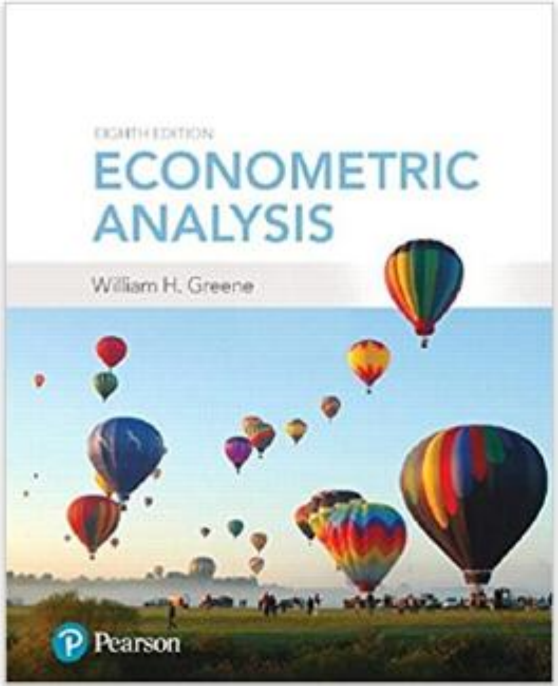
Econometrics I

Professor William Greene

Stern School of Business

Department of Economics



Econometrics I

Part 9 – Asymptotics for the Regression Model

Asymptotic Properties

- Main finite sample results are not necessarily useful
  - **1. Unbiased**: Not very useful if the estimator does not improve as more information is added. (The mean of the first 10 observations in a sample of n observations is unbiased, but not a good estimator.)
  - **2. Variance under narrow assumptions**: Assumptions are often not met in realistic settings, so standard errors might be inaccurate.
  - **3. Gauss-Markov theorem**: Not necessarily interested in the most efficient estimator if we have to make unrealistic assumptions.
  - **4. Normal distribution of estimator**: We would rather not make this assumption if not necessary. It is generally not necessary.
- Overriding principle: Robustness to loosened assumptions.

Asymptotic Properties of Interest

- Consistent Estimator of β
- Asymptotic Normal Distribution
- Appropriate Asymptotic Variance

We develop properties that will hold in large samples then assume they hold acceptably well in finite observed samples. The objective is to relax the assumptions of the model where possible.

Core Results

The least squares estimator is

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})$$

$$=\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$$

$$=\boldsymbol{\beta}+\left(\frac{1}{\mathrm{n}}\sum_{i=1}^{\mathrm{n}}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{\mathrm{n}}\sum_{i=1}^{\mathrm{n}}\mathbf{x}_i\varepsilon_i\right)$$

So, it equals $\boldsymbol{\beta}$ plus the sampling variability = the estimation error

The question for the present is how does this sum of random variables behave in large samples?

Well Behaved Regressors

A crucial assumption: Convergence of the moment matrix **X′X**/n to a positive definite matrix of finite elements, **Q.**

For now, we assume that data on the rows of **X** act like a random sample of observations that are independent of the sample of observations on ε.

Crucial Assumption of the Model

What must be assumed to get $\text{plim}\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon}\right) = \left(\frac{1}{n}\sum_{i=1}^{N}\mathbf{x}_i\varepsilon_i\right) = \mathbf{0}$?

(1) $\mathbf{x}_i$ = a random vector with finite means and variance and identical joint distributions.

(2) ε_i = a random variable with a constant distribution with $E[\varepsilon_i|\mathbf{x}_i]=0$ and $Var[\varepsilon_i|\mathbf{x}_i]=0\ \sigma^2$

(3) $\mathbf{x}_i$ and ε_i statistically independent.

Then, $\mathbf{w}_i = \mathbf{x}_i\varepsilon_i$ = an observation in a random sample, with constant variance matrix and mean vector 0.

$\bar{\mathbf{w}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_i$ converges to its expectation by the law of large numbers.

If $\text{plim}[(1/n)\mathbf{X}'\boldsymbol{\varepsilon}] = \mathbf{0}$, then $\text{plim}\ \mathbf{b} = \beta + \mathbf{Q}^{-1}\mathbf{0} = \beta$.

Mean Square Convergence of b

We use convergence in mean square. Adequate for almost all problems.

$$\mathbf{b} = \beta + \left(\frac{1}{n}\mathbf{X'X}\right)^{-1} \times \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \varepsilon_i\right)$$

$$(\mathbf{b}-\beta)(\mathbf{b}-\beta)' = \left(\frac{1}{n}\mathbf{X'X}\right)^{-1} \times \left[\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \varepsilon_i\right) \times \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i' \varepsilon_i\right)\right]\left(\frac{1}{n}\mathbf{X'X}\right)^{-1}$$

$$= \left(\frac{1}{n}\mathbf{X'X}\right)^{-1} \left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{x}_i \varepsilon_i \mathbf{x}_j' \varepsilon_j\right)\left(\frac{1}{n}\mathbf{X'X}\right)^{-1}$$

In $E[(\mathbf{b}-\beta)(\mathbf{b}-\beta)' \mid \mathbf{X}]$ in the double sum, terms with unequal subscripts have expectation zero.

$$E[(\mathbf{b}-\beta)(\mathbf{b}-\beta)' \mid \mathbf{X}] = \left(\frac{1}{n}\mathbf{X'X}\right)^{-1} \left(\frac{1}{n^2}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \underbrace{E[\varepsilon_i^2 \mid \mathbf{X}]}_{\sigma^2}\right)\left(\frac{1}{n}\mathbf{X'X}\right)^{-1}$$

$$= \frac{\sigma^2}{n}\left(\frac{1}{n}\mathbf{X'X}\right)^{-1}\left(\frac{1}{n}\mathbf{X'X}\right)\left(\frac{1}{n}\mathbf{X'X}\right)^{-1} = \frac{\sigma^2}{n}\left(\frac{1}{n}\mathbf{X'X}\right)^{-1}$$

Mean Square Convergence

E[**b**|**X**]=**β** for any **X**.

