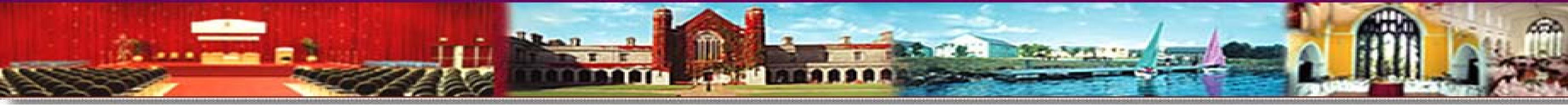


Cornwell and Rupert Data

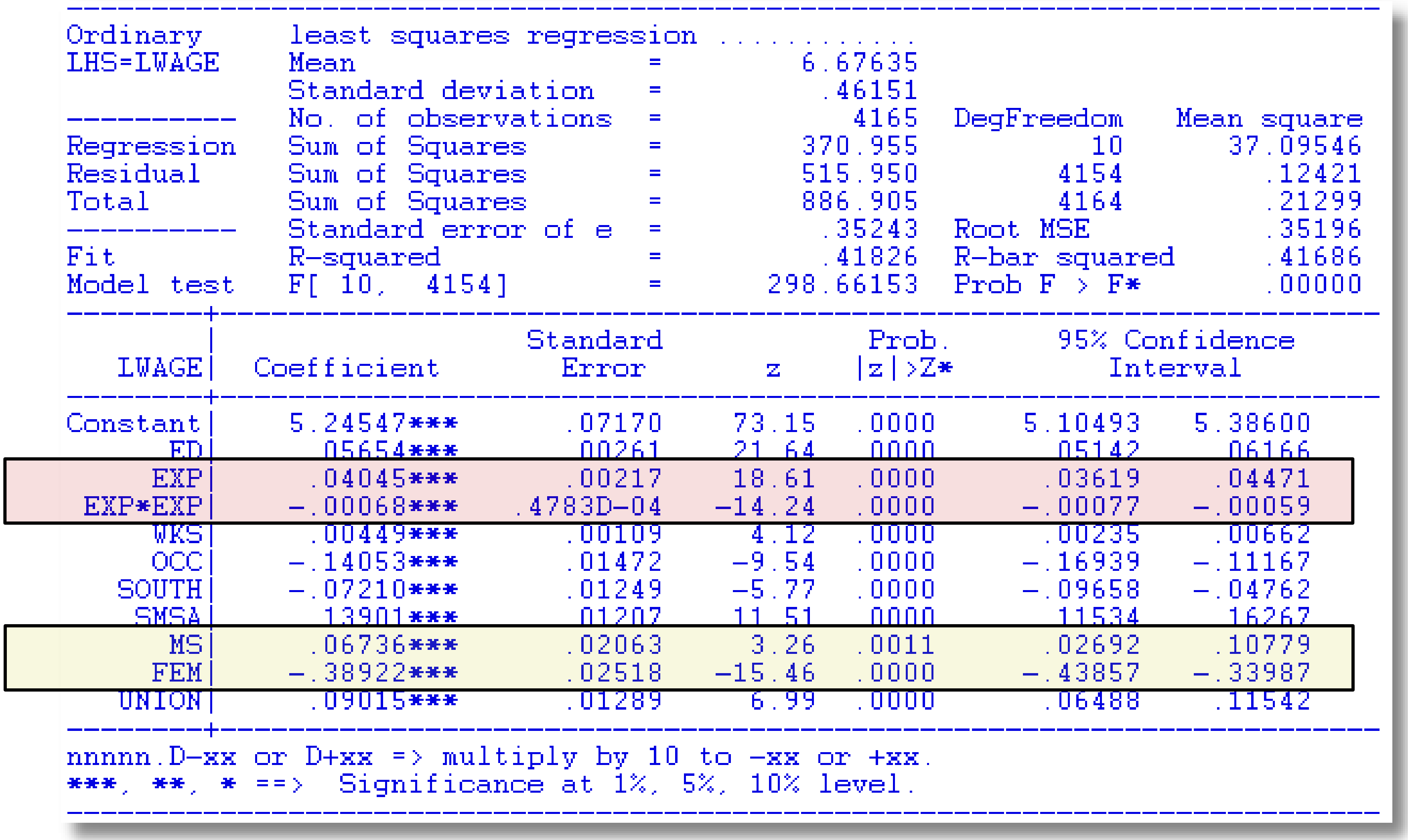
Cornwell and Rupert Returns to Schooling Data, 595 Individuals, 7 Years
Variables in the file are

EXP = work experience
WKS = weeks worked
OCC = occupation, 1 if blue collar,
IND = 1 if manufacturing industry
SOUTH = 1 if resides in south
SMSA = 1 if resides in a city (SMSA)
MS = 1 if married
FEM = 1 if female
UNION = 1 if wage set by union contract
ED = years of education
LWAGE = log of wage = dependent variable in regressions

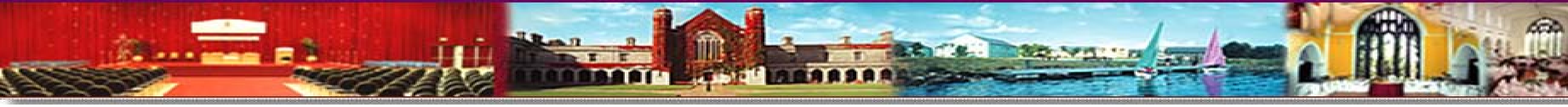
These data were analyzed in Cornwell, C. and Rupert, P., "Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variable Estimators," Journal of Applied Econometrics, 3, 1988, pp. 149-155. See Baltagi, page 122 for further analysis. The data were downloaded from the website for Baltagi's text.



Specification: Quadratic Effect of Experience



```
--------------------------------------------------------------------------------
Ordinary     least squares regression ............
LHS=LWAGE    Mean                   =       6.67635
             Standard deviation     =        .46151
----------   No. of observations    =          4165  DegFreedom   Mean square
Regression   Sum of Squares         =       370.955          10      37.09546
Residual     Sum of Squares         =       515.950        4154        .12421
Total        Sum of Squares         =       886.905        4164        .21299
----------   Standard error of e    =        .35243  Root MSE          .35196
Fit          R-squared              =        .41826  R-bar squared     .41686
Model test   F[ 10,  4154]          =     298.66153  Prob F > F*       .00000
--------+-----------------------------------------------------------------------
        |                  Standard            Prob.      95% Confidence
   LWAGE|  Coefficient       Error       z    |z|>Z*         Interval
--------+-----------------------------------------------------------------------
Constant|    5.24547***      .07170    73.15  .0000      5.10493    5.38600
      ED|     .05654***      .00261    21.64  .0000       .05142     .06166
     EXP|     .04045***      .00217    18.61  .0000       .03619     .04471
 EXP*EXP|    -.00068***   .4783D-04   -14.24  .0000      -.00077    -.00059
     WKS|     .00449***      .00109     4.12  .0000       .00235     .00662
     OCC|    -.14053***      .01472    -9.54  .0000      -.16939    -.11167
   SOUTH|    -.07210***      .01249    -5.77  .0000      -.09658    -.04762
    SMSA|     .13901***      .01207    11.51  .0000       .11534     .16267
      MS|     .06736***      .02063     3.26  .0011       .02692     .10779
     FEM|    -.38922***      .02518   -15.46  .0000      -.43857    -.33987
   UNION|     .09015***      .01289     6.99  .0000       .06488     .11542
--------+-----------------------------------------------------------------------
nnnnn.D-xx or D+xx => multiply by 10 to -xx or +xx.
***, **, * ==>  Significance at 1%, 5%, 10% level.
--------------------------------------------------------------------------------
```

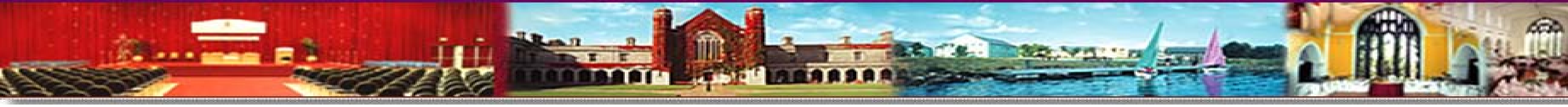


The Effect of Education on LWAGE

$$\mathbf{LWAGE} = \beta_1 + \beta_2\mathbf{EDUC} + \beta_3\mathbf{EXP} + \beta_4\mathbf{EXP}^2 + \ldots\varepsilon +$$

What is ε? Ability, Motivation,... + everything else

$$\mathbf{EDUC} = f(\mathbf{GENDER}, \mathbf{SMSA}, \mathbf{SOUTH}, \text{Ability}, \text{Motivation}, \ldots)$$

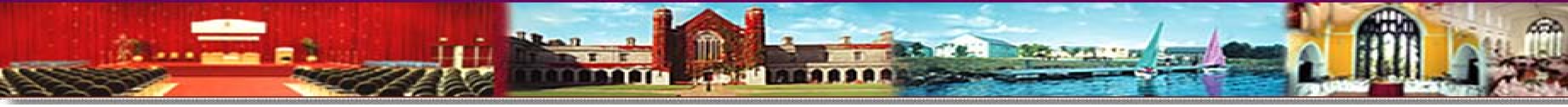


What Influences LWAGE?

$$\textbf{LWAGE} = \beta_1 + \beta_2\textbf{EDUC}(\textbf{X}, \text{Ability}, \text{Motivation}, \ldots)$$
$$+ \beta_3\textbf{EXP} + \beta_4\textbf{EXP}^2 + \ldots$$
$$\varepsilon(\text{Ability}, \text{Motivation})$$

Increased Ability is associated with increases in $\textbf{EDUC}(\textbf{X}, \text{Ability}, \text{Motivation}, \ldots)$ and $\varepsilon(\text{Ability}, \text{Motivation})$

What looks like an effect due to increase in **EDUC** may be an increase in Ability. The estimate of β_2 picks up the effect of **EDUC** and the hidden effect of Ability.



