Large Sample Properties of OLS Estimates

Consistency

Let W_n *be an estimate for the parameter* θ *constructed from a sample size of n.* W_n *is consistent if*

$$\Pr[|W_n - \theta| > \varepsilon] \to 0 \text{ as } n \to \infty$$

[*for* ε *abitrarily small*]

Consistent estimates

$$written\ as\ p\lim(W_n) = \theta$$

Consistency

- Minimum criteria for an estimate. If not consistent in large samples, then usually the estimator stinks
- If $Var(W_0) \to 0$ as $n \to \infty$ and it is an unbiased estimate, then the estimate is consistent
- However, a consistent estimate can be biased

1980 Census PUMS

- 5% sample of US population
- Construct analysis sample of
  - Males 18-64
  - Work full time (30+ hours per week) full year (40+weeks per year)
- Two variables
  - ln(weekly earnings)
  - Years of education

5

- Total of 1,942,028 observations
- Estimate simple bivariate regression model
- $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$
- $Ln(\text{weekly earnings})_i = \beta_0 + EDUC_i\beta_1 + \varepsilon_i$
- Treat as a 'population'
- "Actual" rate of return to education is 0.05135

6

```
. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        educ |   1942028    12.93697    3.026103          0         20
    weekearn |   1942028    379.2444    213.3202   120.0962   1829.268
   weekearnl |   1942028    5.813814    .4863825   4.788293   7.511672

. reg weekearnl educ

      Source |       SS       df       MS              Number of obs = 1942028
-------------+------------------------------           F(  1,1942026) =       .
       Model |  46888.0464     1  46888.0464           Prob > F      =  0.0000
    Residual |  412533.2571942026  .212424168           R-squared     =  0.1021
-------------+------------------------------           Adj R-squared =  0.1021
       Total |  459421.3031942027  .236567928           Root MSE      =  .46089

------------------------------------------------------------------------------
   weekearnl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0513475   .0001093   469.82   0.000     .0511333    .0515618
       _cons |   5.149533   .0014521  3546.32   0.000     5.146687    5.152379
------------------------------------------------------------------------------
```

Rate of return to education is 0.05135
-- treat as the true population value

7

- From population of almost 2 million
- Sample N observations from population
- Estimate OLS model
- Do this 500,000 times
- Look at the distribution of $\hat{\beta}_1$
- Repeat exercise for different sample sizes
  - 50, 500, 5000, and 50,000 observations
- Took 15 hours to do simulation

8

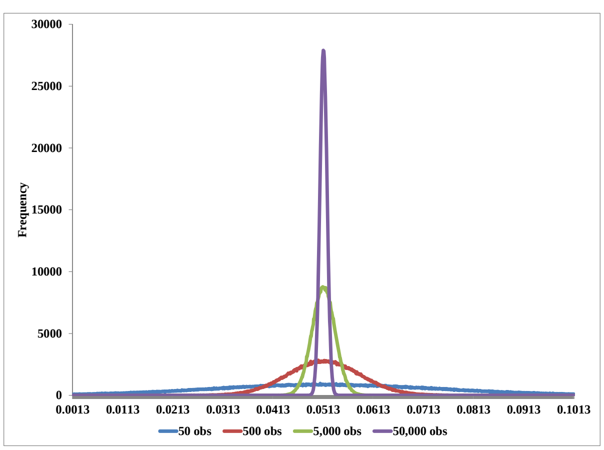
Notice a few things

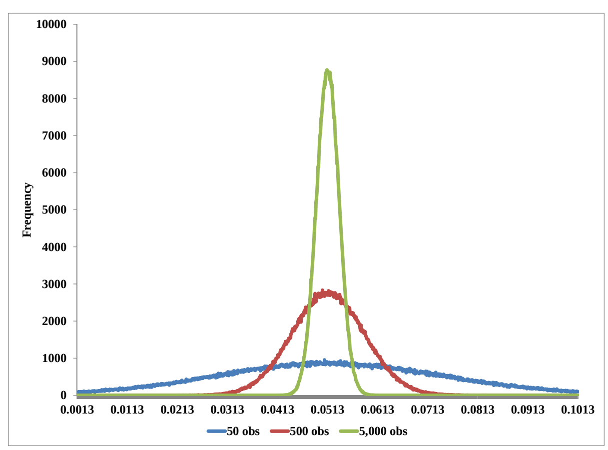
- Every time we draw a new sample of n, we get a different value for parameters
- With smaller n, the variation in the parameter estimates is much larger
- The distribution of the parameter is essentially a normal distribution (for any n)
- As n grows, the variance in the shrinks to zero and the estimate converges to the truth

