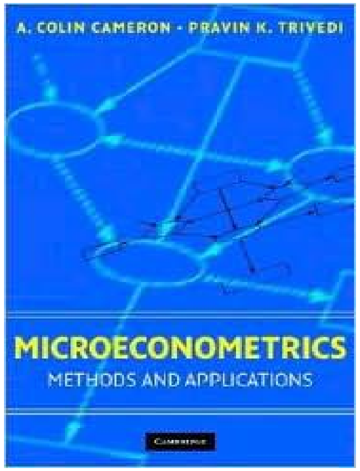
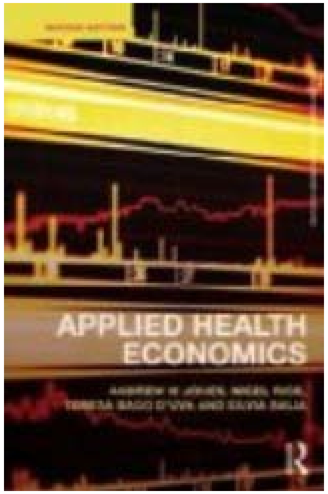
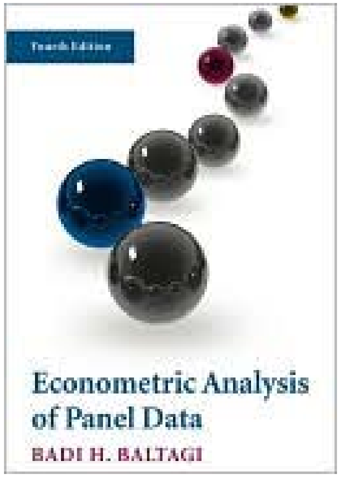
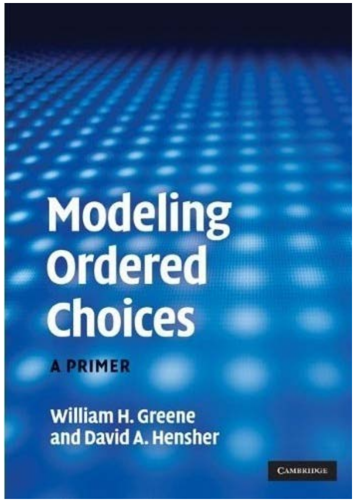
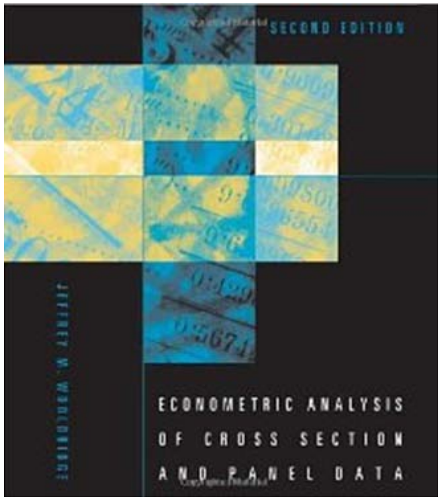


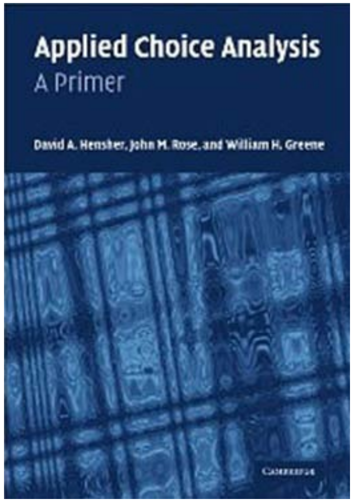




1. Descriptive Tools, Regression, Panel Data



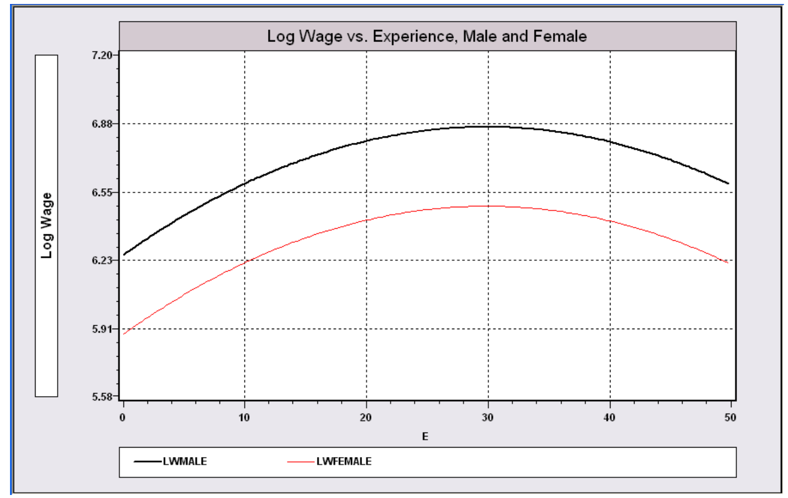




Model Building in Econometrics

- Parameterizing the model
  - Nonparametric analysis
  - Semiparametric analysis
  - Parametric analysis
- Sharpness of inferences follows from the strength of the assumptions