An Exogenous Influence

$$\mathbf{LWAGE} = \beta_1 + \beta_2\mathbf{EDUC}(\mathbf{X}, \mathbf{Z}, \text{Ability}, \text{Motivation}, ...)$$

$$+ \beta_3\mathbf{EXP} + \beta_4\mathbf{EXP}^2 + ...$$

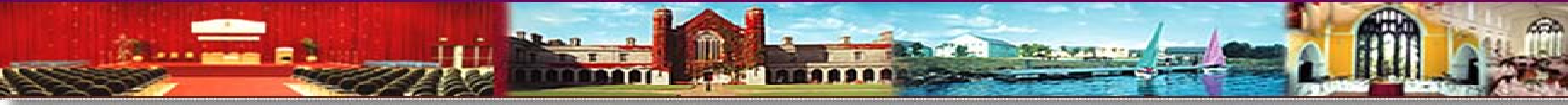
$$\varepsilon(\quad \text{Ability}, \text{Motivation})$$

Increased **Z** is associated with increases in

$\mathbf{EDUC}(\mathbf{X}, \mathbf{Z}, \text{Ability}, \text{Motivation}, ...)$ and not $\varepsilon(\text{Ability}, \text{Motivation})$

An effect due to the effect of an increase **Z** on **EDUC** will only be an increase in **EDUC**. The estimate of β_2 picks up the effect of **EDUC** only.

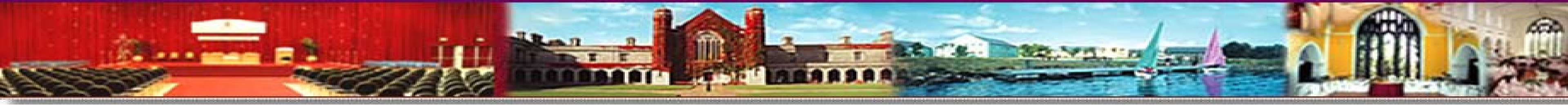
Z is an Instrumental Variable



Instrumental Variables

- Structure
  - LWAGE (**ED,EXP,EXPSQ,WKS,OCC,**
    SOUTH,SMSA,UNION)
  - ED (**MS**, **FEM**)

- Reduced Form:
  LWAGE[**ED** (**MS**, **FEM**),
  EXP,EXPSQ,WKS,OCC,
  SOUTH,SMSA,UNION]

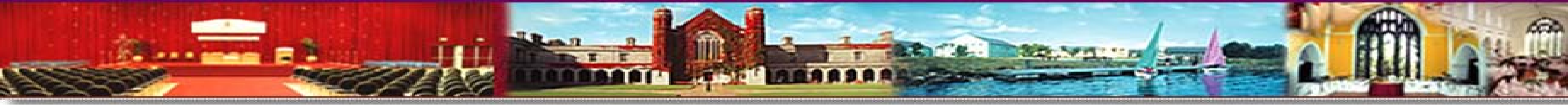


Two Stage Least Squares Strategy

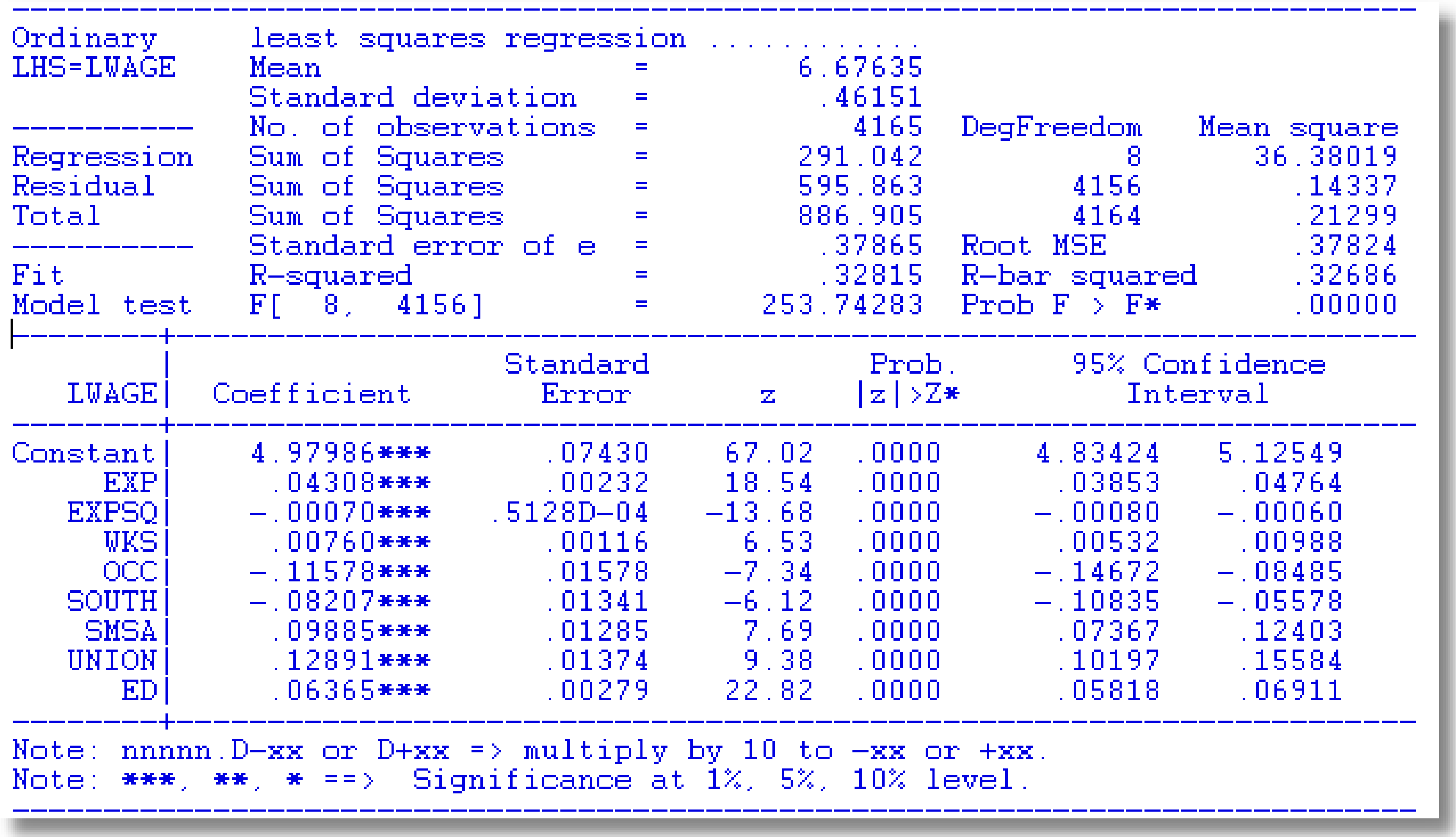
- Reduced Form:

  LWAGE[**ED** (**MS**, **FEM**,**X**),
  EXP,EXPSQ,WKS,OCC,
  SOUTH,SMSA,UNION]

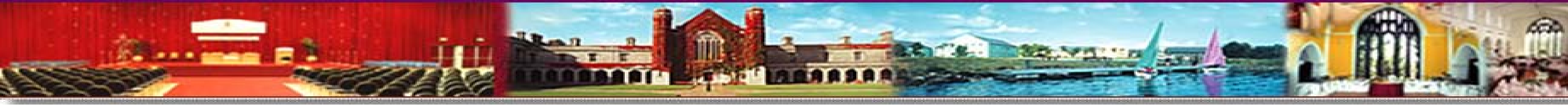
- Strategy
  - (1) Purge ED of the influence of everything but MS, FEM (and the other variables). Predict ED using all exogenous information in the sample (**X** and **Z**).
  - (2) Regress LWAGE on this prediction of ED and everything else.
  - Standard errors must be adjusted for the predicted ED

