A Cost Model

Worksheet 1 ***

↓	C1	C2	C3
	ID	Cost	Output
31	61	13.071	1627
32	62	24.154	3965
33	65	22.996	2682
34	67	17.330	2001
35	68	13.477	2764
36	69	19.464	2487
37	71	45.539	7320
38	75	25.848	3571
39	76	36.305	6837
40	77	22.183	2020
41	78	16.555	2445
42	80	24.833	3981
43	81	38.817	6770
44	82	33.088	4187
45	85	34.164	5643
46	87	27.644	6793

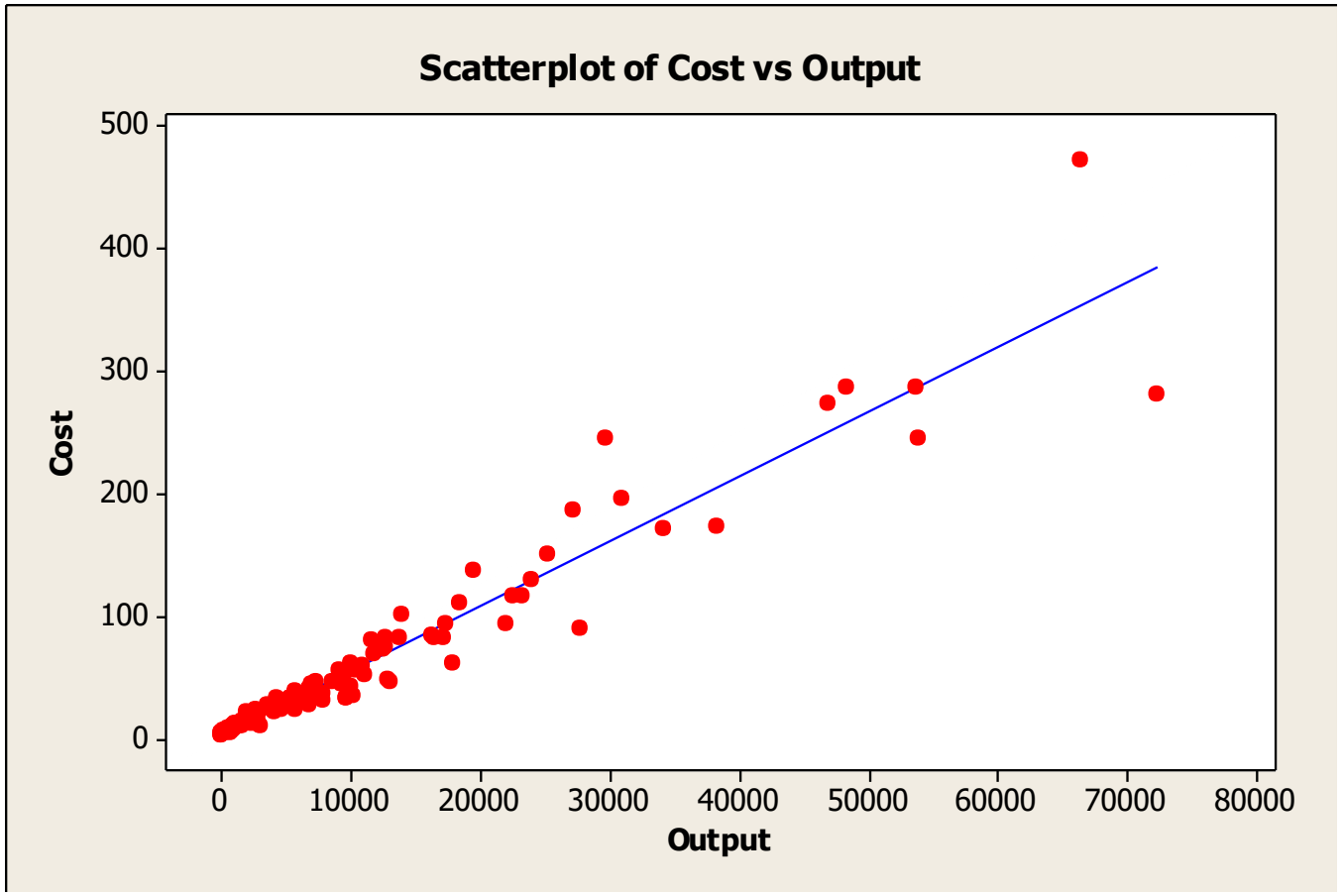
Total cost in $Million

Output in Million KWH

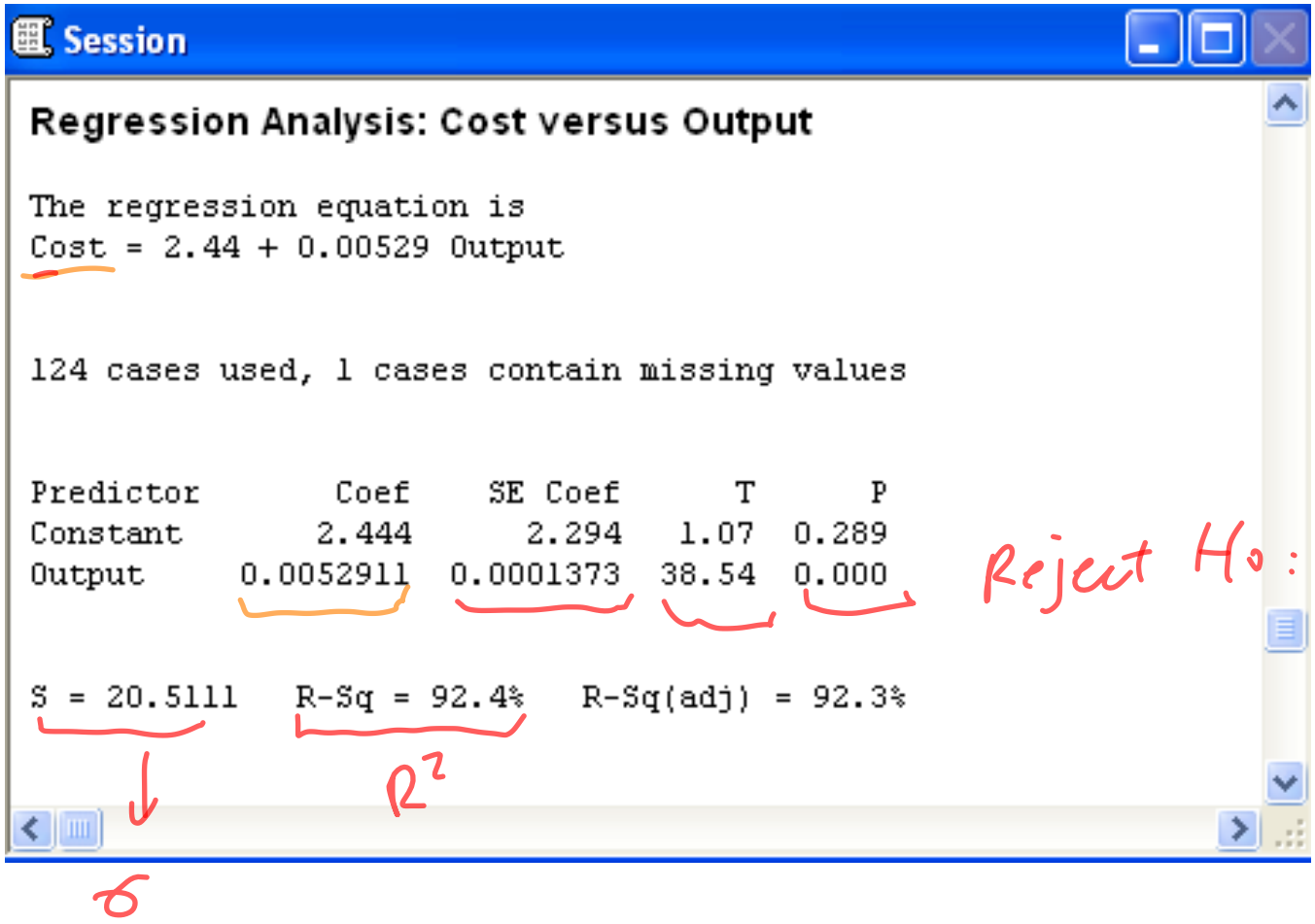
N = 123 American electric utilities

Model: $Cost = \alpha + \beta KWH + \varepsilon$

Cost Relationship



Sample Regression



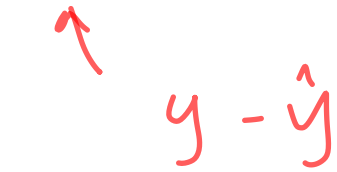
Interpreting the Model

- Cost = 2.44 + 0.00529 Output + e
- Cost is $Million, Output is Million KWH.
- Fixed Cost = Cost when output = 0
  Fixed Cost = $2.44Million
- Marginal cost
  = Change in cost/change in output
  = .00529 * $Million/Million KWH
  = .00529 $/KWH = 0.529 cents/KWH.

Using the Residuals

- How do you know the model is “good?”
- Various diagnostics to be developed over the semester.
- But, the first place to look is at the residuals.

Residuals Can Signal a Flawed Model



- Standard application: Cost function for output of a production process.
- Compare linear equation to a quadratic model (in logs)
- (123 American Electric Utilities)

Electricity Cost Function

```
124 cases used, 1 cases contain missing values


Predictor      Coef  SE Coef      T      P
Constant   -1.3926   0.1819  -7.65  0.000
logg       0.57903  0.02164  26.75  0.000


S = 0.441973   R-Sq = 85.4%   R-Sq(adj) = 85.3%


Analysis of Variance

Source           DF      SS      MS       F      P
Regression        1  139.81  139.81  715.74  0.000
Residual Error  122   23.83    0.20
Total           123  163.64
```

Interpreting Coefficient of Log(x)

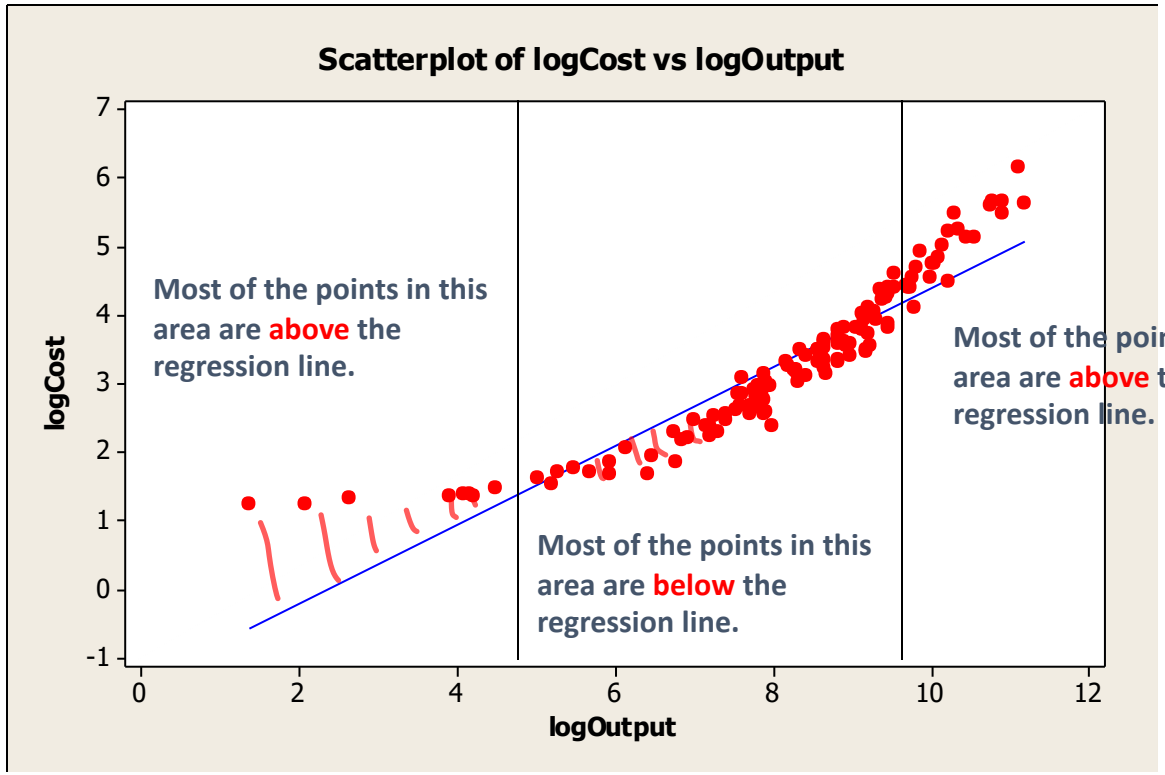
Read “Coefficient Interpretation” in course folder.

