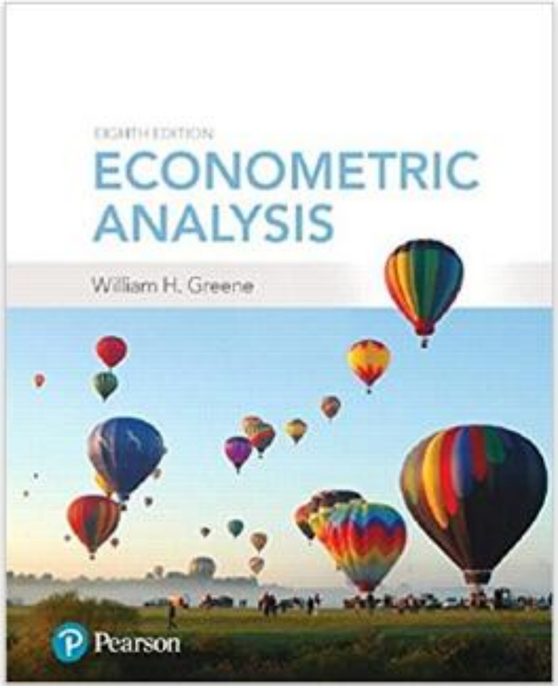
Econometrics I

Professor William Greene

Stern School of Business

Department of Economics



Econometrics I

Part 7 – Finite Sample Properties of Least Squares; Multicollinearity

Terms of Art

- Estimates and estimators
- Properties of an estimator - the sampling distribution
- “Finite sample” properties as opposed to “asymptotic” or “large sample” properties
- Scientific principles behind sampling distributions and ‘repeated sampling’

Application: Health Care Panel Data

German Health Care Usage Data, 7,293 Individuals, Varying Numbers of Periods
Data downloaded from Journal of Applied Econometrics Archive. **There are altogether 27,326 observations. The number of observations per household ranges from 1 to 7. (Frequencies are: 1=1525, 2=2158, 3=825, 4=926, 5=1051, 6=1000, 7=987).**

Variables in the file are

DOCVIS	=	**number of doctor visits in last three months**
HOSPVIS	=	**number of hospital visits in last calendar year**
DOCTOR	=	**1(Number of doctor visits > 0)**
HOSPITAL	=	**1(Number of hospital visits > 0)**
HSAT	=	**health satisfaction, coded 0 (low) - 10 (high)**
PUBLIC	=	**insured in public health insurance = 1; otherwise = 0**
ADDON	=	**insured by add-on insurance = 1; otherswise = 0**
HHNINC	=	**household nominal monthly net income in German marks / 10000**. (4 observations with income=0 were dropped)
HHKIDS	=	**children under age 16 in the household = 1; otherwise = 0**
EDUC	=	**years of schooling**
AGE	=	**age in years**
MARRIED	=	**marital status**

For now, treat this sample as if it were a cross section, and as if it were the full population.

Population Regression of Household Income on Education

```
--------------------------------------------------------------------------------
Ordinary     least squares regression ............
LHS=HHNINC   Mean                   =        .35208
             Standard deviation     =        .17691
----------   No. of observations    =         27326  DegFreedom    Mean square
Regression   Sum of Squares         =       58.8591           1       58.85906
Residual     Sum of Squares         =       796.319       27324         .02914
Total        Sum of Squares         =       855.178       27325         .03130
----------   Standard error of e    =        .17071  Root MSE           .17071
Fit          R-squared              =        .06883  R-bar squared      .06879
Model test   F[  1, 27324]          =    2019.62500  Prob F > F*        .00000
Model was estimated on Jul 21, 2012 at 02:20:01 PM
--------+-----------------------------------------------------------------------
        |                 Standard              Prob.       95% Confidence
  HHNINC|  Coefficient      Error         z    |z|>Z*          Interval
--------+-----------------------------------------------------------------------
Constant|     .12609***     .00513     24.56   .0000       .11603     .13615
    EDUC|     .01996***     .00044     44.94   .0000       .01909     .02083
--------+-----------------------------------------------------------------------
```

The population value of β is +0.020

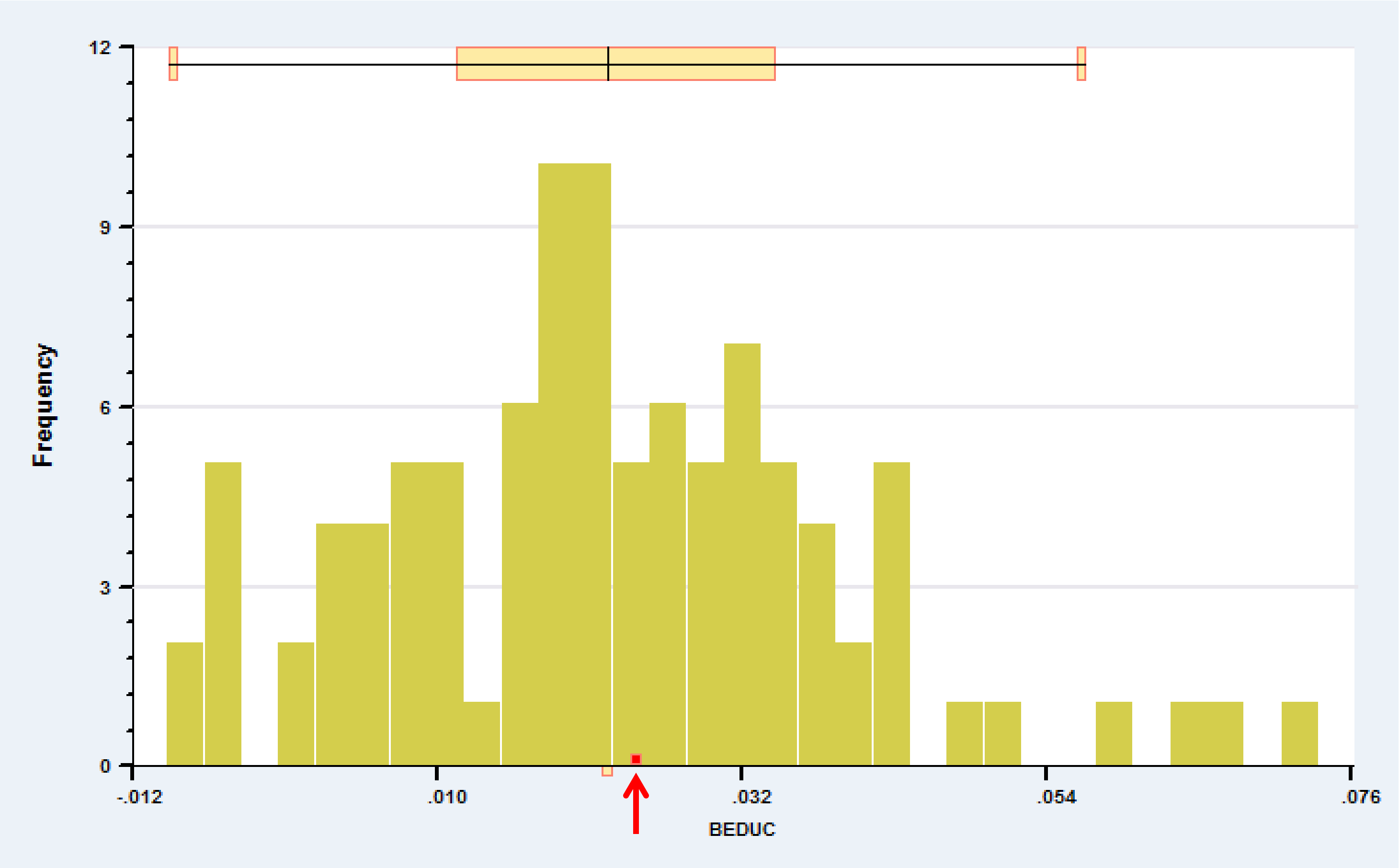
Sampling Distribution

A sampling experiment: Draw 25 observations at random from the population. Compute the regression. Repeat 100 times. Display estimated slopes in a histogram.

<u>Resampling y and x. Sampling variability over y, x, ε</u>

```
matrix ; beduc=init(100,1,0)$
proc$
draw ; n=25 $
regress; quietly ; lhs=hhninc ; rhs = one,educ $
matrix ; beduc(i)=b(2) $
sample;all$
endproc$
execute ; i=1,100 $
histogram;rhs=beduc; boxplot $
```

The least squares estimator is random. In repeated random samples, it varies randomly above and below β.



Sample mean = 0.022

How should we interpret this variation in the regression slope?

The Statistical Context of Least Squares Estimation

The sample of data from the population:
Data generating process is $y = \mathbf{x}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

The stochastic specification of the regression model: Assumptions about the random $\boldsymbol{\varepsilon}$.

Endowment of the stochastic properties of the model upon the least squares estimator. The estimator is a function of the observed (realized) data.

Least Squares as a Random Variable

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

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

$\mathbf{b}$ = The true parameter plus sampling error.

Also

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y} \qquad = (\mathbf{X'X})^{-1}\sum_{i=1}^{n}\mathbf{x}_i y_i$$

$$= \beta + (\mathbf{X'X})^{-1}\mathbf{X'}\varepsilon \quad = \beta + (\mathbf{X'X})^{-1}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i \quad = \beta + \sum_{i=1}^{n}(\mathbf{X'X})^{-1}\mathbf{x}_i\varepsilon_i$$

$$= \beta + \sum_{i=1}^{n}\mathbf{v}_i\varepsilon_i$$

$\mathbf{b}$ = The true parameter plus a linear function of the disturbances.