A Model Relating (Log)Wage to Gender and Experience

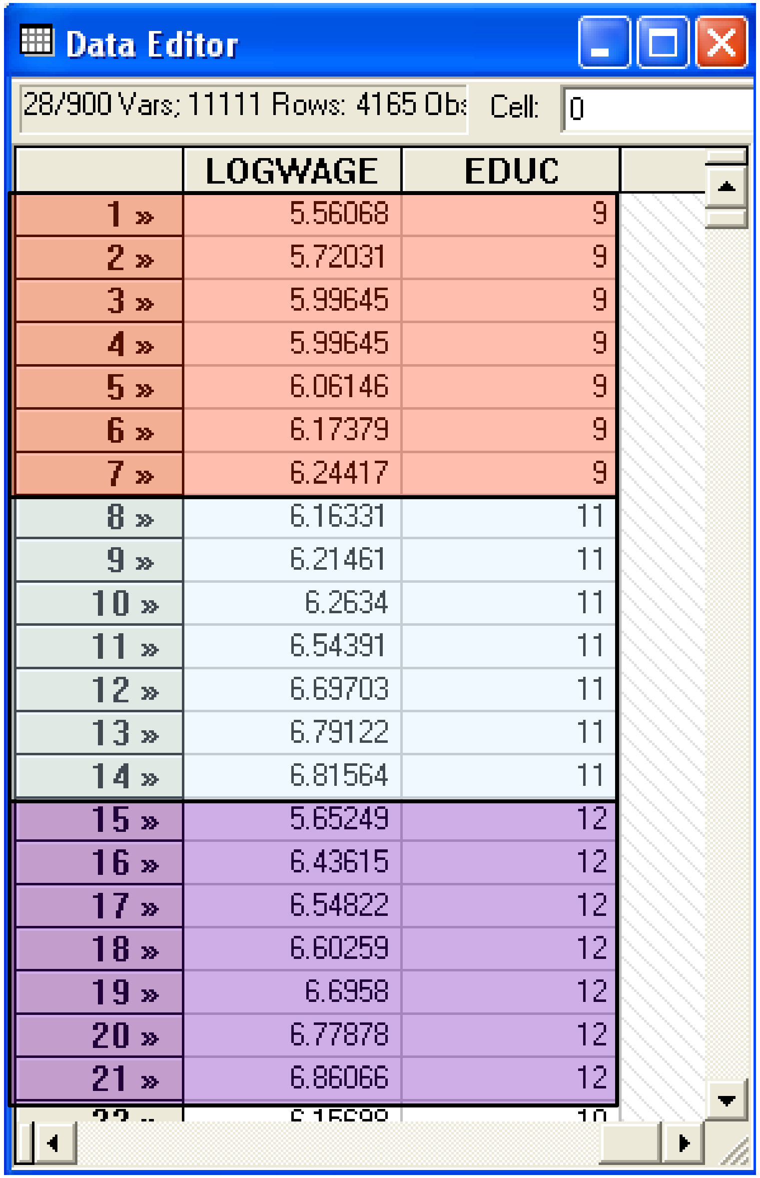


Cornwell and Rupert Panel Data

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

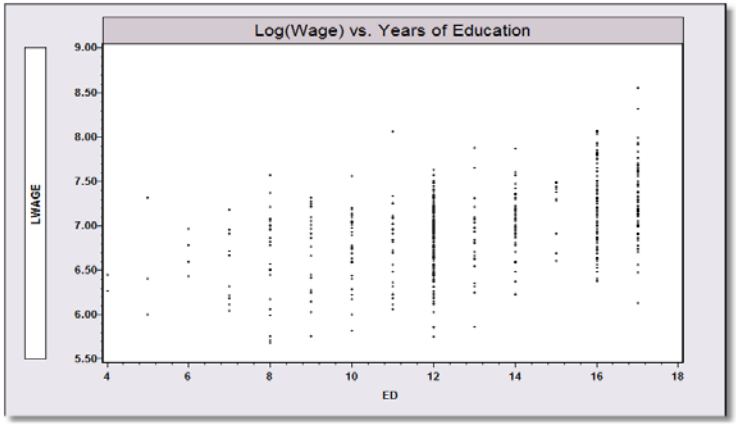
EXP	= work experience
WKS	= weeks worked
OCC	= occupation, 1 if blue collar,
IND	= 1 if manufacturing industry
SOUTH	= 1 if resides in south
SMSA	= 1 if resides in a city (SMSA)
MS	= 1 if married
FEM	= 1 if female
UNION	= 1 if wage set by union contract
ED	= years of education
LWAGE	= log of wage = dependent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

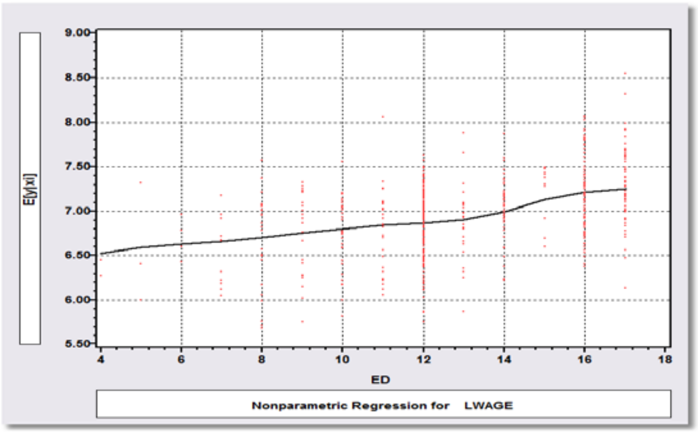


	LOGWAGE	EDUC
1 »	5.56068	9
2 »	5.72031	9
3 »	5.99645	9
4 »	5.99645	9
5 »	6.06146	9
6 »	6.17379	9
7 »	6.24417	9
8 »	6.16331	11
9 »	6.21461	11
10 »	6.2634	11
11 »	6.54391	11
12 »	6.69703	11
13 »	6.79122	11
14 »	6.81564	11
15 »	5.65249	12
16 »	6.43615	12
17 »	6.54822	12
18 »	6.60259	12
19 »	6.6958	12
20 »	6.77878	12
21 »	6.86066	12

Application: Is there a relationship between Log(wage) and Education?

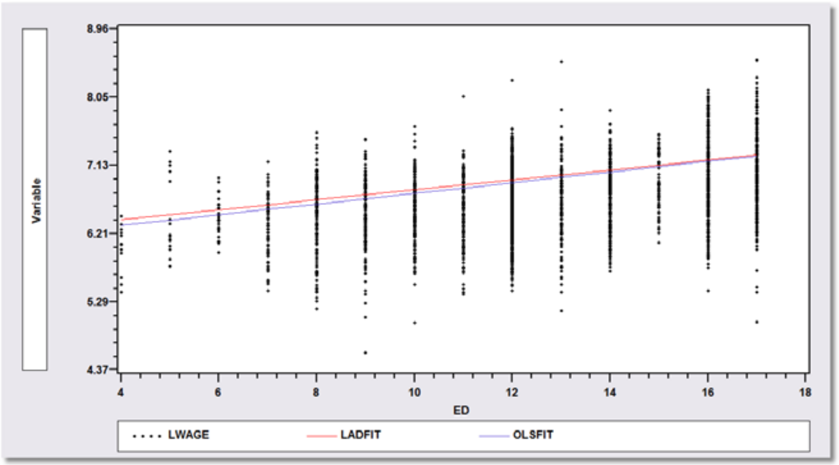


Nonparametric Regression
Kernel regression of y on x



Semiparametric Regression: Least absolute deviations regression of y on x

Parametric Regression: Least squares – maximum likelihood – regression of y on x



A First Look at the Data
Descriptive Statistics

- Basic Measures of Location and Dispersion
- Graphical Devices
  - Box Plots
  - Histogram
  - Kernel Density Estimator