TABLE 2.3 Summary of Functional Forms Involving Logarithms

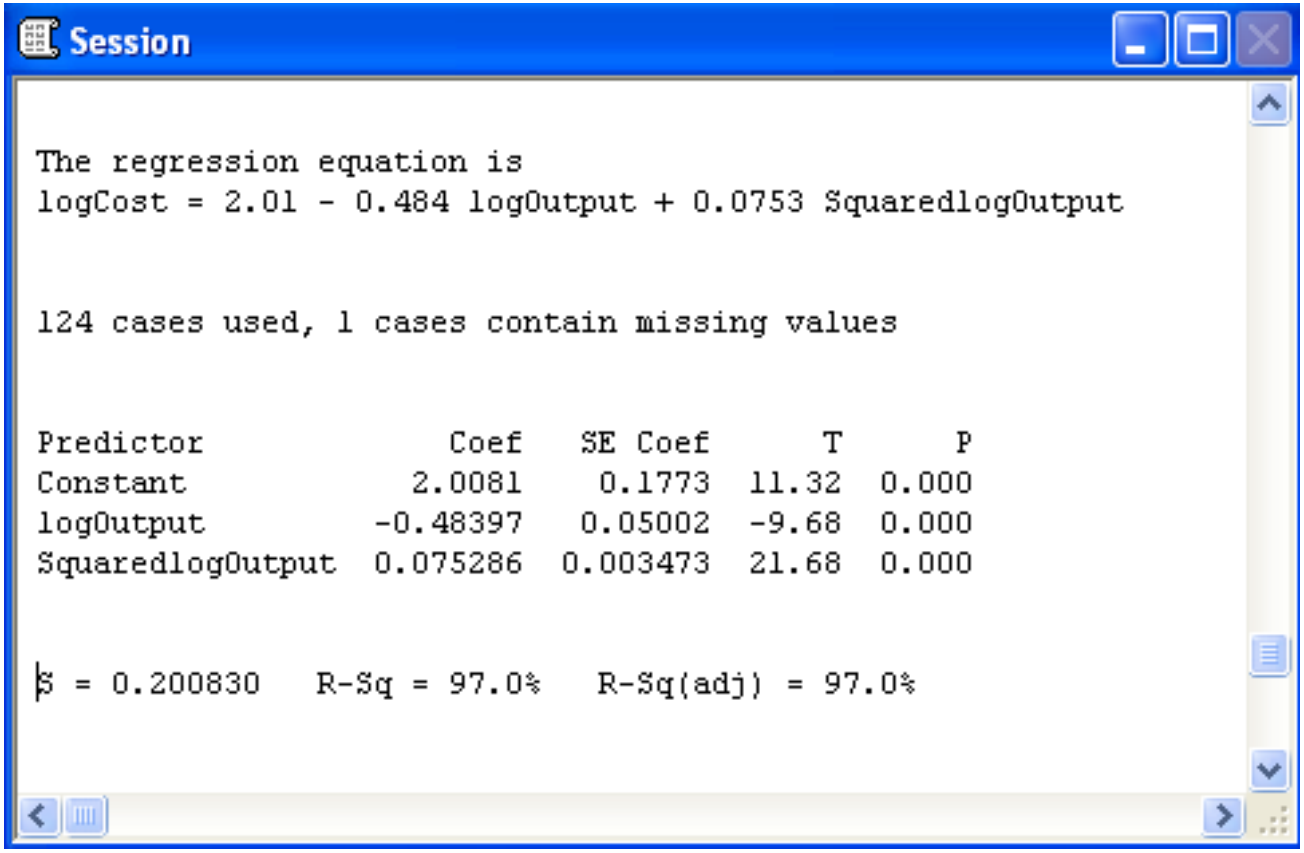
Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Candidate Model for Cost

$$\text{Log } c = a + b \log q + e$$



A Better Model?



$$\text{Log Cost} = \alpha + \beta_1 \text{logOutput} + \beta_2 [\text{logOutput}]^2 + \varepsilon$$

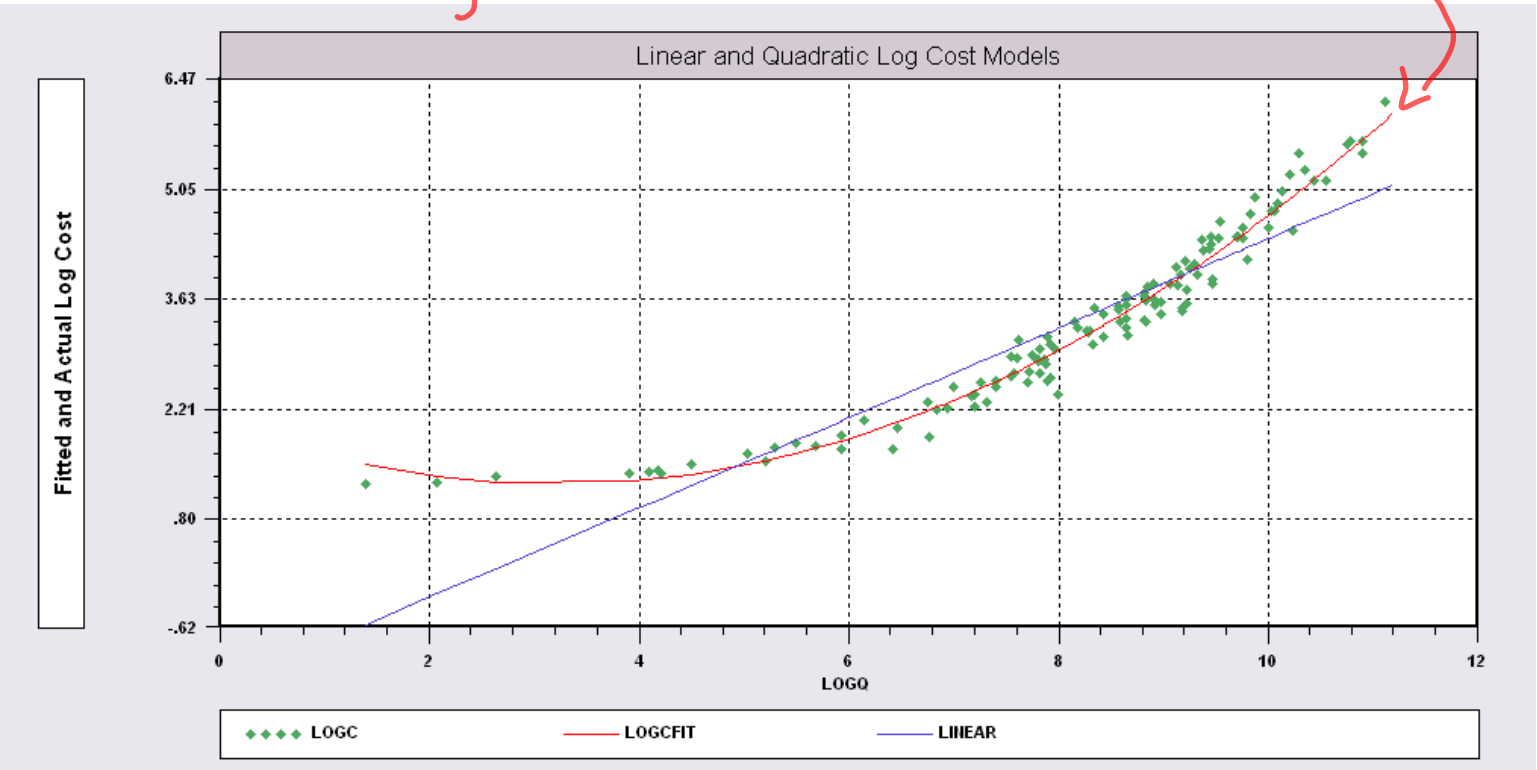
X X^2

Candidate Models for Cost

The quadratic equation is the appropriate model.

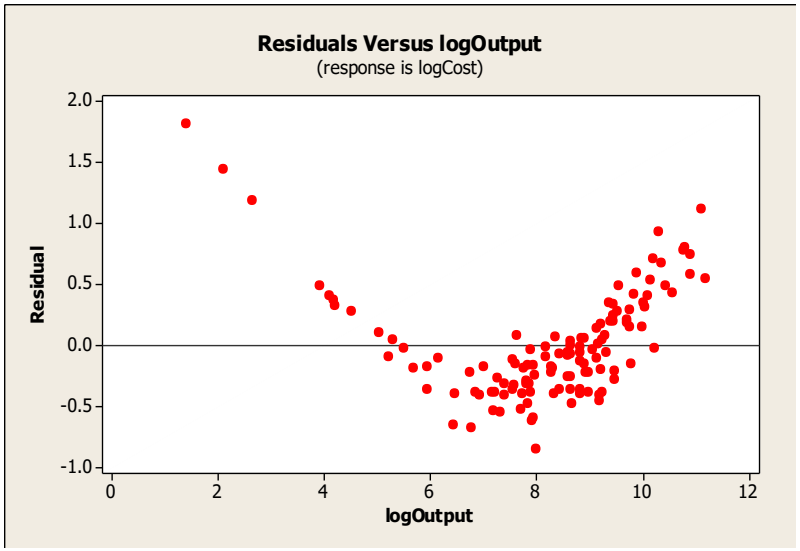
$$Logc = a + b1\ logq + b2\ log^2 q + e$$

