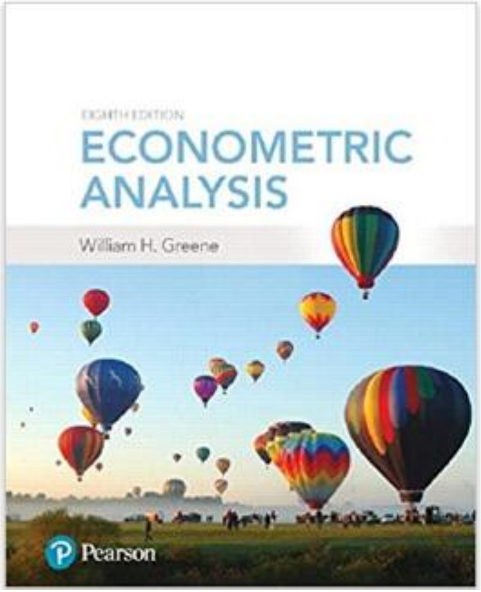
Econometrics I

Professor William Greene
Stern School of Business
Department of Economics



Econometrics I

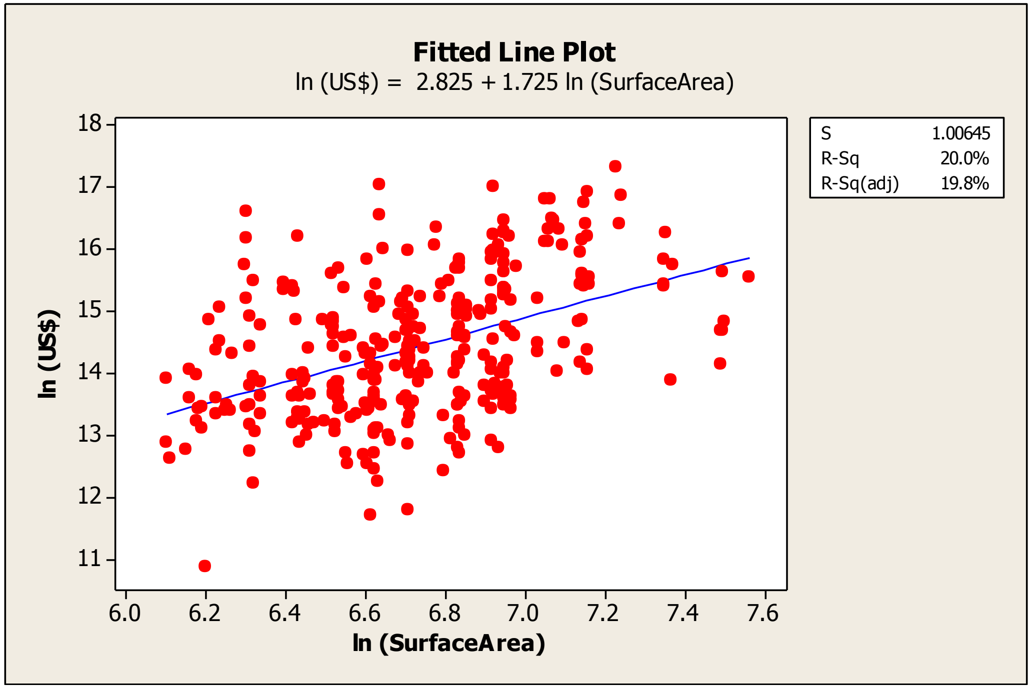
Part 6 – Dummy Variables and Functional Form

Agenda

- Dummy variables
- Interaction
- Categorical variables and transition tables
- Nonlinear functional form
- Differences
- Difference in differences
- Regression discontinuity
- Kinked regression

Monet in Large and Small

Sale prices of 328 signed Monet paintings



Log of $price = a + b log surface area + e

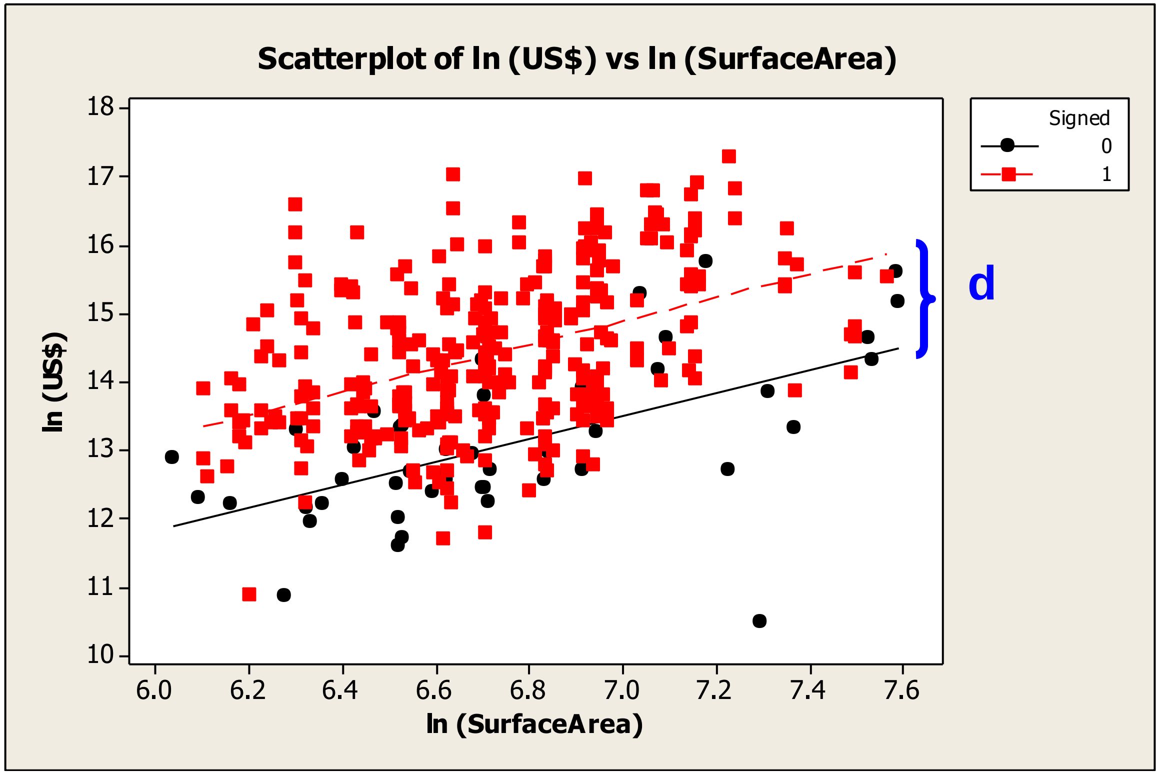
How Much for the Signature?

- The sample also contains 102 unsigned paintings

	Average Sale Price
Signed	$3,364,248
Not signed	$1,832,712

- Average price of a signed Monet is almost twice that of an unsigned one.

A Multiple Regression



Ln Price = a + b × ln Area + d × (0 if unsigned, 1 if signed) + e

Monet Multiple Regression

```
Regression Analysis: ln (US$) versus ln (SurfaceArea), Signed
The regression equation is
ln (US$) = 4.12 + 1.35 ln (SurfaceArea) + 1.26 Signed
Predictor            Coef  SE Coef      T      P
Constant           4.1222   0.5585   7.38  0.000
ln (SurfaceArea)   1.3458   0.08151 16.51  0.000
Signed             1.2618   0.1249  10.11  0.000
S = 0.992509   R-Sq = 46.2%   R-Sq(adj) = 46.0%
```

Interpretation:

(1) Elasticity of price with respect to surface area is 1.3458 – very large

(2) The signature multiplies the price of a painting by exp(1.2618) (about 3.5), for any given size.

A Conspiracy Theory for Art Sales at Auction

Sotheby's and Christies, 1995 to about 2000 conspired on commission rates.

