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; otherwise = 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

The population value of $\beta$ is $+0.020$
A sampling experiment: Draw 25 observations at random from the population. Compute the regression. Repeat 100 times. Display estimated slopes in a histogram.

Resampling y and x. Sampling variability over y, x, ε

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 $\beta$.

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 \( y = \mathbf{x}'\beta + \varepsilon \)

The stochastic specification of the regression model: Assumptions about the random \( \varepsilon \).

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

\[ b = (X'X)^{-1}X'y \]

\[ = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon \]

\( b \) = The true parameter plus sampling error.

Also

\[ b = (X'X)^{-1}X'y \quad = (X'X)^{-1}\sum_{i=1}^{n}x_iy_i \]

\[ = \beta + (X'X)^{-1}X'\varepsilon \quad = \beta + (X'X)^{-1}\sum_{i=1}^{n}x_i\varepsilon_i \quad = \beta + \sum_{i=1}^{n}(X'X)^{-1}x_i\varepsilon_i \]

\[ = \beta + \sum_{i=1}^{n}v_i\varepsilon_i \]

\( b \) = The true parameter plus a linear function of the disturbances.
Deriving the **Properties** of $\mathbf{b}$

$\mathbf{b} = \text{a parameter vector} + \text{a linear combination of the disturbances, each times a vector.}$

Therefore, $\mathbf{b}$ is a vector of random variables.

We do the analysis conditional on an $\mathbf{X}$, then show that results do not depend on the particular $\mathbf{X}$ in hand, so the result must be general – i.e., independent of $\mathbf{X}$. 
Properties of the LS Estimator:
(1) $b$ is unbiased

Expected value and the property of unbiasedness.

\[
E[b|X] = E[\beta + (X'X)^{-1}X'\varepsilon|X] \\
= \beta + (X'X)^{-1}X'E[\varepsilon|X] \\
= \beta + 0 \\
= \beta 
\]

\[
E[b] = E_x\{E[b|X]\} \text{ (The law of iterated expectations.)} \\
= E_x\{\beta\} \\
= \beta. 
\]
A Sampling Experiment: Unbiasedness

X is fixed in repeated samples

**Holding X fixed. Resampling over ε**

draw;n=25 $ \text{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 + \text{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:

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

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? $E[b_1|(y = X_1\beta_1 + X_2\beta_2 + \epsilon)]$

$$b_1 = (X_1'X_1)^{-1}X_1'y$$
The Left Out Variable Formula

“Short” regression means we regress $y$ on $X_1$ when
$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$ and $\beta_2$ is not 0
(This is a VVIR!)

$$b_1 = (X_1'X_1)^{-1}X_1'y$$
$$= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon)$$
$$= (X_1'X_1)^{-1}X_1'X_1\beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2$$
$$\quad + (X_1'X_1)^{-1}X_1'\varepsilon$$

$$E[b_1] = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2$$

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?

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

In time series data, \( \beta_1 < 0, \beta_2 > 0 \) (usually)
\( \text{Cov}[\text{Price}, \text{Income}] > 0 \) 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:

\[ y = X_1\beta_1 + X_2\beta_2 + \varepsilon \]  
but \( \beta_2 \) really is 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!)

\[ E[b_{1.2} \mid \beta_2 = 0] = \beta_1 \]

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:

- $\varepsilon_i$ has zero mean and is uncorrelated with every other $\varepsilon_j$
- $\text{Var}[\varepsilon_i|X] = \sigma^2$. The variance of $\varepsilon_i$ does not depend on any data in the sample.

\[
\text{Var} \begin{pmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\vdots \\
\varepsilon_n
\end{pmatrix} | X = \begin{bmatrix}
\sigma^2 & 0 & \cdots & 0 \\
0 & \sigma^2 & \cdots & 0 \\
0 & 0 & \ddots & 0 \\
0 & 0 & \cdots & \sigma^2
\end{bmatrix} = \sigma^2 I
\]
Conditional Variance

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

Unconditional Variance

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

$$= \text{E} \left \{ \sigma^2 I \right \} + \text{Var} \left \{ \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \right \} = \sigma^2 I.$$
Conditional Variance of the Least Squares Estimator

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

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

\[ \mathbb{E}[\mathbf{b}|\mathbf{X}] = \beta \quad (\text{We established this earlier.}) \]

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

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

\[ = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbb{E}[\varepsilon\varepsilon' | \mathbf{X}] \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \]

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

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

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

\[ = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \]
Unconditional Variance of the Least Squares Estimator

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

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

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

\[ \text{Var}[\mathbf{b}] = \mathbb{E}\{\text{Var}[\mathbf{b} | \mathbf{X}]\} + \text{Var}\{\mathbb{E}[\mathbf{b} | \mathbf{X}]\} \]

\[ = \sigma^2 \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}] + \text{Var}\{\beta\} \]

\[ = \sigma^2 \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}] + \mathbf{0} \]

We will ultimately need to estimate \( \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}] \).
We will use the only information we have, \( \mathbf{X} \), itself.
Variance Implications of Specification Errors: Omitted Variables

Suppose the correct model is
\[ y = X_1\beta_1 + X_2\beta_2 + \varepsilon. \] I.e., two sets of variables.