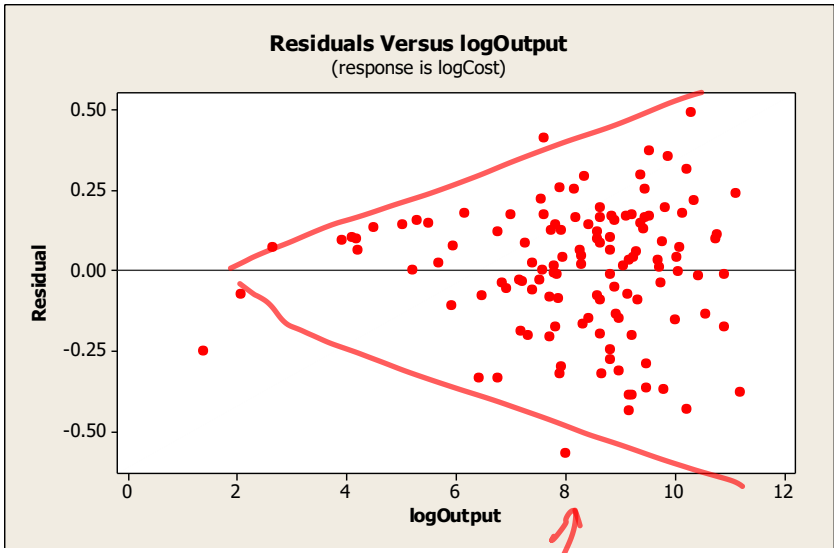
y x x^2



Missing Variable Included

Residuals from the quadratic cost model



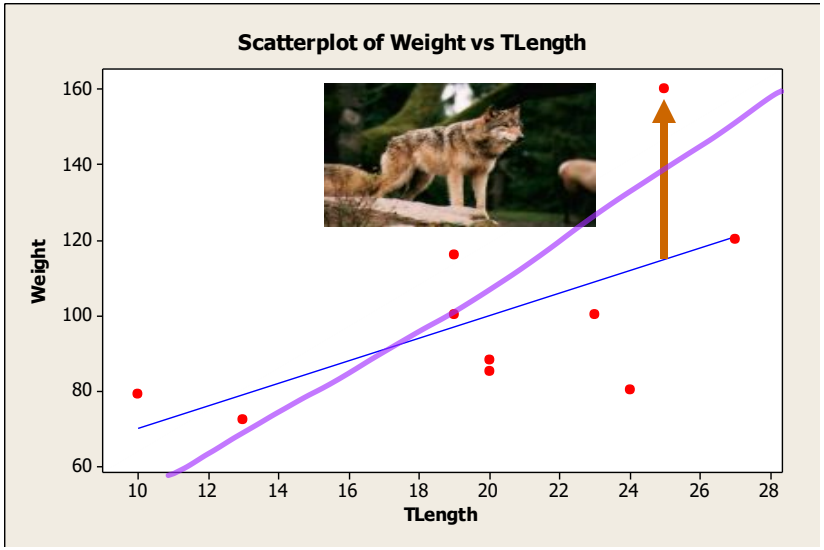


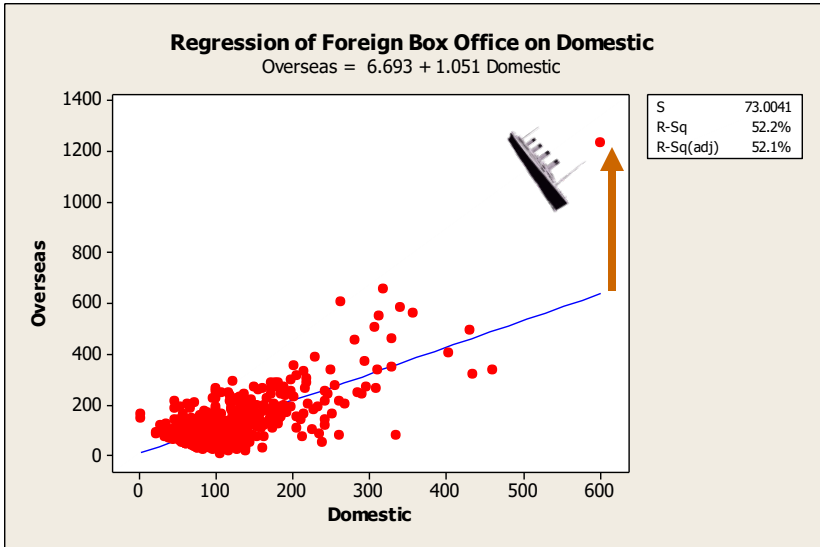
weighted least square

Residuals from the linear cost model

Unusual Data Points

Outliers have (what appear to be) very large disturbances, ε



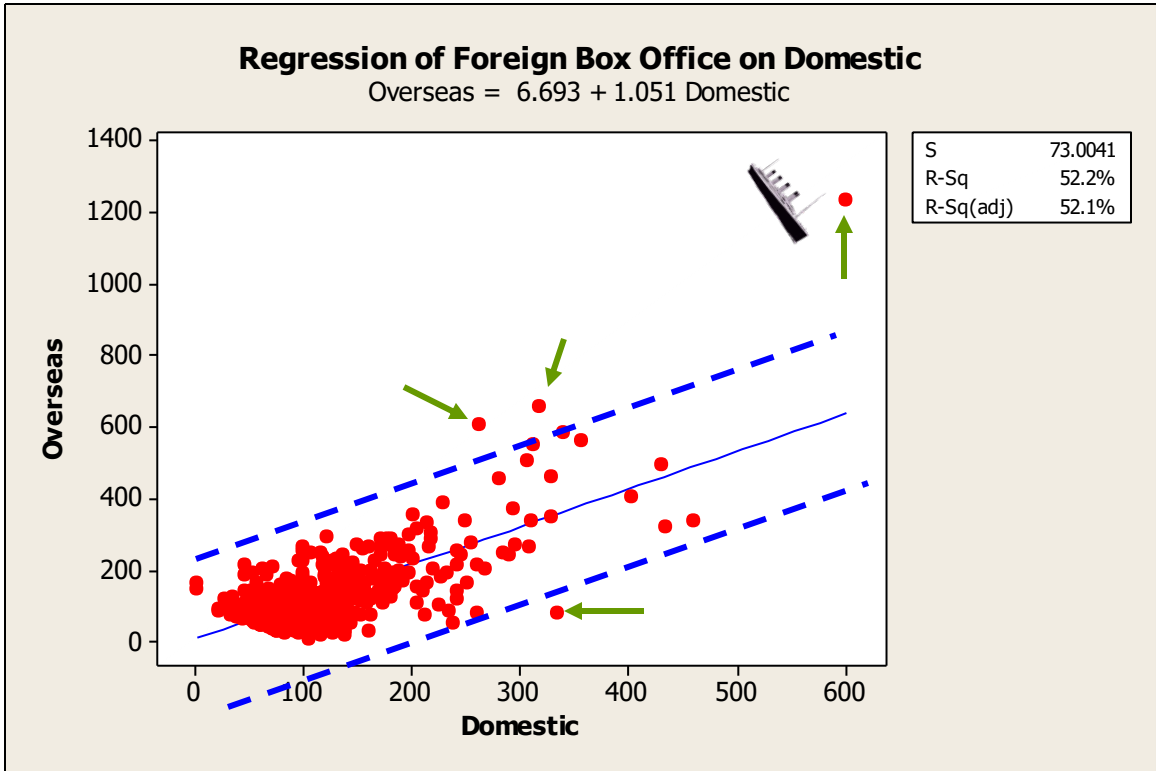


Wolf weight vs. tail length

The 500 most successful movies

Outliers

99.5% of observations will lie within mean ± 3 standard deviations. We show (a+bx) ± $3s_e$ below.)

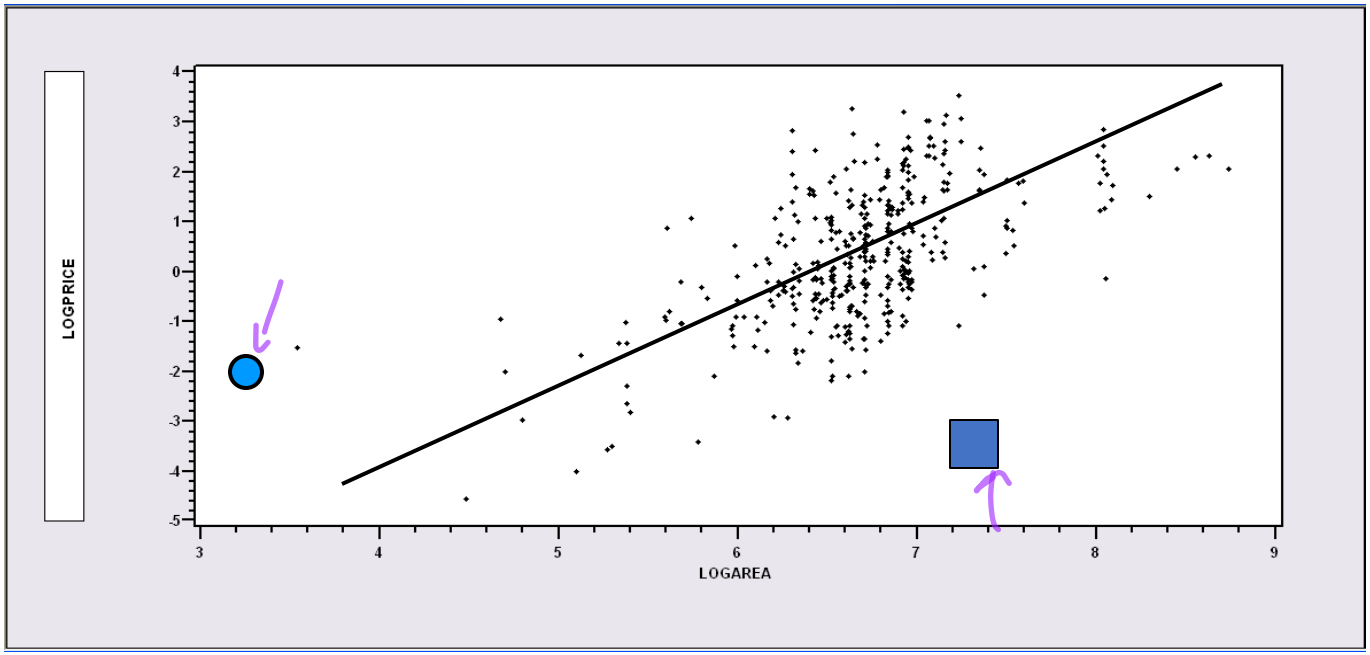


These observations might deserve a close look.

Titanic is 8.1 standard deviations from the regression!

Only 0.86% of the 466 observations lie outside the bounds. (We will refine this later.)

Prices paid at auction for Monet paintings vs. surface area (in logs)



logPrice = a + b logArea + e

- Not an outlier: Monet chose to paint a small painting.
- Possibly an outlier: Why was the price so low?

What to Do About Outliers

$\sum (y-\hat{y})^2$

$\sum |y-\hat{y}|$

(1) Examine the data

(2) Are they due to mismeasurement error or obvious "coding errors?" Delete the observations.

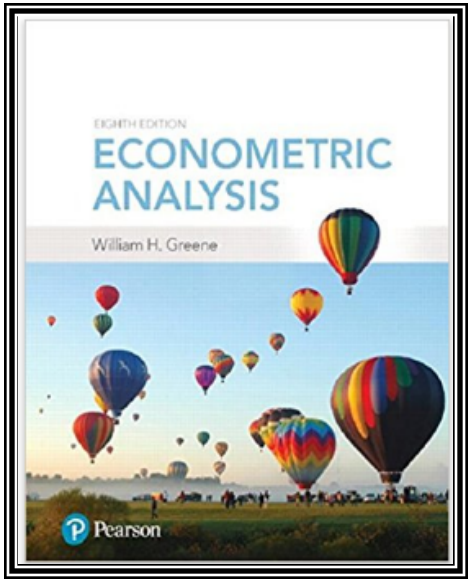
(3) Are they just unusual observations? Do nothing.