Var[**b**|**X**]→0 for well behaved **X**

$$\frac{\sigma^2}{n}\left(\frac{1}{n}\mathbf{X'X}\right)^{-1} \longrightarrow 0 \times \mathbf{Q}^{-1}$$

b converges in mean square to **β**

s^2 **= e'e/**n is a consistent estimator of σ^2. The usual estimator is **e'e**/(n-K)

The *estimator* of the **asymptotic variance** of **b** is $(s^2/n)[\mathbf{X'X}/n]^{-1} = s^2(\mathbf{X'X})^{-1}$

Asymptotic Distribution

$$\mathbf{b} = \beta + \left(\frac{1}{n}\mathbf{X'X}\right)^{-1} \times \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right)$$

The limiting behavior of **b** is the same as that of the statistic that results when the moment matrix is replaced by its limit. We examine the behavior of the modified sum

$$\beta + \mathbf{Q}^{-1} \times \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right)$$

Asymptotic Distribution

- Finding the asymptotic distribution
- **b** → **β** in probability. How to describe the distribution?
  - 'Limiting' distribution
    - Variance → **0**; it is $O(1/n)$
    - Stabilize the variance? $\text{Var}[\sqrt{n}\,\mathbf{b}] \sim \sigma^2\mathbf{Q}^{-1}$ is $O(1)$
    - But, $E[\sqrt{n}\,\mathbf{b}] = \sqrt{n}\,\boldsymbol{\beta}$ which diverges.
  - $\sqrt{n}\,(\mathbf{b} - \boldsymbol{\beta})$ → a random variable with finite mean and variance. (stabilizing transformation)
  - **b** apx. **β** $+1/\sqrt{n}$ times that random variable

The Asymptotic Distribution

Limiting distribution of

$$\sqrt{n}(\mathbf{b}-\beta) = \sqrt{n}\left(\frac{\mathbf{X'X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\varepsilon}{n}\right)$$

is the same as that of $\mathbf{Q}^{-1}\sqrt{n}\ \bar{\mathbf{w}} = \sqrt{n}\ \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right)$

$\sqrt{n}\bar{\mathbf{w}} \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}]$ (By Lindeberg-Feller CLT)

Therefore,

$$\mathbf{Q}^{-1}\sqrt{n}\bar{\mathbf{w}} \xrightarrow{d} N[0,\ \mathbf{Q}^{-1}(\sigma^2\mathbf{Q})\ \mathbf{Q}^{-1}] = N[\mathbf{0},\ \sigma^2\mathbf{Q}^{-1}]$$

Conclude : $\sqrt{n}(\mathbf{b}-\beta) \xrightarrow{d} N[\mathbf{0},\ \sigma^2\mathbf{Q}^{-1}]$

Approximately : $\mathbf{b} \xrightarrow{a} N[\beta,\ (\sigma^2/n)\mathbf{Q}^{-1}]$

Asymptotic Results

- **b** is a consistent estimator of $\boldsymbol{\beta}$
- Asymptotic normal distribution of **b** does not depend on normality of ε; it depends on the Central Limit Theorem
- Estimator of the asymptotic variance $(\sigma^2/n)\mathbf{Q}^{-1}$ is $(s^2/n)\,(\mathbf{X'X}/n)^{-1}$. (Degrees of freedom corrections are irrelevant but conventional.)
- Slutsky theorem and the delta method apply to functions of **b** using Est.Asy.Var[**b**] = $s^2(\mathbf{X'X})^{-1}$.

An Application:
Cornwell and Rupert Labor Market Data
Is Wage Related to Education?

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.

Regression with Conventional Standard Errors

Reported Standard Errors are Square roots of Diagonal Elements of $s^2(\mathbf{X'X})^{-1}$

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

Robustness

- Thus far, the model is a specific set of assumptions.
- Consistency and asymptotic normality of **b**, and the form of the asymptotic covariance matrix follow from the assumptions.
- Robust Estimation and Inference
  - Estimators that "work" even if some of the assumptions are not met.
  - E.g., LS is robust to failure of normality of ε - only the finite sample distribution is incorrect. It is still unbiased and Var[**b**|**X**]=$\sigma^2(\mathbf{X'X})^{-1}$.
  - Is least squares robust to failures of other assumptions?
- Robustness to failures of other assumptions:
  - (**E**) Suppose Cov[$\mathbf{x}_i,\varepsilon_i$] $\neq$ 0? (Endogeneity) No. Unbiasedness fails.
  - (**H**) Heteroscedasticity? Still unbiased. Variance is no longer $\sigma^2(\mathbf{X'X})^{-1}$.
  - (**A**) Autocorrelation? Same.
- We continue to use **b** but look for an alternative to $s^2(\mathbf{X'X})^{-1}$ that will be appropriate in cases **H** and **A**. (Case **E** is not workable yet.)

A Robust Covariance Matrix

- Crucial assumptions about ε
  - Homoscedastic, $Var[\varepsilon_i]=\sigma^2$.
  - Nonautocorrelation, $Cov[\varepsilon_i,\varepsilon_j] = 0$ for all i,j.
- Leading cases
  - Heteroscedasticity; $Var[\varepsilon_i]=\sigma_i^2$.
  - Groupwise correlation; $Cov[\varepsilon_{it},\varepsilon_{is}] \neq 0$ for observations t and s in group i, i = 1,…,C_i.
  - Time series autocorrelation; $Cov[\varepsilon_t,\varepsilon_s] \neq 0$
- General result: **b** remains consistent and asymptotically normally distributed. Asymptotic covariance matrix is incorrect.