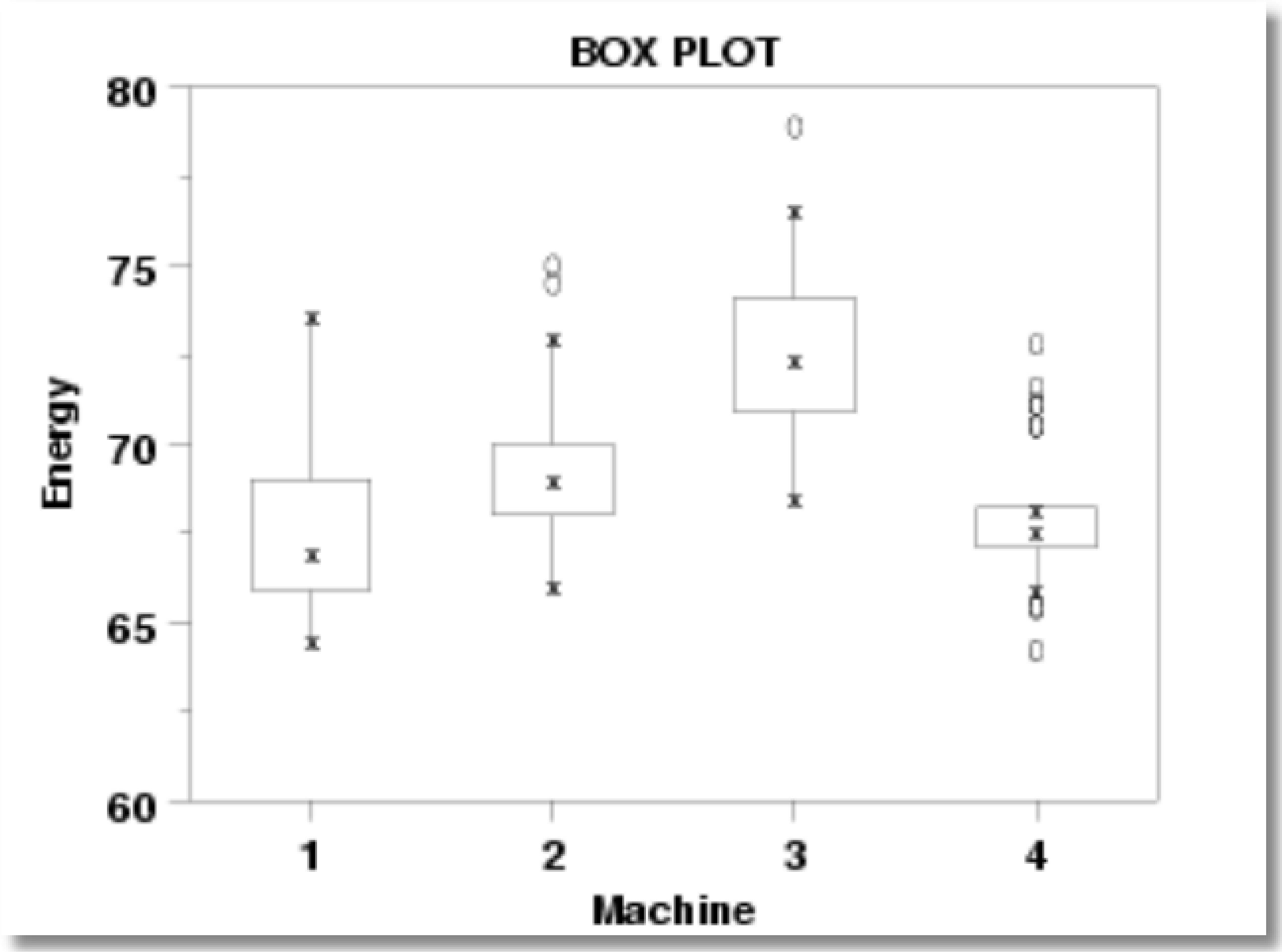
Descriptive Statistics for 11 variables

Variable	Mean	Std.Dev.	Minimum	Maximum	Cases	Missing
EXP	19.85378	10.96637	1.0	51.0	4165	0
WKS	46.81152	5.129098	5.0	52.0	4165	0
OCC	.511164	.499935	0.0	1.0	4165	0
IND	.395438	.489003	0.0	1.0	4165	0
SOUTH	.290276	.453944	0.0	1.0	4165	0
SMSA	.653782	.475821	0.0	1.0	4165	0
MS	.814406	.388826	0.0	1.0	4165	0
FEM	.112605	.316147	0.0	1.0	4165	0
UNION	.363986	.481202	0.0	1.0	4165	0
LWAGE	6.676346	.461512	4.605170	8.537000	4165	0
YEAR	4.0	2.000240	1.0	7.0	4165	0

Descriptive Statistics for LWAGE
Stratification is based on YEAR

Subsample	Mean	Std.Dev.	Cases	Sum of wts	Missing
YEAR = 1	6.375173	.388426	595	595.00	0
YEAR = 2	6.465212	.362702	595	595.00	0
YEAR = 3	6.596717	.446691	595	595.00	0
YEAR = 4	6.696079	.440750	595	595.00	0
YEAR = 5	6.786454	.424013	595	595.00	0
YEAR = 6	6.864045	.424021	595	595.00	0
YEAR = 7	6.950745	.438403	595	595.00	0
Full Sample	6.676346	.461512	4165	4165.00	0

Box Plots



From Jones and Schurer (2011)

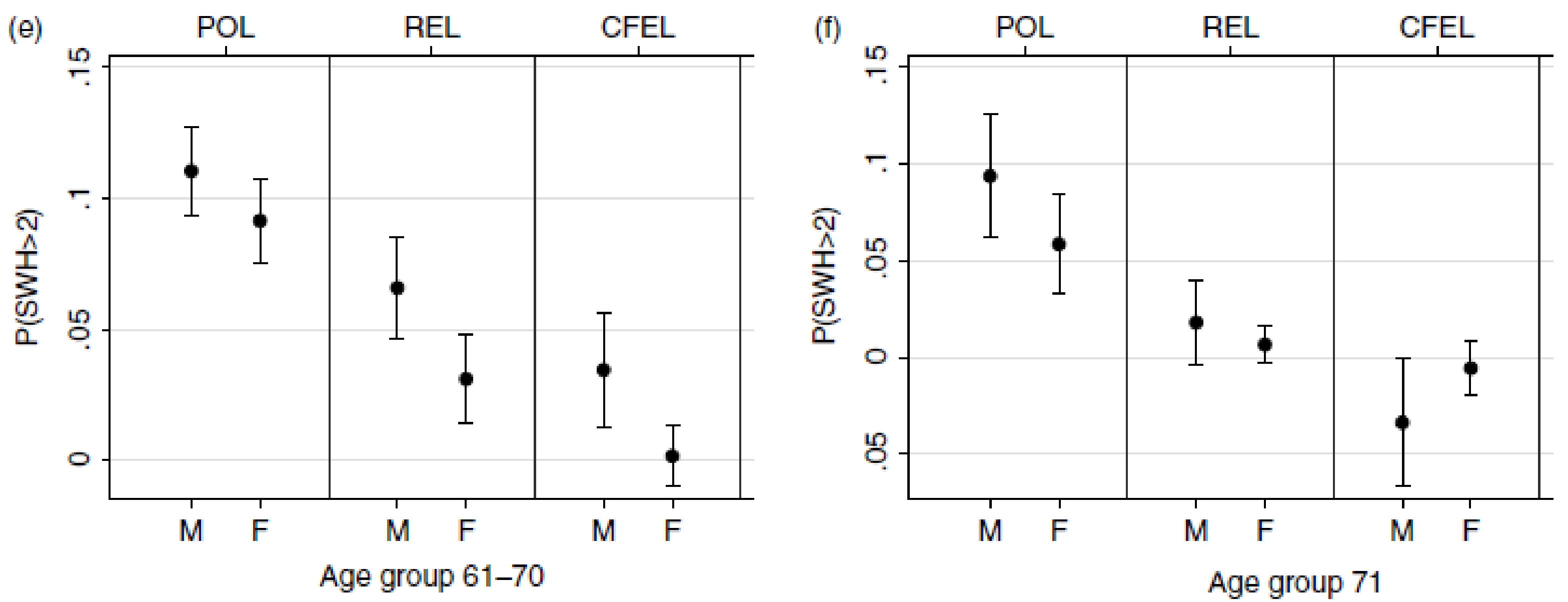
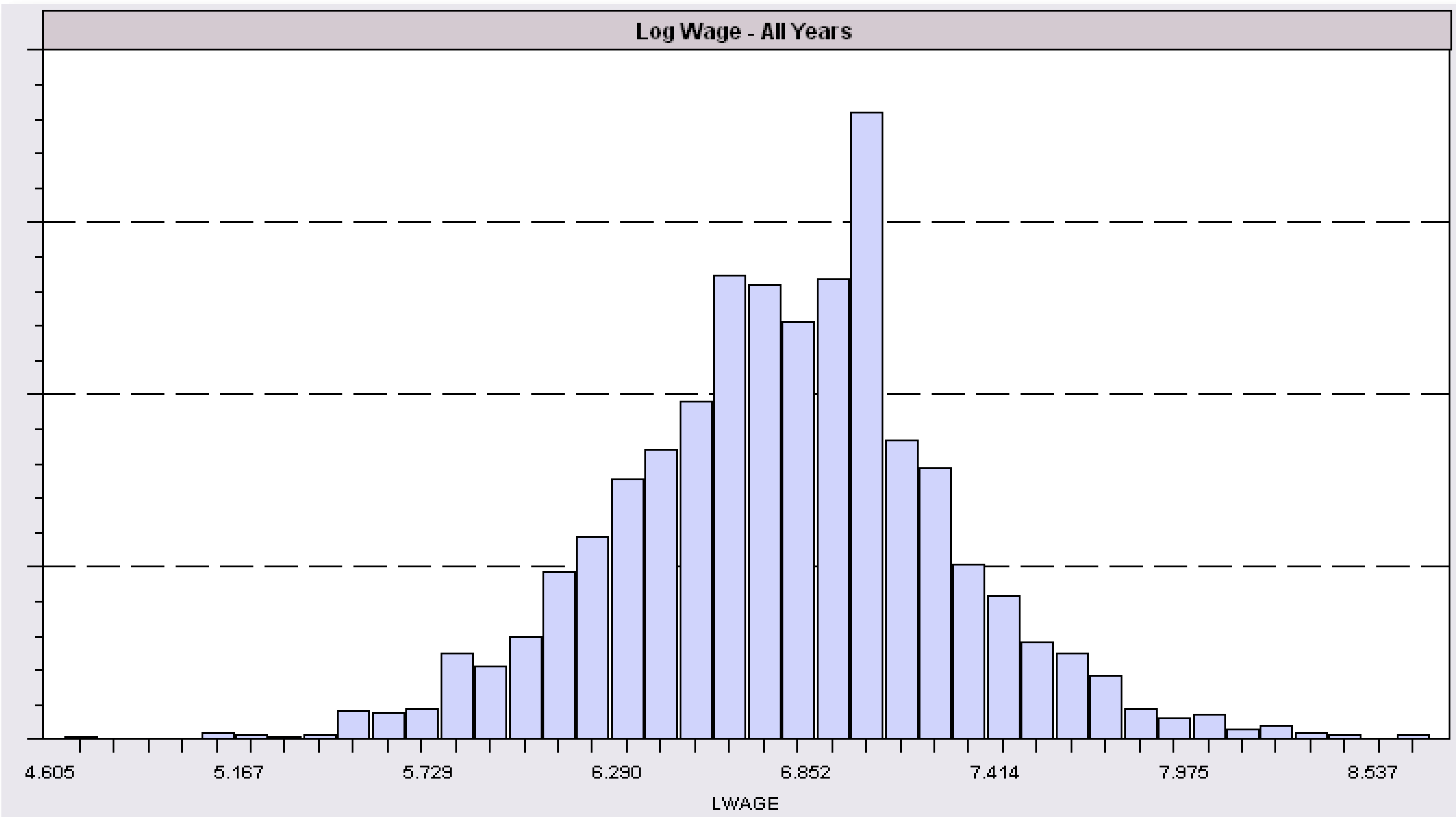