(4) Generally, resist the temptation to remove outliers. Especially if the sample is large. (500 movies is large. 10 wolves is not.)

(5) Question why you think it is an outlier. Is it really?

Multiple Regression

An Enduring Art Mystery



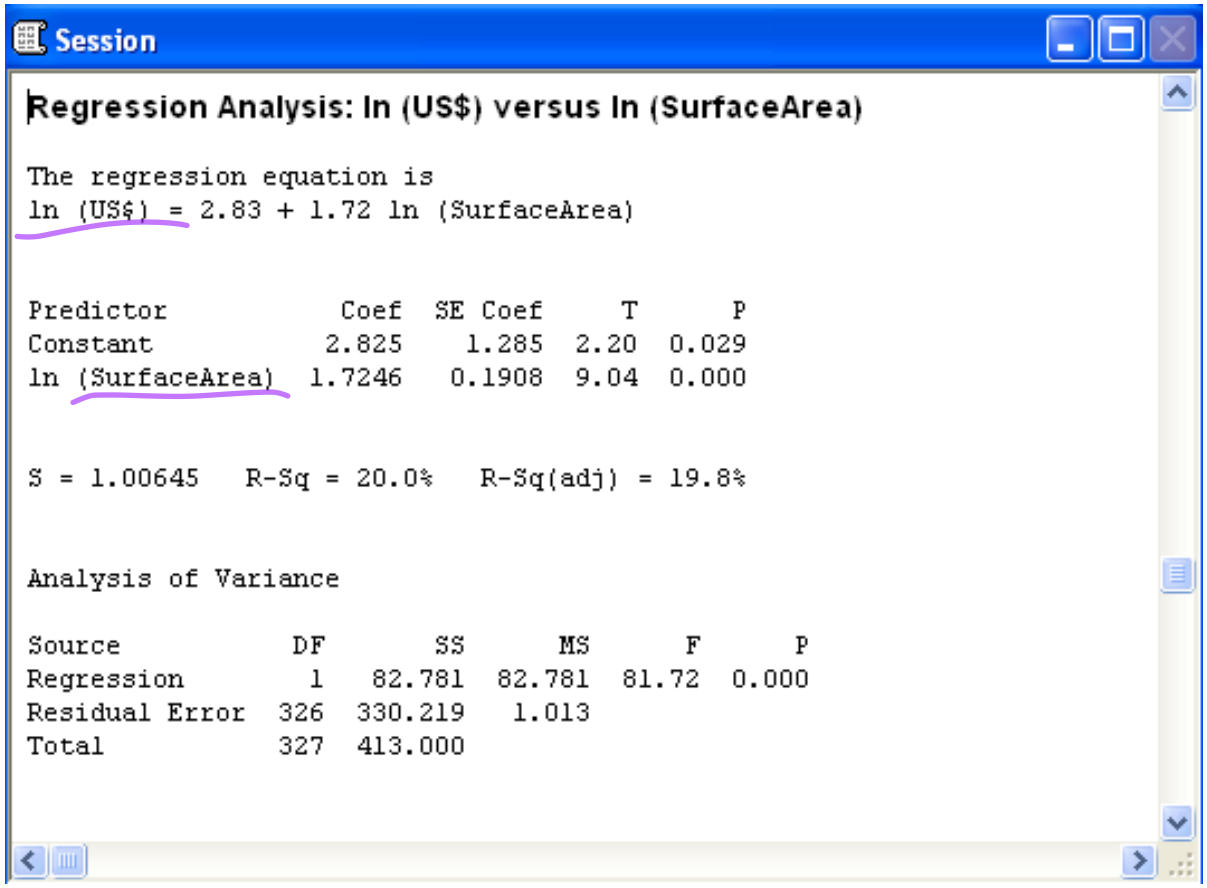
The Persistence of Statistics. Rice, 2007

Graphics show relative sizes of the two works.

The Persistence of Memory. Salvador Dali, 1931

Why do larger paintings command higher prices?

Monet Regression: There seems to be a regression. Is there a theory?



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 signed Monet's is almost twice that of unsigned

Can we separate the two effects?

	Average Prices	
	Small	**Large**
Unsigned	**346,845**	**5,795,000**
Signed	**689,422**	**5,556,490**

What do the data suggest?

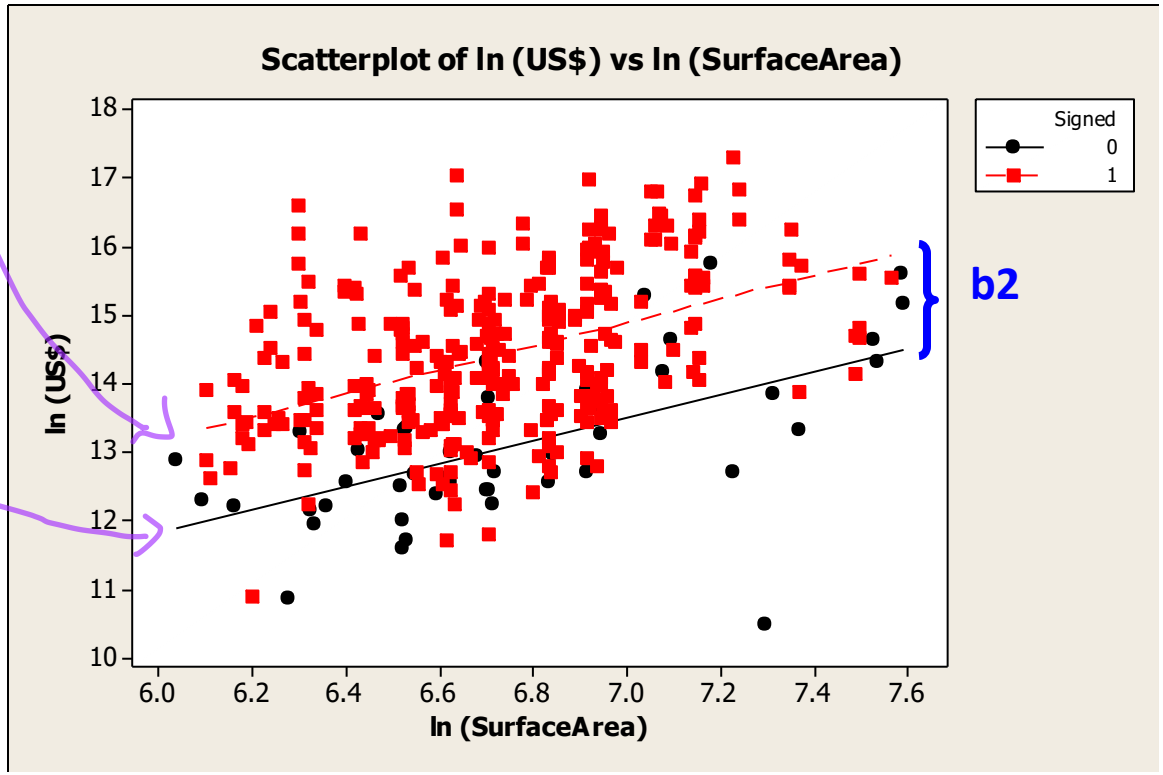
(1) The size effect is huge

(2) The signature effect is confined to the small paintings.

If unsigned: Price = $b_0 + b_1$ Area

Signed: Price = $b_0 + b_1$ Area + b_2

A Multiple Regression



y x_1 x_2

Ln Price = a + b1 ln Area + b2 (0 if unsigned, 1 if signed) + e

β_0 β_1 β_2

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 (to be explored as we develop the topic):

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

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

Classical Linear Regression Model

- The ***model*** is $y = f(x_1,x_2,\ldots,x_K,\beta_1,\beta_2,\ldots\beta_K) + \varepsilon$

  = **a *multiple* regression** model.

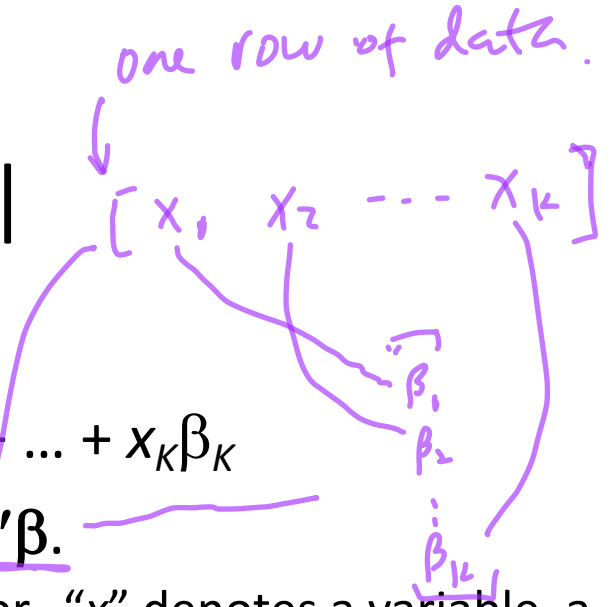
  Important examples:
  - Marginal cost in a multiple output setting
  - Separate age and education effects in an earnings equation.
  - Denote $(x_1,x_2,\ldots,x_K)$ as **x. Boldface symbol = vector.**

- Form of the model – $E[y|\mathbf{x}]$ = a linear function of **x**.

  y X_1 X_2 . . . X_K

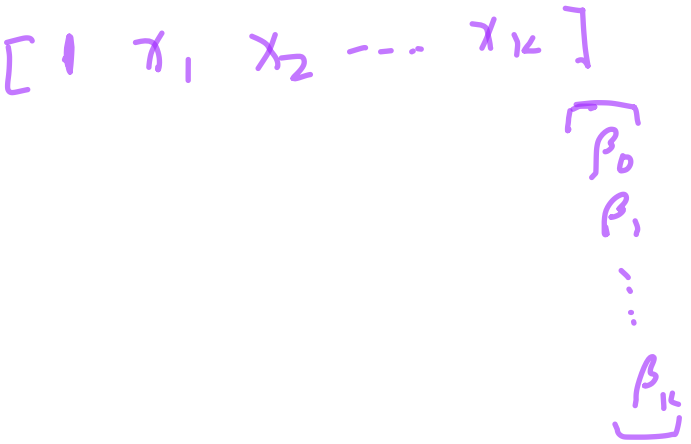
- '**Dependent' and 'independent' variables**.
  - Independent of what? Think in terms of autonomous variation.
  - Can y just 'change?' What 'causes' the change?