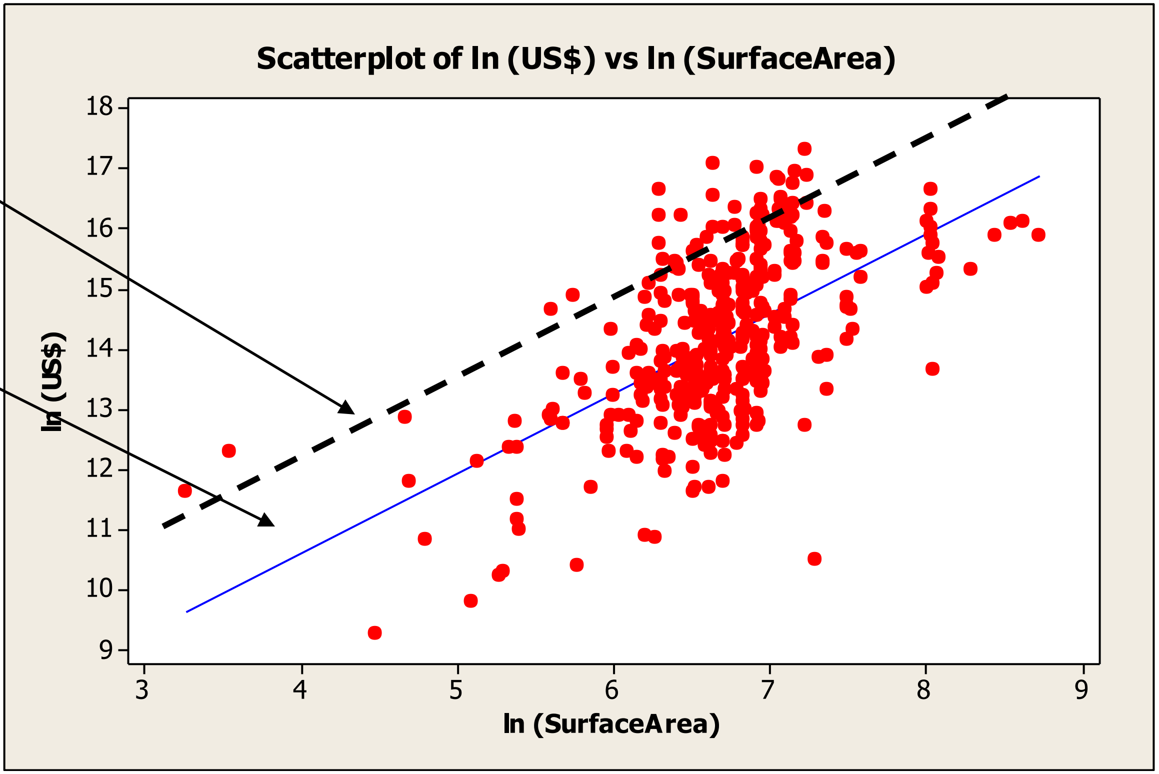
MONET.MTP ***

↓	C19-T
	Auction House
65	Christie's New York: Wednesday, May 9, 2001
66	Christie's New York: Wednesday, May 9, 2001
67	Christie's New York: Wednesday, May 9, 2001
68	Christie's New York: Wednesday, May 9, 2001
69	Phillips, de Pury & Luxembourg New York: Monday, May 7, 2001
70	Christie's London: Wednesday, February 7, 2001
71	Sotheby's London: Tuesday, February 6, 2001
72	Sotheby's London: Monday, February 5, 2001
73	Sotheby's London: Monday, February 5, 2001
74	Sotheby's New York: Thursday, November 9, 2000
75	Sotheby's New York: Thursday, November 9, 2000
76	Sotheby's New York: Thursday, November 9, 2000
77	Sotheby's New York: Thursday, November 9, 2000
78	Christie's New York: Wednesday, November 8, 2000
79	Christie's New York: Wednesday, November 8, 2000
80	Christie's New York: Wednesday, November 8, 2000
81	Christie's New York: Wednesday, November 8, 2000
82	Christie's New York: Wednesday, November 8, 2000
83	Christie's London: Wednesday, June 28, 2000
84	Christie's London: Wednesday, June 28, 2000
85	Sotheby's London: Tuesday, June 27, 2000
86	Sotheby's London: Tuesday, June 27, 2000
87	Sotheby's New York: Wednesday, May 10, 2000
88	Sotheby's New York: Wednesday, May 10, 2000

If the Theory is Correct…

Sold from 1995 to 2000

Sold before 1995 or after 2000



Evidence

```
The regression equation is
ln (US$) = 4.03 + 1.35 ln (SurfaceArea) + 1.28 Signed
                  + 0.201 conspiracy
Predictor            Coef   SE Coef      T      P
Constant           4.0270    0.5585   7.21  0.000
ln (SurfaceArea)  1.34756   0.08122  16.59  0.000
Signed             1.2777    0.1247  10.25  0.000
conspiracy         0.2009    0.1001   2.01  0.045
S = 0.989012   R-Sq = 46.7%   R-Sq(adj) = 46.3%
Analysis of Variance
Source            DF      SS      MS       F      P
Regression         3  365.44  121.81  124.53  0.000
Residual Error   426  416.69    0.98
```

The statistical evidence seems to be consistent with the theory.

Effects on Price	Unsigned	Signed
Not 1995 - 2000	$\exp(0.0000) = 1.0000$	$\exp(1.2777) = 3.5884$
1995 - 2000	$\exp(0.2009) = 1.2225$	$\exp(1.2777 + 0.2009) = 4.3868$

Women appear to assess health satisfaction differently from men.

Descriptive Statistics for HLTHSAT
Stratification is based on FEMALE

Subsample	Mean	Std.Dev.	Cases	Sum of wts	Missing
FEMALE = 0	6.922699	2.251837	14243	14243.00	0
FEMALE = 1	6.633417	2.329590	13083	13083.00	0
Full Sample	6.784198	2.293907	27326	27326.00	0

Least squares regression ....
LHS=HLTHSAT Mean = 6.78420
Standard deviation = 2.29391

				DegFreedom	Mean square
	No. of observations	=	27326		
Regression	Sum of Squares	=	570.655	1	570.65542
Residual	Sum of Squares	=	143214.	27324	5.24132
Total	Sum of Squares	=	143784.	27325	5.26201
	Standard error of e	=	2.28939	Root MSE	2.28931
Fit	R-squared	=	.00397	R-bar squared	.00393
Model test	F[1, 27324]	=	108.87633	Prob F > F*	.00000

HLTHSAT	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	6.92270***	.01918	360.87	.0000	6.88510	6.96030
FEMALE	-.28928***	.02772	-10.43	.0000	-.34362	-.23494

Or do they? Not when other things are held constant

```
----------------------------------------------------------------------------
Least squares regression ............
LHS=HLTHSAT  Mean                    =        6.78420
             Standard deviation      =        2.29391
----------   No. of observations     =          27326  DegFreedom   Mean square
Regression   Sum of Squares          =        10755.6          7   1536.51901
Residual     Sum of Squares          =        133029.      27318      4.86964
Total        Sum of Squares          =        143784.      27325      5.26201
----------   Standard error of e     =        2.20673  Root MSE       2.20640
Fit          R-squared               =         .07480  R-bar squared   .07457
Model test   F[  7, 27318]           =      315.53041  Prob F > F*     .00000
--------+-------------------------------------------------------------------
        |                Standard            Prob.       95% Confidence
 HLTHSAT|  Coefficient     Error       z    |z|>Z*          Interval
--------+-------------------------------------------------------------------
Constant|    7.21588***    .10583    68.19  .0000     7.00846     7.42330
  FEMALE|    -.02248       .02936     -.77  .4438     -.08003      .03506
     AGE|    -.04118***    .00138   -29.82  .0000     -.04389     -.03848
    EDUC|     .07740***    .00617    12.54  .0000      .06531      .08950
  HHNINC|     .48500***    .08169     5.94  .0000      .32490      .64511
 MARRIED|     .07108**     .03509     2.03  .0428      .00230      .13986
  HHKIDS|     .13925***    .03150     4.42  .0000      .07751      .20100
 WORKING|     .31704***    .03269     9.70  .0000      .25297      .38110
--------+-------------------------------------------------------------------
```

Dummy Variable for One Observation

A dummy variable that isolates a single observation. What does this do?

Define **d** to be the dummy variable in question.

Z = all other regressors. **X** = [**Z**,**d**]

Multiple regression of **y** on **X**. We know that

X'e = **0** where **e** = the column vector of

residuals. That means **d'e** = 0, which says that e_j = 0 for that particular residual. The observation will be predicted perfectly.

Fairly important result. Important to know.

I have a simple question for you. Yesterday, I was estimating a regional production function with yearly dummies. The coefficients of the dummies are usually interpreted as a measure of technical change with respect to the base year (excluded dummy variable). However, I felt that it could be more interesting to redefine the dummy variables in such a way that the coefficient could measure technical change from one year to the next. You could get the same result by subtracting two coefficients in the original regression but you would have to compute the standard error of the difference if you want to do inference.