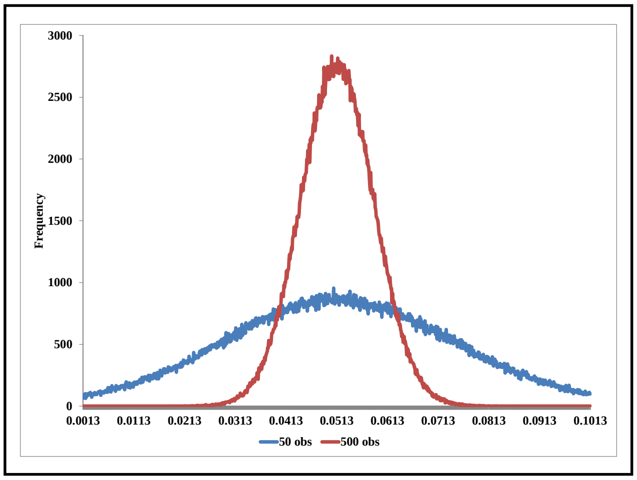
Distribution of $\hat{\beta}_1$

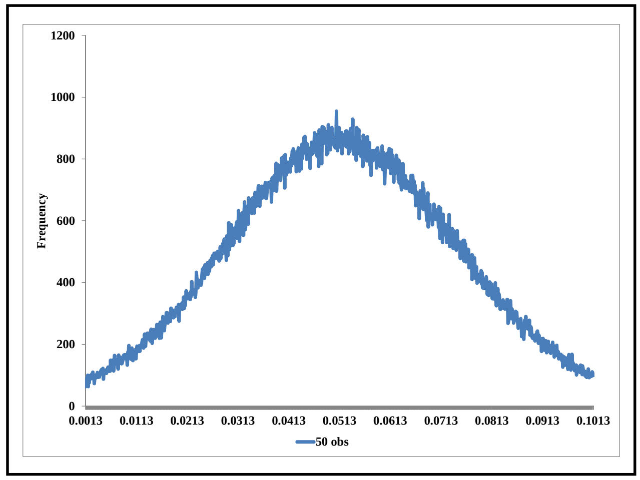
Sample size	Mean	Standard Dev.	Min	Max
50	0.05203	0.02378	-0.071	0.190
500	0.05143	0.00725	0.018	0.087
5,000	0.05135	0.00228	0.041	0.063
50,000	0.05135	0.00071	0.048	0.055