Linearity of the Model



- $f(x_1,x_2,\ldots,x_K,\beta_1,\beta_2,\ldots\beta_K) = x_1\beta_1 + x_2\beta_2 + \ldots + x_K\beta_K$
- **Notation:** $x_1\beta_1 + x_2\beta_2 + \ldots + x_K\beta_K = \mathbf{x}'\boldsymbol{\beta}$.
  - Boldface letter indicates a column vector. "x" denotes a variable, a function of a variable, or a function of a set of variables.
  - There are K "variables" on the right hand side of the conditional mean "function."
  - The first "variable" is usually a constant term. (Wisdom: Models should have a constant term unless the theory says they should not.)
- $E[y|\mathbf{x}] = \beta_1*1 + \beta_2*x_2 + \ldots + \beta_K*x_K$.

  (β_1*1 = the intercept term).



Linearity $\beta_1 x + \beta_2 x^2 + \beta_0$

- **Linearity** means *linear in the parameters*, not in the variables
- $E[y|\mathbf{x}] = \beta_1 f_1(\ldots) + \beta_2 f_2(\ldots) + \ldots + \beta_K f_K(\ldots)$.

  $f_k()$ may be any function of data.
- Examples:
  - Logs and levels in economics
  - Time trends, and time trends in loglinear models – rates of growth
  - Dummy variables
  - Quadratics, power functions, log-quadratic, trig functions, interactions and so on.

Matrix Notation

$$x_{ik} = x_{row,column} = x_{observation,\text{var}iable}$$

Observation 1 $\quad y_1 = \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_K x_{1K} + \varepsilon_1$

Observation 2 $\quad y_2 = \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_K x_{2K} + \varepsilon_2$

Observation 3 $\quad y_3 = \beta_1 x_{31} + \beta_2 x_{32} + \ldots + \beta_K x_{3K} + \varepsilon_3$

...

Observation i $\quad y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} + \varepsilon_i$

...

Observation N $\quad y_1 = \beta_1 x_{N1} + \beta_2 x_{N2} + \ldots + \beta_K x_{NK} + \varepsilon_N$

Notation

n : # of obs
K : # of variables.

Define column vectors of N observations on y and the *K* x variables.

$n \times 1 \quad n \times K \quad K \times 1 \quad n \times 1$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

β_0

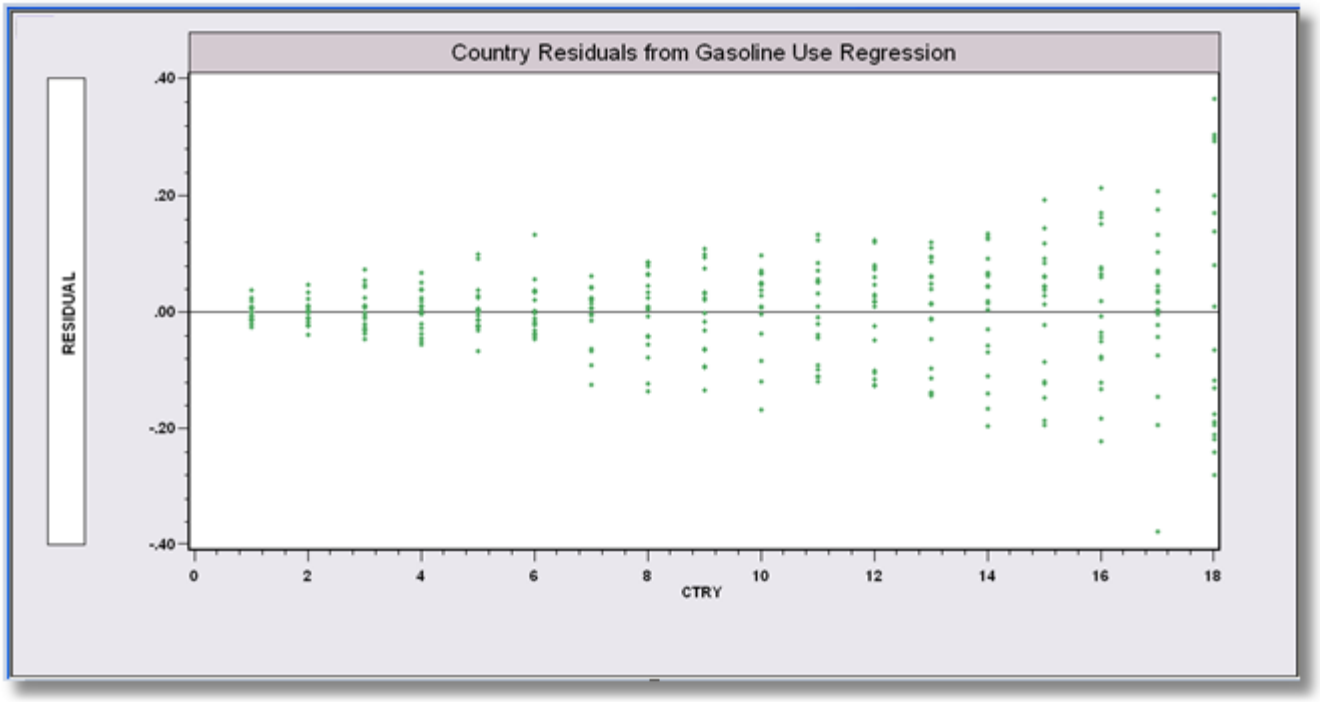
$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

No 2 cols are perfectly correlated.

The assumption means that the rank of the matrix X is ***K*.**
No linear dependencies => FULL COLUMN RANK of the matrix X.

Heteroscedasticity

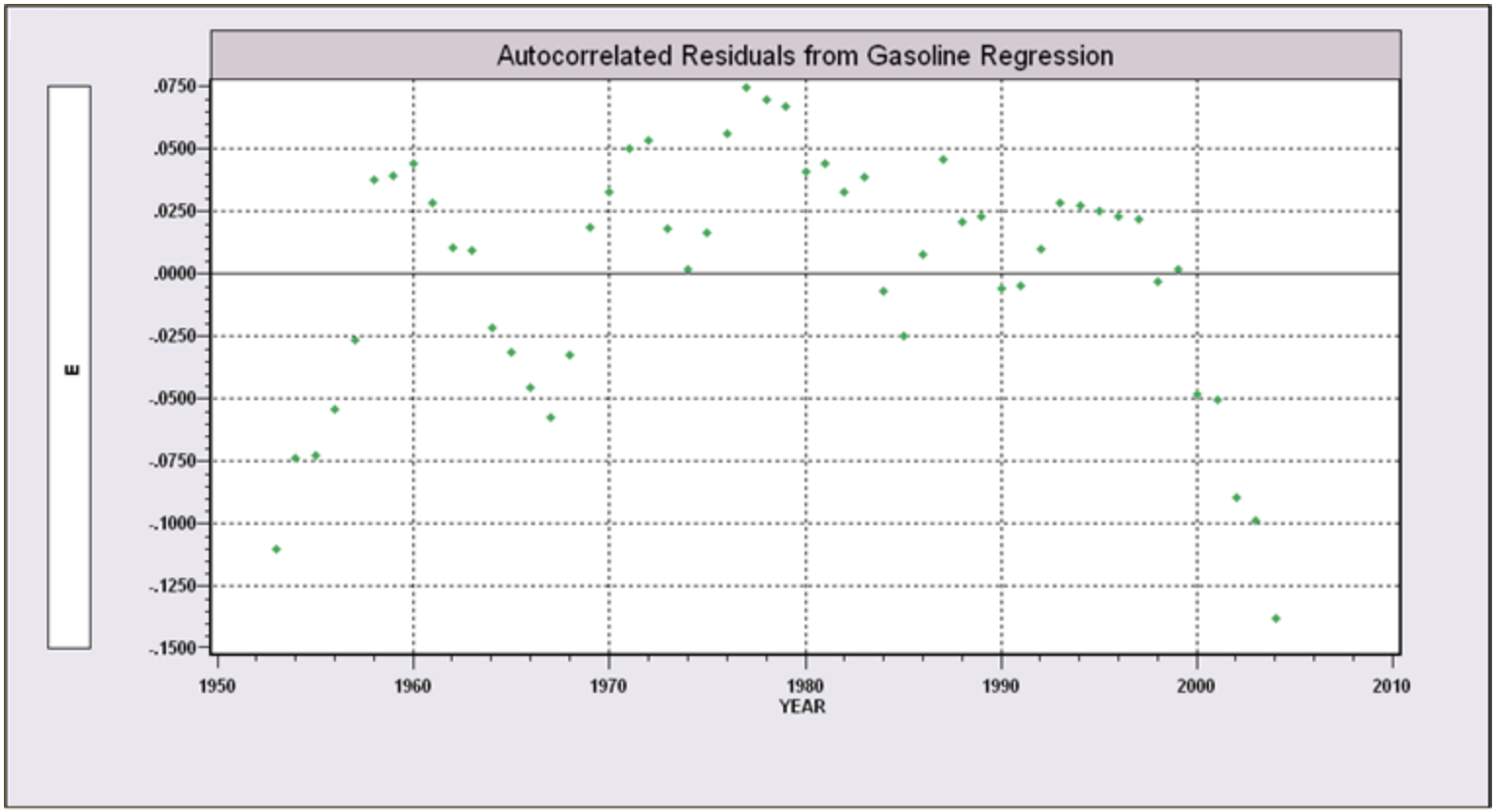
Countries are ordered by the standard deviation of their 19 residuals.



Regression of log of per capita gasoline use on log of per capita income, gasoline price and number of cars per capita for 18 OECD countries for 19 years. The standard deviation varies by country.

Autocorrelation

$\log G = \beta_1 + \beta_2 \log Pg + \beta_3 \log Y + \beta_4 \log Pnc + \beta_5 \log Puc + \varepsilon$



Least Squares

$$\sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

A digression on multivariate calculus.

Matrix and vector derivatives.

- Derivative of a scalar with respect to a vector
- Derivative of a column vector wrt a row vector
- Other derivatives

Matrix Results

$$\mathbf{X'X} = \begin{bmatrix} \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{iK} \\ \sum_{i=1}^{n} x_{i2}x_{i1} & \sum_{i=1}^{n} x_{i2}^2 & \cdots & \sum_{i=1}^{n} x_{i2}x_{iK} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n} x_{iK}x_{i1} & \sum_{i=1}^{n} x_{iK}x_{i2} & \cdots & \sum_{i=1}^{n} x_{iK}^2 \end{bmatrix}$$

$$\mathbf{X'y} = \begin{bmatrix} \sum_{i=1}^{n} x_{i1}y_i \\ \sum_{i=}^{n} x_{i2}y_i \\ \cdots \\ \sum_{i=}^{n} x_{iK}y_i \end{bmatrix}$$

Least Squares Normal Equations

$\min\ (y - Xb)'(y - Xb)$

$$\frac{\partial (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})}{\partial \mathbf{b}} = -2\mathbf{X}'(\mathbf{y} - \mathbf{Xb}) = \mathbf{0}$$

$$\partial(1\times 1) / \partial(\mathbf{k}\times 1) \qquad (-2)(\mathbf{N}\times\mathbf{K})'(\mathbf{N}\times 1)$$

$$= (-2)(\mathbf{K}\times\mathbf{N})(\mathbf{N}\times 1) = \mathbf{K}\times 1$$

Note: Derivative of 1×1 wrt $K\times 1$ is a $K\times 1$ vector.