Is this a well known procedure? YES

Ordinary least squares regression ....
LHS=LWAGE Mean = 6.67635

LWAGE	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	5.53761***	.03107	178.26	.0000	5.47672	5.59849
YEAR2	.09004***	.02188	4.12	.0000	.04716	.13291
YEAR3	.22154***	.02188	10.13	.0000	.17867	.26442
YEAR4	.32091***	.02188	14.67	.0000	.27803	.36378
YEAR5	.41128***	.02188	18.80	.0000	.36840	.45416
YEAR6	.48887***	.02188	22.35	.0000	.44600	.53175
YEAR7	.57557***	.02188	26.31	.0000	.53270	.61845
ED	.06520***	.00210	31.09	.0000	.06109	.06931

Constant	5.53761***	.03107	178.26	.0000	5.47672	5.59849
Q2	.09004***	.02188	4.12	.0000	.04716	.13291
Q3	.13150***	.02188	6.01	.0000	.08863	.17438
Q4	.09936***	.02188	4.54	.0000	.05649	.14224
Q5	.09037***	.02188	4.13	.0000	.04750	.13325
Q6	.07759***	.02188	3.55	.0004	.03472	.12047
Q7	.08670***	.02188	3.96	.0001	.04382	.12958
ED	.06520***	.00210	31.09	.0000	.06109	.06931

Example with 4 Periods

The estimated model with time dummies is

$\mathbf{y} = a + b_2{*}\mathbf{d}_2 + b_3{*}\mathbf{d}_3 + b_4{*}\mathbf{d}_4 + \mathbf{e}$ (possibly some other variables, not needed now).

Estimated least squares coefficients are

$\mathbf{b} = a, b_2, b_3, b_4$

Desired coefficients are

$\mathbf{c} = a, b_2, b_3 - b_2, b_4 - b_3$

The original model is $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$.

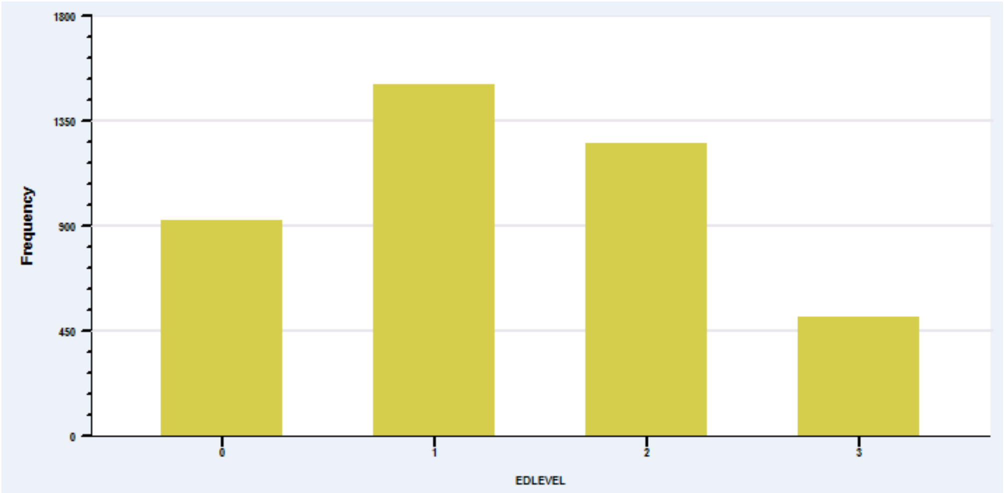
The new model would be $\mathbf{y} = (\mathbf{XC})(\mathbf{C}^{-1}\mathbf{b}) + \mathbf{e} = \mathbf{Qc} + \mathbf{e}$

The transformation of the data is $\mathbf{Q} = \mathbf{XC}$. $\mathbf{c} = \mathbf{C}^{-1}\mathbf{b}$

The transformed $\mathbf{X}$ is $[1, d_2+d_3+d_4, d_3+d_4 . d_4]$

$$\mathbf{C}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

A Categorical Variable




```
--------------------------------------------------------------------------------
Ordinary    least squares regression ............
LHS=LWAGE   Mean                    =      6.67635
            Standard deviation      =       .46151
----------  No. of observations     =         4165  DegFreedom   Mean square
Regression  Sum of Squares          =      122.335           3      40.77838
Residual    Sum of Squares          =      764.570        4161        .18375
Total       Sum of Squares          =      886.905        4164        .21299
----------  Standard error of e     =       .42866  Root MSE          .42845
Fit         R-squared               =       .13793  R-bar squared     .13731
Model test  F[  3,  4161]           =    221.92719  Prob F > F*       .00000
--------+-----------------------------------------------------------------------
        |                Standard            Prob.      95% Confidence
   LWAGE|  Coefficient     Error       z    |z|>Z*         Interval
--------+-----------------------------------------------------------------------
Constant|    6.45177***    .01416   455.78  .0000     6.42402   6.47951
--------+-----------------------------------------------------------------------
 EDLEVEL|  Base = 0
      1 |     .15176***    .01797     8.44  .0000      .11653    .18698
      2 |     .35319***    .01865    18.94  .0000      .31664    .38975
      3 |     .53167***    .02377    22.37  .0000      .48509    .57826
--------+-----------------------------------------------------------------------
```

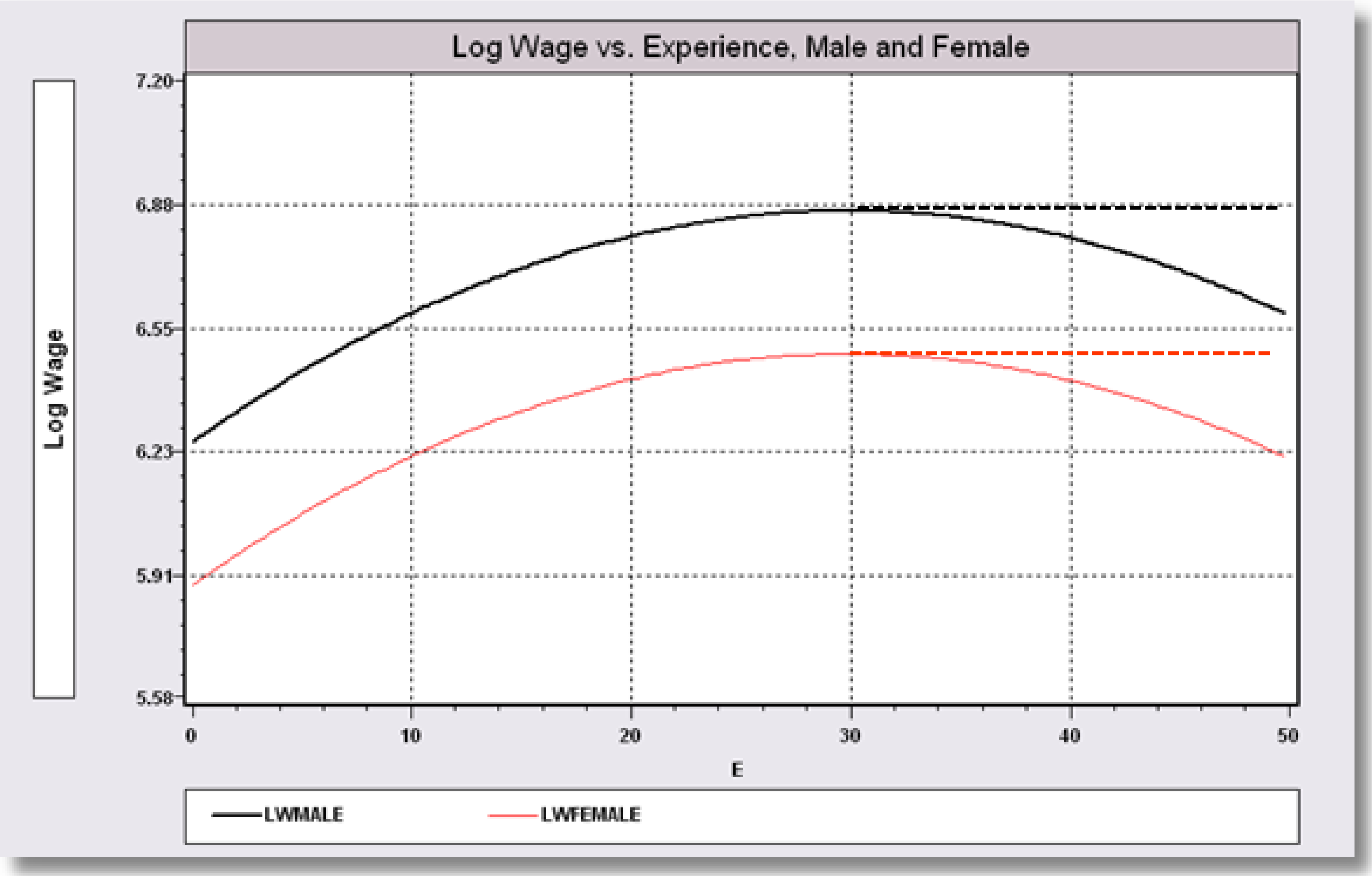
```
--------------------------------------------------------------------
Simulation and partial effects based on categorical variable EDLEVEL
Results computed by setting all observations to category value and
comparing to base value.
Sample proportions apply to full sample before @ settings in command
--------------------------------------------------------------------
Category Dummy  Sample | Fraction  Category
Base value   0     917     .22017  LTHS
             1    1498     .35966  HIGHSCHL
             2    1246     .29916  COLLEGE
             3     504     .12101  GRAD

--------------------------------------------------------------------
Partial Effects  Analysis for Linear Regression Function
--------------------------------------------------------------------
Effects of switches between categories in EDLEV=xx (dummy variables)
Results are computed by average over sample observations
LTHS    = .2202  HIGHSCHL= .3597   COLLEGE = .2992   GRAD    = .1210
--------------------------------------------------------------------
df/dEDLEV=xx        Partial    Standard
From --> To         Effect      Error       |t|  95% Confidence Interval
--------------------------------------------------------------------
LTHS     HIGHSCHL   .15176      .01797     8.44       .11653      .18698
LTHS     COLLEGE    .35319      .01865    18.94       .31664      .38975
LTHS     GRAD       .53167      .02377    22.37       .48509      .57826
HIGHSCHL LTHS      -.15176      .01797     8.44      -.18698     -.11653
HIGHSCHL COLLEGE    .20144      .01644    12.26       .16923      .23365
HIGHSCHL GRAD       .37992      .02207    17.21       .33665      .42318
COLLEGE  LTHS      -.35319      .01865    18.94      -.38975     -.31664
COLLEGE  HIGHSCHL  -.20144      .01644    12.26      -.23365     -.16923
COLLEGE  GRAD       .17848      .02263     7.89       .13413      .22283
GRAD     LTHS      -.53167      .02377    22.37      -.57826     -.48509
GRAD     HIGHSCHL  -.37992      .02207    17.21      -.42318     -.33665
GRAD     COLLEGE   -.17848      .02263     7.89      -.22283     -.13413
--------------------------------------------------------------------
```