Robustness to Heteroscedasticity

$$E[(\mathbf{b}-\beta)(\mathbf{b}-\beta)'\,|\,\mathbf{X}] = \left(\frac{1}{n}\mathbf{X'X}\right)^{-1}\left(\frac{1}{n^2}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\,E[\varepsilon_i^2\,|\,\mathbf{x}_i]\right)\left(\frac{1}{n}\mathbf{X'X}\right)^{-1}$$

$$= \left(\frac{1}{n}\mathbf{X'X}\right)^{-1}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2\mathbf{x}_i\mathbf{x}_i'\right)\left(\frac{1}{n}\mathbf{X'X}\right)^{-1}$$

In large samples,

$$= \frac{1}{n}\mathbf{Q}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\hat{\mathbf{Q}}_i\right)\mathbf{Q}^{-1}$$

We seek a matrix that will mimic this matrix whether σ_i^2 varies across observations or not. (I.e., a robust estimator.)

Heteroscedasticity Robust Covariance Matrix

The White Estimator

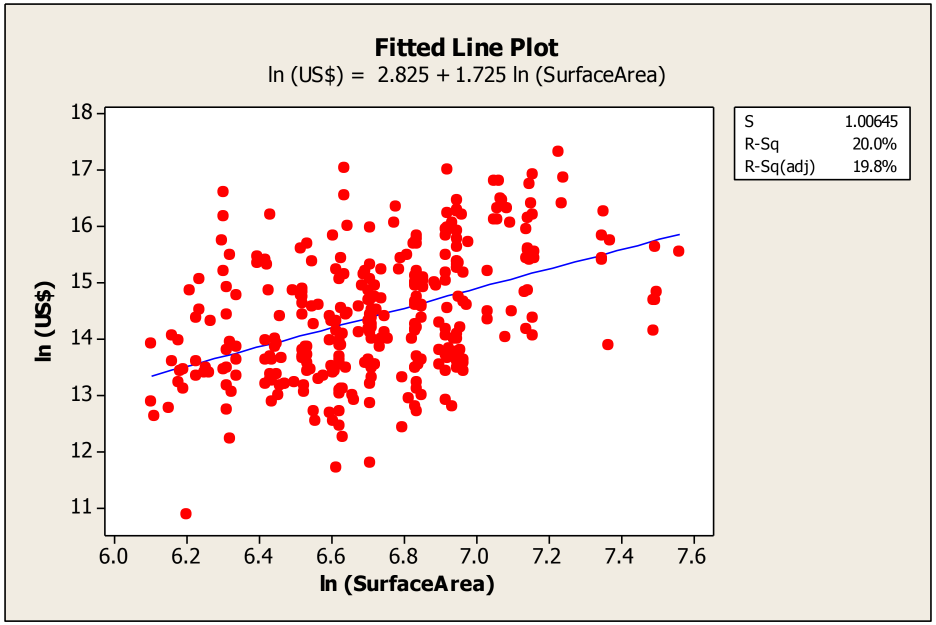
$$\text{Est.Asy.Var}[\mathbf{b}] = (\mathbf{X}'\mathbf{X})^{-1}\left[\sum_i e_i^2 \mathbf{x}_i \mathbf{x}_i'\right](\mathbf{X}'\mathbf{X})^{-1}$$

e_i = least squares residual.

- **Robust standard errors**; (**b** is not "robust")
  - Robust to: Heteroscedasticty
  - Not robust to: (all considered later)
    - Correlation across observations
    - Individual unobserved heterogeneity
    - Incorrect model specification
- **Robust inference** means hypothesis tests and confidence intervals using robust covariance matrices
- **Wisdom**: Robust standard errors are usually larger, but not always.

Monet in Large and Small

Sale prices of 430 signed Monet paintings



Log of \$price = a + b_1 log surface area + b_2 aspect ratio + ε

TABLE 4.4 Robust Standard Errors

Variable	*Estimated Coefficient*	*LS Standard Error*	*Heteroscedasticity Robust Std.Error*
Constant	–8.34237	0.67820	0.73342
lnArea	1.31638	0.09205	0.10598
Aspect Ratio	–0.09623	0.15784	0.16706

Heteroscedasticity 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

Note the conflict: Test favors heteroscedasticity.
Robust VC matrix is essentially the same.

Cluster Robust Covariance Matrix

The least squares estimator is

$$\mathbf{b} = \boldsymbol{\beta} + \left(\sum_{c=1}^{C} \mathbf{X}_c'\mathbf{X}_c\right)^{-1}\left[\sum_{c=1}^{C}\left(\sum_{i=1}^{N} \mathbf{x}_{i,c}\varepsilon_{i,c}\right)\right] = \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left[\sum_{c=1}^{C}\left(\mathbf{X}_c'\boldsymbol{\varepsilon}_c\right)\right],$$

where $\mathbf{X}_c$ is the $N_c \times K$ matrix of exogenous variables for cluster c and $\boldsymbol{\varepsilon}_c$ is the N_c disturbances for the group. Assuming that the clusters are independent,

$$\text{Var}[\mathbf{b}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left[\sum_{c=1}^{C} \mathbf{X}_c\boldsymbol{\Omega}_c\mathbf{X}_c'\right]\left(\mathbf{X}'\mathbf{X}\right)^{-1}.$$