Solution - Least squares normal equations: $\mathbf{X'y = X'Xb}$

Assuming it exists: $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$

$\mathbf{b}$: $K\times 1$; $(\mathbf{X'X})^{-1}$: $K\times K$; $\mathbf{X'}$: $K\times n$; $\mathbf{y}$: $n\times 1$

K: # of col

n: # of rows

Second Order Conditions

$$\frac{\partial(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})}{\partial \mathbf{b}} = -2\mathbf{X}'(\mathbf{y}-\mathbf{Xb})$$

$$\frac{\partial^2(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})}{\partial \mathbf{b}\partial \mathbf{b}'} = \frac{\partial\left(\frac{\partial(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})}{\partial \mathbf{b}}\right)}{\partial \mathbf{b}'} = \frac{\partial \text{ column vector}}{\partial \text{ row vector}}$$

$$= \frac{\partial[-2\mathbf{X}'(\mathbf{y}-\mathbf{Xb})]}{\partial \mathbf{b}'}$$

$$= 2\mathbf{X}'\mathbf{X}$$

Does **b** Minimize **e'e**?

$$\frac{\partial^2 \mathbf{e'e}}{\partial \mathbf{b} \partial \mathbf{b'}} = 2\mathbf{X'X} = \mathbf{2}\begin{bmatrix} \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \dots & \sum_{i=1}^{n} x_{i1}x_{iK} \\ \sum_{i=1}^{n} x_{i2}x_{i1} & \sum_{i=1}^{n} x_{i2}^2 & \dots & \sum_{i=1}^{n} x_{i2}x_{iK} \\ \dots & \dots & \dots & \dots \\ \sum_{i=1}^{n} x_{iK}x_{i1} & \sum_{i=1}^{n} x_{iK}x_{i2} & \dots & \sum_{i=1}^{n} x_{iK}^2 \end{bmatrix}$$

If there were a single b, we would require this to be positive, which it would be; $2\mathbf{x'x} = 2\sum_{i=1}^{n} x_i^2 > 0$.

The matrix counterpart of a positive number is a **positive definite matrix**.

$\hat{y} = X\beta$

$\underset{\beta}{\text{Arg Min}} \sum (\hat{y}_i - y_i)^2$

Perfect Multicollinearity

- If **X** does not have full rank, then at least one column can be written as a linear combination of the other columns.
- **X'X** does not have full rank and cannot be inverted, so **b** cannot be computed.

1. 2 cols are highly correlated
2. (next class) Dummy var Trap.
3. One col has no variation.

$$X = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ 1 & X_1 & X_2 & X_3 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Variance Inflation and Multicollinearity

- When variables are highly but not perfectly correlated, least squares is difficult to compute accurately
- Variances of least squares slopes become very large.
- Variance inflation factors: For each x_k, VIF(k) = 1/[1 – R^2(k)] where R^2(k) is the R^2 in the regression of x_k on all the other x variables in the data matrix
- Check the condition number of **X'X**. Large condition number -> Multicollinearity.

VIF

$\hat{\beta} = (X'X)^{-1}X'y$

condition number is large → Multicollinearity (ill-conditioned)

$\frac{\max \det(A)}{\min \det(A)}$ → ∞

Adjusted R-Squared

- We will discover when we study regression with more than one variable, a researcher can **ALWAYS** increase R^2 just by adding variables to a model, even if those variables do not really explain y or have any real relationship at all.
- To have a fit measure that accounts for this, "Adjusted R^2" is a number that increases with the correlation, but decreases with the number of variables.

sample size

$$\bar{R}^2 = 1 - \frac{N-1}{N-K-1}(1-R^2)$$

K is the number of "x" variables in the equation.

Linear Transformations of Data

- Change units of measurement by dividing every observation – e.g., $ to Millions of $ (see internet buzz regression) by dividing Box by 1000000.
- Change meaning of variables:
  x=(x1=nominal interest=i, x2=inflation=dp, x3=GDP)
  z=(x1-x2 = real interest i-dp, x2=inflation=dp, x3=GDP)
- Change theory of art appreciation:
  x=(x1=logHeight, x2=logWidth, x3=signature)
  z=(x1-x2=logAspectRatio, x2=logHeight, x3=signature)

- Coefficients will change.
- R squared and sum of squared residuals do not change.

Time Trends in Regression

- $y = \alpha + \beta_1 x + \beta_2 t + \varepsilon$

  β_2 is the period to period increase not explained by anything else.

- $\log y = \alpha + \beta_1 \log x + \beta_2 t + \varepsilon$

  (not log t, just t)

  $100\beta_2$ is the period to period **% increase** not explained by anything else.

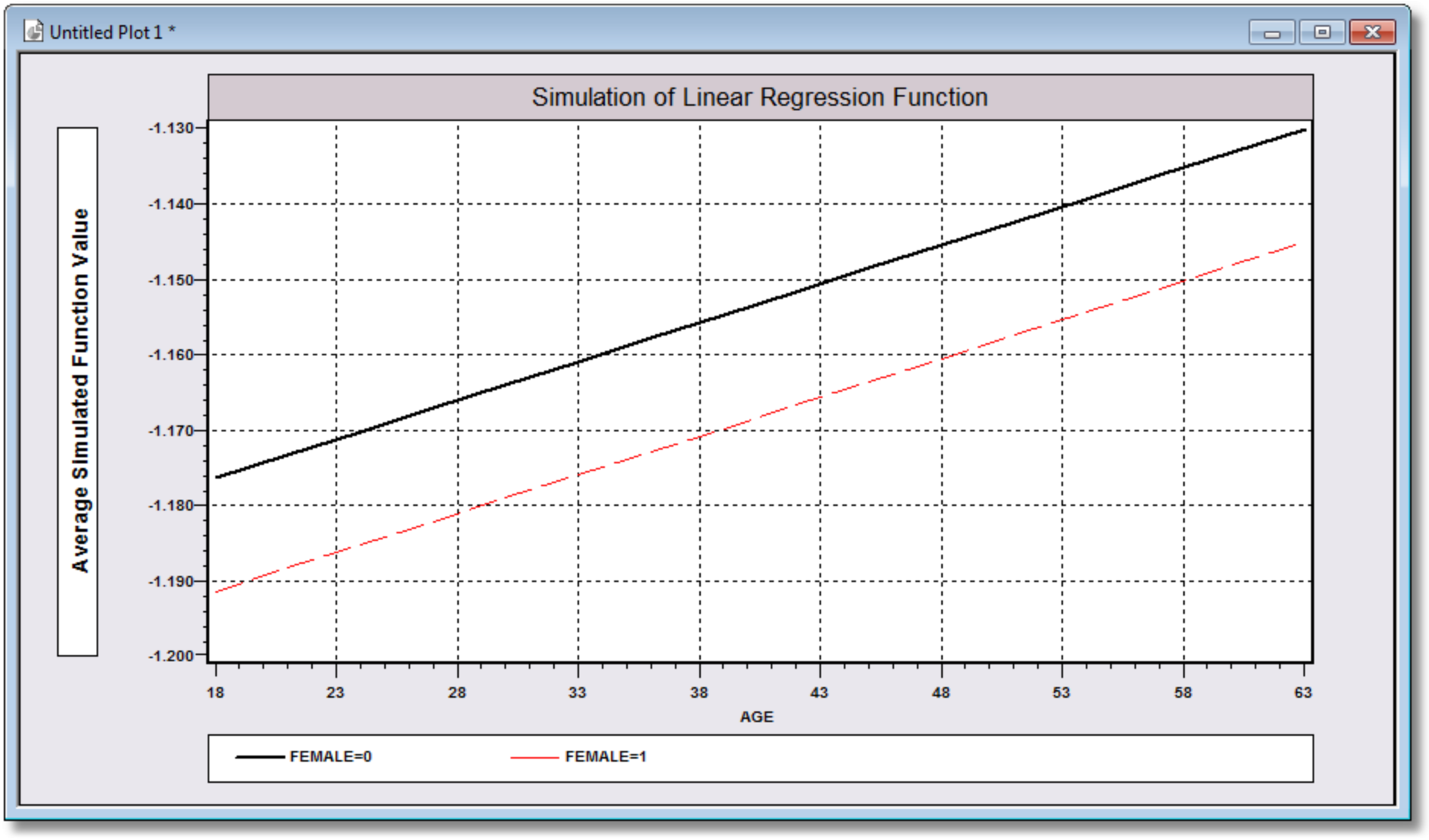
Application: Health Care Data

German Health Care Usage Data, There are altogether 27,326 observations on German households, 1984-1994.

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
INCOME (Y) = household nominal monthly net income in German marks / 10000.
HHKIDS = children under age 16 in the household = 1; otherwise = 0
EDUC = years of schooling
FEMALE = 1(female headed household)
AGE = age in years
MARRIED = marital status
EDUC = years of education

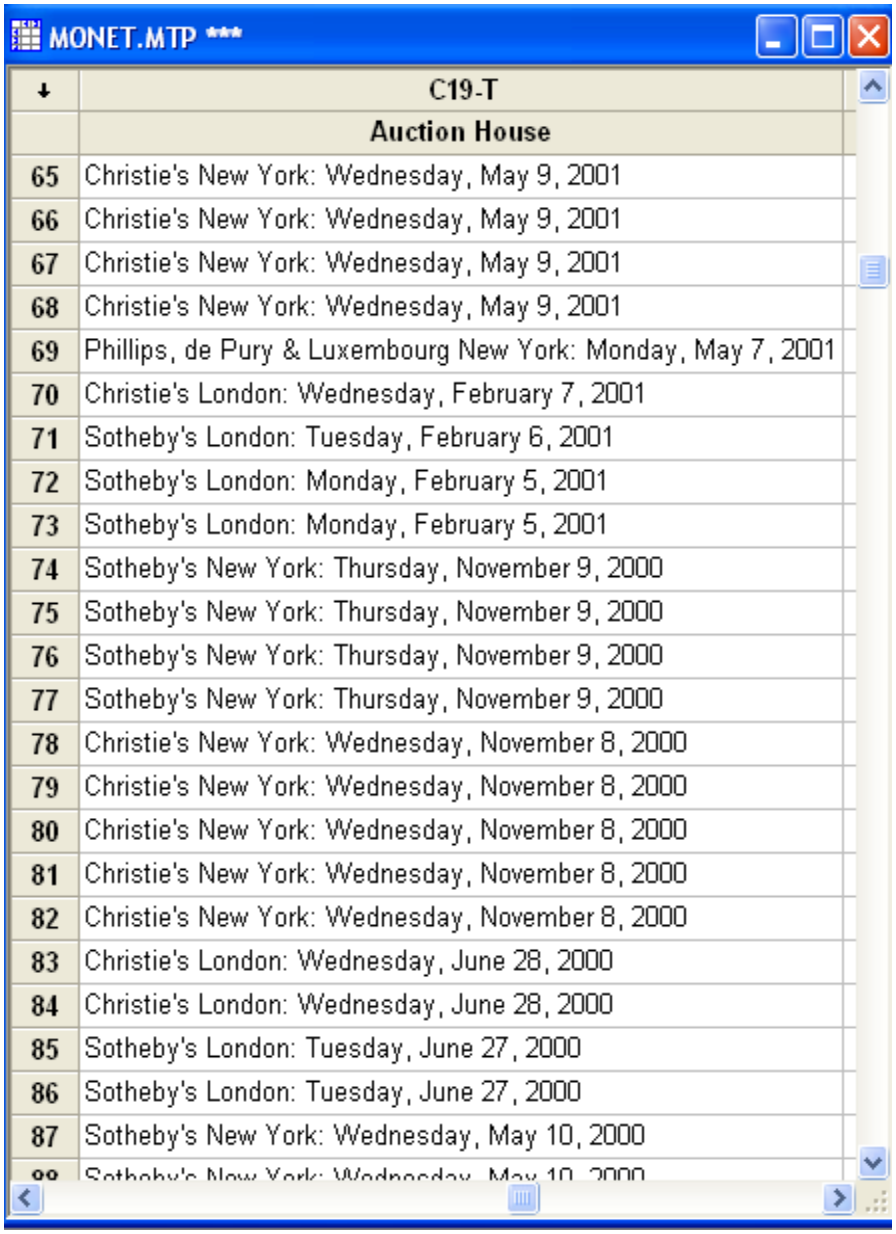
Dummy Variable

- D = 0 in one case and 1 in the other
- X: Age
- Y: Income
- $Y = a + bX + cD + e$
- When D = 0, $E[Y|X] = a + bX$
- When D = 1, $E[Y|X] = a + c + bX$