Compute least squares omitting \( X_2 \). Some easily proved results:

\( \text{Var}[b_1] \) is smaller than \( \text{Var}[b_{1.2}] \). Proof: \( \text{Var}[b_1] = \sigma^2(X_1'X_1)^{-1} \).

\( \text{Var}[b_{1.2}] = \sigma^2(X_1'M_2X_1)^{-1} \). To compare the matrices, we can ignore \( \sigma^2 \). To show that \( \text{Var}[b_1] \) is smaller than \( \text{Var}[b_{1.2}] \), we show that its inverse is bigger. So, is
\[
[(X_1'X_1)^{-1}]^{-1} \text{ larger than } [(X_1'M_2X_1)^{-1}]^{-1}?
\]
Is \( X_1'X_1 \) larger than \( X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1 \)? Obviously.
Variance Implications of Specification Errors: Omitted Variables

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

Omitting $X_2$ amounts to using extra information ($\beta_2 = 0$). **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.

$b_1$ 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.
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 \( X_2 \) are truly superfluous, then \( \beta_2 = 0 \), so \( \mathbb{E}[\beta_{1.2}] = \beta_1 + C\beta_2 = \beta_1 \)
Linear Restrictions

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

**Model:** \( y = X\beta + \varepsilon \)

**Theory (information)** \( R\beta - q = 0 \)

Restricted least squares estimator:

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

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

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

\[
= \text{Var}[b] - \text{a nonnegative definite matrix} < \text{Var}[b]
\]

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

Case 1: Theory is correct: $R\beta - q = 0$
(= the restrictions do hold).
$b^*$ is unbiased
$\text{Var}[b^*]$ is smaller than $\text{Var}[b]$

Case 2: Theory is incorrect: $R\beta - q \neq 0$
(= the restrictions do not hold).
$b^*$ is biased – what does this mean?
$\text{Var}[b^*]$ is still smaller than $\text{Var}[b]$
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 \( \hat{\mathbf{\beta}} + \mathbf{X}' \mathbf{v} \)
2. Unbiased: \( \mathbb{E}[\hat{\mathbf{\beta}} | \mathbf{X}] = \mathbf{\beta} \)

**Theorem:** \( \text{Var}[\hat{\mathbf{\beta}}|\mathbf{X}] - \text{Var}[\hat{\mathbf{\beta}}|\mathbf{X}] \) is nonnegative definite for any other linear and unbiased estimator \( \hat{\mathbf{\beta}}^* \) that is not equal to \( \hat{\mathbf{\beta}} \).

**Definition:** \( \hat{\mathbf{\beta}} \) is **efficient** in this class of estimators.
Implications of Gauss-Markov

- Theorem: \( \text{Var}[b^*|X] - \text{Var}[b|X] \) is nonnegative definite for any other linear and unbiased estimator \( b^* \) that is not equal to \( b \). Implies:
  - \( b_k = \) the kth particular element of \( b \).
  - \( \text{Var}[b_k|X] = \) the kth diagonal element of \( \text{Var}[b|X] \).
  - \( \text{Var}[b_k|X] \leq \text{Var}[b_k^*|X] \) for each coefficient.
- \( c'b = \) any linear combination of the elements of \( b \).
  - \( \text{Var}[c'b|X] \leq \text{Var}[c'b^*|X] \) 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 \quad \text{where} \quad \mathbf{x}_i \quad \text{is row i of \( \mathbf{X} \).} \]

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

If \( \varepsilon_i \) has a normal distribution, denoted \( \sim N[0, \sigma^2] \), then

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

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

Note how \( \mathbf{b} \) inherits its stochastic properties from \( \varepsilon \).
Summary: Finite Sample Properties of $b$

1. Unbiased: $E[b] = \beta$
2. Variance: $\text{Var}[b|X] = \sigma^2(X'X)^{-1}$
3. Efficiency: Gauss-Markov Theorem with all implications
4. Distribution: Under normality, $b|X \sim \text{Normal}[\beta, \sigma^2(X'X)^{-1}]$
   (Without normality, the distribution is generally unknown.)
Estimating the Variance of $b$

The true variance of $b|X$ is $\sigma^2(X'X)^{-1}$. 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 $\sigma^2$

Using the residuals instead of the disturbances:
The natural estimator: $e'e/n$ as a sample surrogate for $E[e'e/n]$
Imperfect observation of $\varepsilon_i$, $e_i = \varepsilon_i - (\beta - b)'x_i$
Downward bias of $e'e/n$.
We obtain the result $E[e'e|X] = (n-K)\sigma^2$
Expectation of $e'e$