Like σ_i^2 before, $\boldsymbol{\Omega}_c$ is not meant to suggest a particular set of population parameters. Rather, $\boldsymbol{\Omega}_c$ represents the possibly unstructured correlations allowed among the N_c disturbances in cluster c. The construction is essentially the same as the White estimator, though $\boldsymbol{\Omega}_c$ is the matrix of variances and covariances for the full vector $\boldsymbol{\varepsilon}_c$. (It would be identical to the White estimator if each cluster contained one observation.) Taking the same approach as before, we obtain the asymptotic variance

$$\text{Asy.Var}[\mathbf{b}] = \frac{1}{C}\mathbf{Q}^{-1}\left[\text{plim}\frac{1}{C}\sum_{c=1}^{C} \mathbf{X}_c\boldsymbol{\Omega}_c\mathbf{X}_c'\right]\mathbf{Q}^{-1}.$$

A feasible estimator of the bracketed matrix based on the least squares residuals is

$$\mathbf{W}_{cluster} = \frac{1}{C}\sum_{c=1}^{C}\left(\mathbf{X}_c'\mathbf{e}_c\right)\left(\mathbf{e}_c'\mathbf{X}_c\right) = \frac{1}{C}\sum_{c=1}^{C}\left(\sum_{i=1}^{N_c} \mathbf{x}_{ic}e_{ic}\right)\left(\sum_{i=1}^{N_c} \mathbf{x}_{ic}e_{ic}\right)'.$$

Then,

$$\text{Est.Asy.Var}[\mathbf{b}] = C(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{W}_{cluster}(\mathbf{X}'\mathbf{X})^{-1}.$$

http://cameron.econ.ucdavis.edu/research/Cameron_Miller_Cluster_Robust_October152013.pdf

A Cluster Estimator

Robust variance estimator for Var[$\mathbf{b}|\mathbf{X}$]

Est.Var[$\mathbf{b}|\mathbf{X}$]

$$= (\mathbf{X}'\mathbf{X})^{-1}\left[\left(\frac{N}{N-1}\right)\Sigma_{i=1}^{N}\left(\Sigma_{t=1}^{T_i}\mathbf{x}_{it}\hat{w}_{it}\right)\left(\Sigma_{t=1}^{T_i}\mathbf{x}_{it}'\hat{w}_{it}\right)\right](\mathbf{X}'\mathbf{X})^{-1}$$

$$\boxed{= (\mathbf{X}'\mathbf{X})^{-1}\left[\left(\frac{N}{N-1}\right)\Sigma_{i=1}^{N}\left(\Sigma_{t=1}^{T_i}\Sigma_{s=1}^{T_i}\hat{w}_{it}\hat{w}_{is}\mathbf{x}_{it}\mathbf{x}_{is}'\right)\right](\mathbf{X}'\mathbf{X})^{-1}}$$

$\hat{w}_{it}$ = a least squares residual.

(If $T_i = 1$, this is the White estimator.)

(The finite population correction $[N/(N-1)]$ is ad hoc.)

(The further finite population correction, $\left(\dfrac{\Sigma_{i=1}^{N}T_i}{\left(\Sigma_{i=1}^{N}T_i\right)-N}\right)$ is also ad hoc.)

Example 4.5 Robust Inference About the Art Market

The Monet paintings examined in Example 4.3 were sold at auction over 1989–2006. Our model thus far is

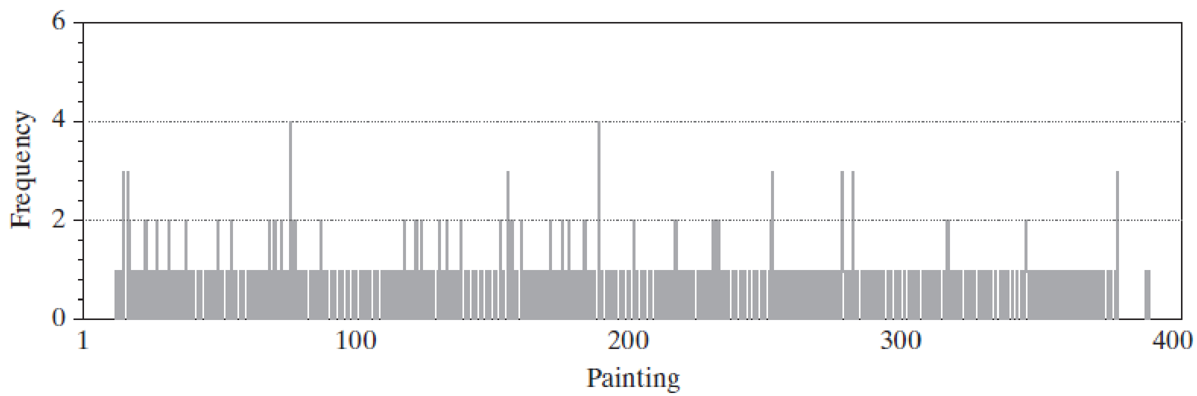
$$\ln Price_{it} = \beta_1 + \beta_2 \ln Area_{it} + \beta_3 AspectRatio_{it} + \varepsilon_{it}$$

The subscript "*it*" uniquely identifies the painting and when it was sold. Prices in open outcry auctions reflect (at least) three elements, the common (public), observable features of the item, the public unobserved (by the econometrician) elements of the asset, and the private unobservable preferences of the winning bidder. For example, it will turn out (in a later example) that whether the painting is signed or not has a large and significant influence on the price. For now, we assume (for sake of the example), that we do not observe whether the painting is signed or not, though, of course, the winning bidders do observe this. It does seem reasonable to suggest that the presence of a signature is uncorrelated with the two attributes we do observe, area and aspect ratio. We respecify the regression as

$$\ln Price_{it} = \beta_1 + \beta_2 \ln Area_{it} + \beta_3 AspectRatio_{it} + w_{it} + u_{it},$$

where w_{it} represents the intrinsic, unobserved features of the painting and u_{it} represents the unobserved preferences of the buyer.