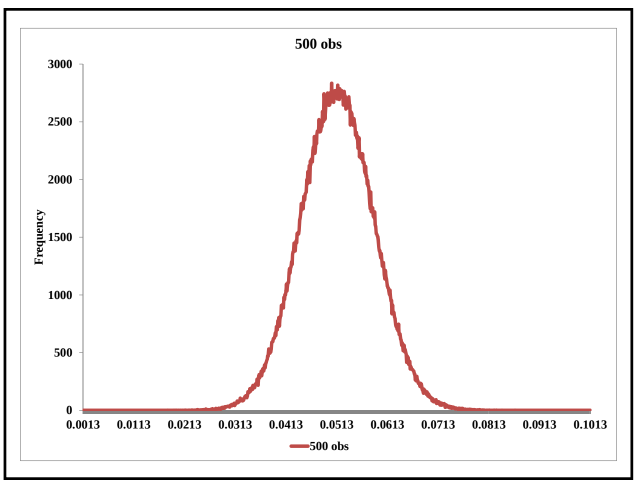
10

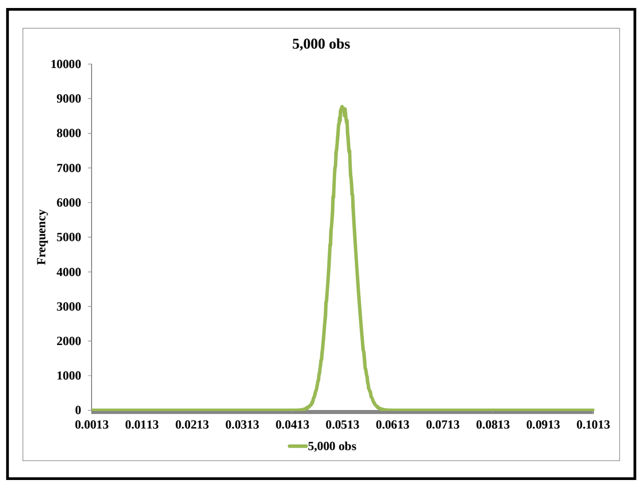


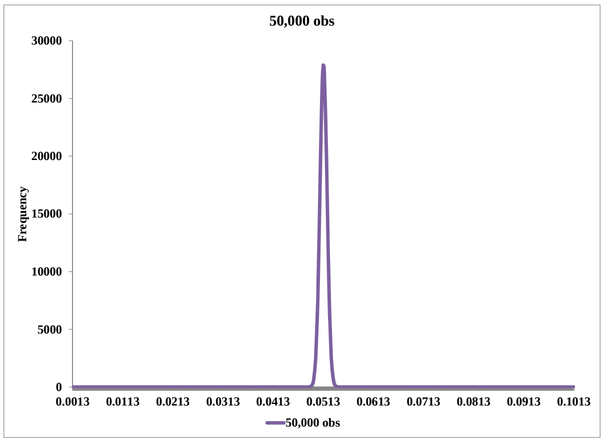



Frequency
3000
2500
2000
1500
1000
500
0
0.0013 0.0113 0.0213 0.0313 0.0413 0.0513 0.0613 0.0713 0.0813 0.0913 0.1013
50 obs
500 obs


Frequency
1200
1000
800
600
400
200
0
0.0013 0.0113 0.0213 0.0313 0.0413 0.0513 0.0613 0.0713 0.0813 0.0913 0.1013
50 obs


500 obs
Frequency
3000
2500
2000
1500
1000
500
0
0.0013 0.0113 0.0213 0.0313 0.0413 0.0513 0.0613 0.0713 0.0813 0.0913 0.1013
500 obs


5,000 obs
Frequency
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0.0013 0.0113 0.0213 0.0313 0.0413 0.0513 0.0613 0.0713 0.0813 0.0913 0.1013
5,000 obs



$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^{n}\varepsilon_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \beta_1 + \frac{\left(\frac{1}{N-1}\right)\sum_{i=1}^{n}\varepsilon_i(x_i - \bar{x})}{\left(\frac{1}{N-1}\right)\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1 + \frac{\hat{\sigma}^2_{x\varepsilon}}{\hat{\sigma}^2_x}$$

- plim$(\hat{\sigma}_{x\varepsilon})$ is $(\sigma_{x\varepsilon})$
- plim $(\hat{\sigma}^2_x)$ is (σ^2_x)

$$p\lim \hat{\beta}_1 = \beta_1 + p\lim\left(\frac{\hat{\sigma}_{x\varepsilon}}{\hat{\sigma}^2_x}\right) = \beta_1 + \frac{\sigma_{x\varepsilon}}{\sigma^2_x}$$

- So long as:
- plim$(\hat{\sigma}_{x\varepsilon})$ is $(\sigma_{x\varepsilon}) = 0$
- plim $(\hat{\beta}_1) = \beta_1$

Notion

- Begin with the assumption that cov(x,ε)=0
- In small samples, can randomly have correlation between x and ε
- But if the assumption is correct, large samples will eliminate small sample problems