Nonlinear 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.
----------------------------------------------------------------------
```

Model Implication: Effect of Experience and Male vs. Female



Partial Effect of Experience: Coefficients do not tell the story

```
-----------------------------------------------------------------------------
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
--------+--------------------------------------------------------------------
```

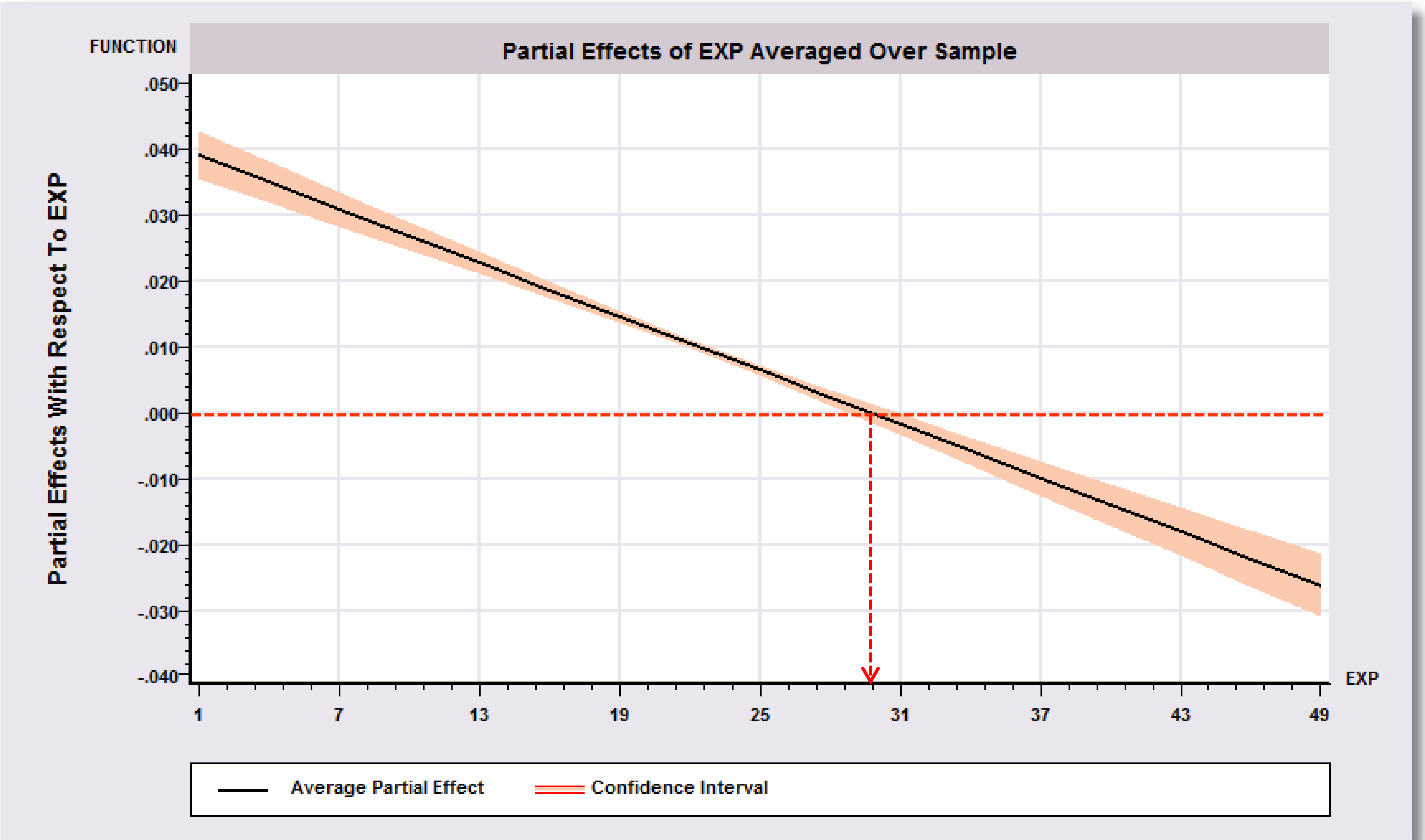
Constant	5.24547***
ED	.05654***
EXP	.04045***
EXP*EXP	-.00068***
WKS	.00449***
OCC	-.14053***
SOUTH	-.07210***
SMSA	.13901***
MS	.06736***
FEM	-.38922***
UNION	.09015***

Education: .05654

Experience: .04045 - 2*.00068*Exp

FEM: -.38922

Effect of Experience = .04045 - 2 * 0.00068*Exp
Positive from 1 to 30, negative after.



Specification and Functional Form: Nonlinearity

Population

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 z + \varepsilon$$

$$\delta_x = \frac{\partial E[y \mid x, z]}{\partial x} = \beta_2 + 2\beta_3 x$$

Estimators

$$\hat{y} = b_1 + b_2 x + b_3 x^2 + b_4 z$$

$$\hat{\delta}_x = b_2 + 2b_3 x$$

Log Income Equation