Deriving the **Properties** of **b**

b = a parameter vector + a linear combination of the disturbances, each times a vector.

Therefore, **b** is a vector of random variables.

We do the analysis conditional on an **X**, then show that results do not depend on the particular **X** in hand, so the result must be general – i.e., independent of **X**.

Properties of the LS Estimator: (1) b is unbiased

Expected value and the property of unbiasedness.

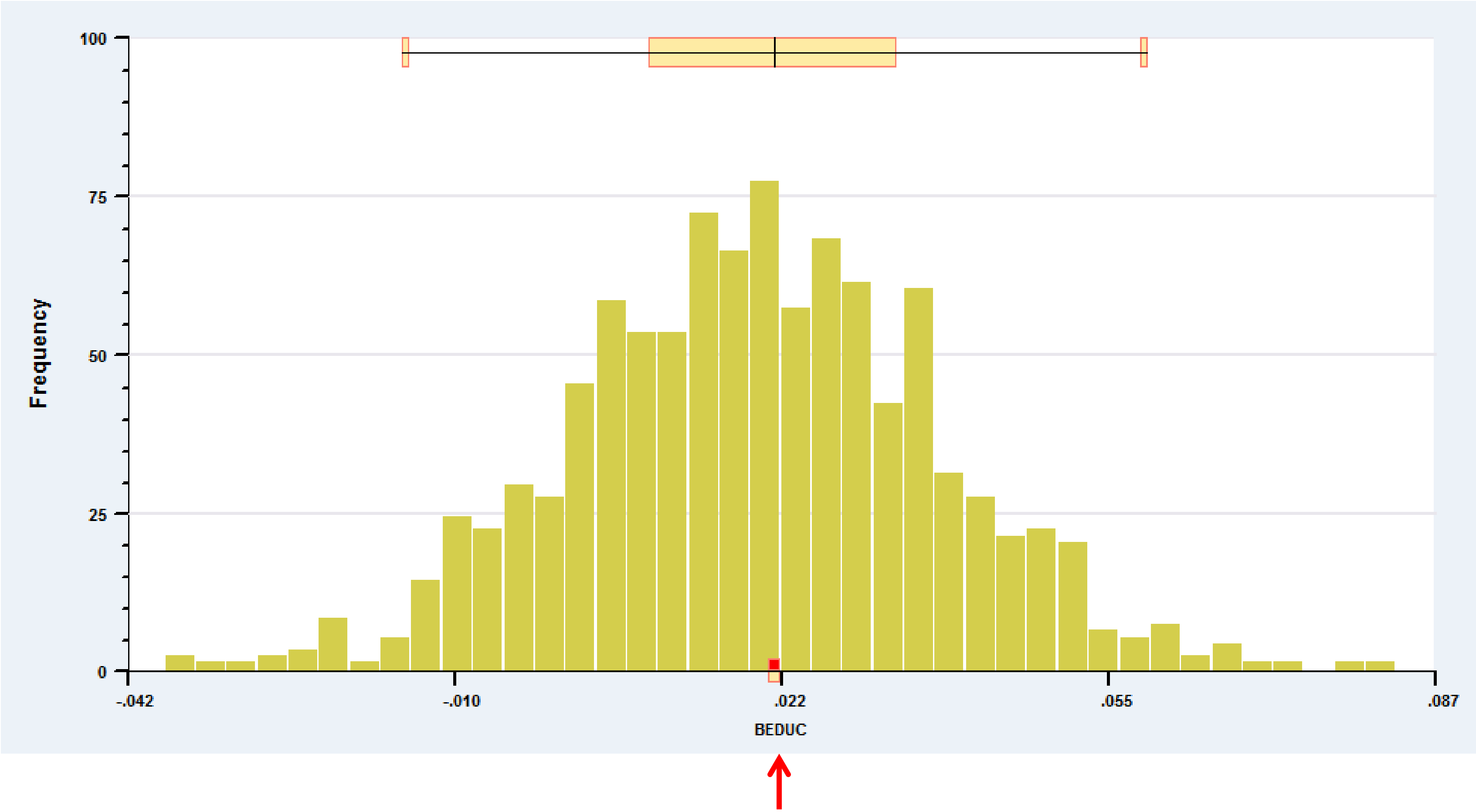
$$
\begin{aligned}
E[\mathbf{b}|\mathbf{X}] &= E[\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon|\mathbf{X}] \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\varepsilon|\mathbf{X}] \\
&= \beta + \mathbf{0} \\
&= \beta \\
E[\mathbf{b}] &= E_X\{E[\mathbf{b}|\mathbf{X}]\} \text{ (The law of iterated expectations.)} \\
&= E_X\{\beta\} \\
&= \beta.
\end{aligned}
$$

A Sampling Experiment: Unbiasedness X is fixed in repeated samples

Holding X fixed.   Resampling over ε

```
draw;n=25  $  Draw a particular sample of 25 observations
matrix   ; beduc = init(1000,1,0)$
proc$
? Reuse X, resample epsilon each time, 1000 samples.
   create ; inc = .12609+.01996*educ + rnn(0,.17071) $
   regress; quietly ; lhs=inc ; rhs = one,educ $
   matrix ; beduc(i)=b(2) $
endproc$
execute ; i=1,1000 $
histogram;rhs=beduc ;boxplot$
```

1000 Repetitions of b|x



Using the Expected Value of **b**
Partitioned Regression

A Crucial Result About Specification:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

Two sets of variables. What if the regression is computed without the second set of variables?

What is the expectation of the "short" regression estimator? $E[\mathbf{b}_1|(\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon})]$

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$$

The Left Out Variable Formula

"Short" regression means we regress **y** on $\mathbf{X}_1$ when

$$y = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \boldsymbol{\varepsilon} \text{ and } \beta_2 \text{ is not } \mathbf{0}$$

(This is a VVIR!)

$$\mathbf{b}_1 = (\mathbf{X_1}'\mathbf{X_1})^{-1}\mathbf{X_1}'\mathbf{y}$$

$$= (\mathbf{X_1}'\mathbf{X_1})^{-1}\mathbf{X_1}'(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \boldsymbol{\varepsilon})$$

$$= (\mathbf{X_1}'\mathbf{X_1})^{-1}\mathbf{X_1}'\mathbf{X}_1\beta_1 + (\mathbf{X_1}'\mathbf{X_1})^{-1}\mathbf{X_1}'\mathbf{X}_2\beta_2 + (\mathbf{X_1}'\mathbf{X_1})^{-1}\mathbf{X_1}'\boldsymbol{\varepsilon})$$

$$E[\mathbf{b}_1] = \beta_1 + (\mathbf{X_1}'\mathbf{X_1})^{-1}\mathbf{X_1}'\mathbf{X}_2\beta_2$$

Omitting relevant variables causes LS to be "biased."

This result educates our general understanding about regression.

Application

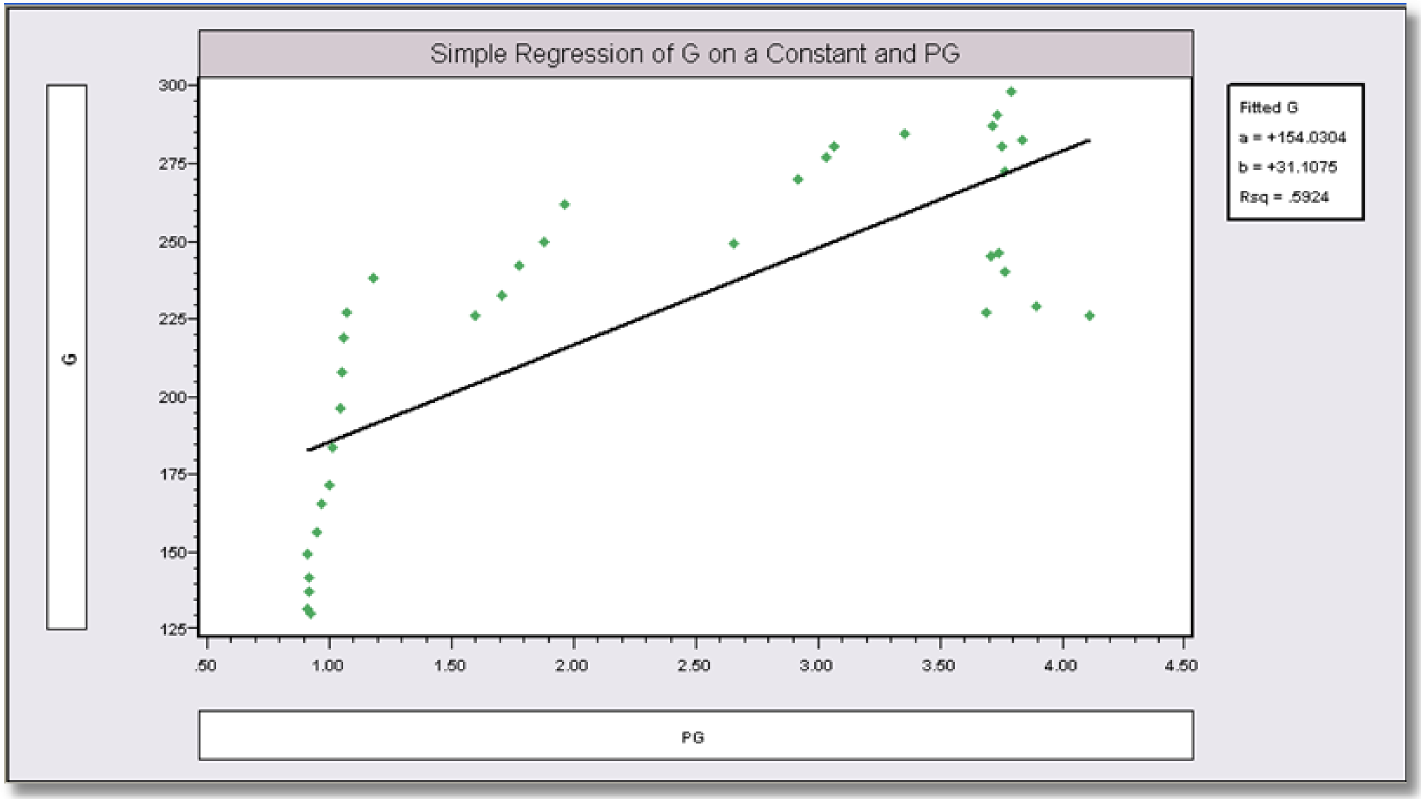
The (truly) short regression estimator is biased.