Figure 4. Marginal effects of income on the probability to report satisfaction with health greater than two for both men (M) and women (W) obtained from three different models: pooled ordered logit (POL), random-effects logit (REL), and conditional fixed-effects logit (CFEL)

Histogram for LWAGE



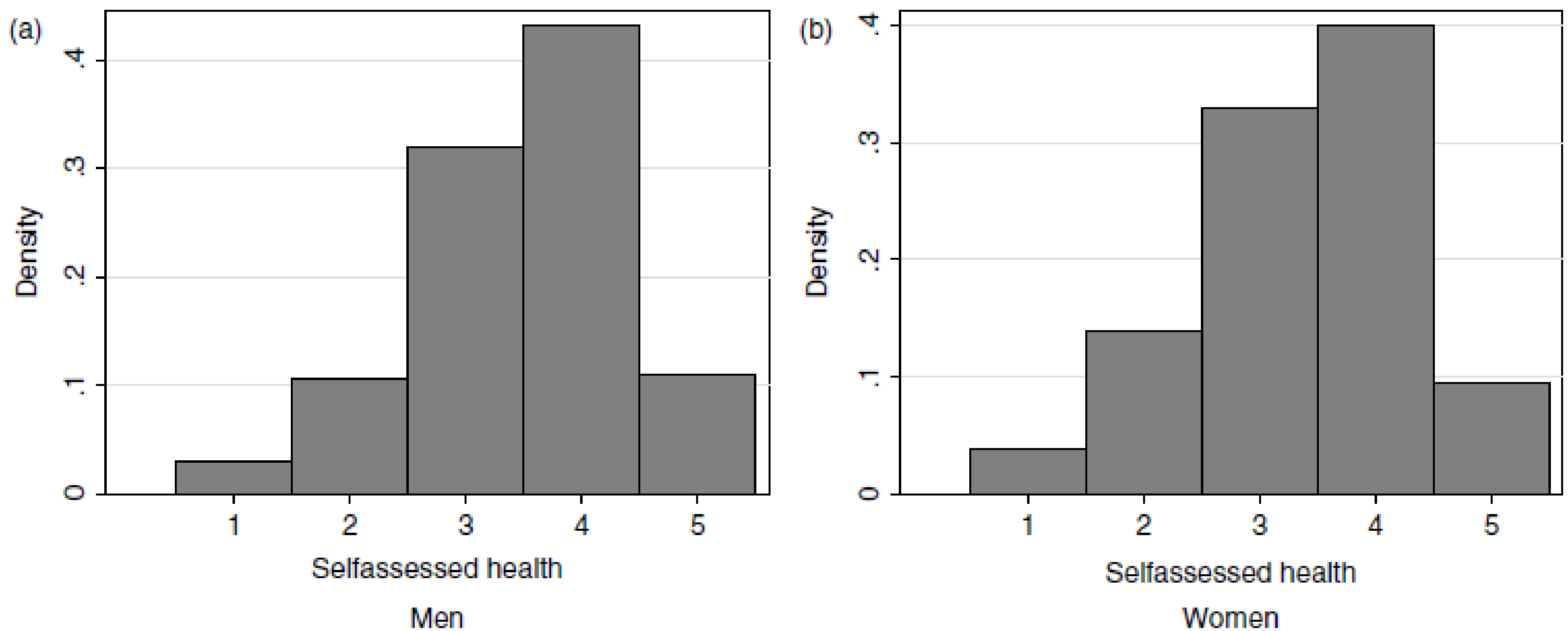
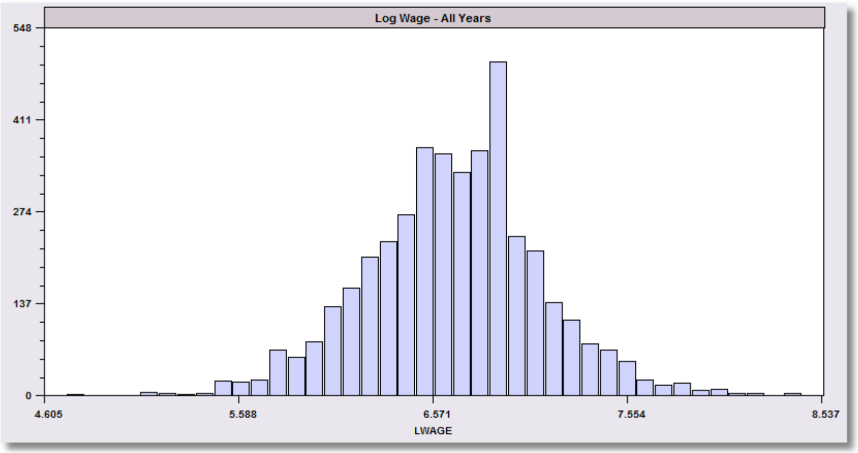


Figure 2. Distribution of health measure: self-assessed health (SAH)

The kernel density estimator is a histogram (of sorts).



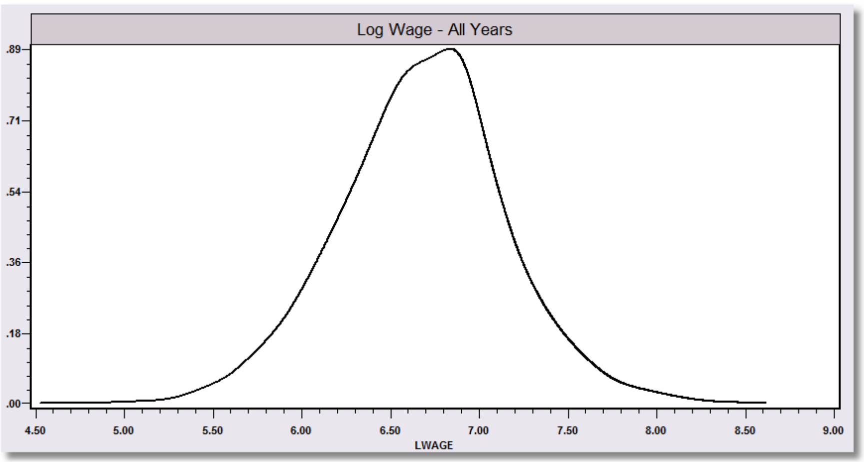
$$\hat{f}(x_m^*) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{B}K\left[\frac{x_i - x_m^*}{B}\right], \text{ for a set of points } x_m^*$$

B = "bandwidth" chosen by the analyst

K = the kernel function, such as the normal or logistic pdf (or one of several others)

x* = the point at which the density is approximated.

This is essentially a histogram with small bins.



Kernel Density Estimator

The curse of dimensionality

$$\hat{f}(x_m^*) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{B}K\left[\frac{x_i - x_m^*}{B}\right], \text{ for a set of points } x_m^*$$

$B =$ "bandwidth"

$K =$ the kernel function

$x^* =$ the point at which the density is approximated.