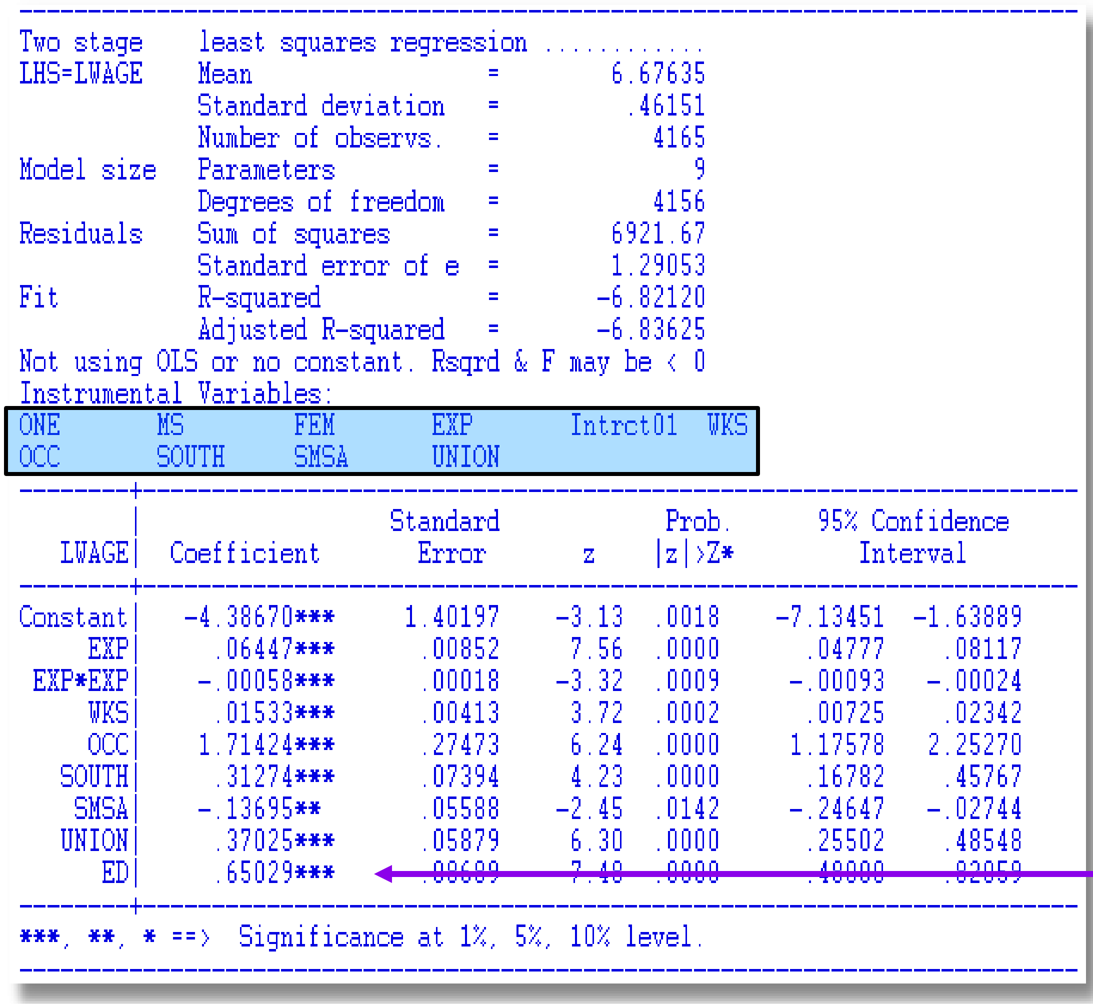


OLS



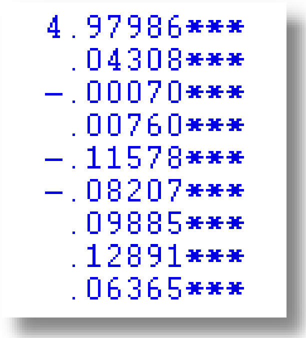
```
----------------------------------------------------------------------------
Ordinary    least squares regression ............
LHS=LWAGE   Mean                 =        6.67635
            Standard deviation   =         .46151
----------  No. of observations  =           4165  DegFreedom   Mean square
Regression  Sum of Squares       =        291.042           8      36.38019
Residual    Sum of Squares       =        595.863        4156        .14337
Total       Sum of Squares       =        886.905        4164        .21299
----------  Standard error of e  =         .37865  Root MSE            .37824
Fit         R-squared            =         .32815  R-bar squared       .32686
Model test  F[  8,  4156]        =      253.74283  Prob F > F*         .00000
--------+-------------------------------------------------------------------
        |              Standard            Prob.      95% Confidence
   LWAGE|  Coefficient       Error       z    |z|>Z*         Interval
--------+-------------------------------------------------------------------
Constant|    4.97986***      .07430   67.02  .0000     4.83424    5.12549
     EXP|     .04308***      .00232   18.54  .0000      .03853     .04764
   EXPSQ|    -.00070***   .5128D-04  -13.68  .0000     -.00080    -.00060
     WKS|     .00760***      .00116    6.53  .0000      .00532     .00988
     OCC|    -.11578***      .01578   -7.34  .0000     -.14672    -.08485
   SOUTH|    -.08207***      .01341   -6.12  .0000     -.10835    -.05578
    SMSA|     .09885***      .01285    7.69  .0000      .07367     .12403
   UNION|     .12891***      .01374    9.38  .0000      .10197     .15584
      ED|     .06365***      .00279   22.82  .0000      .05818     .06911
--------+-------------------------------------------------------------------
Note: nnnnn.D-xx or D+xx => multiply by 10 to -xx or +xx.
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------------
```



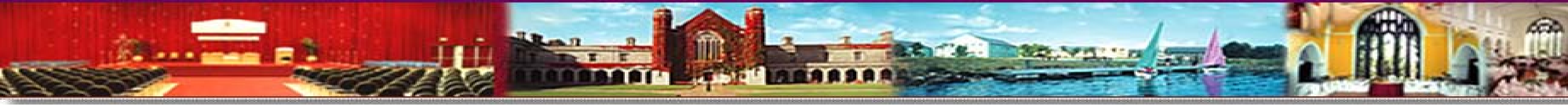


```
----------------------------------------------------------------------------
Two stage    least squares regression ............
LHS=LWAGE    Mean                  =       6.67635
             Standard deviation    =        .46151
             Number of observs.    =          4165
Model size   Parameters            =             9
             Degrees of freedom    =          4156
Residuals    Sum of squares        =       6921.67
             Standard error of e   =       1.29053
Fit          R-squared             =      -6.82120
             Adjusted R-squared    =      -6.83625
Not using OLS or no constant. Rsqrd & F may be < 0
Instrumental Variables:
ONE       MS        FEM       EXP       Intrct01  WKS
OCC       SOUTH     SMSA      UNION
--------+-------------------------------------------------------------------
        |                Standard            Prob.      95% Confidence
   LWAGE|  Coefficient     Error       z    |z|>Z*         Interval
--------+-------------------------------------------------------------------
Constant|   -4.38670***   1.40197    -3.13  .0018    -7.13451   -1.63889
     EXP|     .06447***    .00852     7.56  .0000      .04777     .08117
 EXP*EXP|    -.00058***    .00018    -3.32  .0009     -.00093    -.00024
     WKS|     .01533***    .00413     3.72  .0002      .00725     .02342
     OCC|    1.71424***    .27473     6.24  .0000     1.17578    2.25270
   SOUTH|     .31274***    .07394     4.23  .0000      .16782     .45767
    SMSA|    -.13695**     .05588    -2.45  .0142     -.24647    -.02744
   UNION|     .37025***    .05879     6.30  .0000      .25502     .48548
      ED|     .65029***    .08689     7.48  .0000      .48000     .82059
--------+-------------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------------
```

The weird results for the coefficient on ED happened because the instruments, MS and FEM are dummy variables. There is not enough variation in these variables.

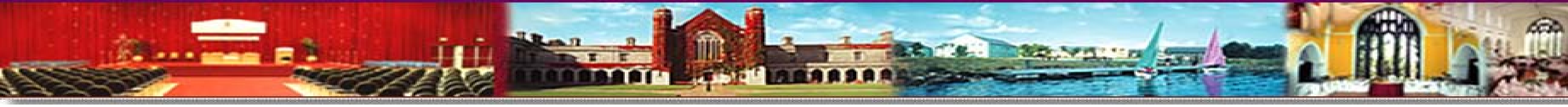


```
 4.97986***
  .04308***
 -.00070***
  .00760***
 -.11578***
 -.08207***
  .09885***
  .12891***
  .06365***
```



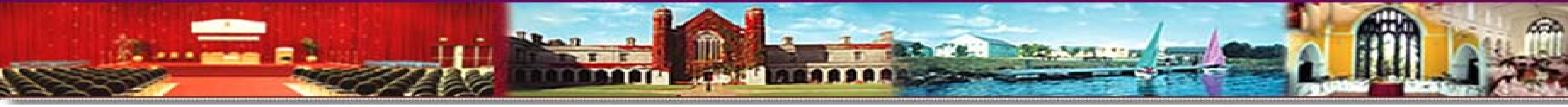
Source of Endogeneity

- **LWAGE** = f(**ED,**
  EXP,EXPSQ,WKS,OCC,
  SOUTH,SMSA,UNION) + ε
- **ED** = f(**MS,FEM,**
  EXP,EXPSQ,WKS,OCC,
  SOUTH,SMSA,UNION) + u