Application:

$$\text{Quantity} = \beta_1 \text{Price} + \beta_2 \text{Income} + \varepsilon$$

If you regress Quantity only on Price and leave out Income. What do you get?

Estimated 'Demand' Equation
Shouldn't the Price Coefficient be Negative?



Application: Left out Variable

Leave out Income. What do you get?

$$E[b_1] = \beta_1 + \left(\frac{\text{Cov[Price,Income]}}{\text{Var[Price]}}\right)\beta_2$$

In time series data, $\beta_1 < 0$, $\beta_2 > 0$ (usually)

Cov[*Price*,*Income*] > 0 in time series data.

So, the short regression will overestimate the price coefficient. It will be pulled toward and even past zero.

Simple Regression of G on a constant and PG

Price Coefficient should be negative.

Multiple Regression of G on Y and PG. The Theory Works!

```
----------------------------------------------------------------------
Ordinary     least squares regression ............
LHS=G        Mean                   =     226.09444
             Standard deviation     =      50.59182
             Number of observs.     =            36
Model size   Parameters             =             3
             Degrees of freedom     =            33
Residuals    Sum of squares         =    1472.79834
             Standard error of e    =       6.68059
Fit          R-squared              =        .98356
             Adjusted R-squared     =        .98256
Model test   F[  2,    33] (prob) =   987.1(.0000)
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  t-ratio  P[|T|>t]   Mean of X
--------+-------------------------------------------------------------
Constant|   -79.7535***      8.67255        -9.196   .0000
       Y|     .03692***       .00132        28.022   .0000     9232.86
      PG|   -15.1224***      1.88034        -8.042   .0000     2.31661
--------+-------------------------------------------------------------
```

The Extra Variable Formula

A Second Crucial Result About Specification:

$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$ but $\boldsymbol{\beta}_2$ really is $\mathbf{0}$.

Two sets of variables. One is superfluous. What if the regression is computed with it anyway?

The Extra Variable Formula: (This is a VIR!)

$E[\mathbf{b}_{1.2} | \boldsymbol{\beta}_2 = \mathbf{0}] = \boldsymbol{\beta}_1$

The long regression estimator in a short regression is unbiased.)

Extra variables in a model do not induce biases. Why not just include them? We will develop this result.

(2) The Sampling Variance of **b**

Assumption about disturbances:

- ε_i has zero mean and is uncorrelated with every other ε_j
- Var[ε_i|**X**] = σ^2. The variance of ε_i does not depend on any data in the sample.

$$\text{Var}\left[\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \dots\\ \varepsilon_n\end{pmatrix} \mid \mathbf{X}\right] = \begin{bmatrix}\sigma^2 & 0 & \dots & 0\\ 0 & \sigma^2 & \dots & 0\\ 0 & 0 & \ddots & 0\\ 0 & 0 & \dots & \sigma^2\end{bmatrix} = \sigma^2\mathbf{I}$$

Conditional Variance

$$\text{Var}\left[\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \ldots\\ \varepsilon_n\end{pmatrix} | \mathbf{X}\right] = \begin{bmatrix}\sigma^2 & 0 & \ldots & 0\\ 0 & \sigma^2 & \ldots & 0\\ 0 & 0 & \ddots & 0\\ 0 & 0 & \ldots & \sigma^2\end{bmatrix} = \sigma^2\mathbf{I}$$

Unconditional Variance

$$\text{Var}\left[\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \ldots\\ \varepsilon_n\end{pmatrix}\right] = \text{E}\left\{\text{Var}\left[\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \ldots\\ \varepsilon_n\end{pmatrix} | \mathbf{X}\right]\right\} + \text{Var}\left\{\text{E}\left[\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \ldots\\ \varepsilon_n\end{pmatrix} | \mathbf{X}\right]\right\}$$

$$= \text{E}\left\{\sigma^2\mathbf{I}\right\} + \text{Var}\left\{\begin{pmatrix}0\\ 0\\ \ldots\\ 0\end{pmatrix}\right\} = \sigma^2\mathbf{I}.$$

Conditional Variance of the Least Squares Estimator

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

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

$E[\mathbf{b}|\mathbf{X}]=\beta$ (We extablished this earlier.)

$$\text{Var}[\mathbf{b} \mid \mathbf{X}] = E[(\mathbf{b}-\beta)(\mathbf{b}-\beta)' \mid \mathbf{X}]$$

$$= E\left[\left\{(\mathbf{X'X})^{-1}\mathbf{X'}\varepsilon\right\}\left\{\varepsilon'\mathbf{X}(\mathbf{X'X})^{-1}\right\} \mid \mathbf{X}\right]$$

$$= (\mathbf{X'X})^{-1}\mathbf{X'}E[\varepsilon\varepsilon' \mid \mathbf{X}]\,\mathbf{X}(\mathbf{X'X})^{-1}$$

$$= (\mathbf{X'X})^{-1}\mathbf{X'}\sigma^2\mathbf{I}\,\mathbf{X}(\mathbf{X'X})^{-1}$$

$$= \sigma^2(\mathbf{X'X})^{-1}\mathbf{X'I}\,\mathbf{X}(\mathbf{X'X})^{-1}$$

$$= \sigma^2(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1}$$

$$= \sigma^2(\mathbf{X'X})^{-1}$$

Unconditional Variance of the Least Squares Estimator

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

$$E[\mathbf{b}|\mathbf{X}] = \beta$$

$$Var[\mathbf{b} \mid \mathbf{X}] = \sigma^2(\mathbf{X'X})^{-1}$$

$$\begin{aligned} Var[\mathbf{b}] &= E\{Var[\mathbf{b} \mid \mathbf{X}]\} + Var\{E[\mathbf{b}|\mathbf{X}]\} \\ &= \sigma^2 E[(\mathbf{X'X})^{-1}] + Var\{\beta\} \\ &= \sigma^2 E[(\mathbf{X'X})^{-1}] + \mathbf{0} \end{aligned}$$

We will ultimately need to estimate $E[(\mathbf{X'X})^{-1}]$.

We will use the only information we have, **X**, itself.

Variance Implications of Specification Errors: Omitted Variables

Suppose the correct model is

$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$. I.e., two sets of variables.

Compute least squares omitting $\mathbf{X}_2$. Some easily proved results:

Var[$\mathbf{b}_1$] is smaller than Var[$\mathbf{b}_{1.2}$]. Proof: Var[$\mathbf{b}_1$] = $\sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$.

Var[$\mathbf{b}_{1.2}$] = $\sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}$. To compare the matrices, we can ignore σ^2. To show that Var[$\mathbf{b}_1$] is smaller than Var[$\mathbf{b}_{1.2}$], we show that its inverse is bigger. So, is

$[(\mathbf{X}_1'\mathbf{X}_1)^{-1}]^{-1}$ larger than $[(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}]^{-1}$**?**

Is $\mathbf{X}_1'\mathbf{X}_1$ larger than $\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1$? Obviously.

Variance Implications of Specification Errors: Omitted Variables

I.e., you get a smaller variance when you omit X_2.

Omitting $\mathbf{X}_2$ amounts to using extra information ($\beta_2 = \mathbf{0}$). ***Even if the information is wrong (see the next result), it reduces the variance.*** (This is an important result.) It may induce a bias, but either way, it reduces variance.

$\mathbf{b}_1$ may be more "precise."

Precision = Mean squared error

= variance + squared bias.

Smaller variance but positive bias. If bias is small, may still favor the short regression.

Specification Errors-2

Including superfluous variables: Just reverse the results.

Including superfluous variables increases variance. (The cost of not using information.)

Does not cause a bias, because if the variables in $\mathbf{X}_2$ are truly superfluous, then $\boldsymbol{\beta}_2 = \mathbf{0}$,
so $E[\mathbf{b}_{1.2}] = \boldsymbol{\beta}_1 + \mathbf{C}\boldsymbol{\beta}_2 = \boldsymbol{\beta}_1$

Linear Restrictions

Context: How do linear restrictions affect the properties of the least squares estimator?

Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Theory (information) $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$

Restricted least squares estimator:

$$\mathbf{b}^* = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})$$

Expected value: $E[\mathbf{b}^*] = \boldsymbol{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\boldsymbol{\beta} - \mathbf{q})$

Variance: $\sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$

$= \text{Var}[\mathbf{b}]$ – a nonnegative definite matrix $< \text{Var}[\mathbf{b}]$

Implication: (As before) **nonsample information reduces the variance of the estimator**.

Interpretation

Case 1: Theory is correct: $\mathbf{R}\beta - \mathbf{q} = \mathbf{0}$
(the restrictions do hold).

b* is unbiased

Var[**b***] is smaller than Var[**b**]

Case 2: Theory is incorrect: $\mathbf{R}\beta - \mathbf{q} \neq \mathbf{0}$
(the restrictions do not hold).

b* is biased – what does this mean?

Var[**b***] is still smaller than Var[**b**]

Restrictions and Information

How do we interpret this important result?

- The theory is "information"
- Bad information leads us away from "the truth"
- Any information, good or bad, makes us more certain of our answer. In this context, any information reduces variance.

What about ignoring the information?

- Not using the correct information does not lead us away from "the truth"
- Not using the information foregoes the variance reduction - i.e., does not use the ability to reduce "uncertainty."

(3) Gauss-Markov Theorem

A theorem of Gauss and Markov: Least Squares is the **minimum variance linear unbiased estimator** (MVLUE)

1. Linear estimator $= \beta + \sum_{i=1}^{n} \mathbf{v}_i \varepsilon_i$
2. Unbiased: E[**b**|**X**] = **β**

Theorem: Var[**b***|**X**] – Var[**b**|**X**] is nonnegative definite for any other linear and unbiased estimator **b*** that is not equal to **b**.

Definition: **b** is **efficient** in this class of estimators.

Implications of Gauss-Markov

- Theorem: Var[**b***|**X**] – Var[**b**|**X**] is nonnegative definite for any other linear and unbiased estimator **b*** that is not equal to **b**. Implies:
- $\mathbf{b}_k$ = the kth particular element of b.
  Var[$\mathbf{b}_k$|**X**] = the kth diagonal element of Var[**b**|**X**]
  Var[$\mathbf{b}_k$|**X**] $\leq$ Var[$\mathbf{b}_k$*|**X**] for each coefficient.
- **c′b** = any linear combination of the elements of b. Var[**c′b**|**X**] $\leq$ Var[**c′b***|**X**] for any nonzero c and **b*** that is not equal to **b**.

Aspects of the Gauss-Markov Theorem

Indirect proof: Any other linear unbiased estimator has a larger covariance matrix.

Direct proof: Find the minimum variance linear unbiased estimator. It will be least squares.

Other estimators

Biased estimation – a minimum mean squared error estimator. Is there a biased estimator with a smaller ‘dispersion’? Yes, always

Normally distributed disturbances – the Rao-Blackwell result. (General observation – for normally distributed disturbances, ‘linear’ is superfluous.)

Nonnormal disturbances - Least Absolute Deviations and other nonparametric approaches may be better in small samples

(4) Distribution

Source of the random behavior of $\mathbf{b} = \beta + \sum_{i=1}^{n} \mathbf{v}_i \varepsilon_i$

$\mathbf{v}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$ where $\mathbf{x}_i$ is row i of $\mathbf{X}$.

We derived E[$\mathbf{b}\,|\,\mathbf{X}$] and Var[$\mathbf{b}\,|\,\mathbf{X}$] earlier. The distribution of $\mathbf{b}\,|\,\mathbf{X}$ is that of the linear combination of the disturbances, ε_i.

If ε_i has a normal distribution, denoted $\sim N[0,\sigma^2]$, then

$\mathbf{b}\,|\,\mathbf{X} = \beta + \mathbf{A}\varepsilon$ where $\varepsilon \sim N[0,\sigma^2\mathbf{I}]$ and $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

$\mathbf{b}\,|\,\mathbf{X} \sim N[\beta, \mathbf{A}\sigma^2\mathbf{I}\mathbf{A}'] = N[\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$.

Note how $\mathbf{b}$ inherits its stochastic properties from ε.

Summary: Finite Sample Properties of b

(1) Unbiased: $E[\mathbf{b}]=\beta$

(2) Variance: $Var[\mathbf{b}|\mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

(3) Efficiency: Gauss-Markov Theorem with all implications

(4) Distribution: Under normality,

$\mathbf{b}|\mathbf{X} \sim Normal[\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$

(Without normality, the distribution is generally unknown.)

Estimating the Variance of **b**

The true variance of **b|X** is $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ **.** We consider how to use the sample data to estimate this matrix. The ultimate objectives are to form interval estimates for regression slopes and to test hypotheses about them. Both require estimates of the variability of the distribution. We then examine a factor which affects how "large" this variance is, multicollinearity.

Estimating σ^2

Using the residuals instead of the disturbances:

The natural estimator: $\mathbf{e}'\mathbf{e}/n$ as a sample surrogate for $E[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/n]$

Imperfect observation of ε_i, $e_i = \varepsilon_i - (\boldsymbol{\beta} - \mathbf{b})'\mathbf{x}_i$

Downward bias of $\mathbf{e}'\mathbf{e}/n$.

We obtain the result $E[\mathbf{e}'\mathbf{e}|\mathbf{X}] = (n-K)\sigma^2$

Expectation of **e′e**

$$
\begin{aligned}
\mathbf{e} &= \mathbf{y} - \mathbf{Xb} \\
&= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y} \\
&= \mathbf{My} = \mathbf{M}(\mathbf{X}\beta + \varepsilon) = \mathbf{MX}\beta + \mathbf{M}\varepsilon = \mathbf{M}\varepsilon \\
\mathbf{e}'\mathbf{e} &= (\mathbf{M}\varepsilon)'(\mathbf{M}\varepsilon) \\
&= \varepsilon'\mathbf{M}'\mathbf{M}\varepsilon = \varepsilon'\mathbf{MM}\varepsilon = \varepsilon'\mathbf{M}\varepsilon
\end{aligned}
$$

Method 1:

$$E[\mathbf{e'e} \mid \mathbf{X}] = E[\varepsilon'\mathbf{M}\varepsilon \mid \mathbf{X}]$$

$$= E[\text{ trace } (\varepsilon'\mathbf{M}\varepsilon \mid \mathbf{X}) \text{] scalar = its trace}$$

$$= E[\text{ trace } (\mathbf{M}\varepsilon\varepsilon' \mid \mathbf{X}) \text{] permute in trace}$$

$$= \text{ [trace E } (\mathbf{M}\varepsilon\varepsilon' \mid \mathbf{X}) \text{] linear operators}$$

$$= \text{ [trace } \mathbf{M} \text{ E } (\varepsilon\varepsilon' \mid \mathbf{X}) \text{] conditioned on X}$$

$$= \text{ [trace } \mathbf{M}\, \sigma^2 \mathbf{I}_n \text{] model assumption}$$

$$= \sigma^2 [\text{trace } \mathbf{M}] \text{ scalar multiplication and } \mathbf{I} \text{ matrix}$$

$$= \sigma^2 \text{trace } [\mathbf{I}_n - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}]$$

$$= \sigma^2 \{\text{trace } [\mathbf{I}_n] - \text{trace}[\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}]\}$$

$$= \sigma^2 \{n - \text{trace}[(\mathbf{X'X})^{-1}\mathbf{X'X}]\} \text{ permute in trace}$$

$$= \sigma^2 \{n - \text{trace}[\mathbf{I}_K]\}$$

$$= \sigma^2 \{n - K\}$$

Notice that $E[\mathbf{e}'\mathbf{e}|\mathbf{X}]$ is not a function of **X**.

Estimating σ^2

The **unbiased estimator** is $s^2 = \mathbf{e}'\mathbf{e}/(n\text{-}K)$.

(n-K) is a “degrees of freedom correction”

Therefore, the *unbiased* estimator of σ^2 is

$$s^2 = \mathbf{e}'\mathbf{e}/(n\text{-}K)$$

Method 2: Some Matrix Algebra

$E[\mathbf{e'e} \mid \mathbf{X}] = \sigma^2 \text{ trace } \mathbf{M}$

What is the trace of **M**? Trace of square matrix = sum of diagonal elements.

(Result A - 108) **M** is idempotent, so its trace equals its rank.

(Theorem A.4) Its rank equals the number of nonzero characeristic roots.

Characteric Roots : Signature of a Matrix = Spectral Decomposition

= Eigen (own) value Decomposition

(Definition A.16) $\mathbf{A} = \mathbf{C}\Lambda\mathbf{C}'$ where

$\mathbf{C}$ = a matrix of columns such that $\mathbf{CC'} = \mathbf{C'C} = \mathbf{I}$

Λ = a diagonal matrix of the characteristic roots

(Elements of Λ may be zero.)

Decomposing **M**

Useful Result: If $\mathbf{A} = \mathbf{C}\Lambda\mathbf{C}'$ is the spectral decomposition, then $\mathbf{A}^2 = \mathbf{C}\Lambda^2\mathbf{C}'$ (just multiply)

$\mathbf{M} = \mathbf{M}^2$, so $\Lambda^2 = \Lambda$. All of the characteristic roots of **M** are 1 or 0. How many of each?

trace($\mathbf{A}$) = trace($\mathbf{C}\Lambda\mathbf{C}'$)=trace($\Lambda\mathbf{C}'\mathbf{C}$)=trace($\Lambda$)

Trace of a matrix equals the sum of its characteristic roots. Since the roots of **M** are all 1 or 0, its trace is just the number of ones, which is n-K as we saw.