FIGURE 4.5 Repeat Sales of Monet Paintings.



Robust Asymptotic Covariance Matrices

TABLE 4.4 Robust Standard Errors

Variable	Estimated Coefficient	LS Standard Error	Heteroscedasticity Robust Std.Error	Cluster Robust Std.Error
Constant	–8.34237	0.67820	0.73342	0.75873
lnArea	1.31638	0.09205	0.10598	0.10932
Aspect Ratio	–0.09623	0.15784	0.16706	0.17776

Alternative OLS Variance Estimators

Cluster correction increases SEs

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]
Constant	5.40159723	.04838934	111.628	.0000
EXP	.04084968	.00218534	18.693	.0000
EXPSQ	-.00068788	.480428D-04	-14.318	.0000
OCC	-.13830480	.01480107	-9.344	.0000
SMSA	.14856267	.01206772	12.311	.0000
MS	.06798358	.02074599	3.277	.0010
FEM	-.40020215	.02526118	-15.843	.0000
UNION	.09409925	.01253203	7.509	.0000
ED	.05812166	.00260039	22.351	.0000
Robust				
Constant	5.40159723	.10156038	53.186	.0000
EXP	.04084968	.00432272	9.450	.0000
EXPSQ	-.00068788	.983981D-04	-6.991	.0000
OCC	-.13830480	.02772631	-4.988	.0000
SMSA	.14856267	.02423668	6.130	.0000
MS	.06798358	.04382220	1.551	.1208
FEM	-.40020215	.04961926	-8.065	.0000
UNION	.09409925	.02422669	3.884	.0001
ED	.05812166	.00555697	10.459	.0000

```
Namelist ; x=one,exp,expsq,occ,smsa,ms,fem,union,ed$
Regress  ; Lhs = lwage ; rhs=x ; cluster=7$
```

ROTTEN APPLES: AN INVESTIGATION OF THE PREVALENCE AND PREDICTORS OF TEACHER CHEATING

Brian A. Jacob
Steven D. Levitt

Working Paper 9413
http://www.nber.org/papers/w9413

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138

Table 7: Patterns of Cheating within Classrooms and Schools

Independent Variables	**Dependent variable = Class suspected of cheating** (Class is above the 95th percentile on both SCORE and ANSWERS on a particular subject test: mean=0.011) Full Sample		Sample of classes and school that existed in the prior year	
Classroom cheated on exactly one other subject this year on this	0.105 (0.008)	0.103 (0.008)	0.101 (0.009)	0.101 (0.009)
Classroom cheated on exactly two other subjects this year	0.289 (0.027)	0.285 (0.027)	0.243 (0.031)	0.243 (0.031)
Number of other subjects this classroom cheated on last year	--	--	0.023 (0.004)	0.018 (0.004)
Cheating in this classroom ever in the past	--	--	--	0.006 (0.002)
Cheating rate among other classrooms in this school in past years	--	--	--	0.090 (0.040)
Full set of grade*subject*year interactions included?	Yes	Yes	Yes	Yes
R-squared	0.090	0.093	0.109	0.109
Number of Observations	165,578	165,578	94,182	94,170

Notes: The dependent variable is an indicator for whether a classroom is above the stated cutoff on ANSWERS and SCORE on a particular subject test. Estimation is done using a linear probability model. Columns that include measures of cheating in prior years, observations where that classroom and/or school does not appear in the data in the prior year are excluded. Standard errors are clustered at the school level to take into account correlations across classroom as well as serial correlation.

Bootstrap Estimation of 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 Algorithm

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

$$\mathbf{V} = \frac{1}{R}\sum_{r=1}^{R} [\mathbf{b}(r)-\bar{\mathbf{b}}][\mathbf{b}(r)-\bar{\mathbf{b}}]'$$

TABLE 4.4 Robust Standard Errors

Variable	*Estimated Coefficient*	*LS Standard Error*	*Heteroscedasticity Robust Std.Error*
Constant	–8.34237	0.67820	0.73342
lnArea	1.31638	0.09205	0.10598
Aspect Ratio	–0.09623	0.15784	0.16706

```
----------------------------------------------------------------------
Ordinary      least squares regression ............
LHS=Y         Mean                  =         .39445
              Standard deviation    =        1.34395
----------    No. of observations   =            430
Regression    Sum of Squares        =        266.785
Residual      Sum of Squares        =        508.079
Total         Sum of Squares        =        774.864
----------    Standard error of e   =        1.09082
Fit           R-squared             =         .34430
Model test    F[  2,   427]         =      112.10576
--------+-------------------------------------------------------------
        |  Estimated        Bootstrap                    Prob.
       Y|  Coefficient      Std.Error         t         |t|>T*
--------+-------------------------------------------------------------
Constant|   -8.34237***        .74193     -11.24      .0000
 LOGAREA|    1.31638***        .10560      12.47      .0000
  ASPECT|    -.09623           .16555       -.58      .5614
--------+-------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
Standard errors based on 500 bootstrap replications.
```