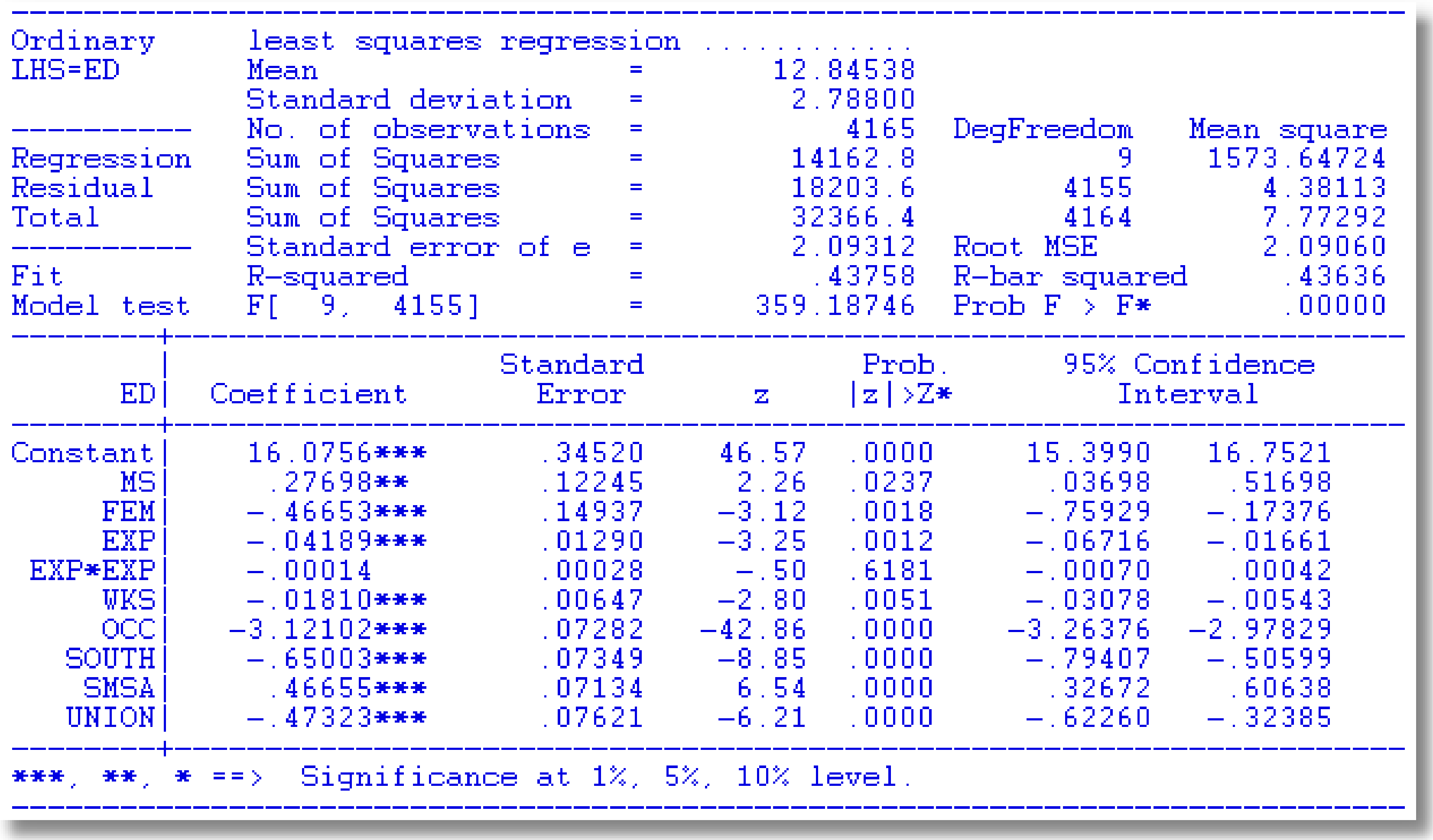


Remove the Endogeneity

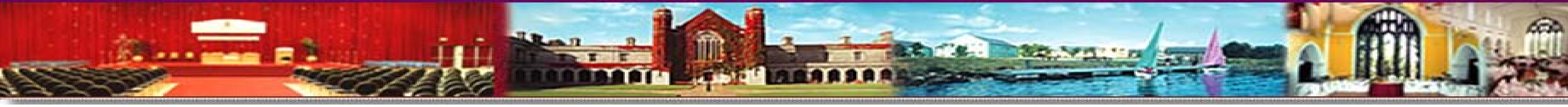
- **LWAGE** = f(**ED,** **EXP,EXPSQ,WKS,OCC,** **SOUTH,SMSA,UNION**) + $\boxed{u + \varepsilon}$
- **LWAGE** = f(**ED,** **EXP,EXPSQ,WKS,OCC,** **SOUTH,SMSA,UNION**) + u + ε
- Strategy
  - □ Estimate u
  - □ Add u to the equation. ED is uncorrelated with ε when u is in the equation.



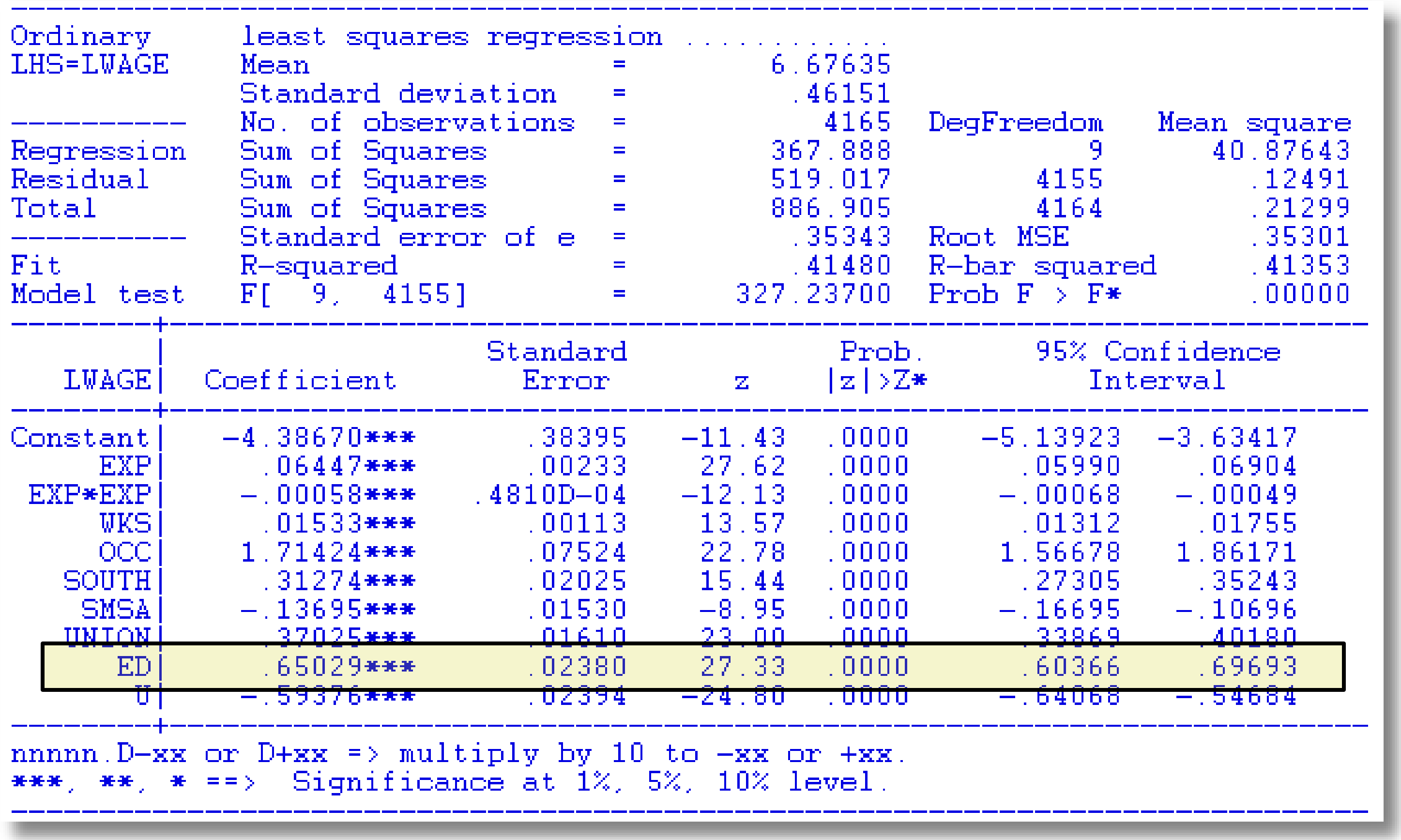
Auxiliary Regression for ED to Obtain Residuals



```
----------------------------------------------------------------------------
Ordinary    least squares regression ............
LHS=ED      Mean                 =       12.84538
            Standard deviation   =        2.78800
----------  No. of observations  =           4165  DegFreedom   Mean square
Regression  Sum of Squares       =        14162.8           9    1573.64724
Residual    Sum of Squares       =        18203.6        4155       4.38113
Total       Sum of Squares       =        32366.4        4164       7.77292
----------  Standard error of e  =        2.09312  Root MSE         2.09060
Fit         R-squared            =         .43758  R-bar squared     .43636
Model test  F[  9,  4155]        =      359.18746  Prob F > F*       .00000
--------+-------------------------------------------------------------------
        |                  Standard            Prob.       95% Confidence
      ED|  Coefficient       Error        z    |z|>Z*         Interval
--------+-------------------------------------------------------------------
Constant|    16.0756***     .34520    46.57   .0000     15.3990    16.7521
      MS|     .27698**      .12245     2.26   .0237      .03698     .51698
     FEM|    -.46653***     .14937    -3.12   .0018     -.75929    -.17376
     EXP|    -.04189***     .01290    -3.25   .0012     -.06716    -.01661
 EXP*EXP|    -.00014        .00028     -.50   .6181     -.00070     .00042
     WKS|    -.01810***     .00647    -2.80   .0051     -.03078    -.00543
     OCC|   -3.12102***     .07282   -42.86   .0000    -3.26376   -2.97829
   SOUTH|    -.65003***     .07349    -8.85   .0000     -.79407    -.50599
    SMSA|     .46655***     .07134     6.54   .0000      .32672     .60638
   UNION|    -.47323***     .07621    -6.21   .0000     -.62260    -.32385
--------+-------------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------------
```