$e = y - Xb$

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

$= [I - X(X'X)^{-1}X']y$

$= My = M(X\beta + \varepsilon) = MX\beta + M\varepsilon = M\varepsilon$

$e'e = (M\varepsilon)'(M\varepsilon)$

$= \varepsilon'M'M\varepsilon = \varepsilon'MM\varepsilon = \varepsilon'M\varepsilon$
Method 1:

\[ E[e'e | X] = E[\varepsilon'M\varepsilon | X] \]

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

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

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

\[ = [ \text{trace} ME(\varepsilon\varepsilon' | X) ] \text{conditioned on } X \]

\[ = [ \text{trace} M \sigma^2 I_n ] \text{model assumption} \]

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

\[ = \sigma^2 \text{trace} [I_n - X(X'X)^{-1}X' ] \]

\[ = \sigma^2 \{\text{trace} [I_n] - \text{trace}[X(X'X)^{-1}X'] \} \]

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

\[ = \sigma^2 \{n - \text{trace}[I_k] \} \]

\[ = \sigma^2 \{n - K \} \]

Notice that \(E[e' e | X]\) is not a function of \(X\).
Estimating $\sigma^2$

The unbiased estimator is $s^2 = e'e/(n-K)$.

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

Therefore, the unbiased estimator of $\sigma^2$ is

$$s^2 = e'e/(n-K)$$
Method 2: Some Matrix Algebra

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

What is the trace of \( \mathbf{M} \)? Trace of square matrix = sum of diagonal elements.

(Result A - 108) \( \mathbf{M} \) is idempotent, so its trace equals its rank.

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

Characteristic Roots: Signature of a Matrix = Spectral Decomposition

\[ \mathbf{A} = \mathbf{C} \Lambda \mathbf{C}' \]

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

\[ \mathbf{C} = \text{a matrix of columns such that } \mathbf{C} \mathbf{C}' = \mathbf{C}' \mathbf{C} = \mathbf{I} \]

\( \Lambda = \text{a diagonal matrix of the characteristic roots } \)

(Elements of \( \Lambda \) may be zero.)
Decomposing $\mathbf{M}$

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 $\mathbf{M}$ are 1 or 0. How many of each? $\text{trace}(\mathbf{A}) = \text{trace}(\mathbf{C}\Lambda\mathbf{C}') = \text{trace}(\Lambda\mathbf{C}'\mathbf{C}) = \text{trace}(\Lambda)$

Trace of a matrix equals the sum of its characteristic roots. Since the roots of $\mathbf{M}$ 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

Note sum = trace = 6.
\[ R = CA^T = \sum_{i=1}^{6} \lambda_i c_i c_i' \]
Gasoline Data (first 20 of 52 observations)

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

![Matrix - XX](image)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>52</td>
<td>193.924</td>
<td>503.093</td>
<td>227.779</td>
<td>213.483</td>
<td>215.381</td>
</tr>
<tr>
<td>2</td>
<td>193.924</td>
<td>746.713</td>
<td>1887.6</td>
<td>864.079</td>
<td>820.842</td>
<td>832.782</td>
</tr>
<tr>
<td>3</td>
<td>503.093</td>
<td>1887.6</td>
<td>4873.57</td>
<td>2211.09</td>
<td>2078.01</td>
<td>2099.16</td>
</tr>
<tr>
<td>4</td>
<td>227.779</td>
<td>864.079</td>
<td>2211.09</td>
<td>1007.49</td>
<td>951.46</td>
<td>962.91</td>
</tr>
<tr>
<td>5</td>
<td>213.483</td>
<td>820.842</td>
<td>2078.01</td>
<td>951.46</td>
<td>904.166</td>
<td>917.147</td>
</tr>
<tr>
<td>6</td>
<td>215.381</td>
<td>832.782</td>
<td>2099.16</td>
<td>962.91</td>
<td>917.147</td>
<td>931.886</td>
</tr>
</tbody>
</table>

---

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

<table>
<thead>
<tr>
<th>Result</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8474.00</td>
</tr>
<tr>
<td>3</td>
<td>40.1984</td>
</tr>
<tr>
<td>4</td>
<td>1.10133</td>
</tr>
<tr>
<td>5</td>
<td>4.03257</td>
</tr>
<tr>
<td>6</td>
<td>.116637</td>
</tr>
<tr>
<td>6</td>
<td>.00102318</td>
</tr>
</tbody>
</table>
Estimating the Covariance Matrix for $\mathbf{b}|\mathbf{X}$

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

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

“Standard errors” of the individual coefficients are the square roots of the diagonal elements.
### Part 7: Finite Sample Properties of LS

#### Key Formulas

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

#### Table

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

#### Notes

- The table above illustrates the computation of matrix inversion and multiplication, which are fundamental in understanding finite sample properties of linear regression (LS) models.
### Standard Regression Results