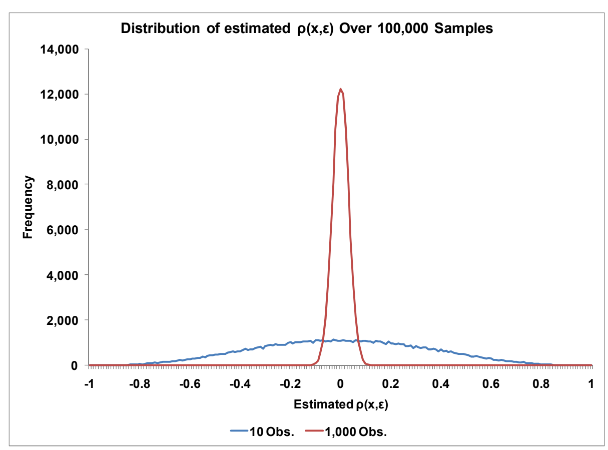
Simulation

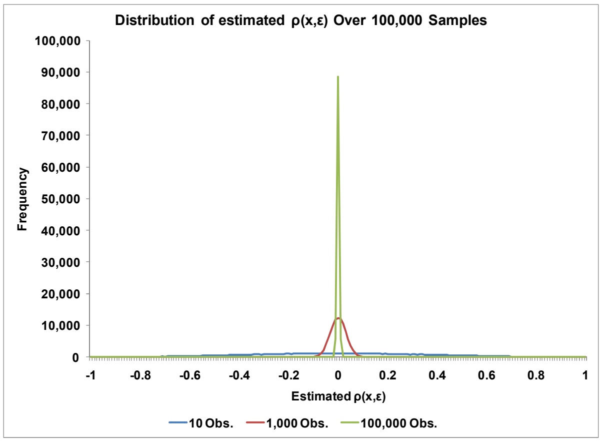
- Draw two independent series: x and ε
- Have a fixed sample size
- For each draw, calculate the sample correlation coefficient
- Repeat 100,000 times
- Three sample sizes
  - n=10;   n=1000;   n=100,000
- What should we see?

Distribution of $\hat{\rho}(x,\varepsilon)$

Sample size	Mean	Stand. Dev.	Min	Max
10	-6.6E-4	0.332	-0.949	0.941
1,000	-7.4E-5	0.032	-0.129	0.134
100,000	4.0E-4	0.0032	-0.014	0.013