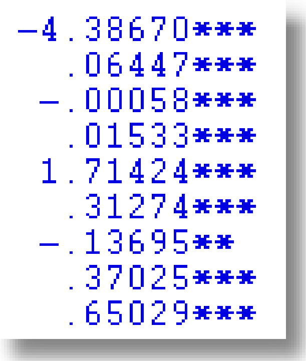


OLS with Residual (Control Function) Added

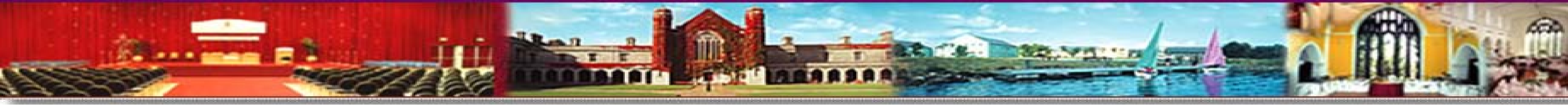


```
----------------------------------------------------------------------------
Ordinary    least squares regression ............
LHS=LWAGE   Mean                 =      6.67635
            Standard deviation   =       .46151
----------  No. of observations  =         4165  DegFreedom   Mean square
Regression  Sum of Squares       =      367.888           9     40.87643
Residual    Sum of Squares       =      519.017        4155       .12491
Total       Sum of Squares       =      886.905        4164       .21299
----------  Standard error of e   =       .35343  Root MSE         .35301
Fit         R-squared            =       .41480  R-bar squared    .41353
Model test  F[  9,  4155]        =    327.23700  Prob F > F*      .00000
--------+-------------------------------------------------------------------
        |                 Standard            Prob.      95% Confidence
   LWAGE|  Coefficient      Error       z    |z|>Z*         Interval
--------+-------------------------------------------------------------------
Constant|   -4.38670***     .38395   -11.43  .0000    -5.13923  -3.63417
     EXP|     .06447***     .00233    27.62  .0000      .05990    .06904
 EXP*EXP|    -.00058***   .4810D-04  -12.13  .0000     -.00068   -.00049
     WKS|     .01533***     .00113    13.57  .0000      .01312    .01755
     OCC|    1.71424***     .07524    22.78  .0000     1.56678   1.86171
   SOUTH|     .31274***     .02025    15.44  .0000      .27305    .35243
    SMSA|    -.13695***     .01530    -8.95  .0000     -.16695   -.10696
   UNION|     .37025***     .01610    23.00  .0000      .33869    .40180
      ED|     .65029***     .02380    27.33  .0000      .60366    .69693
       U|    -.59376***     .02394   -24.80  .0000     -.64068   -.54684
--------+-------------------------------------------------------------------
nnnnn.D-xx or D+xx => multiply by 10 to -xx or +xx.
***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------------
```

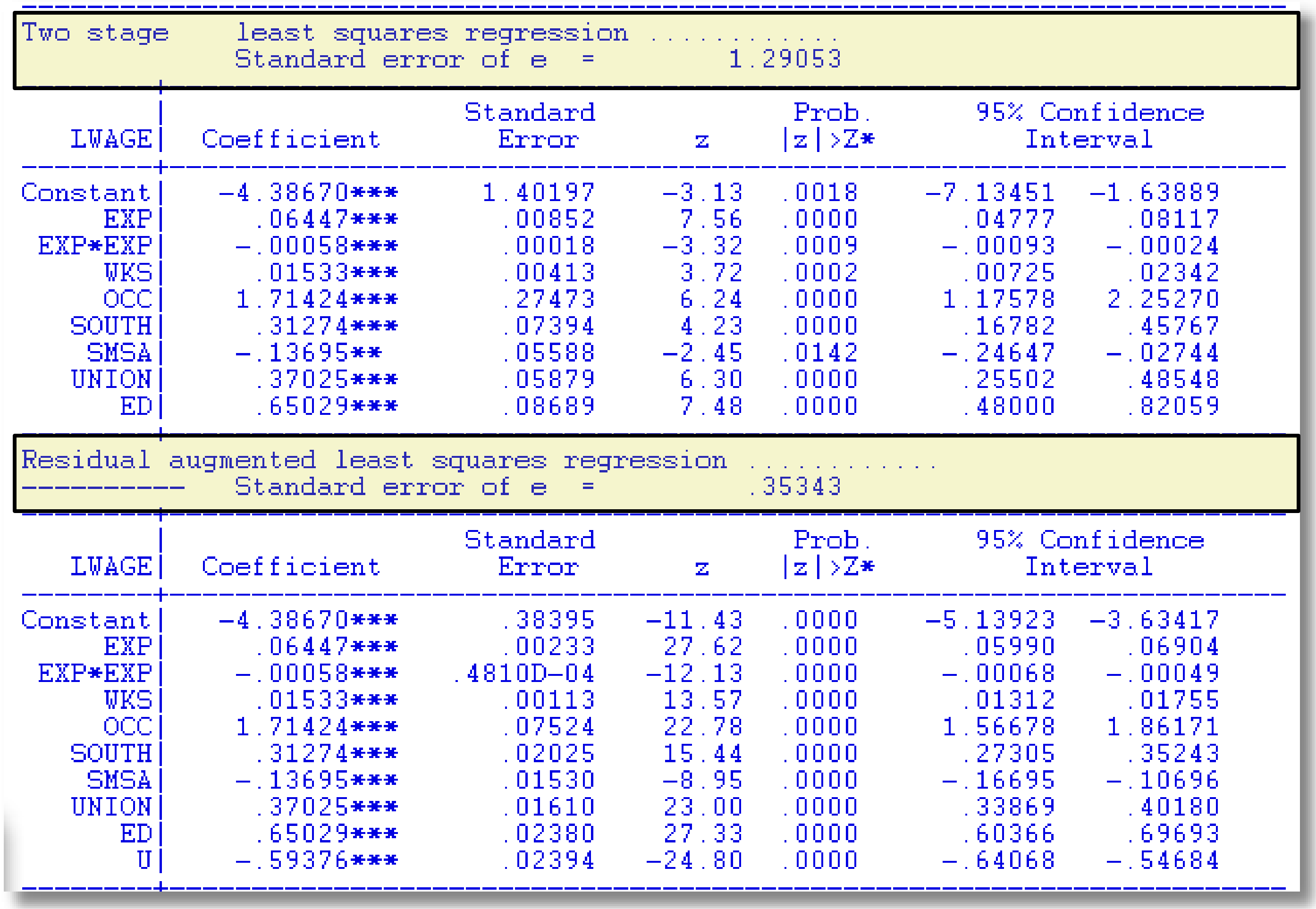
2SLS



```
-4.38670***
  .06447***
 -.00058***
  .01533***
 1.71424***
  .31274***
 -.13695**
  .37025***
  .65029***
```

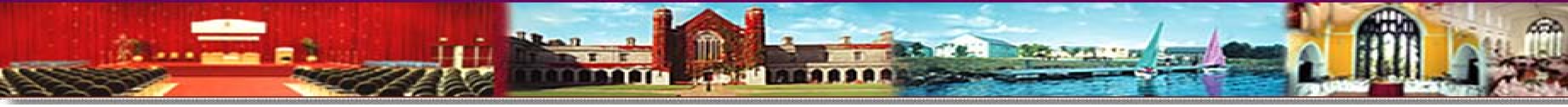


A Warning About Control Functions



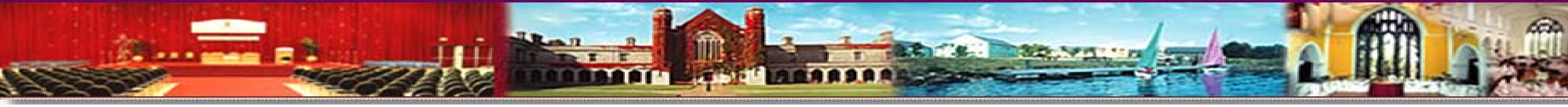
```
-----------------------------------------------------------------------------
Two stage    least squares regression .............
             Standard error of e  =         1.29053
--------+--------------------------------------------------------------------
        |                Standard             Prob.       95% Confidence
   LWAGE|  Coefficient     Error        z    |z|>Z*          Interval
--------+--------------------------------------------------------------------
Constant|   -4.38670***   1.40197     -3.13  .0018     -7.13451  -1.63889
     EXP|     .06447***    .00852      7.56  .0000       .04777    .08117
 EXP*EXP|    -.00058***    .00018     -3.32  .0009      -.00093   -.00024
     WKS|     .01533***    .00413      3.72  .0002       .00725    .02342
     OCC|    1.71424***    .27473      6.24  .0000      1.17578   2.25270
   SOUTH|     .31274***    .07394      4.23  .0000       .16782    .45767
    SMSA|    -.13695**     .05588     -2.45  .0142      -.24647   -.02744
   UNION|     .37025***    .05879      6.30  .0000       .25502    .48548
      ED|     .65029***    .08689      7.48  .0000       .48000    .82059
--------+--------------------------------------------------------------------
Residual augmented least squares regression ............
----------   Standard error of e  =          .35343
-----------------------------------------------------------------------------
        |                Standard             Prob.       95% Confidence
   LWAGE|  Coefficient     Error        z    |z|>Z*          Interval
--------+--------------------------------------------------------------------
Constant|   -4.38670***    .38395    -11.43  .0000     -5.13923  -3.63417
     EXP|     .06447***    .00233     27.62  .0000       .05990    .06904
 EXP*EXP|    -.00058***  .4810D-04   -12.13  .0000      -.00068   -.00049
     WKS|     .01533***    .00113     13.57  .0000       .01312    .01755
     OCC|    1.71424***    .07524     22.78  .0000      1.56678   1.86171
   SOUTH|     .31274***    .02025     15.44  .0000       .27305    .35243
    SMSA|    -.13695***    .01530     -8.95  .0000      -.16695   -.10696
   UNION|     .37025***    .01610     23.00  .0000       .33869    .40180
      ED|     .65029***    .02380     27.33  .0000       .60366    .69693
       U|    -.59376***    .02394    -24.80  .0000      -.64068   -.54684
--------+--------------------------------------------------------------------
```

Sum of squares is not computed correctly because U is in the regression. A general result. Control function estimators usually require a fix to the estimated covariance matrix for the estimator.