Log of $price = a +
b1 log surface area +
b2 aspect ratio + e

Application to Spanish Dairy Farms

$$\ln y = b_1 \ln x_1 + b_2 \ln x_2 + b_3 \ln x_3 + b_4 \ln x_4 + \varepsilon$$

Input	Units	Mean	Std. Dev.	Minimum	Maximum
Y Milk	Milk production (liters)	131,108	92,539	14,110	727,281
X1 Cows	# of milking cows	2.12	11.27	4.5	82.3
X2 Labor	# man-equivalent units	1.67	0.55	1.0	4.0
X3 Land	Hectares of land devoted to pasture and crops.	12.99	6.17	2.0	45.1
X4 Feed	Total amount of feedstuffs fed to dairy cows (tons)	57,941	47,981	3,924.14	376,732

N = 247 farms, T = 6 years (1993-1998)

Example: Bootstrap Replications

Matrix - BOOTSTRP

[500, 5] Cell: 11.5732

	1	2	3	4	5
1	11.5732	0.663365	-0.00172636	0.0274295	0.431544
2	11.5782	0.540323	0.0364048	0.0270598	0.46497
3	11.5756	0.604965	0.0337931	0.00543007	0.434848
4	11.5767	0.607056	0.0139364	0.0120334	0.462593
5	11.5775	0.615707	0.0265037	0.00852363	0.437757
6	11.5797	0.612065	0.0295112	0.0088207	0.441855
7	11.5793	0.585915	0.027172	0.0236685	0.455928
8	11.5831	0.586222	0.0343727	0.0223925	0.448991
9	11.5781	0.592765	0.0124331	0.0169125	0.459289
10	11.5783	0.608996	0.0119008	0.0217884	0.444427
11	11.5833	0.604306	0.0331359	0.0290024	0.439594
12	11.5795	0.557114	0.0405298	0.0354376	0.456092
13	11.5802	0.604786	0.0261206	0.027283	0.437104
14	11.5779	0.581646	0.0314108	0.0289908	0.445211
15	11.5811	0.604491	0.0225034	0.0356162	0.4571
16	11.5761	0.582383	0.0344234	0.0125611	0.456585
17	11.582	0.57254	0.0191405	0.0149456	0.460346
18	11.5785	0.583227	0.0218739	0.040026	0.461734
19	11.5716	0.577647	0.0256428	0.0110496	0.472882
20	11.5692	0.601854	0.00907908	0.0298847	0.451332
21	11.5775	0.627052	0.015275	0.0244473	0.44208

Bootstrapped Regression

Ordinary	least squares regression				
LHS=YIT	Mean	=	11.57749		
	Standard deviation	=	.64344		
----------	No. of observations	=	1482	DegFreedom	Mean square
Regression	Sum of Squares	=	584.056	4	146.01403
Residual	Sum of Squares	=	29.0957	1477	.01970
Total	Sum of Squares	=	613.152	1481	.41401
----------	Standard error of e	=	.14035	Root MSE	.14012
Fit	R-squared	=	.95255	R-bar squared	.95242
Model test	F[4, 1477]	=	7412.18529	Prob F > F*	.00000

YIT	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	11.5775***	.00365	3175.52	.0000	11.5703	11.5846
X1	.59518***	.01958	30.39	.0000	.55679	.63356
X2	.02305**	.01122	2.05	.0400	.00105	.04505
X3	.02319*	.01303	1.78	.0751	-.00235	.04873
X4	.45176***	.01078	41.89	.0000	.43062	.47290

Results of bootstrap estimation of model.
Model has been reestimated 500 times.
Coefficients shown below are the original
model estimates based on the full sample.
Bootstrap samples have 1482 observations.
Estimated parameter vector is B .
Estimated variance matrix saved as VARB.

BootStrp	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
B001	11.5775***	.00365	3174.06	.0000	11.5703	11.5846
B002	.59518***	.02283	26.07	.0000	.55044	.63992
B003	.02305*	.01186	1.94	.0520	-.00020	.04630
B004	.02319*	.01227	1.89	.0588	-.00086	.04724
B005	.45176***	.01221	36.99	.0000	.42782	.47569

Quantile Regression

- $Q(y|\mathbf{x},\alpha) = \beta'\mathbf{x}$, α = quantile
- Estimated by linear programming
- $Q(y|\mathbf{x},.50) = \beta'\mathbf{x}$, .50 ➔ median regression
- Median regression estimated by LAD (estimates same parameters as mean regression if symmetric conditional distribution)
- Why use quantile (median) regression?
  - Semiparametric
  - Robust to some extensions (heteroscedasticity?)
  - Complete characterization of conditional distribution

Estimated Variance for Quantile Regression

Asymptotic Theory Based Estimator of Variance of Q - REG

Model: $y_i = \boldsymbol{\beta}'\mathbf{x}_i + u_i,\ \ Q(y_i \,|\, \mathbf{x}_i, \alpha) = \boldsymbol{\beta}'\mathbf{x}_i,\ Q[u_i \,|\, \mathbf{x}_i, \alpha] = 0$

Residuals: $\hat{u}_i = y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i$

Asymptotic Variance: $\frac{1}{n}\left(\mathbf{A}^{-1}\ \mathbf{C}\ \mathbf{A}^{-1}\right)$

$\mathbf{A} = \mathrm{E}[\mathrm{f}_u(0)\mathbf{xx}']$ Estimated by $\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\mathrm{B}}\frac{1}{2}1\left[|\,\hat{u}_i\,| < \mathrm{B}\right]\mathbf{x}_i\mathbf{x}_i'$

Bandwidth B can be Silverman's Rule of Thumb:

$$\frac{1.06}{n^{.2}} Min\left(s_u, \frac{Q(\hat{u}_i \,|\, .75) - Q(\hat{u}_i \,|\, .25)}{1.349} \right)$$

$\mathbf{C} = \alpha(1-\alpha)E[\mathbf{xx}']$ Estimated by $\frac{\alpha(1-\alpha)}{n}\mathbf{X}'\mathbf{X}$

For α=.5 and normally distributed u, this all simplifies to $\frac{\pi}{2}s_u^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$.