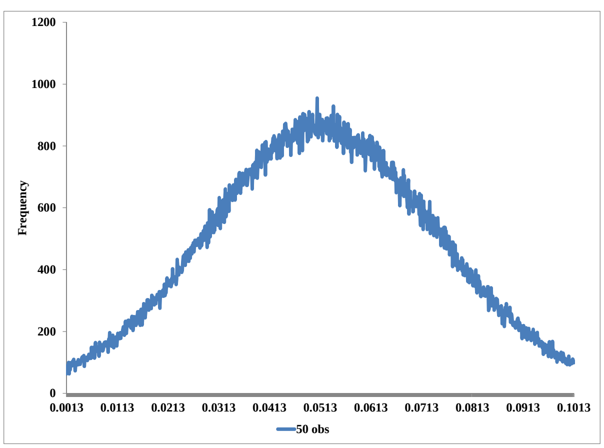
23

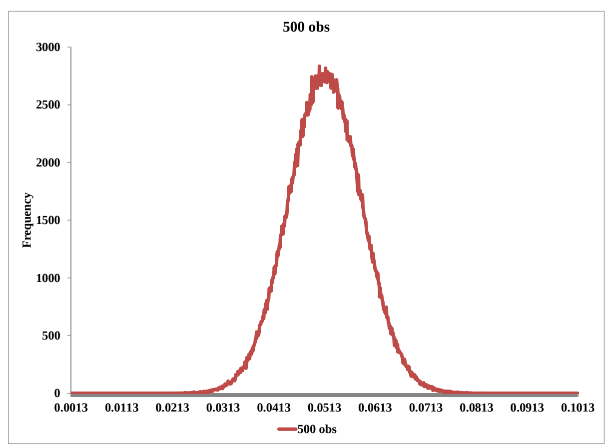


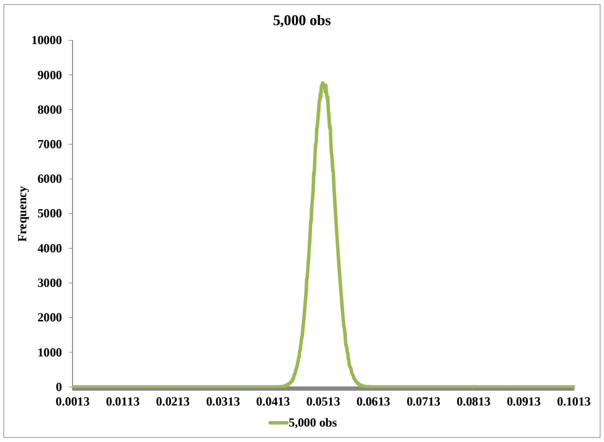


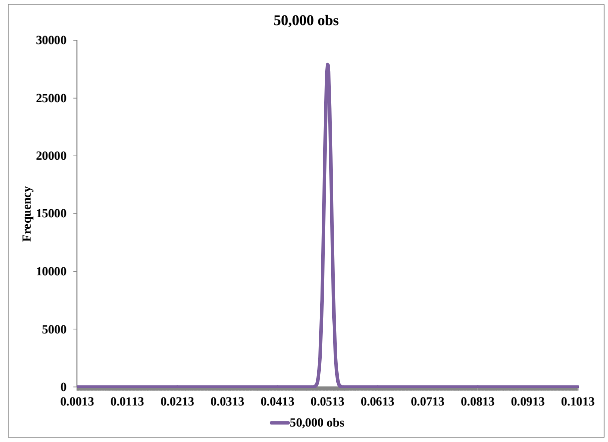
Couple of other notes

- Statistical tests we've used so far are based on the assumption that ε_i is a normal distribution
- At first glance, appears to be a strong assumption
- Look at the underlying distribution of the parameters – notice that for large samples, the distribution approaches a normal