Example: Characteristic Roots of a Correlation Matrix

Matrix - R

[6, 6] Cell: 1

	1	2	3	4	5	6
1	1	0.795578	0.908202	0.924205	0.903905	0.886908
2	0.795578	1	0.928756	0.812462	0.802779	0.791689
3	0.908202	0.928756	1	0.963605	0.954187	0.956742
4	0.924205	0.812462	0.963605	1	0.990628	0.989062
5	0.903905	0.802779	0.954187	0.990628	1	0.987139
6	0.886908	0.791689	0.956742	0.989062	0.987139	1

```
--> matrix;list;root(r)$

Matrix Result    has  6 rows and  1 columns.
               1
        +--------------
       1|     5.53961
       2|      .29845
       3|      .13847
       4|      .01478
       5|      .00608
       6|      .00260
```

Note sum = trace = 6.

Matrix - R

[6, 6] Cell: 1

	1	2	3	4	5	6
1	1	0.795578	0.908202	0.924205	0.903905	0.886908
2	0.795578	1	0.928756	0.812462	0.802779	0.791689
3	0.908202	0.928756	1	0.963605	0.954187	0.956742
4	0.924205	0.812462	0.963605	1	0.990628	0.989062
5	0.903905	0.802779	0.954187	0.990628	1	0.987139
6	0.886908	0.791689	0.956742	0.989062	0.987139	1

$$\mathbf{R} = \mathbf{C\Lambda C'} = \sum_{i=1}^{6} \lambda_i \mathbf{c}_i \mathbf{c}_i'$$

	1	2	3	4	5	6
1	0.399548	-0.121844	-0.895708	-0.0406948	-0.127852	0.0722466
2	0.377099	0.840502	0.067997	0.177137	0.0355656	0.337768
3	0.420955	0.198986	0.132743	-0.413014	-0.104492	-0.764252
4	0.419339	-0.258255	0.101987	0.0247916	0.862514	0.050123
5	0.416351	-0.28231	0.222987	0.750782	-0.325211	-0.166715
6	0.414441	-0.3045	0.339614	-0.481765	-0.348967	0.516048

1	5.53961
2	.29845
3	.13847
4	.01478
5	.00608
6	.00260

Gasoline Data (first 20 of 52 observations)

```
namelist ;| x = one,log(gasp),log(pcincome),log(pnc),log(puc),log(ppt)$
Listing of current sample ---------------------------------------------------
```

Line	Observation	logGASP	logPCINC	logPNC	logPUC	logPPT
1	1	2.81349	9.08273	3.85439	3.28466	2.82138
2	2	2.83492	9.07761	3.83945	3.12236	2.89037
3	3	2.84549	9.12446	3.80221	3.06805	2.91777
4	4	2.87520	9.15377	3.83081	3.03013	2.95491
5	5	2.91761	9.15989	3.88156	3.14415	2.99072
6	6	2.90777	9.15197	3.91202	3.17805	3.03975
7	7	2.92187	9.17833	3.95508	3.28840	3.06805
8	8	2.95032	9.18348	3.94158	3.21888	3.10009
9	9	2.94043	9.20039	3.94158	3.25810	3.14415
10	10	2.94670	9.23279	3.93769	3.34639	3.17805
11	11	2.94428	9.25484	3.93183	3.35690	3.19048
12	12	2.93773	9.31118	3.92986	3.40120	3.20680
13	13	2.97487	9.35824	3.90600	3.39451	3.22684
14	14	2.99763	9.39806	3.88773	3.36730	3.26194
15	15	3.03013	9.43004	3.89792	3.39786	3.31054
16	16	3.04476	9.46436	3.92593	3.42426	3.35690
17	17	3.07713	9.48517	3.94158	3.43076	3.43076
18	18	3.08603	9.51510	3.97029	3.44042	3.56105
19	19	3.09331	9.54688	4.01096	3.49651	3.63231
20	20	3.10620	9.58273	4.00186	3.49953	3.67122

X'X and its Roots

Matrix - XX

[6, 6] Cell:

	1	2	3	4	5	6
1	52	193.924	503.093	227.779	213.483	215.381
2	193.924	746.713	1887.6	864.079	820.842	832.782
3	503.093	1887.6	4873.57	2211.09	2078.01	2099.16
4	227.779	864.079	2211.09	1007.49	951.46	962.91
5	213.483	820.842	2078.01	951.46	904.166	917.147
6	215.381	832.782	2099.16	962.91	917.147	931.886

```
--> matrix;list;root(xx)$

  Result|              1
--------+--------------
       1|        8474.00
       2|        40.1984
       3|        1.10133
       4|        .403257
       5|        .116637
       6|      .00102318
```

Var[**b**|**X**]

Estimating the Covariance Matrix for b|X

The true covariance matrix is $\sigma^2 (\mathbf{X'X})^{-1}$

The natural estimator is $s^2(\mathbf{X'X})^{-1}$

"Standard errors" of the individual coefficients are the square roots of the diagonal elements.

[7, 7] Cell:

X'X

	1	2	3	4	5	6	7
1	36	83.398	332383	630	60.148	84.371	98.815
2	83.398	248.04	838669	1878.67	164.992	251.287	301.047
3	332383	838669	3.18054e+009	6.4692e+006	591999	859749	1.01845e+006
4	630	1878.67	6.4692e+006	14910	1277.71	1972.56	2384.18
5	60.148	164.992	591999	1277.71	114.542	171.935	205.811
6	84.371	251.287	859749	1972.56	171.935	267.306	322.011
7	98.815	301.047	1.01845e+006	2384.18	205.811	322.011	391.845

$(X'X)^{-1}$

	1	2	3	4	5	6	7
1	92.9516	-1.58239	-0.0142015	3.45656	-6.3863	2.85512	-5.3368
2	-1.58239	0.218408	0.000315846	-0.0830075	-0.665387	-0.02755	0.287509
3	-0.0142015	0.000315846	2.25808e-006	-0.000547423	0.000144609	-0.000330383	0.000995983
4	3.45656	-0.0830075	-0.000547423	0.136591	-0.061965	0.0821448	-0.251126
5	-6.3863	-0.665387	0.000144609	-0.061965	8.62577	-1.43238	-1.23058
6	2.85512	-0.02755	-0.000330383	0.0821448	-1.43238	0.940991	-0.360893
7	-5.3368	0.287509	0.000995983	-0.251126	-1.23058	-0.360893	1.00971

$s^2(X'X)^{-1}$

	1	2	3	4	5	6	7
1	2495.92	-42.49	-0.381335	92.8149	-171.484	76.6652	-143.303
2	-42.49	5.86466	0.00848103	-2.2289	-17.8668	-0.739767	7.72013
3	-0.381335	0.00848103	6.06335e-005	-0.0146993	0.003883	-0.00887138	0.026744
4	92.8149	-2.2289	-0.0146993	3.6677	-1.66387	2.20574	-6.74318
5	-171.484	-17.8668	0.003883	-1.66387	231.618	-38.4621	-33.0434
6	76.6652	-0.739767	-0.00887138	2.20574	-38.4621	25.2673	-9.69062
7	-143.303	7.72013	0.026744	-6.74318	-33.0434	-9.69062	27.1126

Standard Regression Results

```
----------------------------------------------------------------------
Ordinary      least squares regression ........
LHS=G         Mean                  =   226.09444
              Standard deviation    =    50.59182
              Number of observs.    =          36
Model size    Parameters            =           7
              Degrees of freedom    =          29
Residuals     Sum of squares        =   778.70227
              Standard error of e   =     5.18187 <= sqr[778.70227/(36 - 7)]
Fit           R-squared             =      .99131
              Adjusted R-squared    =      .98951
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  t-ratio  P[|T|>t]   Mean of X
--------+-------------------------------------------------------------
Constant|   -7.73975        49.95915       -.155   .8780
      PG|  -15.3008***       2.42171      -6.318   .0000      2.31661
       Y|     .02365***       .00779       3.037   .0050      9232.86
   TREND|    4.14359**       1.91513       2.164   .0389      17.5000
     PNC|   15.4387         15.21899       1.014   .3188      1.67078
     PUC|   -5.63438         5.02666      -1.121   .2715      2.34364
     PPT|  -12.4378**        5.20697      -2.389   .0236      2.74486
--------+-------------------------------------------------------------
```

Multicollinearity

Multicollinearity: Short Rank of X

(Not a Monet)

Enhanced Monet Area Effect Model: Height and Width Effects

$$\text{Log(Price)} = \alpha + \beta_1 \log \text{Area} + \beta_2 \log \text{Aspect Ratio} + \beta_3 \log \text{Height} + \beta_4 \text{Signature} + \varepsilon$$

$$= \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

(Aspect Ratio = Width/Height). This is a perfectly respectable theory of art prices. However, it is not possible to learn about the parameters from data on prices, areas, aspect ratios, heights and signatures.

$$x_3 = (1/2)(x_1 - x_2)$$

Multicollinearity: Correlation of Regressors

Not "short rank," which is a deficiency in the model.
Full rank, but columns of X are highly correlated.
A characteristic of the data set which affects the covariance matrix.

Regardless, $\boldsymbol{\beta}$ is unbiased.
Consider one of the unbiased coefficient estimators of β_k. $E[b_k] = \beta_k$

$Var[\mathbf{b}] = \sigma^2(\mathbf{X'X})^{-1}$. The variance of b_k is the kth diagonal element of $\sigma^2(\mathbf{X'X})^{-1}$.