```
----------------------------------------------------------------------
Ordinary     least squares regression ............
LHS=LOGY     Mean                 =      -1.15746
             Standard deviation   =        .49149
             Number of observs.   =         27322
Model size   Parameters           =             7
             Degrees of freedom   =         27315
Residuals    Sum of squares       =    5462.03686
             Standard error of e  =        .44717
Fit          R-squared            =        .17237
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er.  P[|Z|>z]   Mean of X
--------+-------------------------------------------------------------
     AGE|     .06225***       .00213      29.189    .0000     43.5272
   AGESQ|    -.00074***    .242482D-04   -30.576    .0000     2022.99
Constant|   -3.19130***       .04567     -69.884    .0000
 MARRIED|     .32153***       .00703      45.767    .0000      .75869
  HHKIDS|    -.11134***       .00655     -17.002    .0000      .40272
  FEMALE|    -.00491          .00552       -.889    .3739      .47881
    EDUC|     .05542***       .00120      46.050    .0000     11.3202
--------+-------------------------------------------------------------
```

Estimated Cov[b1,b2]

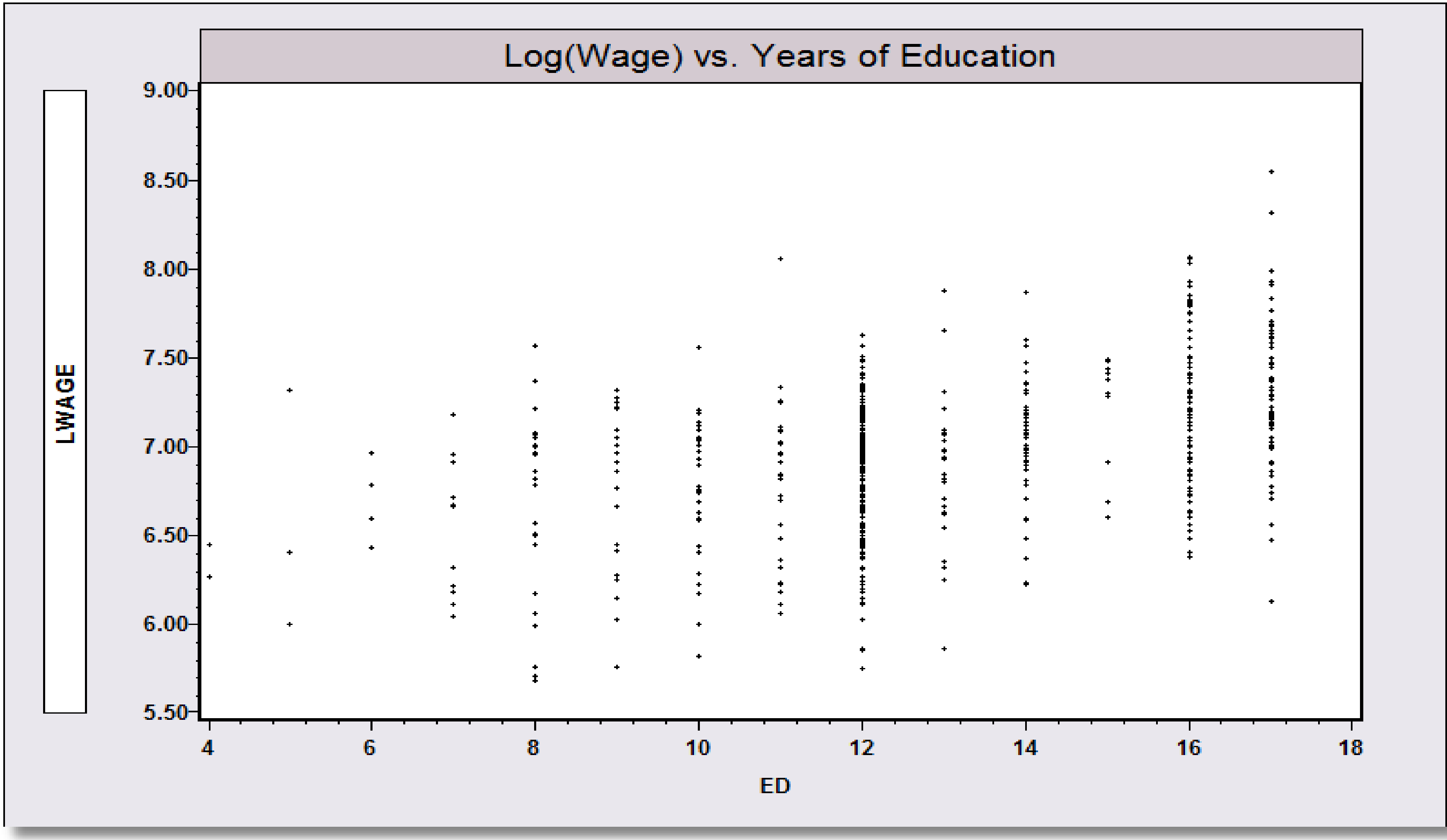
	1	2	
1	4.54799e-006	-5.1285e-008	-9
2	-5.1285e-008	5.87973e-010	9.9
3	-9.0034e-005	9.91407e-007	

Average Age = 43.5272. Estimated Partial effect = .066225 – 2(.00074)43.5272 = .00018.
Estimated Variance 4.54799e-6 + 4(43.5272)^2(5.87973e-10) + 4(43.5272)(-5.1285e-8)
= 7.4755086e-08. Estimated standard error = .00027341.

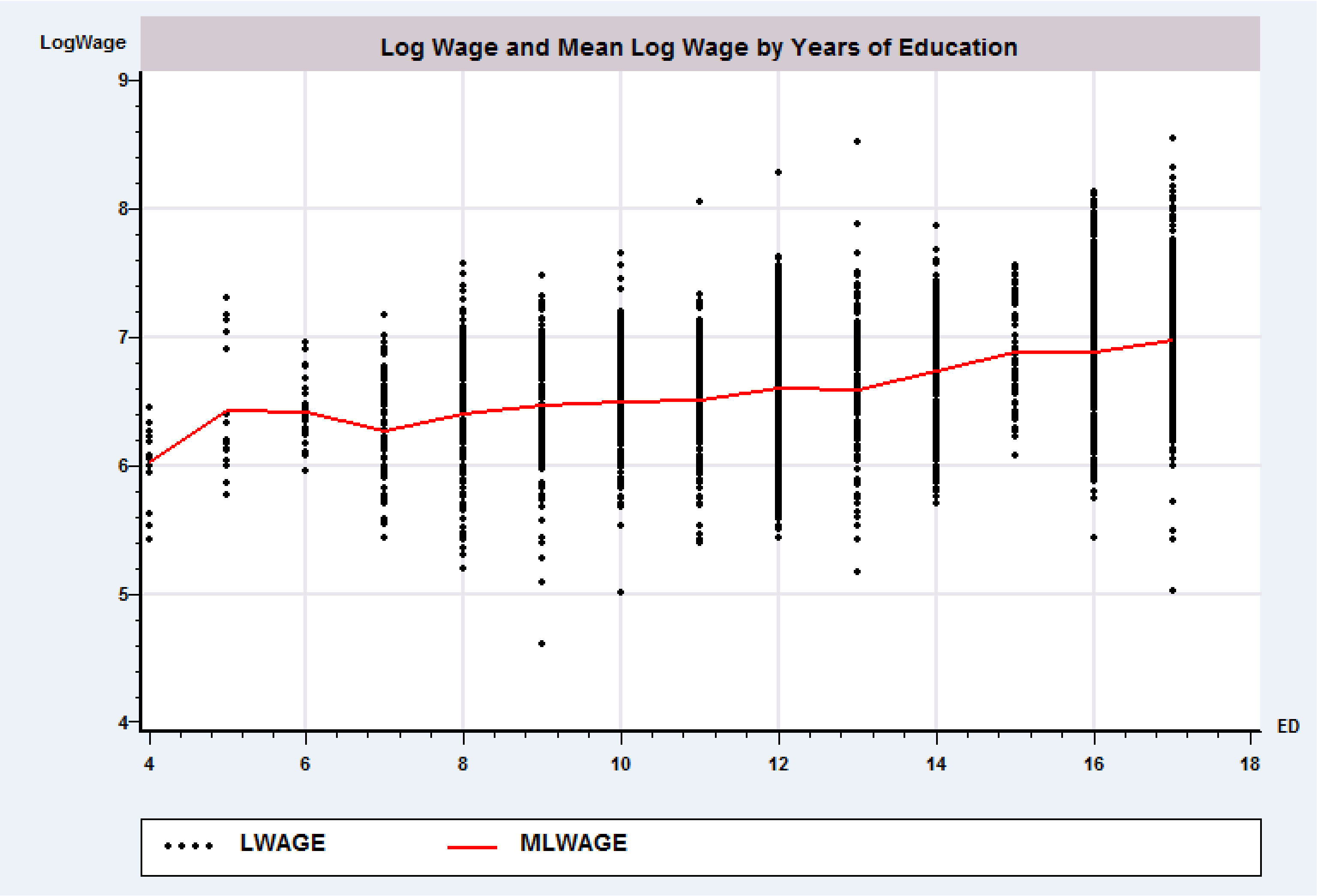
Objective: Impact of Education on (log) Wage

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

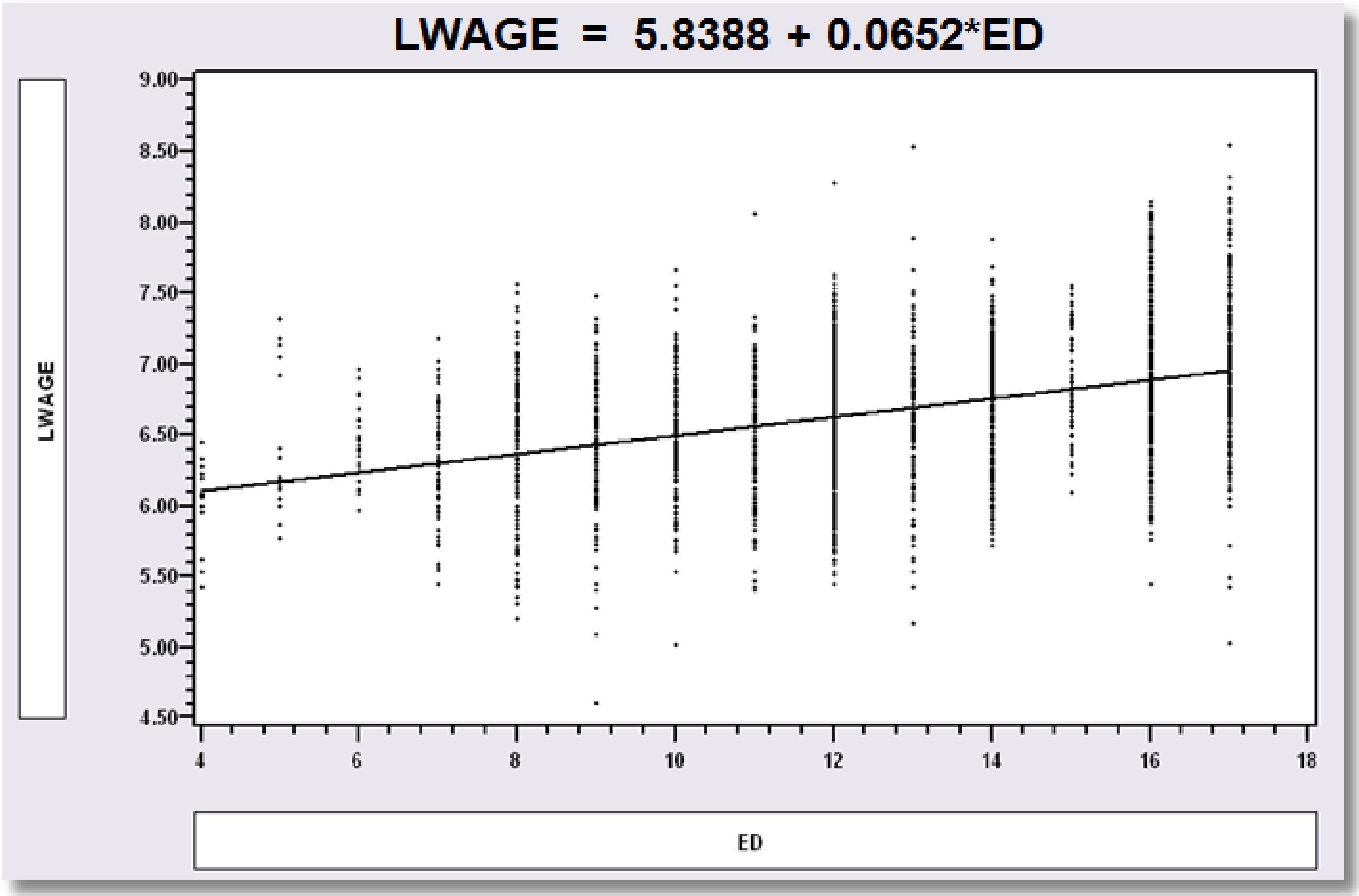
Application: Is there a relationship between (log) Wage and Education?



Group (Conditional) Means (Nonparametric)



Simple Linear Regression (semiparametric)



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
--------+--------------------------------------------------------------------
        |                  Standard            Prob.      95% Confidence
   LWAGE|  Coefficient       Error       z    |z|>Z*         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|      .08568***     .02108     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.
-----------------------------------------------------------------------------
```