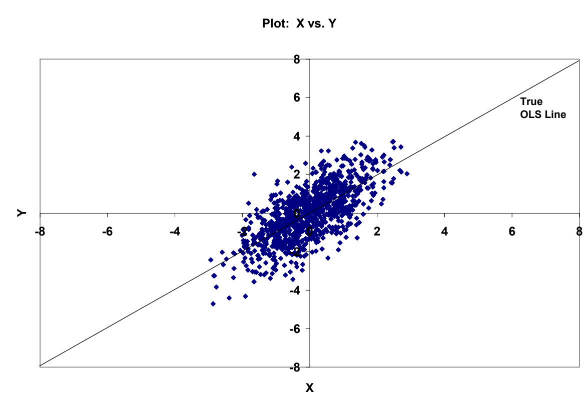


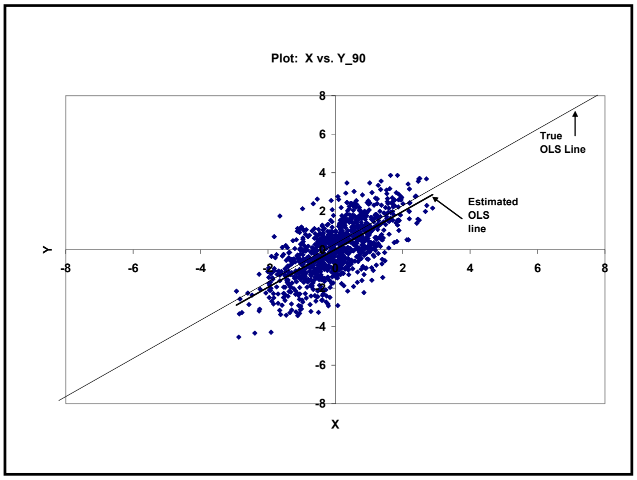


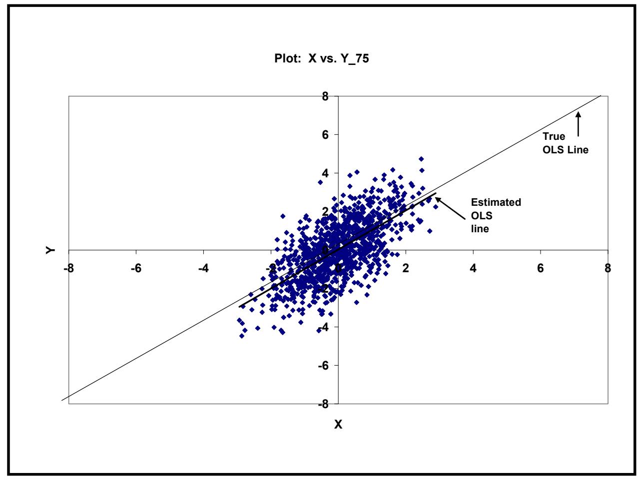
Central Limit Theorem

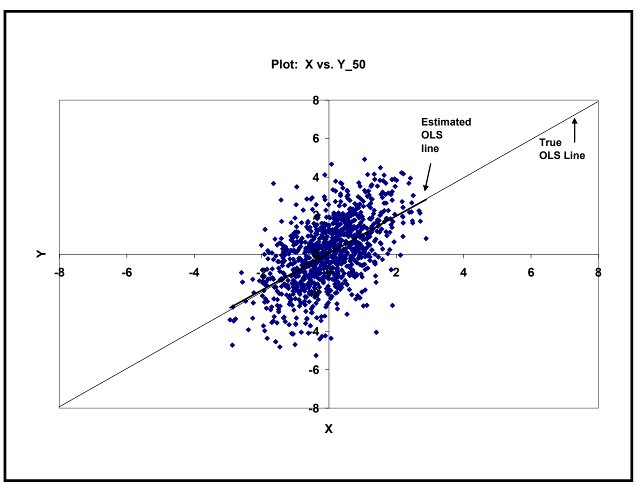
$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + \ldots.x_{ki}\beta_k + \varepsilon_i$$

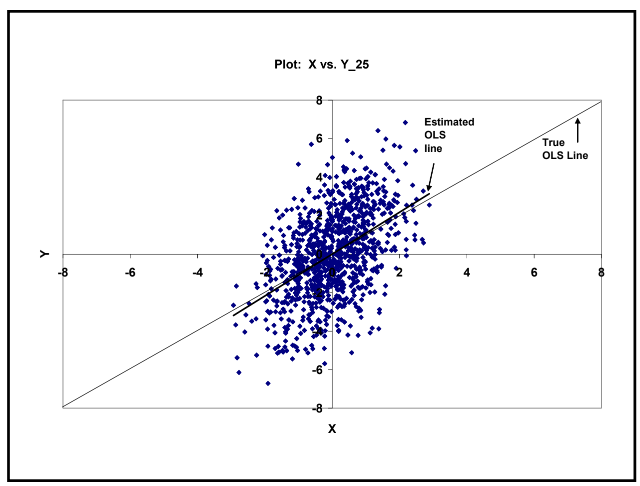
$$as\ n\ gets\ "big"\quad \frac{\left(\hat{\beta}_j - \beta_j\right)}{se(\hat{\beta}_j)} \sim N[0,1]$$