We can isolate this with the result Theorem 3.4, page 39

Let $[\mathbf{X},\mathbf{z}]$ be [Other **x***s*, $\mathbf{x}_k$] = $[\mathbf{X}_1,\mathbf{x}_2]$

The general result is that the diagonal element we seek is $[\mathbf{z}'\mathbf{M}_\mathbf{X}\mathbf{z}]^{-1}$,
the reciprocal of the sum of squared residuals in the regression of **z** on **X**.

Variances of Least Squares Coefficients

$$\text{Model}: \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\gamma + \boldsymbol{\varepsilon}$$

$$\text{Variance of} \begin{pmatrix} \mathbf{b} \\ \text{c} \end{pmatrix} = \sigma^2 \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{z} \\ \mathbf{z}'\mathbf{X} & \mathbf{z}'\mathbf{z} \end{bmatrix}^{-1}$$

Variance of c is the lower right element of this matrix.

$$\text{Var[c]} = \sigma^2[\mathbf{z}'\mathbf{M}_X\mathbf{z}]^{-1} = \frac{\sigma^2}{\mathbf{z}'{*}\mathbf{z}{*}}$$

where **z*** = the vector of residuals from the regression of **z** on **X**.

The R^2 in that regression is $R^2_{\mathbf{z}|X} = 1 - \dfrac{\mathbf{z}'{*}\mathbf{z}{*}}{\sum_{i=1}^{n}(z_i - \overline{z})^2}$, so

$$\mathbf{z}'{*}\mathbf{z}{*} = \left(1 - R^2_{\mathbf{z}|X}\right)\sum_{i=1}^{n}(z_i - \overline{z})^2. \text{ Therefore,}$$

$$\text{Var[c]} = \sigma^2[\mathbf{z}'\mathbf{M}_X\mathbf{z}]^{-1} = \frac{\sigma^2}{\left(1 - R^2_{\mathbf{z}|X}\right)\sum_{i=1}^{n}(z_i - \overline{z})^2}$$

Multicollinearity

$$\text{Var}[c] = \sigma^2[\mathbf{z}'\mathbf{M}_X\mathbf{z}]^{-1} = \frac{\sigma^2}{\left(1-R^2_{\mathbf{z}|\mathbf{X}}\right)\sum_{i=1}^{n}(z_i-\overline{z})^2}$$

All else constant, the variance of the coefficient on **z** rises as the fit in the regression of **z** on the other variables goes up. If the fit is perfect, the variance becomes infinite.

"Detecting" multicollinearity?

Variance inflation factor: $\text{VIF}(z) = \dfrac{1}{\left(1-R^2_{\mathbf{z}|\mathbf{X}}\right)}$.

Regression Analysis: Expenditure versus Year, GasPrice, Income, P_NewCars, ...

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	9	168558	18728.7	5355.77	0.000
Year	1	42	41.7	11.91	0.001
GasPrice	1	1348	1347.7	385.39	0.000
Income	1	91	90.6	25.91	0.000
P_NewCars	1	30	30.0	8.57	0.006
P_UsedCars	1	47	47.5	13.57	0.001
P_PublicTrans	1	0	0.1	0.03	0.865
P_Durables	1	188	187.6	53.65	0.000
P_Nondurables	1	1	1.3	0.37	0.544
P_Services	1	6	5.6	1.60	0.212
Error	42	147	3.5		
Total	51	168705			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.87000	99.91%	99.89%	99.83%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1596	467	3.42	0.001	
Year	-0.840	0.243	-3.45	0.001	198.49
GasPrice	1.3404	0.0683	19.63	0.000	64.62
Income	0.004522	0.000888	5.09	0.000	354.84
P_NewCars	0.645	0.220	2.93	0.006	974.93
P_UsedCars	0.3079	0.0836	3.68	0.001	265.78
P_PublicTrans	0.0142	0.0830	0.17	0.865	481.06
P_Durables	-1.494	0.204	-7.32	0.000	820.66
P_Nondurables	0.132	0.216	0.61	0.544	1614.88
P_Services	0.174	0.137	1.27	0.212	1229.94

The Longley Data

Y,X1,X2,X3,X4,X5,X6						
60323	83.0	234289	2356	1590	107608	1947
61122	88.5	259426	2325	1456	108632	1948
60171	88.2	258054	3682	1616	109773	1949
61187	89.5	284599	3351	1650	110929	1950
63221	96.2	328975	2099	3099	112075	1951
63639	98.1	346999	1932	3594	113270	1952
64989	99.0	365385	1870	3547	115094	1953
63761	100.0	363112	3578	3350	116219	1954
66019	101.2	397469	2904	3048	117388	1955
67857	104.6	419180	2822	2857	118734	1956
68169	108.4	442769	2936	2798	120445	1957
66513	110.8	444546	4681	2637	121950	1958
68655	112.6	482704	3813	2552	123366	1959
69564	114.2	502601	3931	2514	125368	1960
69331	115.7	518173	4806	2572	127852	1961
70551	116.9	554894	4007	2827	130081	1962

TABLE 4.9 Longley Results: Dependent Variable Is Employment

	1947–1961	*Variance Inflation*	*1947–1962*
Constant	1,459,415		1,169,087
Year	−721.756	143.4638	−576.464
GNP Deflator	−181.123	75.6716	−19.7681
GNP	0.0910678	132.467	0.0643940
Armed Forces	−0.0749370	1.55319	−0.0101453

Condition Number and Variance Inflation Factors

```
Characteristic Roots of X'X
  Result|               1
--------+---------------
       1|         8471.26
       2|         40.1922
       3|         1.10146
       4|         .401673
       5|         .116978
       6|      .00104601
Condition Number = sqr(8471.26/.00104601)
                 = 2845.8111

VIFI     =      52.6069923
VIFPG    =      17.6982507
VIFPNC   =     171.7227200
VIFPUC   =     115.3714230
VIFPPT   =     225.7317614
```

Condition number larger than 30 is ‘large.’

What does this mean?

Variance Inflation in Gasoline Market

Regression Analysis:
logG versus logIncome, logPG

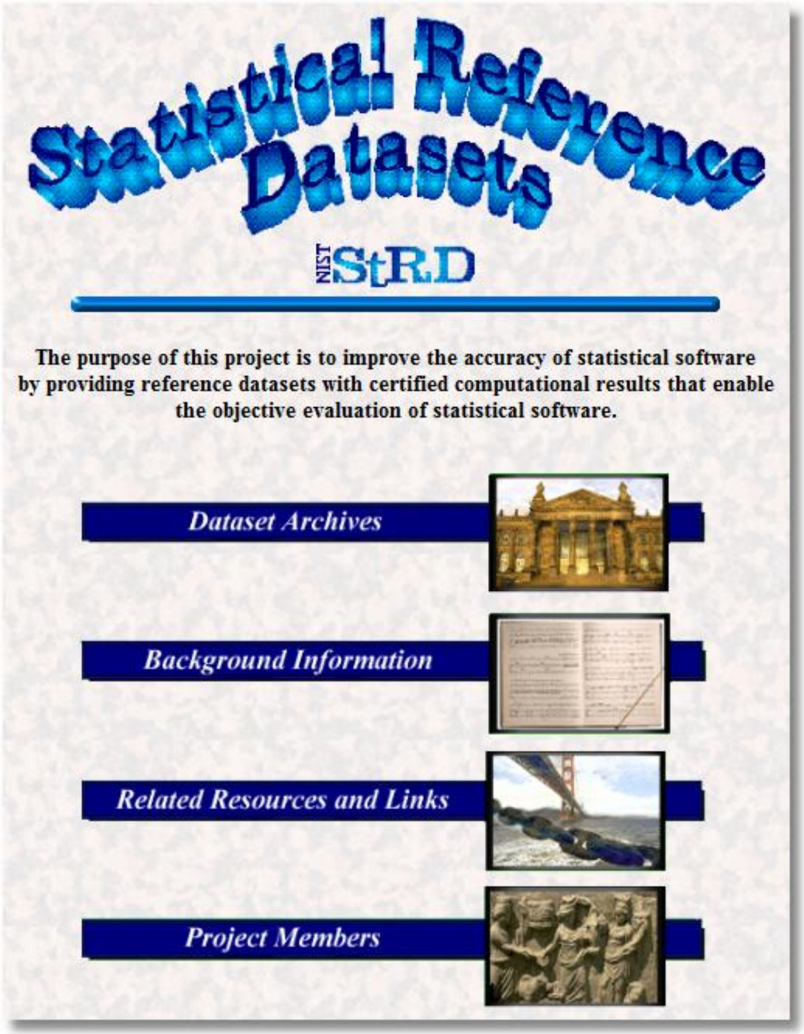
```
The regression equation is
logG = - 0.468 + 0.966 logIncome - 0.169 logPG
Predictor       Coef   SE Coef      T      P
Constant    -0.46772  0.08649  -5.41  0.000
logIncome    0.96595  0.07529  12.83  0.000
logPG       -0.16949  0.03865  -4.38  0.000
S = 0.0614287   R-Sq = 93.6%   R-Sq(adj) = 93.4%
Analysis of Variance
Source          DF      SS      MS       F      P
Regression       2  2.7237  1.3618  360.90  0.000
Residual Error  49  0.1849  0.0038
Total           51  2.9086
```