Interaction Effect
Gender Difference in 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         =       347.213          10      34.72132
Residual     Sum of Squares         =       539.692        4154        .12992
Total        Sum of Squares         =       886.905        4164        .21299
----------   Standard error of e    =        .36045  Root MSE          .35997
Fit          R-squared              =        .39149  R-bar squared     .39002
Model test   F[ 10,   4154]         =     267.24949  Prob F > F*       .00000
--------+-------------------------------------------------------------------
        |                Standard            Prob.      95% Confidence
   LWAGE|  Coefficient     Error       z    |z|>Z*         Interval
--------+-------------------------------------------------------------------
Constant|   5.47075***     .07256    75.39  .0000     5.32853    5.61298
      ED|    .05458***     .00275    19.81  .0000      .04918     .05998
     EXP|    .01035***     .00054    19.29  .0000      .00930     .01140
     WKS|    .00528***     .00111     4.74  .0000      .00310     .00746
     OCC|   -.14659***     .01506    -9.73  .0000     -.17611    -.11707
   SOUTH|   -.07176***     .01278    -5.61  .0000     -.09682    -.04671
    SMSA|    .13351***     .01234    10.82  .0000      .10932     .15770
      MS|    .08392***     .02107     3.98  .0001      .04263     .12520
     FEM|   -.67961***     .09456    -7.19  .0000     -.86495    -.49427
   UNION|    .09496***     .01325     7.17  .0000      .06899     .12093
  ED*FEM|    .02350***     .00703     3.34  .0008      .00971     .03729
--------+-------------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------------
```

Partial Effect of a Year of Education

$\partial E[\log Wage]/\partial ED = \beta_{ED} + \beta_{ED*FEM} * FEM$

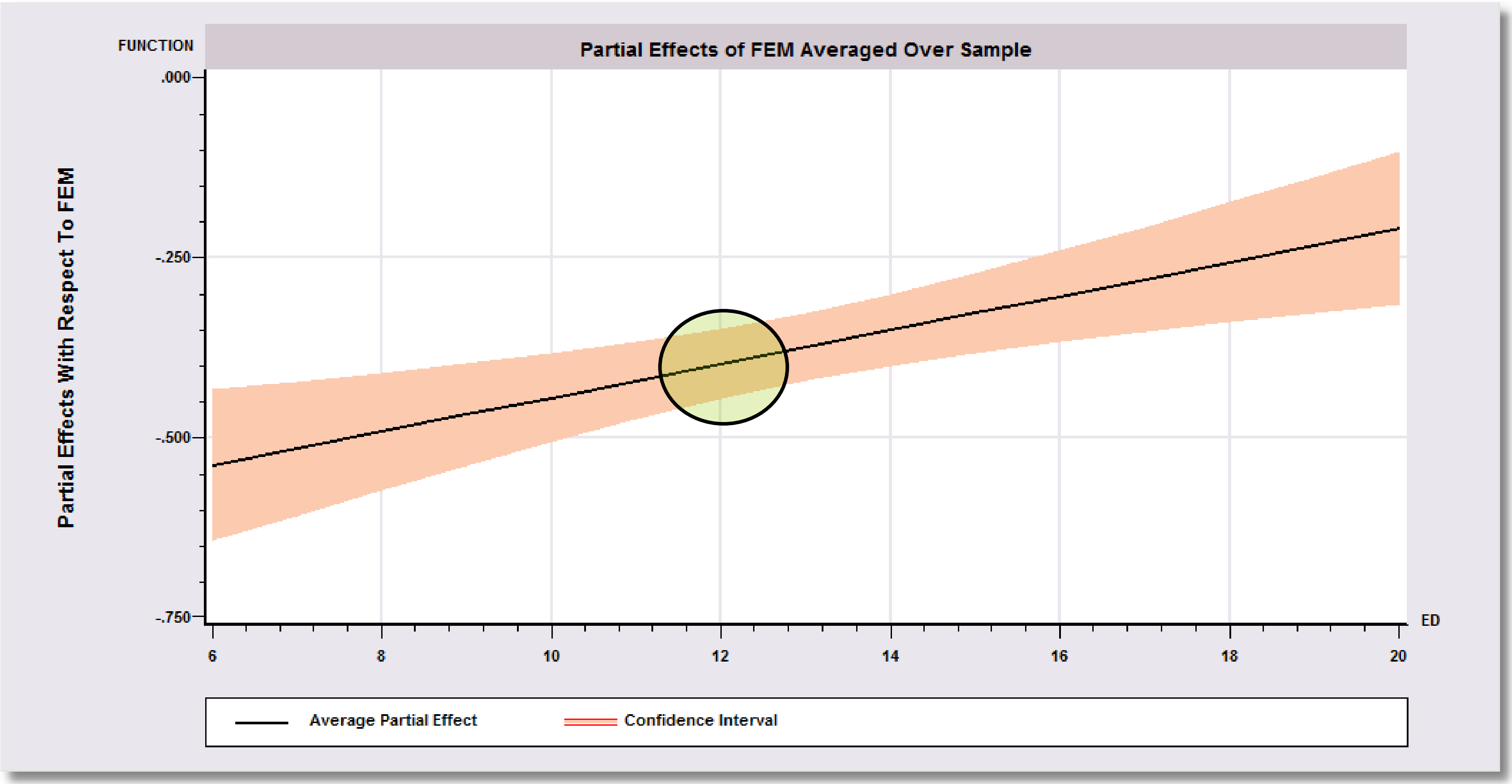
Note, the effect is positive.

Effect is larger for women.