**Ordinary least squares regression**

<table>
<thead>
<tr>
<th>LHS=G</th>
<th>Mean = 226.09444</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation = 50.59182</td>
<td></td>
</tr>
<tr>
<td>Number of observs. = 36</td>
<td></td>
</tr>
</tbody>
</table>

**Model size**

| Parameters = 7 |
| Degrees of freedom = 29 |

**Residuals**

| Sum of squares = 778.70227 |
| Standard error of e = 5.18187 <= \( \sqrt{\frac{778.70227}{36 - 7}} \) |

**Fit**

| R-squared = 0.99131 |
| Adjusted R-squared = 0.98951 |

| Variable | Coefficient | Standard Error | t-ratio | P[|T|>t] | Mean of X |
|----------|-------------|----------------|---------|---------|-----------|
| Constant| -7.73975    | 49.95915       | -0.155  | 0.8780  |           |
| PG      | -15.3008*** | 2.42171        | -6.318  | 0.0000  | 2.31661   |
| Y       | 0.02365***  | 0.00779        | 3.037   | 0.0050  | 9232.86   |
| TREND   | 4.14359**   | 1.91513        | 2.164   | 0.0389  | 17.5000   |
| PNC     | 15.4387     | 15.21899       | 1.014   | 0.3188  | 1.67078   |
| PUC     | -5.63438    | 5.02666        | -1.121  | 0.2715  | 2.34364   |
| PPT     | -12.4378**  | 5.20697        | -2.389  | 0.0236  | 2.74486   |
Multicollinearity
Multicollinearity: Short Rank of X

Enhanced Monet Area Effect Model: Height and Width Effects

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

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

(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, \( \beta \) is unbiased.
Consider one of the unbiased coefficient estimators of \( \beta_k \). \( E[b_k] = \beta_k \)

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

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

Let \([X,z]\) be \([\text{Other } x_s, x_k] = [X_1, x_2]\)

The general result is that the diagonal element we seek is \( [z'M_xz]^{-1} \), the reciprocal of the sum of squared residuals in the regression of \( z \) on \( X \).
Variance of Least Squares Coefficients

Model: \( y = X\beta + z\gamma + \epsilon \)

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

Variance of \( c \) is the lower right element of this matrix.

\( \text{Var}[c] = \sigma^2 [z'M_xz]^{-1} = \frac{\sigma^2}{z^*z^*} \)

where \( z^* = \) the vector of residuals from the regression of \( z \) on \( X \).

The \( R^2 \) in that regression is \( R^2_{z|x} = 1 - \frac{z'^*z'^*}{\sum_{i=1}^{n} (z_i - \bar{z})^2} \), so

\( z'^*z'^* = (1 - R^2_{z|x}) \sum_{i=1}^{n} (z_i - \bar{z})^2 \). Therefore,

\( \text{Var}[c] = \sigma^2 [z'M_xz]^{-1} = \frac{\sigma^2}{(1 - R^2_{z|x}) \sum_{i=1}^{n} (z_i - \bar{z})^2} \)
Multicollinearity

\[
\text{Var}[c] = \sigma^2 [z'M_x z]^{-1} = \frac{\sigma^2}{(1 - R^2_{z|x}) \sum_{i=1}^{n} (z_i - \bar{z})^2}
\]

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: \( \text{VIF}(z) = \frac{1}{(1 - R^2_{z|x})} \).
Regression Analysis: Expenditure versus Year, GasPrice, Income, P_NewCars, ...

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Adj SS</th>
<th>Adj MS</th>
<th>F-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>9</td>
<td>168558</td>
<td>18728.7</td>
<td>5355.77</td>
<td>0.000</td>
</tr>
<tr>
<td>Year</td>
<td>1</td>
<td>42</td>
<td>41.7</td>
<td>11.91</td>
<td>0.001</td>
</tr>
<tr>
<td>GasPrice</td>
<td>1</td>
<td>1348</td>
<td>1347.7</td>
<td>386.39</td>
<td>0.000</td>
</tr>
<tr>
<td>Income</td>
<td>1</td>
<td>91</td>
<td>90.6</td>
<td>25.91</td>
<td>0.000</td>
</tr>
<tr>
<td>P_NewCars</td>
<td>1</td>
<td>30</td>
<td>30.0</td>
<td>8.57</td>
<td>0.006</td>
</tr>
<tr>
<td>P_UsedCars</td>
<td>1</td>
<td>47</td>
<td>47.5</td>
<td>13.57</td>
<td>0.001</td>
</tr>
<tr>
<td>P_PublicTrans</td>
<td>1</td>
<td>0</td>
<td>0.1</td>
<td>0.03</td>
<td>0.865</td>
</tr>
<tr>
<td>P_Durables</td>
<td>1</td>
<td>188</td>
<td>187.6</td>
<td>53.65</td>
<td>0.000</td>
</tr>
<tr>
<td>P_Nondurables</td>
<td>1</td>
<td>1</td>
<td>1.3</td>
<td>0.37</td>
<td>0.544</td>
</tr>
<tr>
<td>P_Services</td>
<td>1</td>
<td>6</td>
<td>5.6</td>
<td>1.60</td>
<td>0.212</td>
</tr>
<tr>
<td>Error</td>
<td>42</td>
<td>147</td>
<td>3.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>168705</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Model Summary