Endogenous Dummy Variable

- $Y = x\beta + \delta T + \varepsilon$ (unobservable factors)
- T = a dummy variable (treatment)
- T = 0/1 depending on:
  - x and z
  - The same unobservable factors
- T is endogenous – same as ED



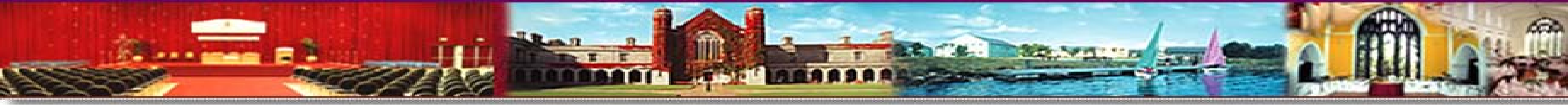
Application: Health Care Panel Data

German Health Care Usage Data, 7,293 Individuals, Varying Numbers of Periods

Variables in the file are

Data downloaded from Journal of Applied Econometrics Archive. This is an unbalanced panel with 7,293 individuals. They can be used for regression, count models, binary choice, ordered choice, and bivariate binary choice. **This is a large data set. There are altogether 27,326 observations. The number of observations ranges from 1 to 7. (Frequencies are: 1=1525, 2=2158, 3=825, 4=926, 5=1051, 6=1000, 7=987).** Note, the variable NUMOBS below tells how many observations there are for each person. This variable is repeated in each row of the data for the person. (Downloaded from the JAE Archive)

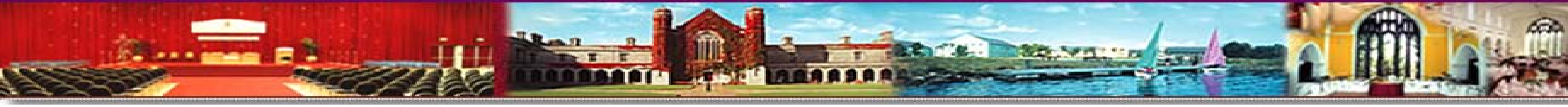
➡ **DOCTOR = 1(Number of doctor visits > 0)**
HOSPITAL = 1(Number of hospital visits > 0)
HSAT = health satisfaction, coded 0 (low) - 10 (high)
DOCVIS = number of doctor visits in last three months
HOSPVIS = number of hospital visits in last calendar year
➡ **PUBLIC = insured in public health insurance = 1; otherwise = 0**
➡ **ADDON = insured by add-on insurance = 1; otherswise = 0**
HHNINC = household nominal monthly net income in German marks / 10000.
(4 observations with income=0 were dropped)
HHKIDS = children under age 16 in the household = 1; otherwise = 0
EDUC = years of schooling
AGE = age in years
MARRIED = marital status
EDUC = years of education



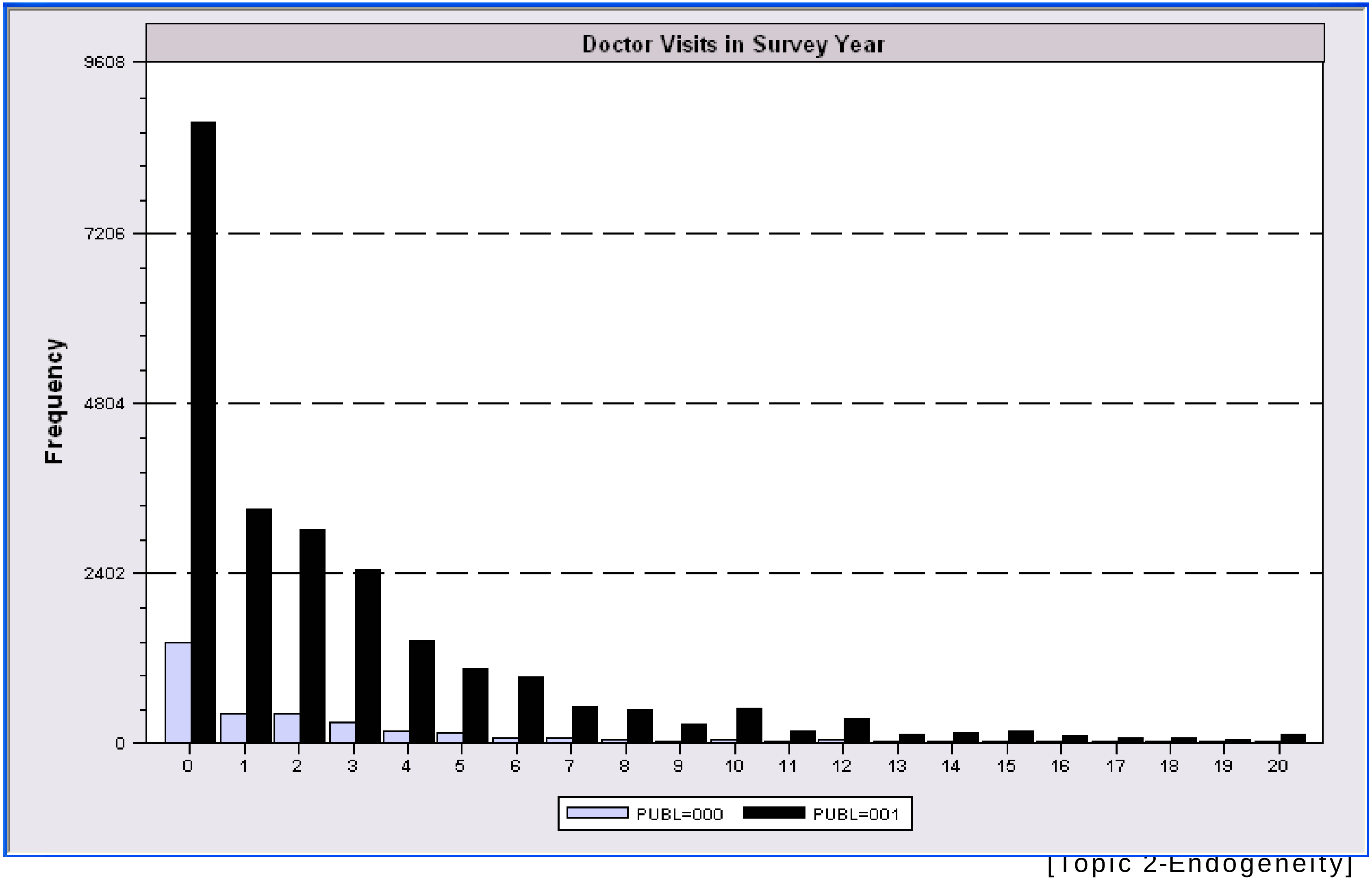
A study of moral hazard
Riphahn, Wambach, Million: “Incentive Effects in the Demand for Healthcare”
Journal of Applied Econometrics, 2003

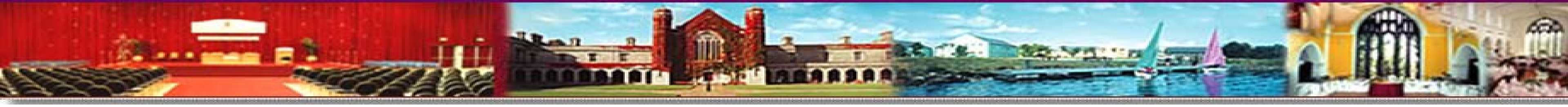
Did the presence of the ADDON insurance influence the demand for health care – doctor visits and hospital visits?

For a simple example, we examine the PUBLIC insurance (89%) instead of ADDON insurance (2%).



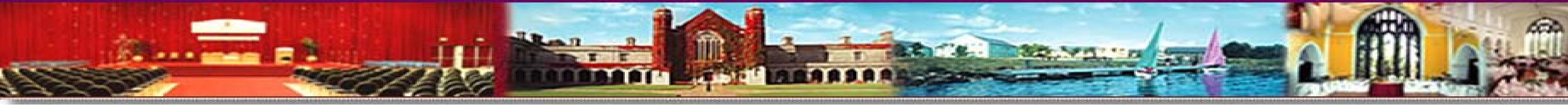
Evidence of Moral Hazard?





Regression Study

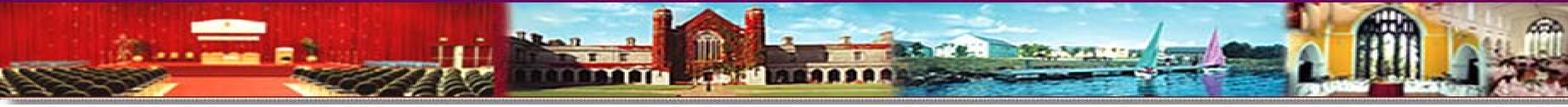
```
------------------------------------------------------------------
Ordinary     least squares regression ............
LHS=DOCVIS   Mean                 =          3.18352
             Standard deviation   =          5.68969
             Number of observs.   =            27326
Model size   Parameters           =                6
             Degrees of freedom   =            27320
Residuals    Sum of squares       =     853326.41135
             Standard error of e  =          5.58878
Fit          R-squared            =           .03533
             Adjusted R-squared   =           .03516
Model test   F[  5, 27320] (prob) =     200.1(.0000)
--------+---------------------------------------------------------
        |                  Standard            Prob.      Mean
  DOCVIS| Coefficient        Error         z    z>|Z|      of X
--------+---------------------------------------------------------
Constant|     .43660         .29014      1.50  .1324
     AGE|     .06754***      .00304     22.25  .0000    43.5257
  HHNINC|   -1.54898***      .19956     -7.76  .0000     .35208
  FEMALE|     .94128***      .06895     13.65  .0000     .47877
    EDUC|    -.05549***      .01624     -3.42  .0006    11.3206
  PUBLIC|     .59843***      .11370      5.26  .0000     .88571
--------+---------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
------------------------------------------------------------------
```



Endogenous Dummy Variable

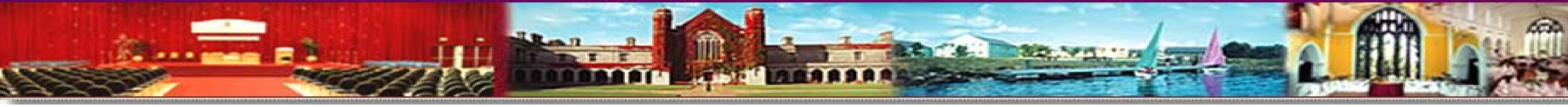
- Doctor Visits = f(Age, Educ, Health,
  Presence of Insurance,
  Other unobservables)