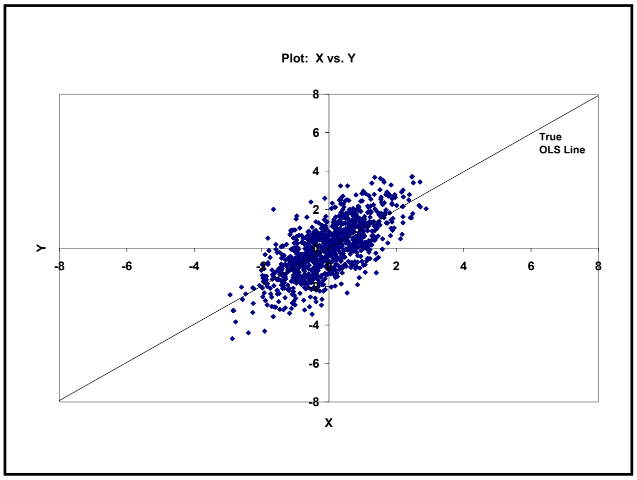


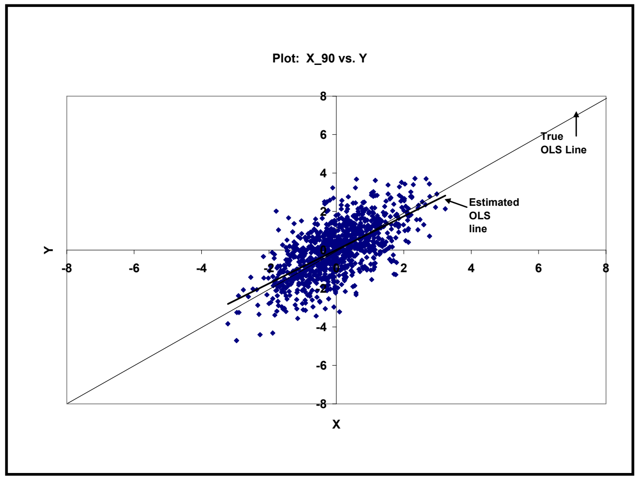

Plot: X vs. Y_90
True
OLS Line
Estimated
OLS
line
Y
X

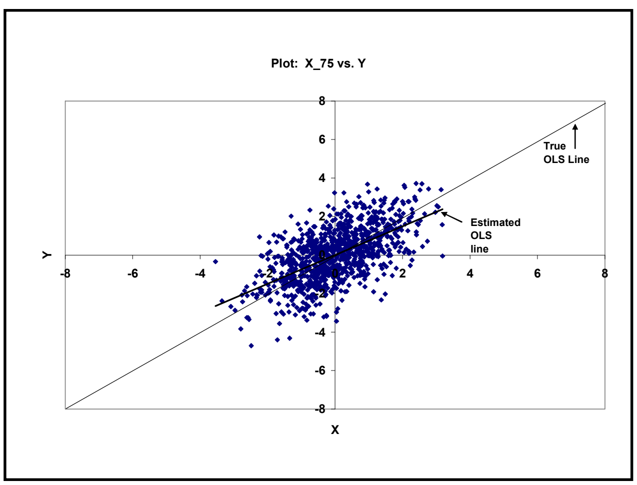

Plot: X vs. Y_75
True
OLS Line
Estimated
OLS
line
Y
X

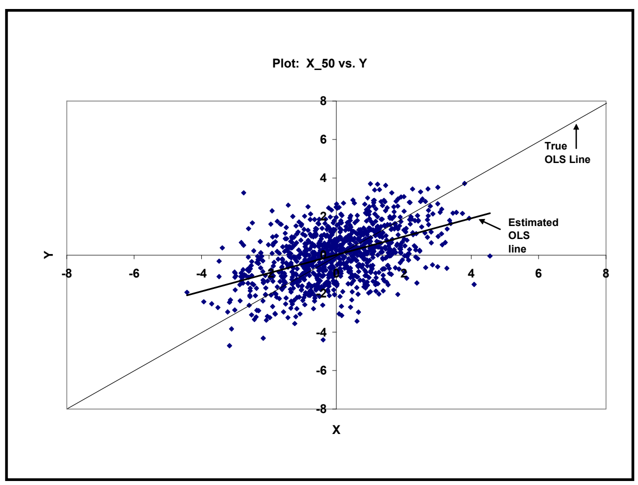

Plot: X vs. Y_50
Estimated
OLS
line
True
OLS Line
Y
X

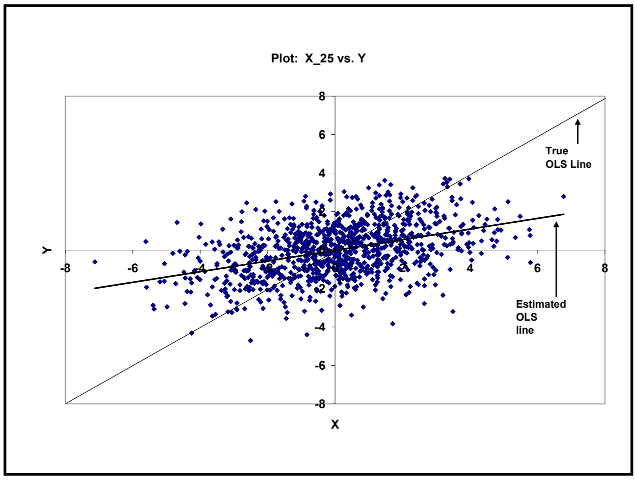

Plot: X vs. Y_25
Estimated
OLS
line
True
OLS Line
Y
X


Plot: X vs. Y
True
OLS Line
Y
X


Plot: X_90 vs. Y
True
OLS Line
Estimated
OLS
line
Y
X


Plot: X_75 vs. Y
True
OLS Line
Estimated
OLS
line
Y
X


Plot: X_50 vs. Y
True
OLS Line
Estimated
OLS
line
Y
X



Plot: X_25 vs. Y
True
OLS Line
Estimated
OLS
line
Y
X