\[
\begin{array}{cccc}
S & R-sq & R-sq(adj) & R-sq(pred) \\
1.87000 & 99.91\% & 99.89\% & 99.83\% \\
\end{array}
\]

Coefficients

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1596</td>
<td>467</td>
<td>3.42</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>-0.840</td>
<td>0.243</td>
<td>-3.45</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>GasPrice</td>
<td>1.3404</td>
<td>0.0683</td>
<td>19.63</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>0.004522</td>
<td>0.000888</td>
<td>5.09</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>P_NewCars</td>
<td>0.645</td>
<td>0.220</td>
<td>2.93</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>P_UsedCars</td>
<td>0.3079</td>
<td>0.0836</td>
<td>3.68</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>P_PublicTrans</td>
<td>0.0142</td>
<td>0.0830</td>
<td>0.17</td>
<td>0.865</td>
<td></td>
</tr>
<tr>
<td>P_Durables</td>
<td>-1.494</td>
<td>0.204</td>
<td>-7.32</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>P_Nondurables</td>
<td>0.132</td>
<td>0.216</td>
<td>0.61</td>
<td>0.544</td>
<td></td>
</tr>
<tr>
<td>P_Services</td>
<td>0.174</td>
<td>0.137</td>
<td>1.27</td>
<td>0.212</td>
<td></td>
</tr>
</tbody>
</table>
The Longley Data

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

TABLE 4.9 Longley Results: Dependent Variable Is Employment

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1,459,415</td>
<td></td>
<td>1,169,087</td>
</tr>
<tr>
<td>Year</td>
<td>−721.756</td>
<td>143.4638</td>
<td>−576.464</td>
</tr>
<tr>
<td>GNP Deflator</td>
<td>−181.123</td>
<td>75.6716</td>
<td>−19.7681</td>
</tr>
<tr>
<td>GNP</td>
<td>0.0910678</td>
<td>132.467</td>
<td>0.0643940</td>
</tr>
<tr>
<td>Armed Forces</td>
<td>−0.0749370</td>
<td>1.55319</td>
<td>−0.0101453</td>
</tr>
</tbody>
</table>
Condition Number and Variance Inflation Factors

Condition number larger than 30 is ‘large.’

What does this mean?

Characteristic Roots of $X'X$

<table>
<thead>
<tr>
<th>Result</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8471.26</td>
</tr>
<tr>
<td>2</td>
<td>40.1922</td>
</tr>
<tr>
<td>3</td>
<td>1.10146</td>
</tr>
<tr>
<td>4</td>
<td>0.401673</td>
</tr>
<tr>
<td>5</td>
<td>0.116978</td>
</tr>
<tr>
<td>6</td>
<td>0.00104601</td>
</tr>
</tbody>
</table>

Condition Number = $\sqrt{\frac{8471.26}{0.00104601}}$

= 2845.8111

VIFI  = 52.6069923
VIFPG = 17.6982507
VIFPNC = 171.7227200
VIFPUC = 115.3714230
VIFPPT = 225.7317614
Variance Inflation in Gasoline Market

Regression Analysis:
logG versus logIncome, logPG

The regression equation is
logG = - 0.468 + 0.966 logIncome - 0.169 logPG

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.46772</td>
<td>0.08649</td>
<td>-5.41</td>
<td>0.000</td>
</tr>
<tr>
<td>logIncome</td>
<td>0.96595</td>
<td>0.07529</td>
<td>12.83</td>
<td>0.000</td>
</tr>
<tr>
<td>logPG</td>
<td>-0.16949</td>
<td>0.03865</td>
<td>-4.38</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 0.0614287  R-Sq = 93.6%  R-Sq(adj) = 93.4%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>2.7237</td>
<td>1.3618</td>
<td>360.90</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>49</td>
<td>0.1849</td>
<td>0.0038</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>2.9086</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Gasoline Market