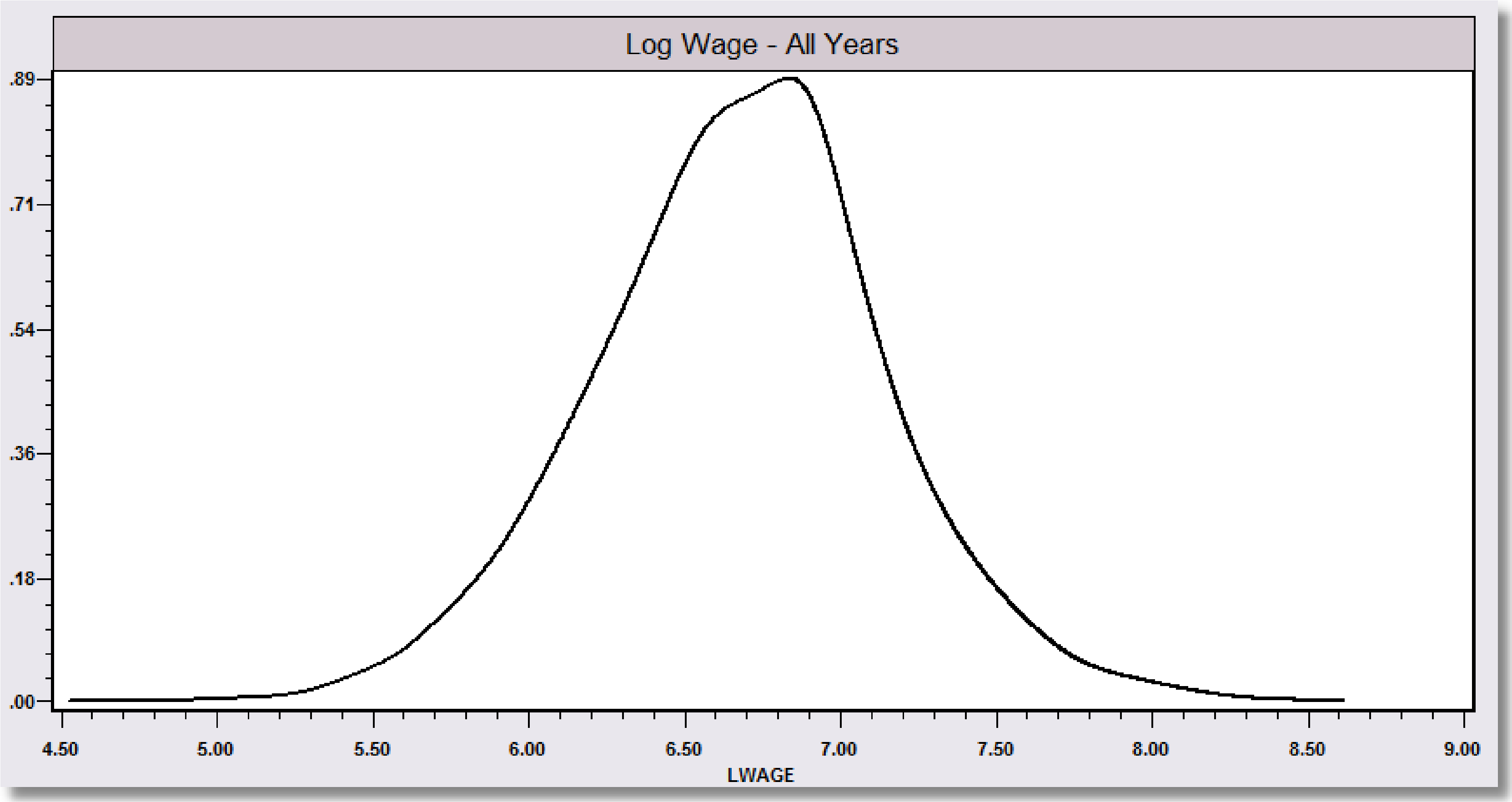
$\hat{f}(x^*)$ is an estimator of $f(x^*)$

$$\frac{1}{n}\sum_{i=1}^{n}Q(x_i \mid x^*) = \bar{Q}(x^*).$$

But, $\text{Var}[\bar{Q}(x^*)] \neq \frac{1}{N} \times \text{Something}$. Rather, $\text{Var}[\bar{Q}(x^*)] = \frac{1}{N^{3/5}} * \text{Something}$

I.e., $\hat{f}(x^*)$ does not converge to $f(x^*)$ at the same rate as a mean converges to a population mean.

Kernel Estimator for LWAGE



From Jones and Schurer (2011)

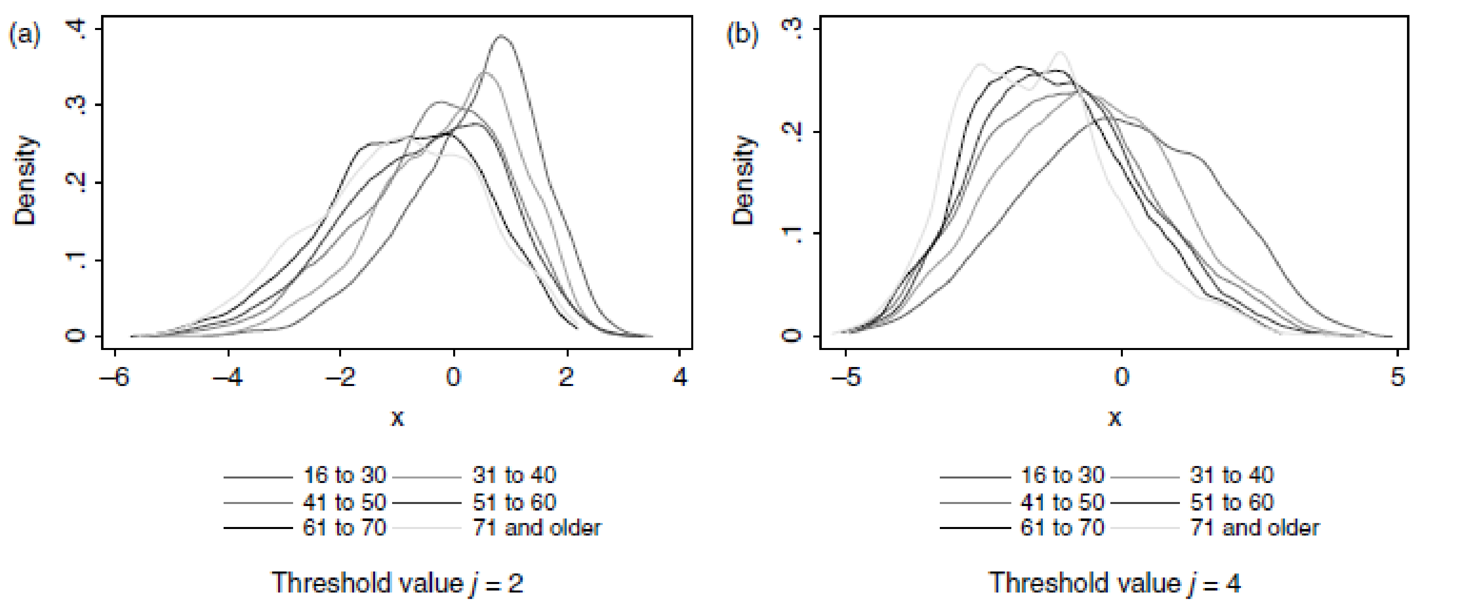
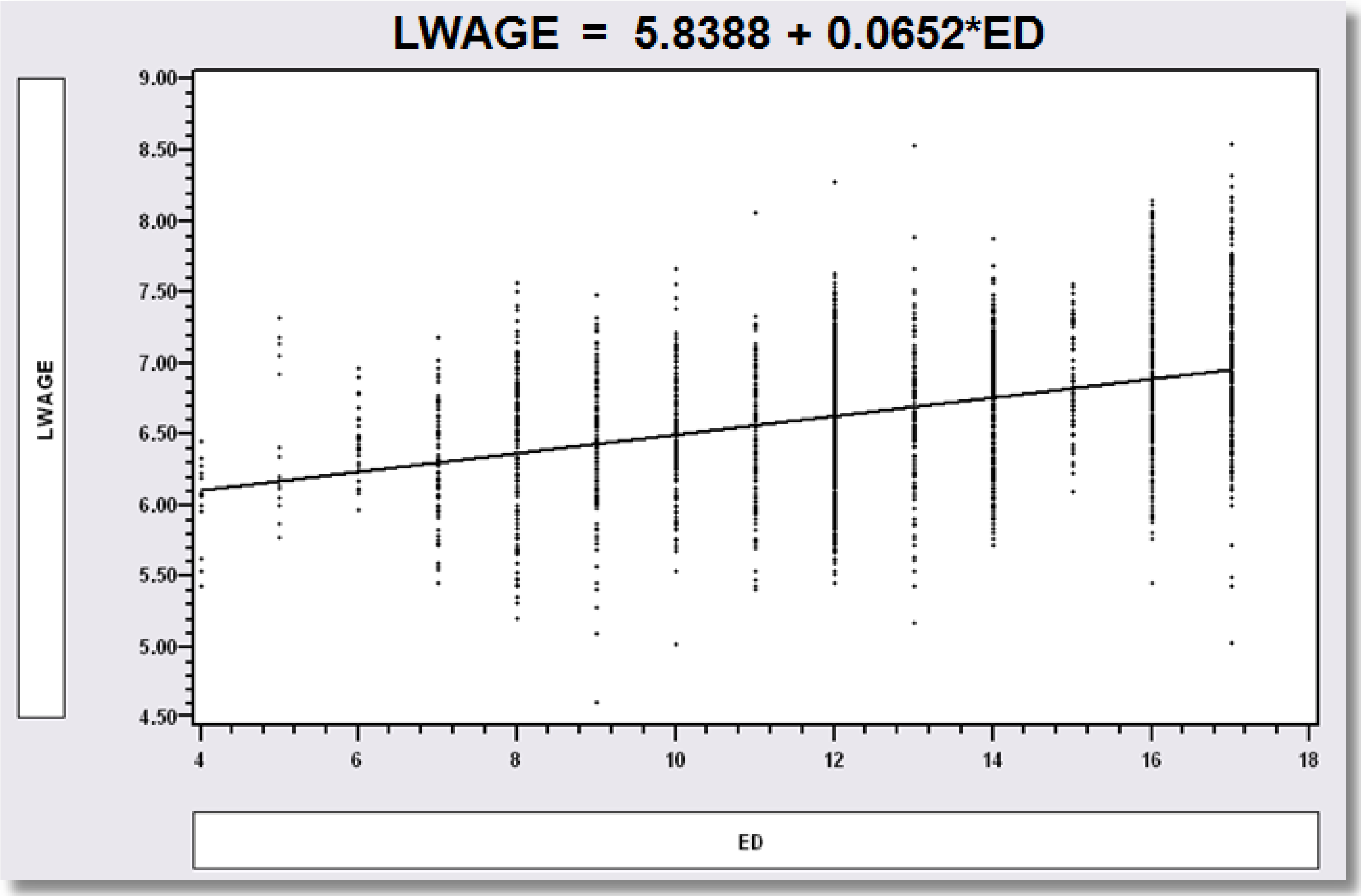