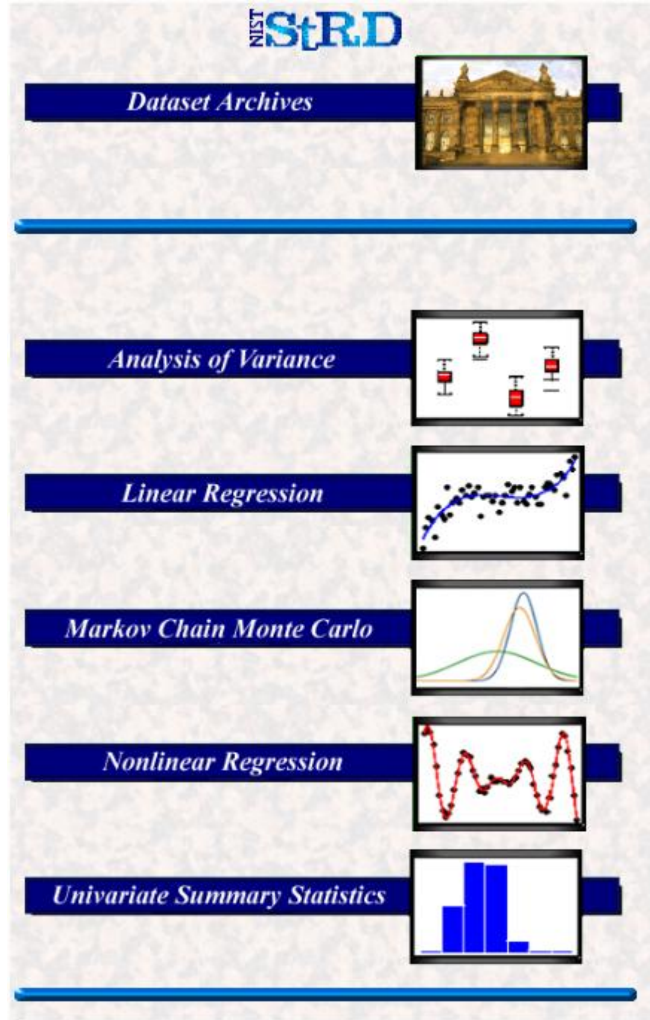
Gasoline Market

Regression Analysis: logG versus logIncome, logPG, ...

```
The regression equation is
logG = - 0.558 + 1.29 logIncome - 0.0280 logPG
               - 0.156 logPNC + 0.029 logPUC - 0.183 logPPT
Predictor        Coef   SE Coef      T      P
Constant      -0.5579   0.5808  -0.96  0.342
logIncome      1.2861   0.1457   8.83  0.000
logPG        -0.02797  0.04338  -0.64  0.522
logPNC        -0.1558   0.2100  -0.74  0.462
logPUC         0.0285   0.1020   0.28  0.781
logPPT        -0.1828   0.1191  -1.54  0.132
S = 0.0499953   R-Sq = 96.0%   R-Sq(adj) = 95.6%
Analysis of Variance
Source          DF       SS        MS       F      P
Regression       5  2.79360  0.55872  223.53  0.000
Residual Error  46  0.11498  0.00250
Total           51  2.90858
```

The standard error on logIncome doubles when the three variables are added to the equation while the coefficient only changes slightly.


Statistical Reference Datasets
NIST StRD
The purpose of this project is to improve the accuracy of statistical software by providing reference datasets with certified computational results that enable the objective evaluation of statistical software.
Dataset Archives
Background Information
Related Resources and Links
Project Members


NIST StRD
Dataset Archives
Analysis of Variance
Linear Regression
Markov Chain Monte Carlo
Nonlinear Regression
Univariate Summary Statistics

NIST Longley Solution

```
               Observed Data
Model:         Polynomial Class
               7 Parameters (B0,B1,...,B7)
               y = B0 + B1*x1 + B2*x2 + B3*x3 + B4*x4 + B5*x5 + B6*x6 + e
               Certified Regression Statistics
                                          Standard Deviation
     Parameter        Estimate               of Estimate
        B0       -3482258.63459582        890420.383607373
        B1        15.0618722713733        84.9149257747669
        B2       -0.358191792925910E-01   0.334910077722432E-01
        B3       -2.02022980381683        0.488399681651699
        B4       -1.03322686717359        0.214274163161675
        B5       -0.511041056535807E-01   0.226073200069370
        B6        1829.15146461355        455.478499142212
```

Y	Coefficient	Standard Error	t	Prob. |t|>T*	95% Confidence Interval	
Constant	-.34823D+07***	890420.4	-3.91	.0036	-.54965D+07	-.14680D+07
X1	15.0619	84.91493	.18	.8631	-177.0290	207.1528
X2	-.03582	.03349	-1.07	.3127	-.11158	.03994
X3	-2.02023***	.48840	-4.14	.0025	-3.12507	-.91539
X4	-1.03323***	.21427	-4.82	.0009	-1.51795	-.54851
X5	-.05110	.22607	-.23	.8262	-.56252	.46031
X6	1829.15***	455.4785	4.02	.0030	798.79	2859.52

Excel Longley Solution

SUMMARY OUTPUT

Regression Statistics	
Multiple F	0.997737
R Square	0.995479
Adjusted I	0.992465
Standard E	304.8541
Observati	16

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	6	1.84E+08	30695400	330.2853	4.98E-10
Residual	9	836424.1	92936.01		
Total	15	1.85E+08			

	Coefficient	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-3482259	890420.4	-3.9108	0.00356	-5496529	-1467988	-5496529	-1467988
X Variable	15.06187	84.91493	0.177376	0.863141	-177.029	207.1528	-177.029	207.1528
X Variable	-0.03582	0.033491	-1.06952	0.312681	-0.11158	0.039943	-0.11158	0.039943
X Variable	-2.02023	0.4884	-4.13643	0.002535	-3.12507	-0.91539	-3.12507	-0.91539
X Variable	-1.03323	0.214274	-4.82199	0.000944	-1.51795	-0.54851	-1.51795	-0.54851
X Variable	-0.0511	0.226073	-0.22605	0.826212	-0.56252	0.460309	-0.56252	0.460309
X Variable	1829.151	455.4785	4.01589	0.003037	798.7875	2859.515	798.7875	2859.515

Estimate
-3482258.63459582
15.0618722713733
-0.358191792925910E-01
-2.02022980381683
-1.03322686717359
-0.511041056535807E-01
1829.15146461355

The NIST Filipelli Problem

```
READ;NOBS=82;NVAR=2;NAMES=Y,X$
0.8116 -6.860120914
0.9072 -4.324130045
0.9052 -4.358625055
0.9039 -4.358426747
0.8053 -6.955852379
0.8377 -6.661145254
0.8667 -6.355462942
0.8809 -6.118102026
0.7975 -7.115148017
0.8162 -6.815308569
. . .
remaining 72 observations
CREATE; X1=X ; X2=X*X ; X3=X2*X ; X4=X3*X ; X5=X4*X ;X6=X5*X
                ; X7=X6*X ; X8=X7*X ; X9=X8*X ; X10=X9*X$
REGRESS;LHS=Y;RHS=ONE,X1,X2,X3,X4,X5,X6,X7,X8,X9,X10$
```

Certified Filipelli Results

Certified Regression Statistics

Parameter	Estimate	Standard Deviation of Estimate
B0	-1467.48961422980	298.084530995537
B1	-2772.17959193342	559.779865474950
B2	-2316.37108160893	466.477572127796
B3	-1127.97394098372	227.204274477751
B4	-354.478233703349	71.6478660875927
B5	-75.1242017393757	15.2897178747400
B6	-10.8753180355343	2.23691159816033
B7	-1.06221498588947	0.221624321934227
B8	-0.670191154593408E-01	0.142363763154724E-01
B9	-0.246781078275479E-02	0.535617408889821E-03
B10	-0.402962525080404E-04	0.896632837373868E-05

Residual Standard Deviation 0.334801051324544E-02

R-Squared 0.996727416185620

Certified Analysis of Variance Table

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Squares
Regression	10	0.242391619837339	0.242391619837339E-01
Residual	71	0.795851382172941E-03	0.112091743968020E-04

Minitab Filipelli Results

Regression Analysis: y versus x1, x2, x3, x4, x5, x6, x7, x8, x9, x10

```
* WARNING * x3 is highly correlated with other predictors.
* WARNING * x4 is highly correlated with other predictors.
* WARNING * x5 is highly correlated with other predictors.
* WARNING * x6 is highly correlated with other predictors.
* WARNING * x7 is highly correlated with other predictors.
* WARNING * x8 is highly correlated with other predictors.
* WARNING * x9 is highly correlated with other predictors.

The regression equation is
y = - 1467 - 2772 x1 - 2316 x2 - 1128 x3 - 354 x4 - 75.1 x5 - 10.9 x6 - 1.06 x7
    - 0.0670 x8 - 0.00247 x9 - 0.000040 x10
```

Predictor	Coef	SE Coef	T	P
Constant	-1467.5	298.1	-4.92	0.000
x1	-2772.1	559.8	-4.95	0.000
x2	-2316.3	466.5	-4.97	0.000
x3	-1128.0	227.2	-4.96	0.000
x4	-354.47	71.65	-4.95	0.000
x5	-75.12	15.29	-4.91	0.000
x6	-10.875	2.237	-4.86	0.000
x7	-1.0622	0.2216	-4.79	0.000
x8	-0.06702	0.01424	-4.71	0.000
x9	-0.0024678	0.0005356	-4.61	0.000
x10	-0.00004030	0.00000897	-4.49	0.000