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

The regression equation is

\[
\log G = -0.558 + 1.29 \log\text{Income} - 0.0280 \log\text{PG} \\
- 0.156 \log\text{PNC} + 0.029 \log\text{PUC} - 0.183 \log\text{PPT}
\]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.5579</td>
<td>0.5808</td>
<td>-0.96</td>
<td>0.342</td>
</tr>
<tr>
<td>logIncome</td>
<td>1.2861</td>
<td>0.1457</td>
<td>8.83</td>
<td>0.000</td>
</tr>
<tr>
<td>logPG</td>
<td>-0.02797</td>
<td>0.04338</td>
<td>-0.64</td>
<td>0.522</td>
</tr>
<tr>
<td>logPNC</td>
<td>-0.1558</td>
<td>0.2100</td>
<td>-0.74</td>
<td>0.462</td>
</tr>
<tr>
<td>logPUC</td>
<td>0.0285</td>
<td>0.1020</td>
<td>0.28</td>
<td>0.781</td>
</tr>
<tr>
<td>logPPT</td>
<td>-0.1828</td>
<td>0.1191</td>
<td>-1.54</td>
<td>0.132</td>
</tr>
</tbody>
</table>

S = 0.0499953  R-Sq = 96.0%  R-Sq(adj) = 95.6%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>5</td>
<td>2.79360</td>
<td>0.55872</td>
<td>223.53</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>46</td>
<td>0.11498</td>
<td>0.00250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>2.90858</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

**Observed Data**

**Polynomial Class**

7 Parameters \((B_0, B_1, \ldots, B_7)\)

\[
y = B_0 + B_1 x_1 + B_2 x_2 + B_3 x_3 + B_4 x_4 + B_5 x_5 + B_6 x_6 + \varepsilon
\]

**Certified Regression Statistics**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Deviation of Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B_0</strong></td>
<td>-3482258.63459582</td>
<td>890420.383607373</td>
</tr>
<tr>
<td><strong>B_1</strong></td>
<td>15.0618722713733</td>
<td>84.9149257747669</td>
</tr>
<tr>
<td><strong>B_2</strong></td>
<td>-0.358191792925910E-01</td>
<td>0.334910077722432E-01</td>
</tr>
<tr>
<td><strong>B_3</strong></td>
<td>-2.0202980381683</td>
<td>0.488399681651699</td>
</tr>
<tr>
<td><strong>B_4</strong></td>
<td>-1.03322686717359</td>
<td>0.214274163161675</td>
</tr>
<tr>
<td><strong>B_5</strong></td>
<td>-0.511041056535807E-01</td>
<td>0.226073200069370</td>
</tr>
<tr>
<td><strong>B_6</strong></td>
<td>1829.15146461355</td>
<td>455.478499142212</td>
</tr>
</tbody>
</table>

**Y**

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Standard Error</th>
<th>t</th>
<th>Prob.</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>(-.34823D+07***)</td>
<td>890420.4</td>
<td>-3.91</td>
<td>.0036 (-.54965D+07) (-.14680D+07)</td>
</tr>
<tr>
<td>X1</td>
<td>15.0619</td>
<td>84.91493</td>
<td>.18</td>
<td>.8631 (-177.0290) (207.1528)</td>
</tr>
<tr>
<td>X2</td>
<td>-.03582</td>
<td>.03349</td>
<td>-1.07</td>
<td>.3127 (-.11158) (.03994)</td>
</tr>
<tr>
<td>X3</td>
<td>-2.02023***</td>
<td>.48840</td>
<td>-4.14</td>
<td>.0025 (-3.12507) (-.91539)</td>
</tr>
<tr>
<td>X4</td>
<td>-1.03323***</td>
<td>.21427</td>
<td>-4.82</td>
<td>.0009 (-1.51795) (-.54851)</td>
</tr>
<tr>
<td>X5</td>
<td>-.05110</td>
<td>.22607</td>
<td>-2.23</td>
<td>.8262 (-.56252) (.46031)</td>
</tr>
<tr>
<td>X6</td>
<td>1829.15***</td>
<td>455.4785</td>
<td>4.02</td>
<td>.0030 (798.79) (2859.52)</td>
</tr>
</tbody>
</table>
Excel Longley Solution

SUMMARY OUTPUT

Regression Statistics
Multiple R 0.997737
R Square 0.995479
Adjusted R Square 0.992465
Standard Error 304.8541
Observations 16

ANOVA
df SS MS F Significance F
Regression 6 1.84E+08 30695400 330.2853 4.98E-10
Residual 9 836424.1 92936.01
Total 15 1.85E+08

Coefficient Standard Error t Stat P-value Lower 95% Upper 95% Lower 95.0% Upper 95.0%
Intercept -3482258.63459582 -3.9108 0.00356 -5496529 -1467988 -5496529 -1467988
X Variable 1 15.06187 84.91493 0.177376 0.863141 -177.029 207.1528 -177.029 207.1528
X Variable 2 -0.03582 0.033491 -1.06952 0.312681 -0.11158 0.039943 -0.11158 0.039943
X Variable 3 -2.02023 0.4884 -4.13643 0.002535 -3.12507 -0.91539 -3.12507 -0.91539
X Variable 4 -1.03323 0.214274 -4.82199 0.000944 -1.51795 -0.54851 -1.51795 -0.54851
X Variable 5 0.0511 0.226073 -0.22605 0.826212 -0.56252 0.460309 -0.56252 0.460309
X Variable 6 1829.151 455.4785 4.01589 0.003037 798.7875 2859.515 798.7875 2859.515