But, this is an ideal application for bootstrapping.

Quantile Regressions

```
Quantile Regression Model. Quantile =         .250000
Linear Programming estimation method
LHS=YIT      Mean                  =         11.57749
             Standard deviation    =           .64322
             Number of observs.    =             1482
             Minimum               =          9.57568
             t= .25000 quantile    =         11.12856
             Maximum               =         13.49707
Model size   Parameters            =                5
             Degrees of freedom    =             1477
Residuals    Sum of squares        =         40.53555
             Standard error of e   =           .14075
Fit          R-squared             =           .95212
             PseudoR2=1-F(0)/F(b)  =           .78027
Not using OLS or no constant. Rsquared may be <= 0
Functions F= Sum r(t)[y(i)-x(i)b] =          67.52015
          F0=Sum r(t)[y(i)-Qy(t)] =         307.29243
             r(t)[u]=t*u-u*[u<0].t=           .250000
Asymptotic cov. matrix based on  kernel estimator.
Heteroscedasticity test, Chi2[ 4] = 42.44 P = .000
```

$\alpha = .25$

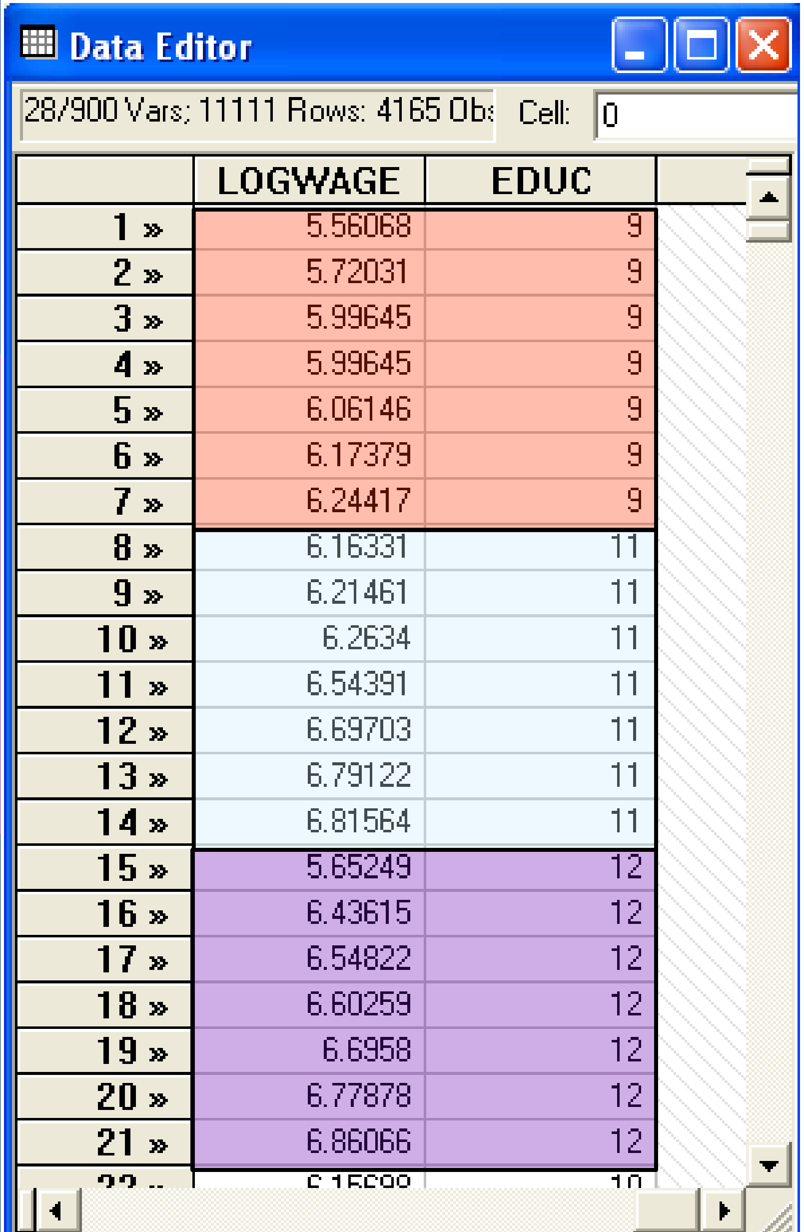
YIT	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	11.4906***	.00511	2248.82	.0000	11.4806	11.5007
X1	.60321***	.02732	22.08	.0000	.54967	.65676
X2	.01440	.01652	.87	.3836	-.01799	.04678
X3	.03818**	.01857	2.06	.0397	.00179	.07457
X4	.46244***	.01455	31.78	.0000	.43391	.49096

$\alpha = .50$

Constant	11.5897***	.00447	2595.57	.0000	11.5809	11.5984
X1	.57412***	.02398	23.94	.0000	.52711	.62113
X2	.03547***	.01374	2.58	.0099	.00853	.06241
X3	.00140	.01596	.09	.9301	-.02988	.03268
X4	.46230***	.01321	35.00	.0000	.43641	.48819

$\alpha = .75$

Constant	11.6738***	.00409	2855.24	.0000	11.6658	11.6818
X1	.58170***	.02245	25.91	.0000	.53769	.62570
X2	.04573***	.01168	3.92	.0001	.02285	.06862
X3	.02409*	.01389	1.73	.0828	-.00313	.05131
X4	.43929***	.01258	34.92	.0000	.41463	.46394



Data Editor

28/900 Vars; 11111 Rows: 4165 Obs Cell: 0

	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

The bootstrap replication must account for panel data nature of the data set.