Untitled Plot 1 *
Simulation of Linear Regression Function
Average Simulated Function Value
-1.130
-1.140
-1.150
-1.160
-1.170
-1.180
-1.190
-1.200
18
23
28
33
38
43
48
53
58
63
AGE
FEMALE=0
FEMALE=1

A Conspiracy Theory for Art Sales at Auction

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

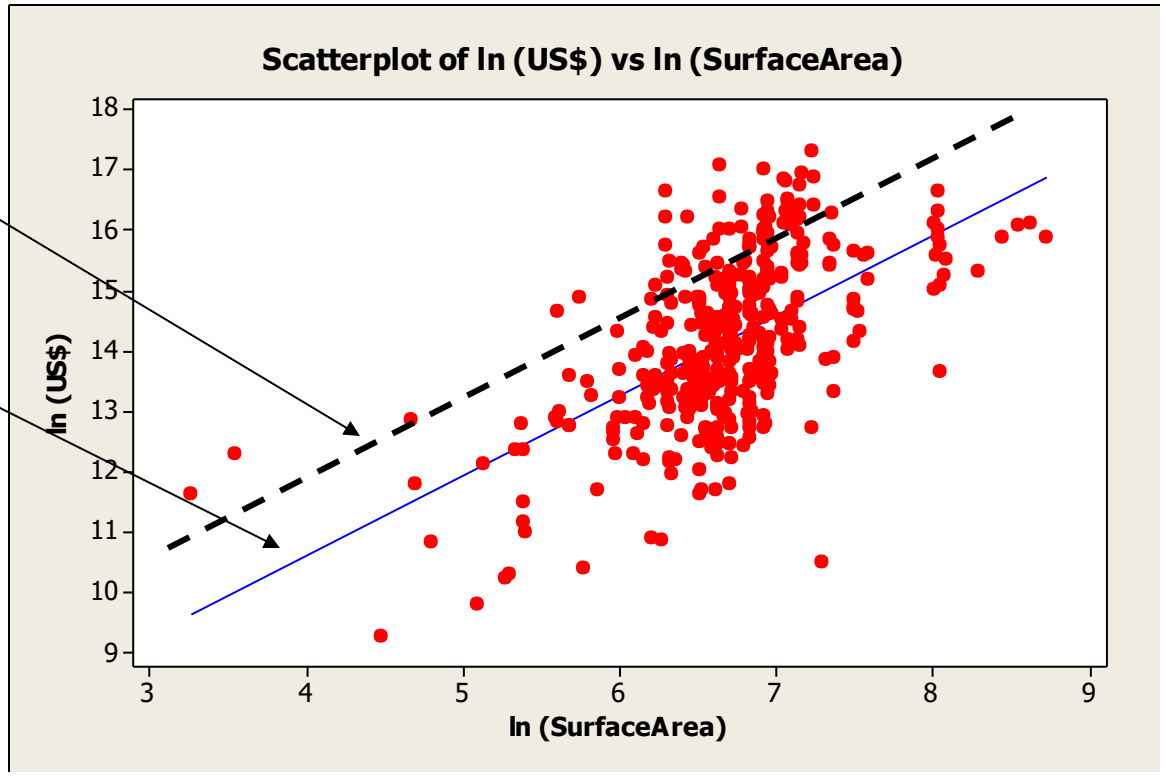


	C19-T
	Auction House
65	Christie's New York: Wednesday, May 9, 2001
66	Christie's New York: Wednesday, May 9, 2001
67	Christie's New York: Wednesday, May 9, 2001
68	Christie's New York: Wednesday, May 9, 2001
69	Phillips, de Pury & Luxembourg New York: Monday, May 7, 2001
70	Christie's London: Wednesday, 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

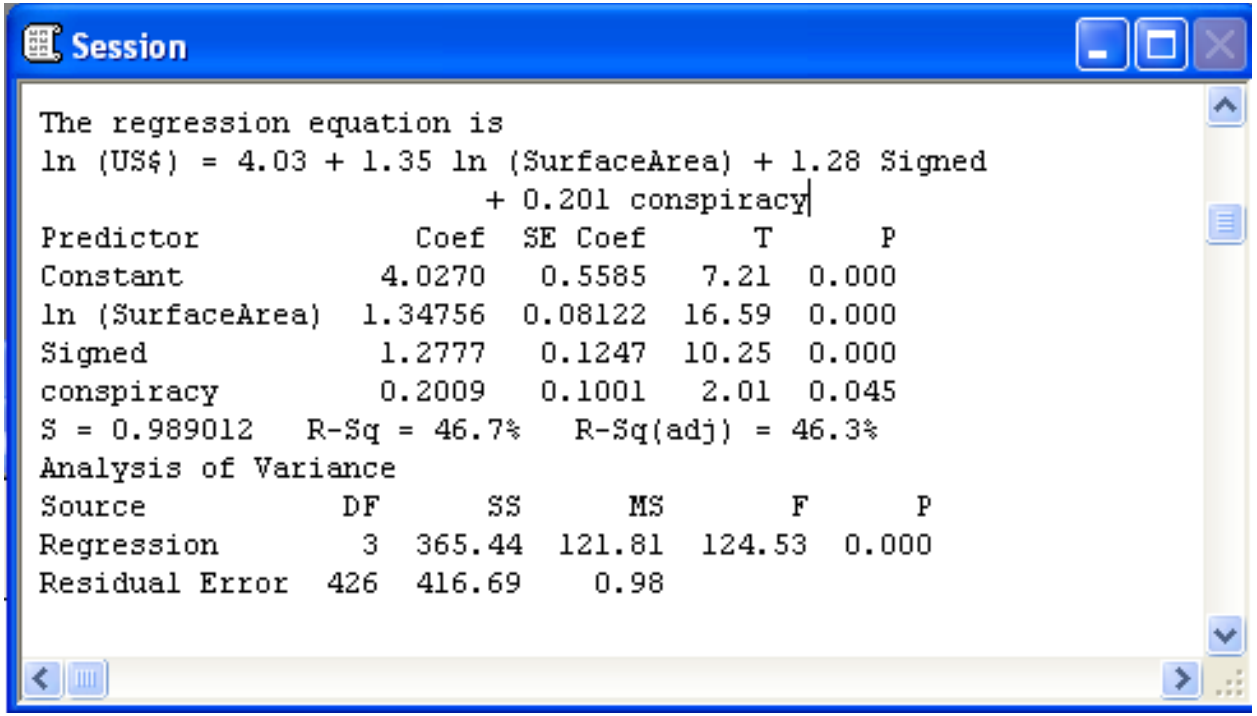
If the Theory is Correct…

Sold from 1995 to 2000

Sold before 1995 or after 2000



Evidence: Two Dummy Variables Signature and Conspiracy Effects



The statistical evidence seems to be consistent with the theory.

Set of Dummy Variables

One-Hot coding

- Usually, Z = Type = 1,2,…,K
- $Y = a + bX + d_1$ if Type=1
  $+ d_2$ if Type=2
  …
  $+ d_K$ if Type=K

A Set of Dummy Variables

- Complete set of dummy variables divides the sample into groups.
- Fit the regression with “group” effects.
- Need to drop one (any one) of the variables to compute the regression. (Avoid the “dummy variable trap.”)

Rankings of 132 U.S. Colleges' Econ Department

Region: North = N, South = S, Midwest = M, West = W

$$Reputation = \alpha + \beta_1 Religious + \beta_2 GenderEcon + \beta_3 EconFac + \beta_4 Region + \varepsilon.$$ How?

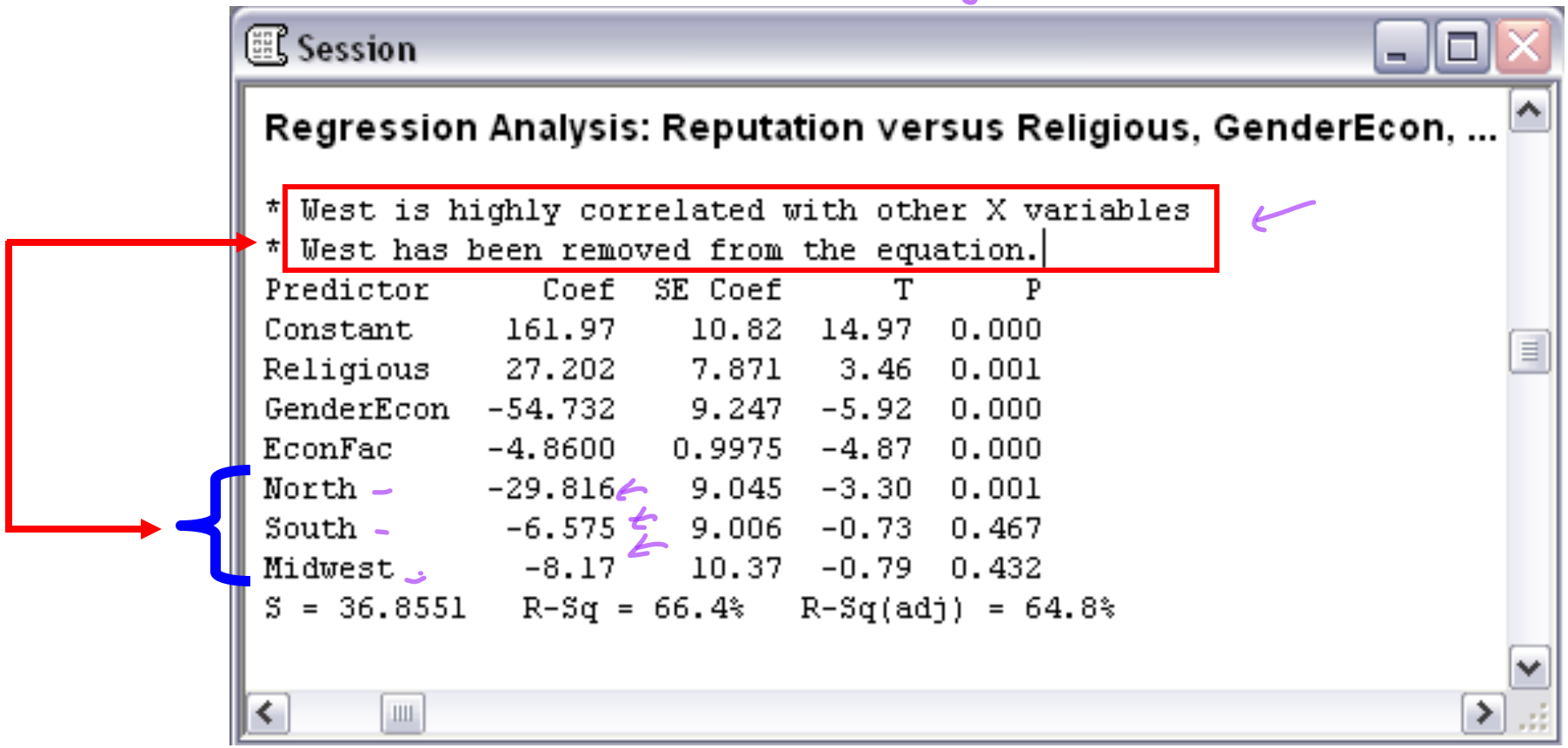
$$Reputation = \alpha + \beta_1 Religious + \beta_2 GenderEcon + \beta_3 EconFac + \beta_4 North + \beta_5 South + \beta_6 Midwest + \beta_7 West + \varepsilon$$