- Insurance = f(Expected Doctor Visits,
  Other unobservables)



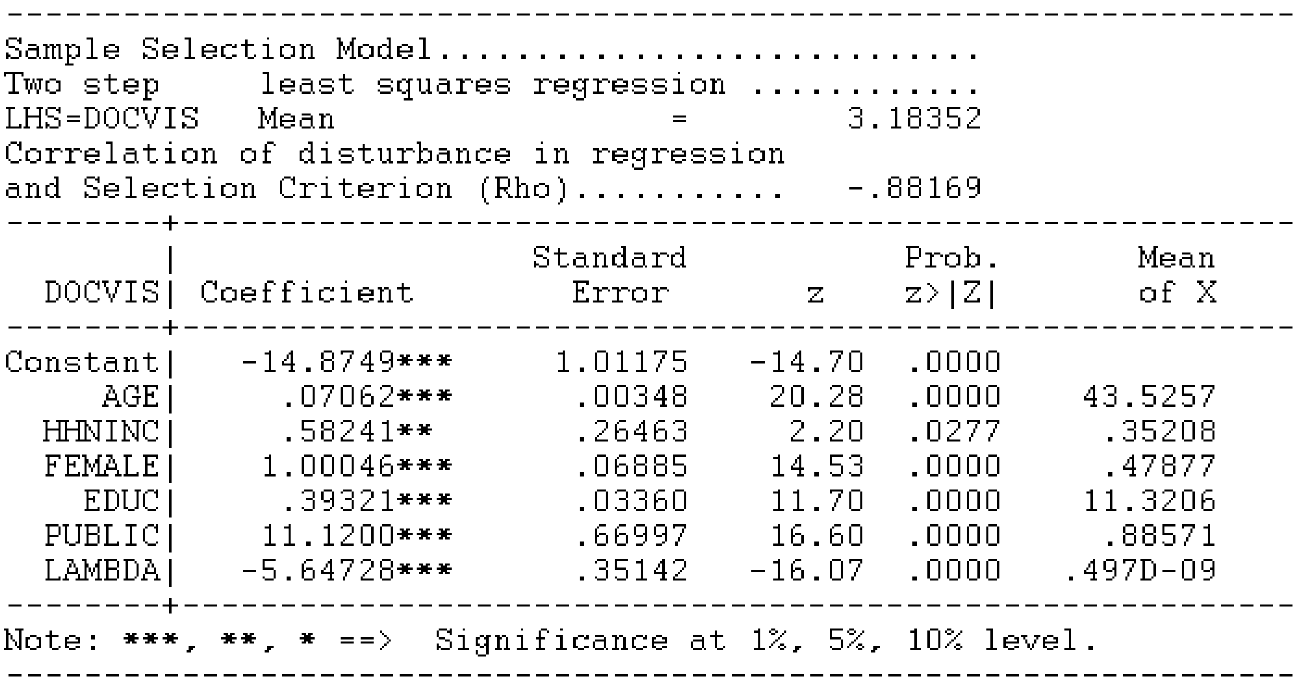
Approaches

- (Parametric) Control Function: Build a structural model for the two variables (Heckman)
- (Semiparametric) Instrumental Variable: Create an instrumental variable for the dummy variable (Barnow/Cain/ Goldberger, Angrist, Current generation of researchers)
- (?) Propensity Score Matching (Heckman et al., Becker/Ichino, Many recent researchers)



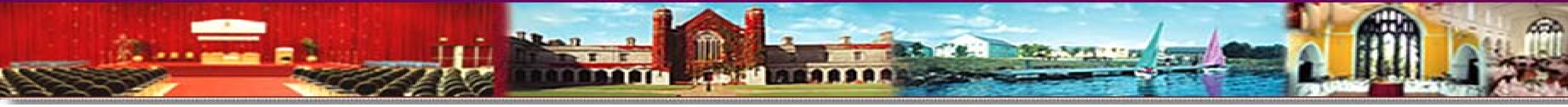
Heckman's Control Function Approach

- $Y = x\beta + \delta T + E[\varepsilon|T] + \{\varepsilon - E[\varepsilon|T]\}$
- $\lambda = E[\varepsilon|T]$, computed from a model for whether T = 0 or 1



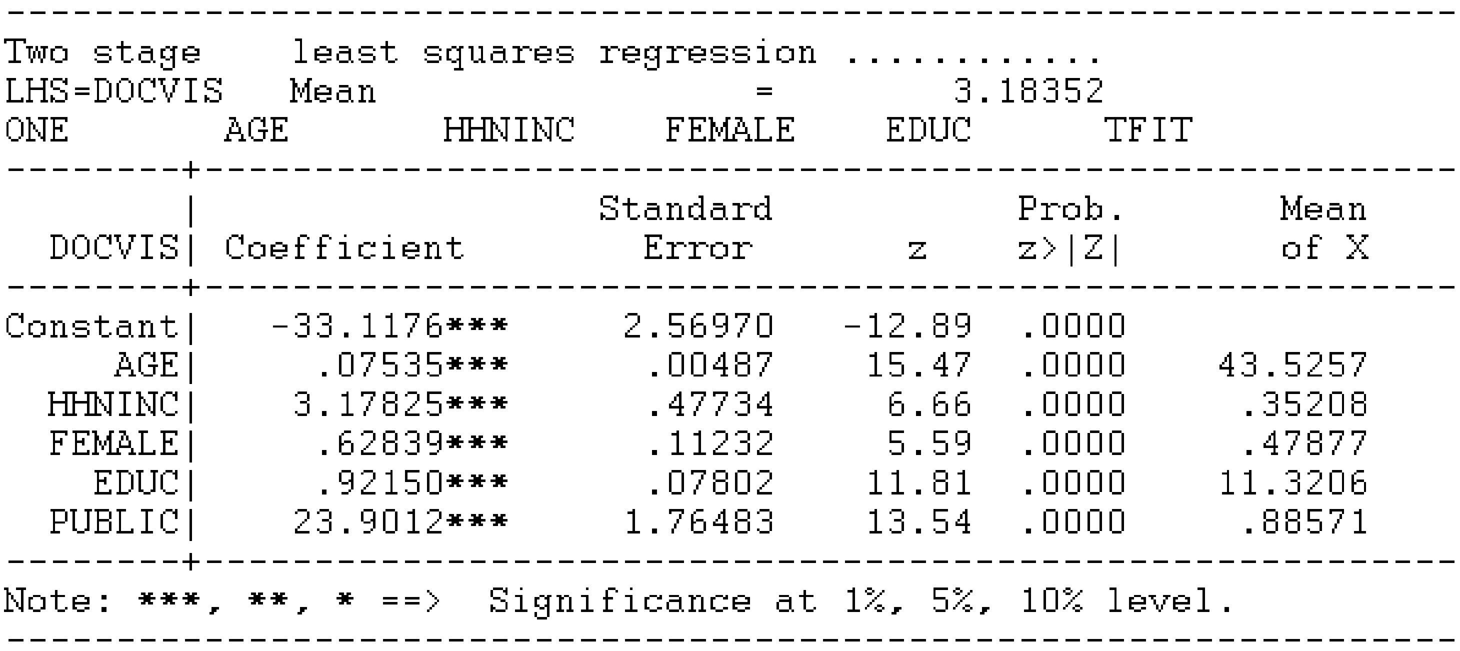
```
-----------------------------------------------------------------------------
Sample Selection Model............................
Two step     least squares regression ...........
LHS=DOCVIS   Mean                    =        3.18352
Correlation of disturbance in regression
and Selection Criterion (Rho)..........   -.88169
--------+--------------------------------------------------------------------
        |                 Standard             Prob.        Mean
  DOCVIS| Coefficient       Error        z     z>|Z|        of X
--------+--------------------------------------------------------------------
Constant|   -14.8749***    1.01175   -14.70  .0000
     AGE|     .07062***     .00348    20.28  .0000      43.5257
  HHNINC|     .58241**      .26463     2.20  .0277       .35208
  FEMALE|    1.00046***     .06885    14.53  .0000       .47877
    EDUC|     .39321***     .03360    11.70  .0000      11.3206
  PUBLIC|    11.1200***     .66997    16.60  .0000       .88571
  LAMBDA|   -5.64728***     .35142   -16.07  .0000     .497D-09
--------+--------------------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
-----------------------------------------------------------------------------
```

Magnitude = 11.1200 is nonsensical in this context.



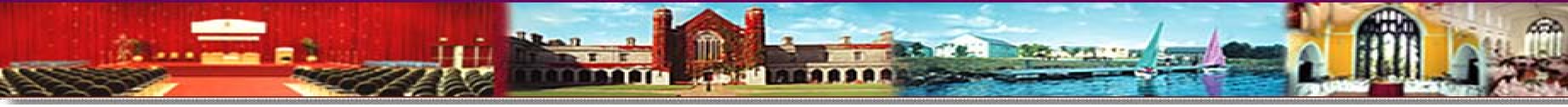
Instrumental Variable Approach

- Construct a prediction for T using only the exogenous information
- Use 2SLS using this instrumental variable.