Estimate
-3482258.63459582
15.06187
-0.03582
-2.02023
-1.03323
0.0511
1829.151

1829.151

15.06187
-0.03582
-2.02023
-1.03323
0.0511
1829.151
The NIST Filipelli Problem

```plaintext
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

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

- Residual Standard Deviation: 0.334801051324544E-02
- R-Squared: 0.996727416185620

### Certified Analysis of Variance Table

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of Freedom</th>
<th>Sums of squares</th>
<th>Mean Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>10</td>
<td>0.242391619837339</td>
<td>0.242391619837339E-01</td>
</tr>
<tr>
<td>Residual</td>
<td>71</td>
<td>0.795851382172941E-03</td>
<td>0.112091743968020E-04</td>
</tr>
</tbody>
</table>
Regression Analysis: $y$ versus $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$

* WARNING * $x_3$ is highly correlated with other predictors.
* WARNING * $x_4$ is highly correlated with other predictors.
* WARNING * $x_5$ is highly correlated with other predictors.
* WARNING * $x_6$ is highly correlated with other predictors.
* WARNING * $x_7$ is highly correlated with other predictors.
* WARNING * $x_8$ is highly correlated with other predictors.
* WARNING * $x_9$ is highly correlated with other predictors.

The regression equation is

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

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-1467.5</td>
<td>298.1</td>
<td>-4.92</td>
<td>0.000</td>
</tr>
<tr>
<td>$x_1$</td>
<td>-2772.1</td>
<td>559.8</td>
<td>-4.95</td>
<td>0.000</td>
</tr>
<tr>
<td>$x_2$</td>
<td>-2316.3</td>
<td>466.5</td>
<td>-4.97</td>
<td>0.000</td>
</tr>
<tr>
<td>$x_3$</td>
<td>-1128.0</td>
<td>227.2</td>
<td>-4.96</td>
<td>0.000</td>
</tr>
<tr>
<td>$x_4$</td>
<td>-354.47</td>
<td>71.65</td>
<td>-4.95</td>
<td>0.000</td>
</tr>
<tr>
<td>$x_5$</td>
<td>-75.12</td>
<td>15.29</td>
<td>-4.91</td>
<td>0.000</td>
</tr>
<tr>
<td>$x_6$</td>
<td>-10.875</td>
<td>2.237</td>
<td>-4.86</td>
<td>0.000</td>
</tr>
<tr>
<td>$x_7$</td>
<td>-1.0622</td>
<td>0.2216</td>
<td>-4.79</td>
<td>0.000</td>
</tr>
<tr>
<td>$x_8$</td>
<td>-0.06702</td>
<td>0.01424</td>
<td>-4.71</td>
<td>0.000</td>
</tr>
<tr>
<td>$x_9$</td>
<td>-0.00024678</td>
<td>0.0005356</td>
<td>-4.61</td>
<td>0.000</td>
</tr>
<tr>
<td>$x_{10}$</td>
<td>-0.00004030</td>
<td>0.000000897</td>
<td>-4.49</td>
<td>0.000</td>
</tr>
</tbody>
</table>

$S = 0.00334800$  \hspace{1em} $R$-Sq = 99.7\%  \hspace{1em} $R$-Sq(adj) = 99.6\%
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.
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
Regression Sum of Squares = 27339.2
Residual Sum of Squares = .515124E-10
Total Sum of Squares = 27339.2

Standard error of e = .00000
Root MSE = .00000
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

<table>
<thead>
<tr>
<th>X2</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>t</th>
<th>Prob.</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-.63802***</td>
<td>.00419</td>
<td>-152.40</td>
<td>.0000</td>
<td>-.64623 - .62982</td>
</tr>
<tr>
<td>X1</td>
<td>-1.19955***</td>
<td>.00394</td>
<td>-304.78</td>
<td>.0000</td>
<td>-1.20726 - -1.19184</td>
</tr>
<tr>
<td>X3</td>
<td>-.48688***</td>
<td>.00159</td>
<td>-305.76</td>
<td>.0000</td>
<td>-.49000 - -.48376</td>
</tr>
<tr>
<td>X4</td>
<td>-.15336***</td>
<td>.00100</td>
<td>-153.37</td>
<td>.0000</td>
<td>-.15532 - -.15140</td>
</tr>
<tr>
<td>X5</td>
<td>-.03267***</td>
<td>.00032</td>
<td>-102.67</td>
<td>.0000</td>
<td>-.03329 - -.03204</td>