```
--------------------------------------------------------------------------------
Partial Effects  Analysis for Linear Regression Function
--------------------------------------------------------------------------------
Effects on function with respect to ED
Results are computed by average over sample observations
Partial effects for continuous ED        computed by differentiation
Effect is computed as derivative      = df(.)/dx
--------------------------------------------------------------------------------
df/dED                 Partial     Standard
(Delta method)         Effect       Error      |t|   95% Confidence Interval
--------------------------------------------------------------------------------
APE. Function           .05723      .00267    21.40      .05199        .06247
--------------------------------------------------------------------------------
FEM      =    .00 ----------------------------------------------------------------
Average effect          .05458      .00275    19.81      .04918        .05998
--------------------------------------------------------------------------------
FEM      =   1.00 ----------------------------------------------------------------
Average effect          .07808      .00690    11.32      .06456        .09161
```

Gender Effect Varies by Years of Education

-0.67961 is misleading



Difference in Differences

With two periods,

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

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.

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

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,

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

Difference in Differences

$$
\begin{aligned}
y_{it} &= \beta_0 + \beta_1 D_{it} + \beta_2 T_t + \beta_3 D_{it} T_t + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it}, t = 1,2 \\
\Delta y_{it} &= \beta_2 + \beta_3 D_{i2} + \Delta(\boldsymbol{\beta}'\mathbf{x}_{it}) + \Delta\varepsilon_{it} \\
&= \beta_2 + \beta_3 D_{i2} + \boldsymbol{\beta}'(\Delta\mathbf{x}_{it}) + u_i
\end{aligned}
$$

$$
\begin{aligned}
&\left(\Delta y_{it} \mid D = 1\right) - \left(\Delta y_{it} \mid D = 0\right) \\
&\quad = \beta_3 + \boldsymbol{\beta}'\left[(\Delta\mathbf{x}_{it} \mid D = 1) - (\Delta\mathbf{x}_{it} \mid D = 0)\right]
\end{aligned}
$$

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.

SAT Tests

Example 6.8 *SAT Scores*

Each year, about 1.7 million American high school students take the SAT test. Students who are not satisfied with their performance have the opportunity to retake the test. Some students take an SAT prep course, such as Kaplan or Princeton Review, before the second attempt in the hope that it will help them increase their scores. An econometric investigation might consider whether these courses are effective in increasing scores. The investigation might examine a sample of students who take the SAT test twice, with scores y_{i0} and y_{i1}. The time dummy variable T_t takes value $T_0 = 0$ "before" and $T_1 = 1$ "after." The treatment dummy variable is $D_i = 1$ for those students who take the prep course and 0 for those who do not. The applicable model would be (6-3),

$$SAT\ Score_{i,t} = \beta_1 + \beta_2\, 2ndTest_t + \beta_3\, PrepCourse_i + \delta\, 2ndTest_t \times PrepCourse_i + \varepsilon_{i,t}.$$

The estimate of δ would, in principle, be the treatment, or prep course effect.

Using least squares,

$$d_3 = (\overline{Score_2} - \overline{Score_1})_{TestPrep=1} - (\overline{Score_2} - \overline{Score_1})_{TestPrep=0}$$

Potential **x** = Income, Parents' Education, GPA

Abrupt Effect on Regression at a Specific Level of x

Figure 6.6 Regression Discontinuity.

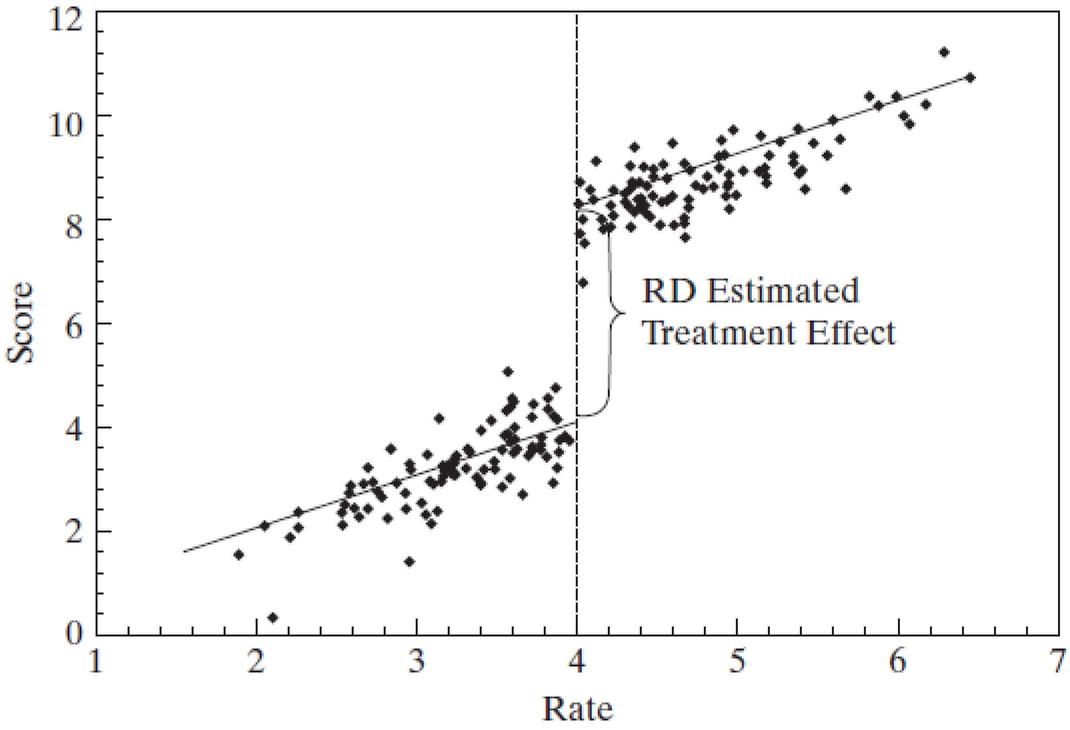