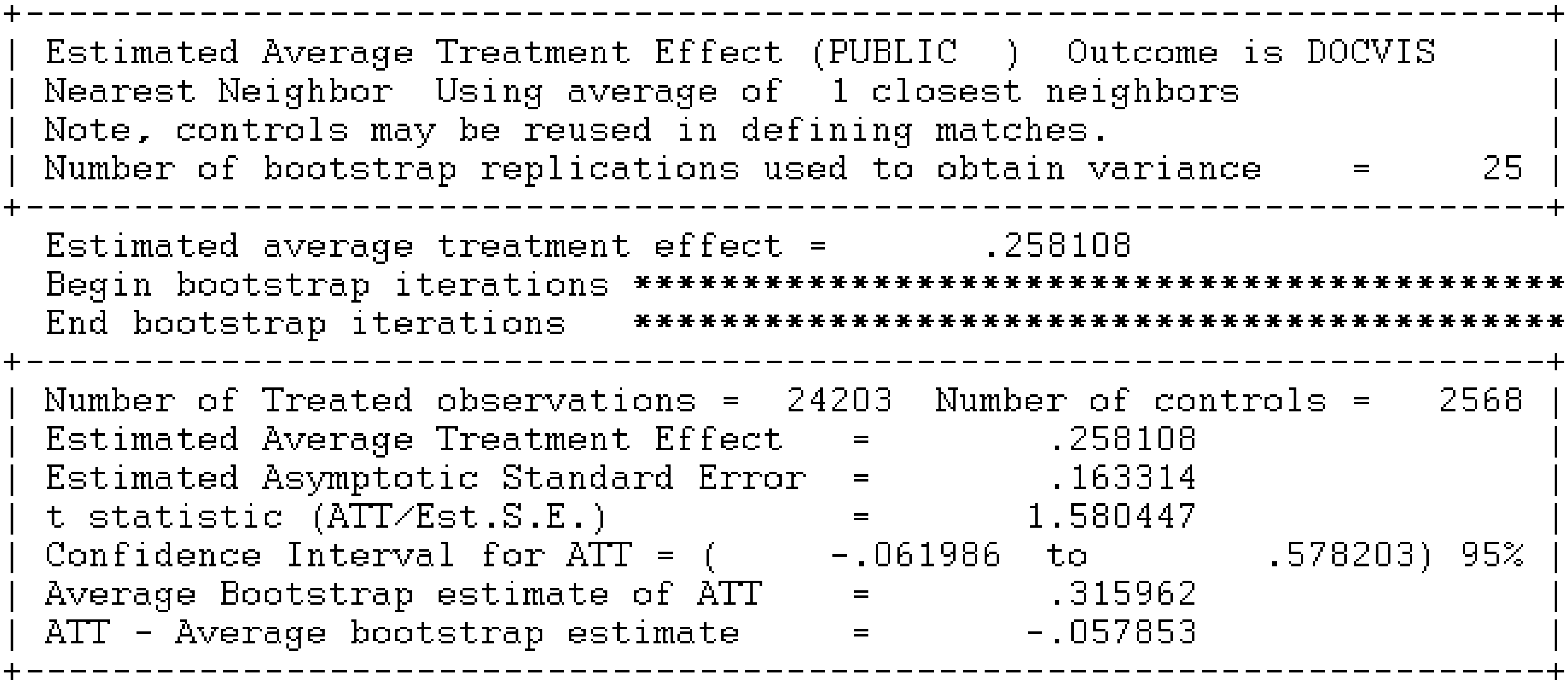
```
------------------------------------------------------------------
Two stage    least squares regression ............
LHS=DOCVIS   Mean                    =        3.18352
ONE        AGE        HHNINC     FEMALE     EDUC      TFIT
--------+---------------------------------------------------------
        |                 Standard            Prob.         Mean
  DOCVIS| Coefficient       Error         z   z>|Z|         of X
--------+---------------------------------------------------------
Constant|    -33.1176***    2.56970   -12.89  .0000
     AGE|      .07535***     .00487    15.47  .0000      43.5257
  HHNINC|     3.17825***     .47734     6.66  .0000       .35208
  FEMALE|      .62839***     .11232     5.59  .0000       .47877
    EDUC|      .92150***     .07802    11.81  .0000      11.3206
  PUBLIC|     23.9012***    1.76483    13.54  .0000       .88571
--------+---------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
------------------------------------------------------------------
```

Magnitude = 23.9012 is also nonsensical in this context.

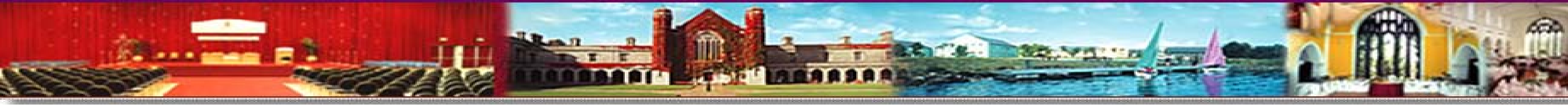


Propensity Score Matching

- Create a model for T that produces probabilities for T=1: “Propensity Scores”
- Find people with the same propensity score – some with T=1, some with T=0
- Compare number of doctor visits of those with T=1 to those with T=0.



```
+----------------------------------------------------------------------+
| Estimated Average Treatment Effect (PUBLIC  )  Outcome is DOCVIS     |
| Nearest Neighbor  Using average of  1 closest neighbors              |
| Note, controls may be reused in defining matches.                    |
| Number of bootstrap replications used to obtain variance   =     25 |
+----------------------------------------------------------------------+
  Estimated average treatment effect =       .258108
  Begin bootstrap iterations *****************************************
  End bootstrap iterations   *****************************************
+----------------------------------------------------------------------+
| Number of Treated observations =  24203  Number of controls =   2568 |
| Estimated Average Treatment Effect   =         .258108               |
| Estimated Asymptotic Standard Error  =         .163314               |
| t statistic (ATT/Est.S.E.)           =        1.580447               |
| Confidence Interval for ATT = (     -.061986  to       .578203) 95% |
| Average Bootstrap estimate of ATT    =         .315962               |
| ATT - Average bootstrap estimate     =        -.057853               |
+----------------------------------------------------------------------+
```



Difference in Differences

With two periods,

$\Delta y_{it} = y_{i2} - y_{i1} = \delta_0 + (\mathbf{x}'_{i2}\boldsymbol{\beta}\mathbf{x}'_{i1}) + u_i$

Consider a "treatment, D_i," that takes place between time 1 and time 2 for some of the individuals

$\Delta y_i = \delta_0 + (\Delta\mathbf{x}_i\boldsymbol{\beta})' + \delta_1 D_i + u_i$

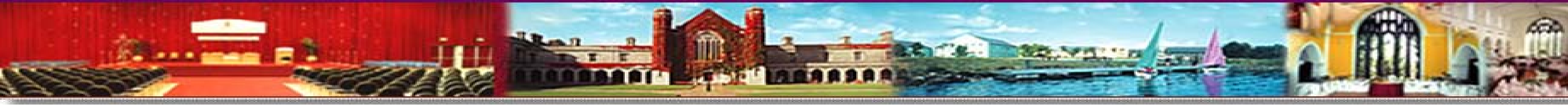
D_i = the "treatment dummy"

This is a linear regression model. If there are no regressors,

$$\hat{\delta}_1 = \overline{\Delta y} \mid \text{treatment} - \overline{\Delta y} \mid \text{control}$$

= "difference in differences" estimator.

$\hat{\delta}_0 =$ Average change in y_i for the "treated"



Difference-in-Differences Model

With two periods and strict exogeneity of D and T,

$$y_{it} = \beta_0 + \beta_1 D_{it} + \beta_2 T_t + \beta_3 T_t D_{it} + \varepsilon_{it}$$

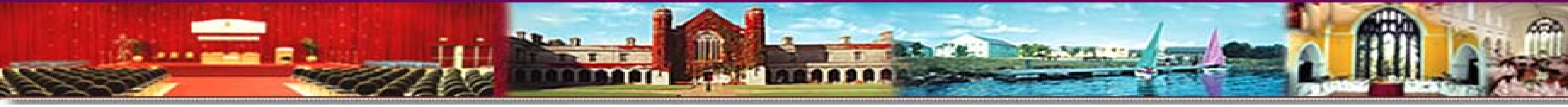
D_{it} = dummy variable for a treatment that takes place between time 1 and time 2 for some of the individuals,

T_t = a time period dummy variable, 0 in period 1, 1 in period 2.

This is a linear regression model. If there are no regressors,

Using least squares,

$$b_3 = (\overline{y}_2 - \overline{y}_1)_{D=1} - (\overline{y}_2 - \overline{y}_1)_{D=0}$$



Difference in Differences

$$\mathrm{y}_{\mathrm{it}} = \beta_0 + \beta_1 \mathrm{D}_{it} + \beta_2 \mathrm{T}_t + \beta_3 \mathrm{D}_{it}\mathrm{T}_t + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it}, t = 1, 2$$

$$\Delta \mathrm{y}_{\mathrm{it}} = \beta_2 + \beta_3 \mathrm{D}_{i2} + \Delta(\boldsymbol{\beta}'\mathbf{x}_{it}) + \Delta\varepsilon_{it}$$

$$= \beta_2 + \beta_3 \mathrm{D}_{i2} + \boldsymbol{\beta}'(\Delta\mathbf{x}_{it}) + \mathrm{u}_i$$

$$\left(\Delta \mathrm{y}_{\mathrm{it}} \mid \mathrm{D} = 1\right) - \left(\Delta \mathrm{y}_{\mathrm{it}} \mid \mathrm{D} = 0\right)$$

$$= \beta_3 + \boldsymbol{\beta}'\left[(\Delta\mathbf{x}_{it} \mid \mathrm{D} = 1) - (\Delta\mathbf{x}_{it} \mid \mathrm{D} = 0)\right]$$

If the same individual is observed in both states, the second term is zero. If the effect is estimated by averaging individuals with D = 1 and different individuals with D=0, then part of the 'effect' is explained by change in the covariates, not the treatment.