$W = 1 - N - S - M$

Worksheet 1 ***

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
	Region	Reputation	Religious	WomenStudies	GenderEcon	EconFac	PctWecn	PctWFac	North	South	Midwest	West
3	1	3	0	1	1	10	0.2000	0.3032	1	0	0	0
4	1	3	0	1	1	16	0.3750	0.5509	1	0	0	0
5	4	8	0	1	1	8	0.1250	0.3469	0	0	1	0
6	1	8	0	1	1	10	0.2000	0.3389	1	0	0	0
7	1	5	0	1	1	14	0.0710	0.2845	1	0	0	0
8	1	15	0	1	1	11	0.0900	0.2675	1	0	0	0
9	1	5	0	1	0	6	0.1660	0.3186	1	0	0	0
10	1	8	0	1	1	16	0.3100	0.4552	1	0	0	0
11	1	8	0	1	0	6	0.5000	0.4594	1	0	0	0
12	3	5	0	1	0	9	0.1100	0.2933	0	0	0	1
13	1	13	0	1	0	9	0.2200	0.4200	1	0	0	0
14	3	13	0	1	1	8	0.2500	0.3730	0	0	0	1
15	1	20	0	1	1	11	0.2700	0.3125	1	0	0	0

Controlling for other factors, colleges in North on average has -29.8 less reputation compared with a college in WEST.



West: Baseline.

Too many dummy variables cause perfect multicollinearity

- If we us all four region dummies
  - Reputation = a + b1*N + … if north
  - Reputation = a + b2*M + … if midwest
  - Reputation = a + b3*S + … if south
  - Reputation = a + b4*W + … if west
- Only three are needed – so drop west
  - Reputation = a + b1*N + … if north
  - Reputation = a + b2*M + … if midwest
  - Reputation = a + b3*S + … if south
  - Reputation = a + … if west

Unordered Categorical Variables

Meta-data (description) !!

Worksheet 1 ***

	C1	C2	C3	C4
	Size	Bedrooms	Price	Style
1	3327.78	3	504838	2
2	2982.63	4	592657	1
3	2996.07	5	635688	2
4	2733.43	4	768217	4
5	3134.90	3	515287	2
6	3025.76	4	733606	3
7	2340.44	3	559340	1
8	3538.86	3	658829	3
9	1633.01	2	594030	3
10	3356.19	3	623363	4
11	2307.83	3	513236	2
12	3492.65	3	629132	1
13	2752.99	4	820569	4
14	1878.12	3	599217	1
15	3260.11	3	754366	3
16	3536.82	3	751722	4

House price data (fictitious)

Type 1 = Split level
Type 2 = Ranch
Type 3 = Colonial
Type 4 = Tudor

Use 3 dummy variables for this kind of data. (Not all 4)

Using variable STYLE in the model makes no sense. You could change the numbering scale any way you like. 1,2,3,4 are just labels.

Worksheet 1 ***

↓	C1	C2	C3	C4	C5	C6	C7	C8
	Size	Bedrooms	Price	Style	Split	Ranch	Colonial	Tudor
1	3327.78	3	504838	2	0	1	0	0
2	2982.63	4	592657	1	1	0	0	0
3	2996.07	5	635688	2	0	1	0	0
4	2733.43	4	768217	4	0	0	0	1
5	3134.90	3	515287	2	0	1	0	0
6	3025.76	4	733606	3	0	0	1	0
7	2340.44	3	559340	1	1	0	0	0
8	3538.86	3	658829	3	0	0	1	0
9	1633.01	2	594030	3	0	0	1	0
10	3356.19	3	623363	4	0	0	0	1
11	2307.83	3	513236	2	0	1	0	0
12	3492.65	3	629132	1	1	0	0	0
13	2752.99	4	820569	4	0	0	0	1
14	1878.12	3	599217	1	1	0	0	0
15	3260.11	3	754366	3	0	0	1	0
16	3536.82	3	751722	4	0	0	0	1

. ANOVA

Hedonic House Price Regression

Session

X

Regression Analysis: Price versus Size, Bedrooms, Ranch, Colonial, Tudor

The regression equation is
Price = 339820 + 18.0 Size + 65298 Bedrooms
- 74369 Ranch + 87116 Colonial + 121564 Tudor

Predictor	Coef	SE Coef	T	P
Constant	339820	32047	10.60	0.000
Size	17.975	9.411	1.91	0.063
Bedrooms	65298	8556	7.63	0.000
Ranch	-74369	16535	-4.50	0.000
Colonial	87116	16072	5.42	0.000
Tudor	121564	16957	7.17	0.000

S = 40278.2 R-Sq = 84.8% R-Sq(adj) = 83.1%

Each of these is relative to a Split Level, since that is the omitted category. E.g., the price of a Ranch house is $74,369 less than a Split Level of the same size with the same number of bedrooms.

Interaction Effect

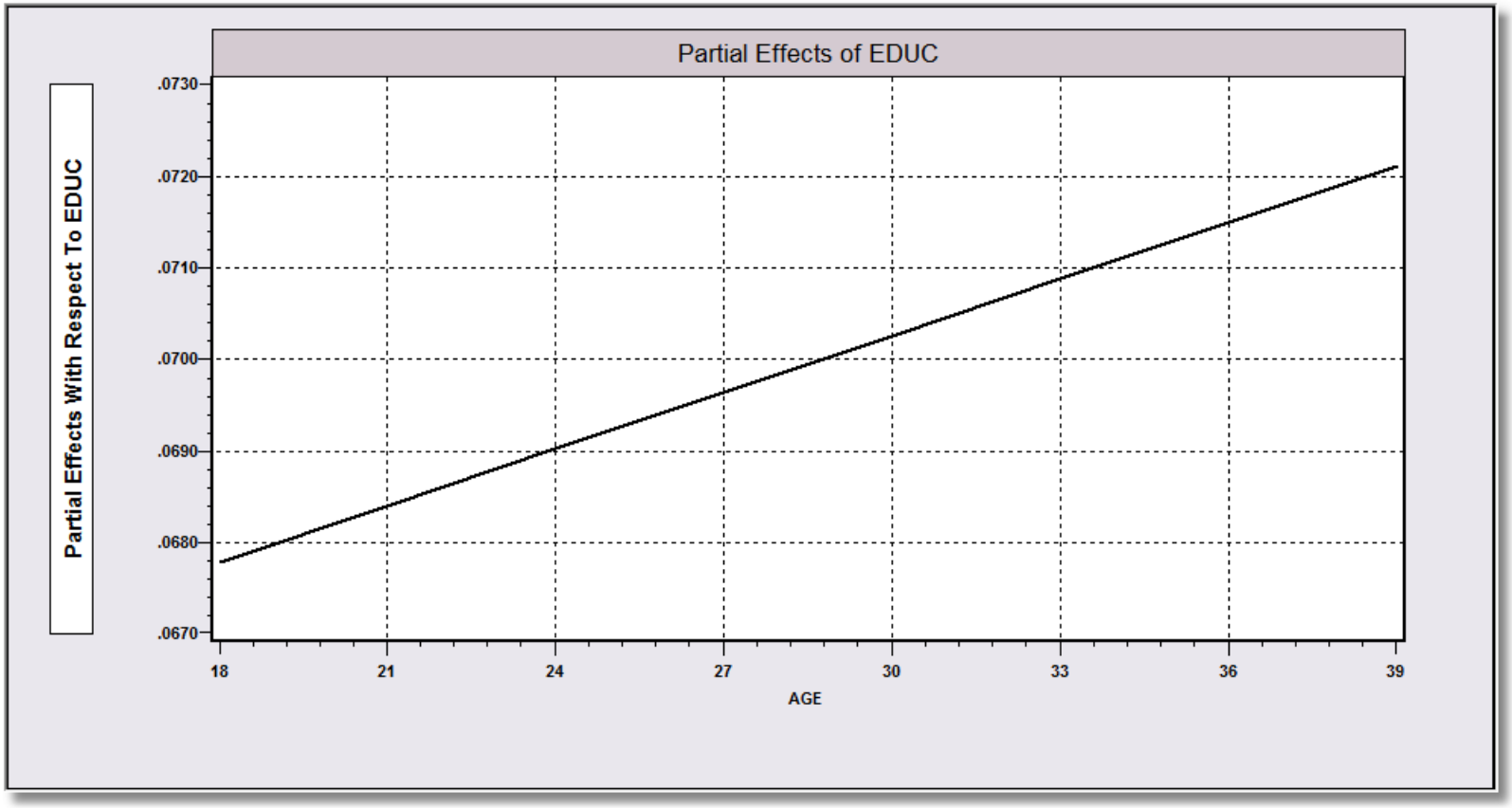
- $Y = a + b_1X + b_2Z + b_3X*Z + e$
- E.g., the benefit of a year of education depends on how old one is.
- $Log(income) = a + b_1*Edu + b_2*Edu*Age + e$
- $d_{logIncome}/d_{Edu} = b_1 + b_2*Age$

interaction.

b_2 pos: edu is worth more when older

effect of edu on income

Effect of an additional year of education increases from about 6.8% at age 20 to 7.2% at age 40



Male: wage = $\beta_0 + \beta_1 edu$ Female: wage = $\beta_0 + \delta_0 + \beta_1 edu + \delta_1 edu$ = $\beta_0 + \delta_0 + (\beta_1 + \delta_1) edu$

Interaction Effect between Continuous and Dummy Variables

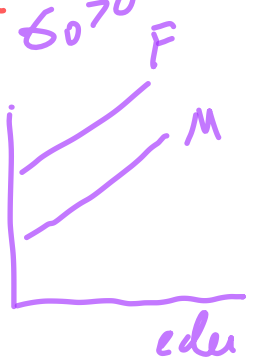
$$\log(wage) = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \cdot educ + u.$$

0/1

1 if female

→ Assuming $\beta_1 > 0$, which of the two plots indicates $\delta_1 > 0$?

effect of edu on wage is stronger for females.



w/o interaction