Figure 7. Probability distribution of individual fixed effect obtained from the conditional fixed-effects logit (CFEL), when the threshold values are $j = 2$ and $j = 4$ for the sample of men

Objective: Impact of Education on (log) Wage

- **Specification**: What is the right model to use to analyze this association?
- **Estimation**
- **Inference**
- **Analysis**

Simple Linear Regression



Multiple Regression

Ordinary least squares regression ....

LHS=LWAGE	Mean	=	6.67635			
	Standard deviation	=	.46151			
----------	No. of observations	=	4165	DegFreedom		Mean square
Regression	Sum of Squares	=	345.763	9		38.41812
Residual	Sum of Squares	=	541.142	4155		.13024
Total	Sum of Squares	=	886.905	4164		.21299
----------	Standard error of e	=	.36089	Root MSE		.36045
Fit	R-squared	=	.38985	R-bar squared		.38853
Model test	F[9, 4155]	=	294.98231	Prob F > F*		.00000

LWAGE	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	5.44028***	.07208	75.48	.0000	5.29902	5.58155
ED	.05682***	.00267	21.25	.0000	.05158	.06207
EXP	.01040***	.00054	19.37	.0000	.00935	.01145
WKS	.00525***	.00111	4.71	.0000	.00306	.00743
OCC	-.14867***	.01507	-9.87	.0000	-.17819	-.11914
SOUTH	-.07024***	.01279	-5.49	.0000	-.09530	-.04517
SMSA	.13241***	.01235	10.72	.0000	.10820	.15663
MS	.08560***	.02100	4.06	.0000	.04435	.12700
FEM	-.37561***	.02577	-14.58	.0000	-.42611	-.32511
UNION	.09995***	.01318	7.58	.0000	.07411	.12579

***, **, * ==> Significance at 1%, 5%, 10% level.

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
```

LWAGE	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence 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.

Partial Effects

```
----------------------------------------------------------------------------
Ordinary      least squares regression ............
LHS=LWAGE     Mean                    =      6.67635
              Standard deviation      =       .46151
----------    No. of observations     =         4165  DegFreedom   Mean square
Regression    Sum of Squares          =      378.218          11      34.38347
Residual      Sum of Squares          =      508.687        4153        .12249
Total         Sum of Squares          =      886.905        4164        .21299
----------    Standard error of e     =       .34998  Root MSE            .34948
Fit           R-squared               =       .42645  R-bar squared       .42493
Model test    F[ 11,   4153]          =    280.71214  Prob F > F*         .00000
--------+-------------------------------------------------------------------
```

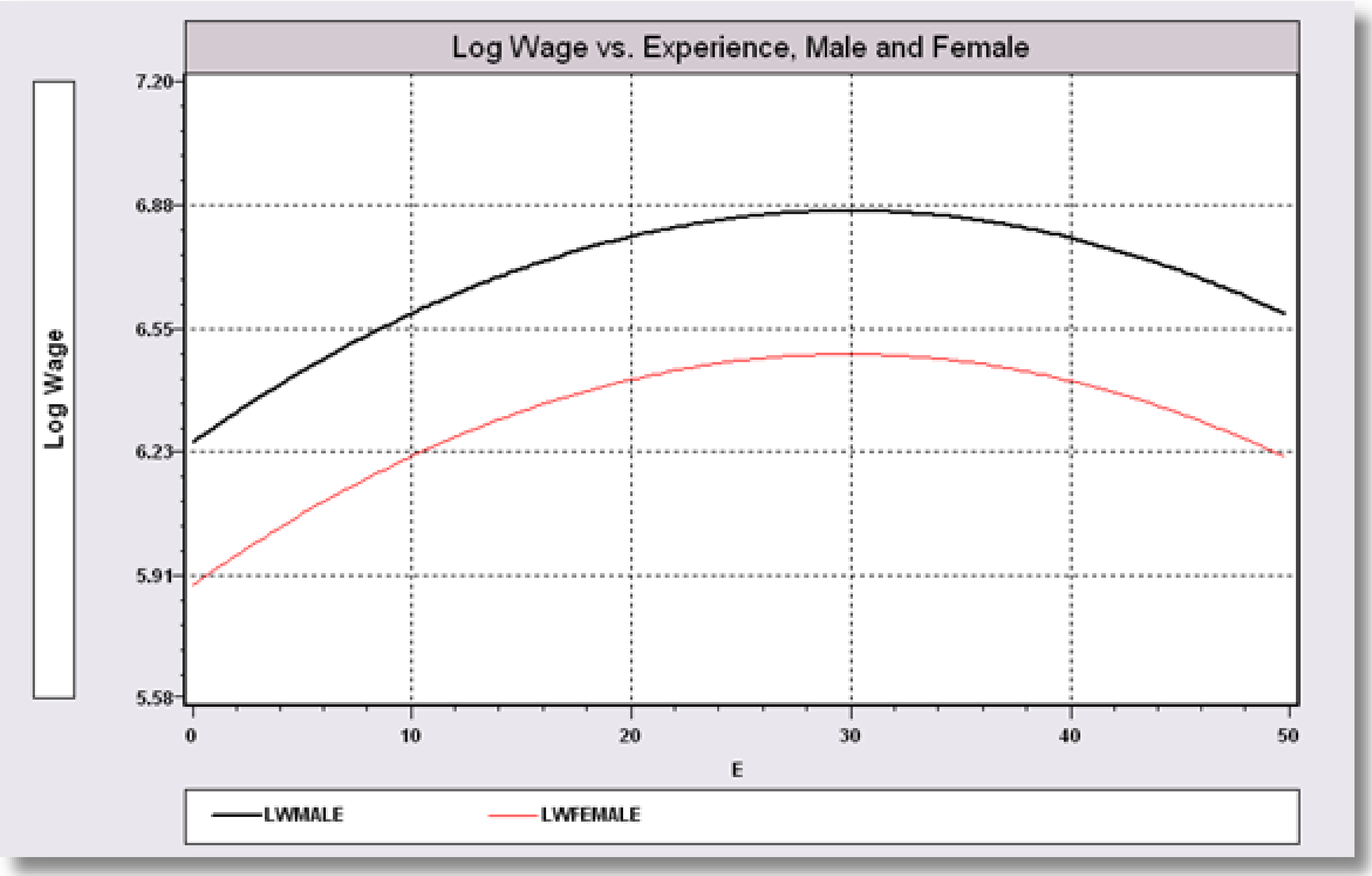
```
Constant|      5.24547***
      ED|       .05654***
     EXP|       .04045***
 EXP*EXP|     -.00068***
     WKS|       .00449***
     OCC|     -.14053***
   SOUTH|     -.07210***
    SMSA|       .13901***
      MS|       .06736***
     FEM|     -.38922***
   UNION|       .09015***
```

Education: .05654

Experience .04045 - 2*.00068*Exp

FEM -.38922

Model Implication: Effect of Experience and Male vs. Female



Hypothesis Test About Coefficients

- Hypothesis
  - Null: Restriction on **β**: **Rβ** – **q** = **0**
  - Alternative: Not the null
- Approaches
  - Fitting Criterion: R^2 decrease under the null?
  - Wald: **Rb – q** close to **0** under the alternative?

Hypotheses