Figure 6.8 Regression Discontinuity Design for Mortgage Demand.

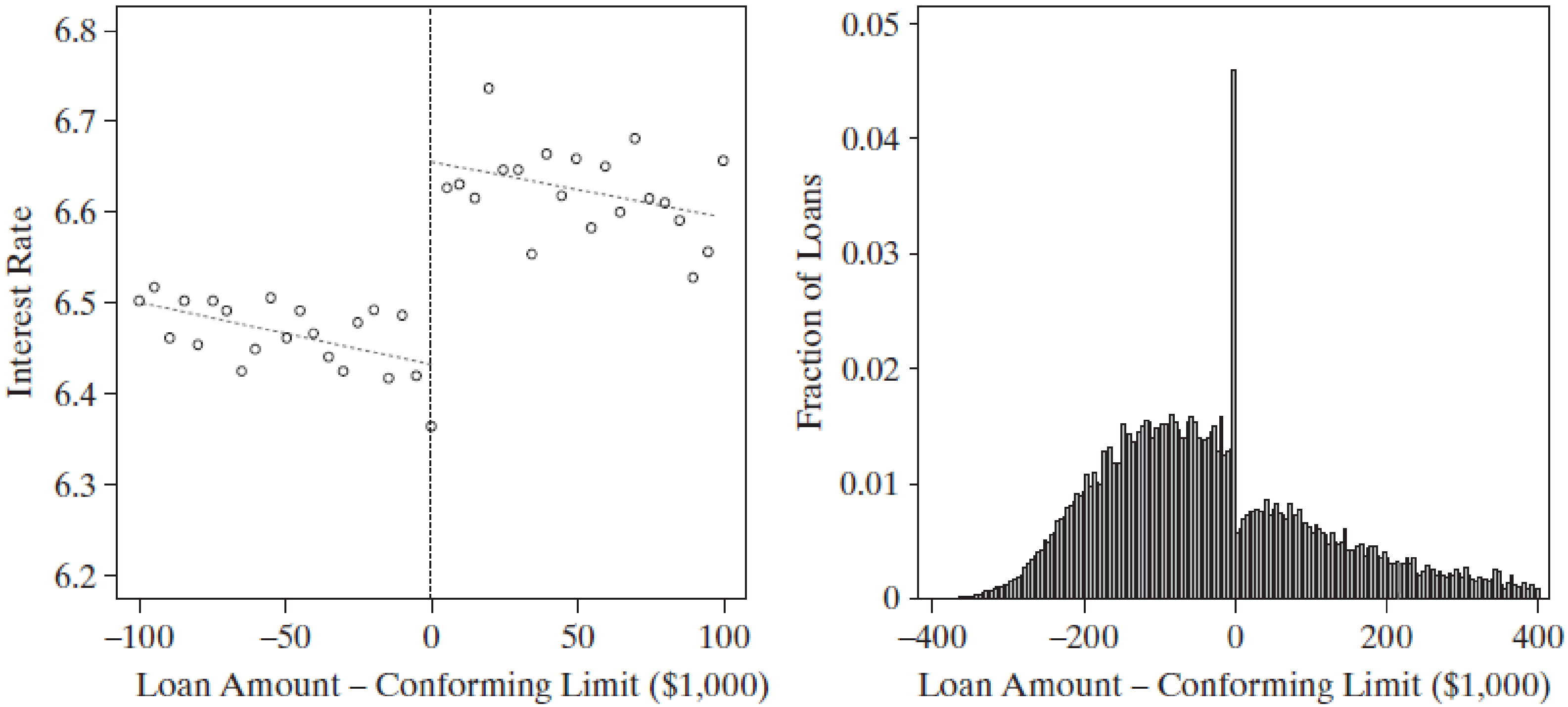
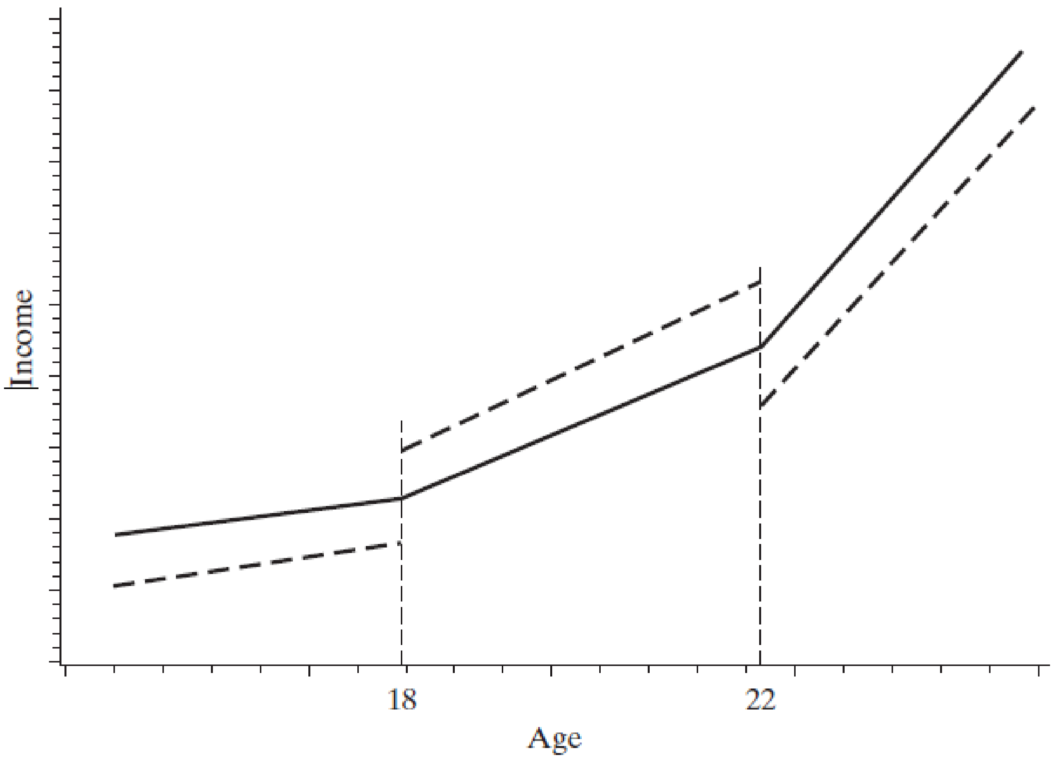


FIG. 2.—Mean Interest Rate Relative to the Conforming Limit, Fixed-Rate Mortgages Only (2006). This figure plots the mean interest rate for fixed rate mortgages originated in 2006 as a function of the loan amount relative to the conforming limit. Each dot represents the mean interest rate within a given $5,000 bin relative to the limit. The dashed lines are predicted values from a regression fit to the binned data allowing for changes in the slope and intercept at the conforming limit. Sample includes all loans in the LPS fixed-rate sample that fall within $100,000 of the conforming limit. See text for details on sample construction.

FIG. 3.—Loan Size Distribution Relative to the Conforming Limit. This figure plots the fraction of all loans that are in any given $5,000 bin relative to the conforming limit. Data are pooled across years and each loan is centered at the conforming limit in effect at the date of origination, so that a value of 0 represents a loan at exactly the conforming limit. Sample includes all transactions in the primary DataQuick sample that fall within $400,000 of the conforming limit. See text for details on sample construction.

Useful Functional Form: Kinked Regression

Figure 6.4 Piecewise Linear Regression.



effect. The function we wish to estimate is

$$
\begin{aligned}
E[income \mid age] = \alpha^0 + \beta^0\, age &\quad \text{if } age < 18,\\
\alpha^1 + \beta^1\, age &\quad \text{if } age \geq 18 \text{ and } age < 22,\\
\alpha^2 + \beta^2\, age &\quad \text{if } age \geq 22.
\end{aligned}
$$

Let

$$
\begin{aligned}
d_1 &= 1 \quad \text{if } age \geq t_1^*,\\
d_2 &= 1 \quad \text{if } age \geq t_2^*,
\end{aligned}
$$

where $t_1^* = 18$ and $t_2^* = 22$. To combine the three equations, we use

$$income = \beta_1 + \beta_2\, age + \gamma_1 d_1 + \delta_1 d_1\, age + \gamma_2 d_2 + \delta_2 d_2\, age + \varepsilon.$$

This produces the dashed function Figure 6.4. The slopes in the three segments are $\beta_2, \beta_2 + \delta_1$, and $\beta_2 + \delta_1 + \delta_2$. To make the function *continuous*, we require that the segments join at the thresholds—that is,

$$
\begin{gathered}
\beta_1 + \beta_2 t_1^* = (\beta_1 + \gamma_1) + (\beta_2 + \delta_1) t_1^* \text{ and}\\
(\beta_1 + \gamma_1) + (\beta_2 + \delta_1) t_2^* = (\beta_1 + \gamma_1 + \gamma_2) + (\beta_2 + \delta_1 + \delta_2) t_2^*.
\end{gathered}
$$

These are linear restrictions on the coefficients. The first one is

$$\gamma_1 + \delta_1 t_1^* = 0 \quad \text{or} \quad \gamma_1 = -\delta_1 t_1^*.$$

Doing likewise for the second, we obtain

$$income = \beta_1 + \beta_2\, age + \delta_1 d_1 (age - t_1^*) + \delta_2 d_2 (age - t_2^*) + \varepsilon.$$

Constrained least squares estimates are obtainable by multiple regression, using a constant and the variables

$$
\begin{aligned}
x_1 &= age,\\
x_2 &= age - 18 \quad \text{if } age \geq 18 \text{ and } 0 \text{ othewise},\\
x_3 &= age - 22 \quad \text{if } age \geq 22 \text{ and } 0 \text{ othewise}.
\end{aligned}
$$

Kinked Regression and Policy Analysis: Unemployment Insurance

Example 6.12 Policy Analysis Using Kinked Regressions

Discontinuities such as those in Figure 6.4 can be used to help identify policy effects. Card, Lee, Pei, and Weber (2012) examined the impact of unemployment insurance (UI) on the duration of joblessness in Austria using a regression kink design. The policy lever, UI, has a sharply defined benefit schedule level tied to base year earnings that can be traced through to its impact on the duration of unemployment. Figure 6.5 [from Card et al. (2012, p. 48)]

Figure 6.5 Regression Kink Design.