</tr>
<tr>
<td>X6</td>
<td>-0.00477***</td>
<td>.6159D-04</td>
<td>-77.40</td>
<td>.0000</td>
<td>-.00489 - -.00465</td>
</tr>
<tr>
<td>X7</td>
<td>0.00047***</td>
<td>.7558D-05</td>
<td>-62.28</td>
<td>.0000</td>
<td>-.00049 - -.00046</td>
</tr>
<tr>
<td>X8</td>
<td>-.30124D-04***</td>
<td>.5766D-06</td>
<td>-52.25</td>
<td>.0000</td>
<td>-.31254D-04 - -.28994D-04</td>
</tr>
<tr>
<td>X9</td>
<td>-.11284D-05***</td>
<td>.2501D-07</td>
<td>-45.11</td>
<td>.0000</td>
<td>-.11775D-05 - -.10794D-05</td>
</tr>
<tr>
<td>X10</td>
<td>0.000***</td>
<td>.4725D-09</td>
<td>-39.78</td>
<td>.0000</td>
<td>-.19725D-07 - -.17872D-07</td>
</tr>
</tbody>
</table>

Note: nnnnn.D-xx or D+xx => multiply by 10 to -xx or +xx.
Note: ***. **. * => Significance at 1%, 5%, 10% level.

$\rightarrow$ calc : peek ; 1 -.515124e-10/27339.2$
\text{[CALC]} = .99999999999999810D+00$
Using QR Decomposition

### Ordinary least squares regression

<table>
<thead>
<tr>
<th>LHS=Y</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>DegFreedom</th>
<th>Mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>= .84958</td>
<td>= .05479</td>
<td>= 82</td>
<td>= .242392</td>
</tr>
</tbody>
</table>

#### Regression

<table>
<thead>
<tr>
<th>Sum of Squares</th>
<th>= .795851E-03</th>
<th>= 71</th>
<th>= .00001</th>
</tr>
</thead>
</table>

#### Residual

| Sum of Squares | = .243187 | = 81 | = .00300 |

#### Total

| Standard error of e | = .00335 | R_root MSE | = .00312 |

| R-squared | = 0.99673 | R-bar squared | = 0.99627 |

**Model test** $F[10, 71] = 2162.43959$  
Prob $F > F^*$ = 0.00000

<table>
<thead>
<tr>
<th>Y</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>t</th>
<th>Prob</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-1467.49***</td>
<td>298.0845</td>
<td>-4.92</td>
<td>0.0000</td>
<td>-1467.48961422980</td>
</tr>
<tr>
<td>X1</td>
<td>-2772.18***</td>
<td>559.7799</td>
<td>-4.95</td>
<td>0.0000</td>
<td>-2772.17959193342</td>
</tr>
<tr>
<td>X2</td>
<td>-2316.37***</td>
<td>466.4776</td>
<td>-4.97</td>
<td>0.0000</td>
<td>-2316.37108160893</td>
</tr>
<tr>
<td>X3</td>
<td>-1127.97***</td>
<td>227.2043</td>
<td>-4.96</td>
<td>0.0000</td>
<td>-1127.97394098372</td>
</tr>
<tr>
<td>X4</td>
<td>-354.478***</td>
<td>71.64787</td>
<td>-4.95</td>
<td>0.0000</td>
<td>-354.478233703349</td>
</tr>
<tr>
<td>X5</td>
<td>-75.1242***</td>
<td>15.28972</td>
<td>-4.91</td>
<td>0.0000</td>
<td>-75.1242017393757</td>
</tr>
<tr>
<td>X6</td>
<td>-10.8753***</td>
<td>2.23691</td>
<td>-4.86</td>
<td>0.0000</td>
<td>-10.8753180355343</td>
</tr>
<tr>
<td>X7</td>
<td>-1.06222***</td>
<td>0.22162</td>
<td>-4.79</td>
<td>0.0000</td>
<td>-1.06221498588947</td>
</tr>
<tr>
<td>X8</td>
<td>-0.06702***</td>
<td>0.01424</td>
<td>-4.71</td>
<td>0.0000</td>
<td>-0.0670191154593408E-01</td>
</tr>
<tr>
<td>X9</td>
<td>-0.00247***</td>
<td>0.00054</td>
<td>-4.61</td>
<td>0.0000</td>
<td>-0.00246781078275479E-02</td>
</tr>
<tr>
<td>X10</td>
<td>-40296D-04***</td>
<td>.8966D-05</td>
<td>-4.49</td>
<td>0.0000</td>
<td>-0.402962525080404E-04</td>
</tr>
</tbody>
</table>
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

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

\[ + \gamma_{11} \log^2 x_1 + \gamma_{12} \frac{1}{2} \log x_1 \log x_2 + \gamma_{13} \frac{1}{2} \log x_1 \log x_3 + \]

\[ + \gamma_{22} \log^2 x_2 + \gamma_{23} \frac{1}{2} \log x_2 \log x_3 + \]

\[ + \gamma_{33} \log^2 x_3 \]

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 values 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!!!