```
----------------------------------------------------------------------
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.
----------------------------------------------------------------------
```

All Coefficients = 0?

R = [0 | I] q = [0]

ED Coefficient = 0?

R = 0,1,0,0,0,0,0,0,0,0,0

q = 0

No Experience effect?

R = 0,0,1,0,0,0,0,0,0,0,0
0,0,0,1,0,0,0,0,0,0,0

q = 0
0

Hypothesis Test Statistics

Subscript 0 = the model under the null hypothesis

Subscript 1 = the model under the alternative hypothesis

1. Based on the Fitting Criterion R^2

$$F = \frac{(R_1^2 - R_0^2) / J}{(1 - R_1^2) / (N - K_1)} = F[J, N - K_1]$$

2. Based on the Wald Distance : Note, for linear models, W = JF.

$$\text{Chi Squared} = (\mathbf{Rb} - \mathbf{q})' \left[\mathbf{R} \left(s^2 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \right) \mathbf{R}' \right]^{-1} (\mathbf{Rb} - \mathbf{q})$$

Hypothesis: All Coefficients Equal Zero

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

LWAGE	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence 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.
----------------------------------------------------------------------------
```

All Coefficients = 0?

R = [0 | I] q = [0]

R_1^2 = .41826
R_0^2 = .00000

F = 298.7 with [10,4154]

Wald = $b_{2\text{-}11}[V_{2\text{-}11}]^{-1}b_{2\text{-}11}$
= 2988.3355

Note that Wald = JF
= 10(298.7)
(some rounding error)

Hypothesis: Education Effect = 0

```
-----------------------------------------------------------------------------
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.
-----------------------------------------------------------------------------
```

ED Coefficient = 0?

R = 0,1,0,0,0,0,0,0,0,0,0

q = 0

R_1^2 = .41826
R_0^2 = .35265 (not shown)

F = 468.29

Wald = $(.05654-0)^2/(.00261)^2$
= 468.29

Note F = t^2 and Wald = F

For a single hypothesis about 1 coefficient.

Hypothesis: Experience Effect = 0

```
--------------------------------------------------------------------------------
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.
--------------------------------------------------------------------------------
```

No Experience effect?

R = $\begin{bmatrix} 0,0,1,0,0,0,0,0,0,0,0 \\ 0,0,0,1,0,0,0,0,0,0,0 \end{bmatrix}$

q = $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$

R_0^2 **= .33475,** R_1^2 **= .41826**

F = 298.15

Wald = 596.3 (W* = 5.99)

	1	2	3	4	
1	0.0050797	-0.000108601	-3.80795e-005	6.45458e-007	-5.
2	-0.000108601	6.74903e-006	2.7876e-007	9.39971e-010	1.
3	-3.80795e-005	2.7876e-007	4.66184e-006	-9.95114e-008	-8.
4	6.45458e-007	9.39971e-010	-9.95114e-008	2.25567e-009	2.
5	-5.61401e-005	1.23728e-007	-8.17968e-008	2.51799e-009	1.
6	-0.000369037	2.09061e-005	1.68653e-006	-2.68766e-008	-5.
7	-0.000120975	4.14423e-006	2.20649e-007	6.31692e-009	-3.
8	-2.8469e-005	-3.35864e-006	3.55984e-007	-2.17099e-008	-7.
9	-0.000297658	-1.56903e-006	-4.25998e-006	6.04884e-008	-3.

Built In Test

```
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
Wald Test:    Chi-squared [   2]        =       596.303  Prob C2 > C2* =  .00000
F Test:       F ratio[ 2, 4154]         =       298.152  Prob F  > F*  =  .00000
```

LWAGE	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence 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.

Robust Covariance Matrix

The White Estimator

$$\text{Est.Var}[\mathbf{b}] = (\mathbf{X}'\mathbf{X})^{-1}\left[\sum\nolimits_i e_i^2 \mathbf{x}_i \mathbf{x}_i'\right](\mathbf{X}'\mathbf{X})^{-1}$$

- What does robustness mean?
- Robust to: Heteroscedasticty
- Not robust to:
  - Autocorrelation
  - Individual heterogeneity
  - The wrong model specification
- ‘Robust inference’

Robust Covariance Matrix

```
----------------------------------------------------------------------------
Ordinary      least squares regression ............
LHS=LWAGE     Mean                    =       6.67635
              Standard deviation      =        .46151
              Number of observs.      =          4165
Model size    Parameters              =            11
              Degrees of freedom      =          4154
Residuals     Sum of squares          =       515.950
              Standard error of e     =        .35243
Fit           R-squared               =        .41826
              Adjusted R-squared      =        .41686
Model test    F[ 10,   4154] (prob) =    298.7(.0000)
White heteroscedasticity robust covariance matrix.
Br./Pagan LM Chi-sq [ 10]  (prob) = 105.71 (.0000)
```

LWAGE	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	5.24547***	.07567	69.32	.0000	5.09715	5.39379
ED	.05654***	.00273	20.71	.0000	.05119	.06189
EXP	.04045***	.00219	18.46	.0000	.03616	.04474
EXP*EXP	-.00068***	.4893D-04	-13.92	.0000	-.00078	-.00059
WKS	.00449***	.00116	3.85	.0001	.00220	.00677
OCC	-.14053***	.01508	-9.32	.0000	-.17009	-.11098
SOUTH	-.07210***	.01274	-5.66	.0000	-.09707	-.04714
SMSA	.13901***	.01200	11.59	.0000	.11550	.16252
MS	.06736***	.02099	3.21	.0013	.02622	.10849
FEM	-.38922***	.02395	-16.25	.0000	-.43617	-.34227
UNION	.09015***	.01246	7.23	.0000	.06572	.11458

nnnnn.D-xx or D+xx => multiply by 10 to -xx or +xx.
***, **, * ==> Significance at 1%, 5%, 10% level.

Uncorrected

Standard Error	z
.07170	73.15
.00261	21.64
.00217	18.61
.4783D-04	-14.24
.00109	4.12
.01472	-9.54
.01249	-5.77
.01207	11.51
.02063	3.26
.02518	-15.46
.01289	6.99

Bootstrapping

Estimating the Asymptotic Variance of an Estimator

- Known form of asymptotic variance: Compute from known results
- Unknown form, known generalities about properties: Use bootstrapping
  - Root N consistency
  - Sampling conditions amenable to central limit theorems
  - Compute by resampling mechanism within the sample.

Bootstrapping

Method:

1. Estimate parameters using full sample: → **b**
2. Repeat R times:

   Draw n observations from the n, <u>with replacement</u>

   Estimate β with **b**(r).
3. Estimate variance with

   $$\mathbf{V} = (1/R)\Sigma_r\, [\mathbf{b}(r) - \mathbf{b}][\mathbf{b}(r) - \mathbf{b}]'$$

(Some use mean of replications instead of **b**. Advocated (without motivation) by original designers of the method.)