Bootstrap variance for a panel data estimator

- $\boxed{\text{Panel Bootstrap = Block Bootstrap}}$
- Data set is N groups of size T_i
- Bootstrap sample is N groups of size T_i drawn with replacement.

LWAGE	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval		
							OLS
Constant	5.66098***	.04686	120.81	.0000	5.56914	5.75282	
OCC	-.11220***	.01464	-7.66	.0000	-.14090	-.08350	
SMSA	.15504***	.01234	12.57	.0000	.13086	.17922	
MS	.09569***	.02133	4.49	.0000	.05387	.13751	
FEM	-.39478***	.02603	-15.16	.0000	-.44581	-.34376	
ED	.05688***	.00268	21.24	.0000	.05163	.06213	
EXP	.01044***	.00054	19.26	.0000	.00938	.01150	
							Bootstrap
B001	5.66098***	.04683	120.89	.0000	5.56920	5.75276	
B002	-.11220***	.01326	-8.46	.0000	-.13820	-.08620	Assumes no
B003	.15504***	.01205	12.87	.0000	.13143	.17866	correlation
B004	.09569***	.01953	4.90	.0000	.05742	.13396	within groups
B005	-.39478***	.01863	-21.19	.0000	-.43129	-.35827	
B006	.05688***	.00325	17.52	.0000	.05052	.06324	
B007	.01044***	.00053	19.67	.0000	.00940	.01148	
							Cluster
Constant	5.66098***	.10026	56.46	.0000	5.46447	5.85750	
OCC	-.11220***	.02653	-4.23	.0000	-.16421	-.06020	Accounts for
SMSA	.15504***	.02540	6.10	.0000	.10526	.20483	within group
MS	.09569**	.04657	2.05	.0399	.00442	.18696	correlation
FEM	-.39478***	.05319	-7.42	.0000	-.49904	-.29052	
ED	.05688***	.00568	10.01	.0000	.04574	.06802	
EXP	.01044***	.00132	7.93	.0000	.00786	.01302	
							Block Bootstrap
B001	5.66098***	.09497	59.61	.0000	5.47484	5.84712	
B002	-.11220***	.02617	-4.29	.0000	-.16349	-.06092	Mimics results
B003	.15504***	.02351	6.60	.0000	.10897	.20112	of panel
B004	.09569***	.03542	2.70	.0069	.02627	.16511	correction
B005	-.39478***	.04287	-9.21	.0000	-.47880	-.31077	
B006	.05688***	.00536	10.61	.0000	.04637	.06739	
B007	.01044***	.00138	7.57	.0000	.00774	.01314	

```
PROC
REGRESS ; Quiet ; Lhs=lwage ; Rhs=x $
ENDPROC
EXEC ; N = 100 ; Bootstrap = b $
EXEC ; N = 100 ; Bootstrap = b ; pds=7 $
```

```
--------------------------------------------------------------------------------
Ordinary      least squares regression ............
LHS=YIT       Mean                     =       11.57749
              Standard deviation       =         .64344
----------    No. of observations      =           1482  DegFreedom   Mean square
Regression    Sum of Squares           =        584.056           4     146.01403
Residual      Sum of Squares           =        29.0957        1477        .01970
Total         Sum of Squares           =        613.152        1481        .41401
----------    Standard error of e      =         .14035  Root MSE          .14012
Fit           R-squared                =         .95255  R-bar squared     .95242
Model test    F[  4,  1477]            =     7412.18529  Prob F > F*       .00000
```

Conventional

```
--------+-----------------------------------------------------------------------
        |               Standard            Prob.       95% Confidence
     YIT|  Coefficient    Error       z    |z|>Z*          Interval
--------+-----------------------------------------------------------------------
Constant|    11.5775***   .00365  3175.52  .0000     11.5703    11.5846
      X1|     .59518***   .01958    30.39  .0000      .55679     .63356
      X2|     .02305**    .01122     2.05  .0400      .00105     .04505
      X3|     .02319*     .01303     1.78  .0751     -.00235     .04873
      X4|     .45176***   .01078    41.89  .0000      .43062     .47290
--------+-----------------------------------------------------------------------
```

Cluster Robust

```
--------+-----------------------------------------------------------------------
        |               Clustered           Prob.       95% Confidence
     YIT|  Coefficient  Std.Error     z    |z|>Z*          Interval
--------+-----------------------------------------------------------------------
Constant|    11.5775***   .00754  1534.61  .0000     11.5627    11.5923
      X1|     .59518***   .04147    14.35  .0000      .51389     .67646
      X2|     .02305      .02101     1.10  .2725     -.01812     .06422
      X3|     .02319      .02258     1.03  .3044     -.02107     .06745
      X4|     .45176***   .02312    19.54  .0000      .40644     .49707
--------+-----------------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
Standard errors clustered on FARM     (  247 clusters)
```

Block Bootstrap

```
--------+-----------------------------------------------------------------------
        |  Estimated    Bootstrap            Prob.       95% Confidence
     YIT|  Coefficient  Std.Error     z     |z|>Z*          Interval
--------+-----------------------------------------------------------------------
Constant|    11.5775***   .00723  1602.20   .0000     11.5633    11.5916
      X1|     .59518***   .04249    14.01   .0000      .51189     .67846
      X2|     .02305      .02188     1.05   .2920     -.01983     .06593
      X3|     .02319      .02568      .90   .3664     -.02713     .07352
      X4|     .45176***   .02567    17.60   .0000      .40144     .50207
--------+-----------------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
Standard errors based on 100 bootstrap replications.
Bootstrap replications use  247 draws.
--------------------------------------------------------------------------------
```