S = 0.00334800   R-Sq = 99.7%   R-Sq(adj) = 99.6%

Estimate
-1467.48961422980
-2772.17959193342
-2316.37108160893
-1127.97394098372
-354.478233703349
-75.1242017393757
-10.8753180355343
-1.06221498588947
-0.670191154593408E-01
-0.246781078275479E-02
-0.402962525080404E-04

Stata Filipelli Results

```
      Source |       SS       df       MS              Number of obs =      82
-------------+------------------------------           F(  8,    73) = 2059.23
       Model |  .242114595     8  .030264324           Prob > F      =  0.0000
    Residual |  .001072876    73  .000014697           R-squared     =  0.9956
-------------+------------------------------           Adj R-squared =  0.9951
       Total |  .243187471    81  .003002314           Root MSE      =  .00383

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|
-------------+----------------------------------------------------------------
          x1 |   9.585386   1.609771     5.95   0.000
          x2 |  (dropped)
          x3 |  -1.419962   .2137125    -6.64   0.000
          x4 |  (dropped)
          x5 |    .305533   .0417248     7.32   0.000
          x6 |   .1216212   .0159331     7.63   0.000
          x7 |   .0228691   .0028893     7.92   0.000
          x8 |   .0023607   .0002892     8.16   0.000
          x9 |   .0001291   .0000154     8.37   0.000
         x10 |   2.94e-06   3.44e-07     8.55   0.000
       _cons |   13.83021    2.29365     6.03   0.000
------------------------------------------------------------------------------
```

Estimate
-1467.48961422980
-2772.17959193342
-2316.37108160893
-1127.97394098372
-354.478233703349
-75.1242017393757
-10.8753180355343
-1.06221498588947
-0.670191154593408E-01
-0.246781078275479E-02
-0.402962525080404E-04

In the Filippelli test, Stata found two coefficients so collinear that it dropped them from the analysis. Most other statistical software packages have done the same thing, and most authors have interpreted this result as acceptable for this test.

Y	Coefficient	Standard Error	t	Prob. |t|>T*	95% Confidence Interval	
Constant	13.5586***	2.28650	5.93	.0000	9.0016	18.1156
X1	9.39316***	1.60468	5.85	.0000	6.19504	12.59129
X3	-1.39413***	.21302	-6.54	.0000	-1.81868	-.96957
X5	.30045***	.04159	7.22	.0000	.21757	.38334
X6	.11968***	.01588	7.54	.0000	.08803	.15133
X7	.02252***	.00288	7.82	.0000	.01678	.02826
X8	.00233***	.00029	8.07	.0000	.00175	.00290
X9	.00013***	.1537D-04	8.28	.0000	.00010	.00016
X10	.28946D-05***	.3425D-06	8.45	.0000	.22121D-05	.35771D-05

x1	9.585386	1.609771
x2	(dropped)	
x3	-1.419962	.2137125
x4	(dropped)	
x5	.305533	.0417248
x6	.1216212	.0159331
x7	.0228691	.0028893
x8	.0023607	.0002892
x9	.0001291	.0000154
x10	2.94e-06	3.44e-07
_cons	13.83021	2.29365

Even after dropping two (random columns), results are only correct to 1 or 2 digits.

Regression of x_2 on all other variables

```
----------------------------------------------------------------------------
Ordinary     least squares regression ............
LHS=X2       Mean                 =        40.05875
             Standard deviation   =        18.37174
----------   No. of observations  =              82  DegFreedom   Mean square
Regression   Sum of Squares       =         27339.2           9    3037.68778
Residual     Sum of Squares       =     .515124E-10          72        .00000
Total        Sum of Squares       =         27339.2          81     337.52086
----------   Standard error of e  =          .00000  Root MSE          .00000
Fit          R-squared            =         1.00000  R-bar squared    1.00000
Model test   F[  9,    72]        =****************  Prob F > F*      .00000
Model was estimated on Jul 21, 2012 at 09:02:49 PM
--------+-------------------------------------------------------------------
        |                Standard            Prob.      95% Confidence
      X2|  Coefficient     Error       t    |t|>T*         Interval
--------+-------------------------------------------------------------------
Constant|     -.63802***    .00419  -152.40  .0000     -.64623    -.62982
      X1|    -1.19955***    .00394  -304.78  .0000    -1.20726   -1.19184
      X3|     -.48688***    .00159  -305.76  .0000     -.49000    -.48376
      X4|     -.15336***    .00100  -153.37  .0000     -.15532    -.15140
      X5|     -.03267***    .00032  -102.67  .0000     -.03329    -.03204
      X6|     -.00477***  .6159D-04  -77.40  .0000     -.00489    -.00465
      X7|     -.00047***  .7558D-05  -62.28  .0000     -.00049    -.00046
      X8|-.30124D-04***  .5766D-06   -52.25  .0000 -.31254D-04  -.28994D-04
      X9|-.11284D-05***  .2501D-07   -45.11  .0000 -.11775D-05  -.10794D-05
     X10|        0.0***  .4725D-09   -39.78  .0000 -.19725D-07  -.17872D-07
--------+-------------------------------------------------------------------
Note: nnnnn.D-xx or D+xx => multiply by 10 to -xx or +xx.
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------------

|-> calc ; peek ; 1 -.515124e-10/27339.2$
[CALC]        =    .99999999999999810D+00
```

Using QR Decomposition

```
Ordinary     least squares regression ............
LHS=Y        Mean                    =      .84958
             Standard deviation      =      .05479
----------   No. of observations     =          82   DegFreedom   Mean square
Regression   Sum of Squares          =     .242392           10        .02424
Residual     Sum of Squares          =  .795851E-03          71        .00001
Total        Sum of Squares          =     .243187           81        .00300
----------   Standard error of e     =      .00335   Root MSE              .00312
Fit          R-squared               =      .99673   R-bar squared         .99627
Model test   F[ 10,    71]           =  2162.43959   Prob F > F*           .00000
```

Y	Coefficient	Standard Error	t	Prob. |t|>T*
Constant	-1467.49***	298.0845	-4.92	.0000
X1	-2772.18***	559.7799	-4.95	.0000
X2	-2316.37***	466.4776	-4.97	.0000
X3	-1127.97***	227.2043	-4.96	.0000
X4	-354.478***	71.64787	-4.95	.0000
X5	-75.1242***	15.28972	-4.91	.0000
X6	-10.8753***	2.23691	-4.86	.0000
X7	-1.06222***	.22162	-4.79	.0000
X8	-.06702***	.01424	-4.71	.0000
X9	-.00247***	.00054	-4.61	.0000
X10	-.40296D-04***	.8966D-05	-4.49	.0000

Estimate
-1467.48961422980
-2772.17959193342
-2316.37108160893
-1127.97394098372
-354.478233703349
-75.1242017393757
-10.8753180355343
-1.06221498588947
-0.670191154593408E-01
-0.246781078275479E-02
-0.402962525080404E-04

Multicollinearity

There is no "cure" for collinearity. Estimating something else is not helpful (principal components, for example).

There are "measures" of multicollinearity, such as the condition number of **X** and the variance inflation factor.

Best approach: Be cognizant of it. Understand its implications for estimation.

What is better: Include a variable that causes collinearity, or drop the variable and suffer from a biased estimator?

- Mean squared error would be the basis for comparison.
- Some generalities. Assuming **X** has full rank, regardless of the condition,
  - **b** is still unbiased
  - Gauss-Markov still holds

How (not) to deal with multicollinearity in a Translog Production Function

$$\log y = \alpha + \beta_1 \log x_1 + \beta_2 \log x_2 + \beta_3 \log x_3 +$$

$$\gamma_{11} \log^2 x_1 + \gamma_{12} \tfrac{1}{2} \log x_1 \log x_2 + \gamma_{13} \tfrac{1}{2} \log x_1 \log x_3 +$$
$$\gamma_{22} \log^2 x_2 + \gamma_{23} \tfrac{1}{2} \log x_2 \log x_3 +$$
$$\gamma_{33} \log^2 x_3$$

1. Checking for variance inflation factor (VIF) and ensuring that it is less than 10 therefore, if VIF > 10, eliminate the variables in a step-wise way?

2. Maintain either the squares or the cross products depending on which fits data best. However, this might not be useful since most of the time the full model is a better fit.

3. Standardize the variables by the mean and estimating again. If there are still VIF>10, eliminate step-wise by VIF?

How do I deal with the issue of multicollinearity in my dataset?
I know that translog is a better fit than Cobb-Douglas in my data but am faced with the multicollinearity challenge. What would be a way forward in such cases?

I have a sample of 24025 observations in a logit model. Two predictors are highly collinear (pairwaise corr .96; p<.001); vif are about 12 for each of them; average vif is 2.63; condition number is 10.26; determinant of correlation matrix is 0.0211; the two lowest eigen vales are 0.0792 and 0.0427. Centering/standardizing variables does not change the story.

Note: most obs are zeros for these two variables; I only have approx 600 non-zero obs for these two variables on a total of 24.025 obs.

Both variable coefficients are significant and must be included in the model (as per specification).

-- Do I have a problem of multicollinearity??
-- Does the large sample size attenuate this concern, even if I have a correlation of .96?
-- What could I look at to ascertain that the consequences of multi-collinearity are not a problem?
-- Is there any reference I might cite, to say that given the sample size, it is not a problem?

I hope you might help, because I am really in trouble!!!