Application: Correlation between Age and Education

Current Command

Command:

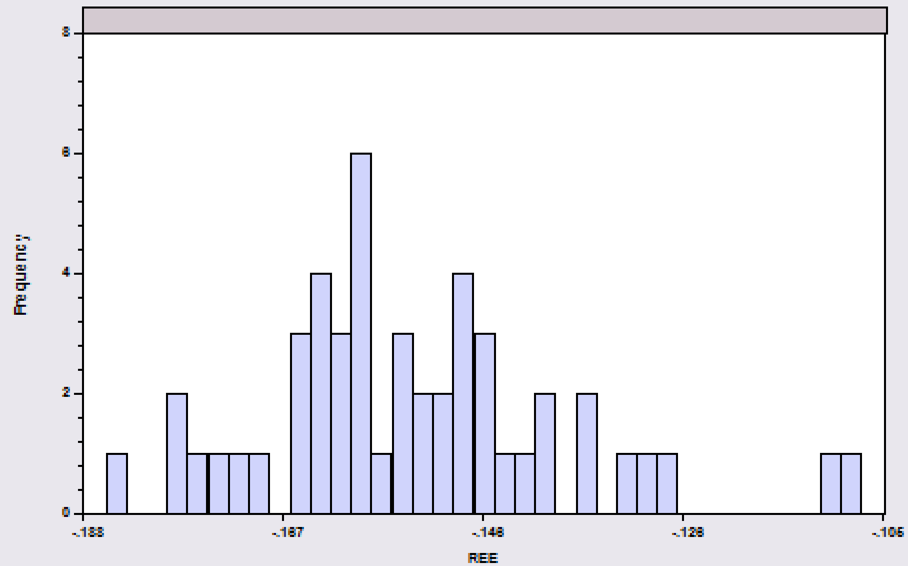
```
--------+--------------------------------------------------
     REE|    -.13459***       .01943     -6.9
--------+--------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5
-----------------------------------------------------------


Maximum repetitions of PROC

|-> Exec ; n = 50 ; Bootstrap = ree ; histo
Completed    50 bootstrap iterations.

-----------------------------------------------------------
Results of bootstrap estimation of model.
Model has been reestimated       50 times.
The statistics shown below are centered
around the  original estimate  based on
the original full sample of observations.
Result is REE       =          -.15299
Bootstrap samples have 3377 observations.
Estimate  RtMnSqDev  Skewness   Kurtosis
 -.15299     .01819     .90437    4.10740
Minimum =    -.18790  Maximum =    -.09821
--------+-----------------------------------------------------------------
        |                 Standard            Prob.      95% Confidence
BootStrp|  Coefficient      Error       z     |z|>Z*        Interval
--------+-----------------------------------------------------------------
     REE|    -.15299***       .01819     -8.41  .0000      -.18863   -.11734
--------+-----------------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
--------------------------------------------------------------------------
```

Untitled Plot 2 *



Untitled 1 *

Insert Name:

```
Proc$
Calc ; ree = Cor(age,educ) $
EndProc$
Exec ; n = 50 ; Bootstrap = ree ; histogram$
```

Bootstrap Regression - Replications

namelist;x=one,y,pg$	**Define X**
regress;lhs=g;rhs=x$	**Compute and display b**
proc	**Define procedure**
regress;quietly;lhs=g;rhs=x$	**... Regression (silent)**
endproc	**Ends procedure**
execute;n=20;bootstrap=b$	**20 bootstrap reps**
matrix;list;bootstrp $	**Display replications**

Results of Bootstrap Procedure

```
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  t-ratio  P[|T|>t]   Mean of X
--------+-------------------------------------------------------------
Constant|   -79.7535***      8.67255       -9.196   .0000
       Y|      .03692***      .00132       28.022   .0000      9232.86
      PG|   -15.1224***      1.88034       -8.042   .0000      2.31661
--------+-------------------------------------------------------------
Completed    20 bootstrap iterations.
----------------------------------------------------------------------
Results of bootstrap estimation of model.
Model has been reestimated    20 times.
Means shown below are the means of the
bootstrap estimates. Coefficients shown
below are the original estimates based
on the full sample.
bootstrap samples have   36 observations.
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+-------------------------------------------------------------
    B001|   -79.7535***      8.35512       -9.545   .0000     -79.5329
    B002|      .03692***      .00133       27.773   .0000       .03682
    B003|   -15.1224***      2.03503       -7.431   .0000     -14.7654
--------+-------------------------------------------------------------
```

Bootstrap Replications

Matrix - Result

[21, 3] Cell:

	1	2	3
1	-79.7535	0.0369204	-15.1224
2	-79.7751	0.0372034	-15.8164
3	-74.4476	0.0362466	-13.7959
4	-95.5803	0.0398037	-20.0141
5	-71.3427	0.0357651	-13.5814
6	-73.1011	0.0356458	-13.1219
7	-72.5021	0.0351552	-11.5075
8	-76.4406	0.0362488	-14.164
9	-77.2569	0.0361277	-13.5284
10	-100.156	0.0399487	-18.7463
11	-75.267	0.0361851	-13.6539
12	-79.4569	0.0366386	-14.0377
13	-82.6841	0.0379192	-18.0799
14	-74.2405	0.0357758	-12.9962
15	-80.2597	0.0369627	-15.2569
16	-75.3873	0.0366071	-14.9952
17	-74.066	0.0359726	-13.6492
18	-69.3163	0.0357294	-15.186
19	-86.3477	0.0376877	-14.9584
20	-95.5345	0.0388132	-15.1778
21	-77.4944	0.0359977	-13.0415

Full sample result

Bootstrapped sample results

Multiple Imputation for Missing Data

The template application of MI can be drawn with reference to a model

$$y = f(x1, x2|\boldsymbol{\beta})$$

where $\boldsymbol{\beta}$ is the parameter vector to be estimated. We suppose that there are n observations in the sample, $n_{c,1}$ complete observations on $x1$, $n_{m,1}$ missing values for $x1$, and $n_{c,2}$ and $n_{m,2}$ complete and missing observations on $x2$. The missing and complete observations on $x1$ and $x2$ need not coincide. We suppose as well that there is additional information in the sample, $\mathbf{Z}$, for which there are observations present for at least some observations when there are missing observations on $x1$ or $x2$.

Imputed Covariance Matrix

The overall approach of MI is to use available information on $x2$ and $\mathbf{Z}$ to predict missing values of $x1$ and available information on $x1$ and $\mathbf{Z}$ to predict missing values of $x2$. It is assumed that the missing values are 'missing at random,' that is, that the data on $x2$ and $\mathbf{Z}$ do not contain information on the probability that $x1$ is missing, and likewise for $x1$ and $\mathbf{Z}$ for $x2$. The three steps listed above are carried out as follows:

Step 1. Construct imputation equations $\hat{x}1 = h_1(x2, Z, \hat{\delta}_1)$ and $\hat{x}2 = h_2(x1, Z, \hat{\delta}_2)$ using available complete observations on relevant variables.

Step 2. (M repetitions): Simulate missing values of $x1$ from the conditional model h_1 and missing values of x_2 from the conditional model h_2. For each repetition, we obtain estimates of the parameters, $\hat{\beta}_m$ and the asymptotic covariance matrix $\hat{\Sigma}_m$.

Step 3. (Aggregation). The estimator of $\boldsymbol{\beta}$ is $\overline{\mathbf{b}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m$. The variance estimator is

$$\overline{\mathbf{S}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\Sigma}_m + \left(1 + \frac{1}{M}\right)\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}_m - \overline{\mathbf{b}}\right)\left(\hat{\beta}_m - \overline{\mathbf{b}}\right)'$$

Implementation

- SAS, Stata: Create full data sets with imputed values inserted. M = 5 is the familiar standard number of imputed data sets.
- NLOGIT/LIMDEP
  - Create an internal map of the missing values and a set of engines for filling missing values
  - Loop through imputed data sets during estimation.
  - M may be arbitrary – memory usage and data storage are independent of M.