Topic 6:
Multiple Regression

Major Topics:

- Matrix Notation
- Assumptions
- Estimators
- Output
- ANOVA
- Adjusted R^2

Multiple Regression
Introduction

- Expanded model
  - Our model now must account for several factors

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_p X_{ip} + e_i$$

  - Now have *i* equations with $p > 1$ independent variables plus one constant
    - Have $p+1$ unknown parameters
    - Still have *n* observations

Multiple Regression
Notation

- To economize on writing equations, use matrix notation
  - Model is now written as

$$\boxed{\mathbf{Y} = \mathrm{X}b + e}$$

where

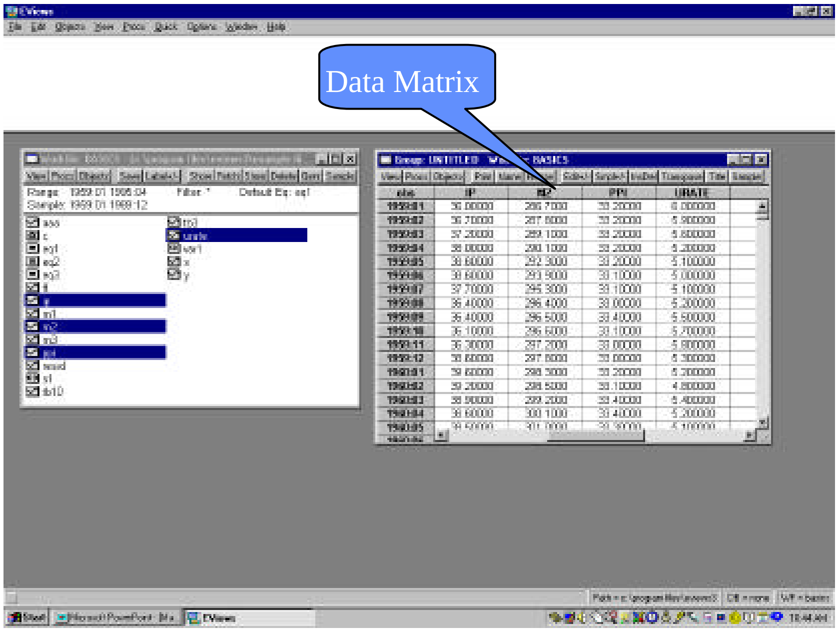
$$\mathbf{Y} = \begin{vmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{vmatrix} \qquad \mathbf{X} = \begin{vmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & \ldots & X_{2p} \\ & & \ldots & & \\ 1 & X_{n1} & X_{n2} & \ldots & X_{np} \end{vmatrix}$$

n X 1 n X (p+1)

Multiple Regression
Notation (Continued)

$$b = \begin{vmatrix} b_0 \\ b_1 \\ b_2 \\ \ldots \\ b_p \end{vmatrix} \qquad e = \begin{vmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{vmatrix}$$

(p+1) X 1 nX1



Multiple Regression
Classical Assumptions

- Same classical assumptions as for simple regression model, but...
  - New assumption:
    - The *p* independent variables are linearly independent

Multiple Regression
Classical Assumptions (Continued)

- New assumption interpretation
  - Cannot write one independent variable as a linear combination of the other *p - 1* variables
    - Example: cannot write

$$\mathbf{X}_1 = a\mathbf{X}_2 + g\mathbf{X}_3$$

  - If could write one variable as linear combination, then that variable is redundant
  - Example

$$\mathrm{RU_t} = b_0 + b_1\mathrm{GDP_t} + b_2\mathrm{C_t} + b_3\mathrm{I_t} + b_4\mathrm{G_t} + b_5\mathrm{X_t} + e_t$$

but

$$\mathrm{GDP} = \mathrm{C} + \mathrm{I} + \mathrm{G} + \mathrm{X}$$

Multiple Regression
Classical Assumptions (Continued)

- Writing any independent variable as linear combination of one or more of the other independent variables is called ***multicollinearity***
  - Serious problems arise if multicollinearity exists
  - Yet this is a common problem with economic data
  - We will discuss this extensively in a separate lecture

Multiple Regression
Digression on Inverses

- Suppose we have the two simultaneous equations

$$-2 = 4b_1 - 10b_2$$
$$13 = 3b_1 + 7b_2$$

- In matrix form, this is

$$\begin{bmatrix} -2 \\ 13 \end{bmatrix} = \begin{bmatrix} 4 & -10 \\ 3 & 7 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

Econometrics 322

Multiple Regression
Digression on Inverses (Continued)

- Solve these two simultaneous equations for b_1

$$(-2)(7) = (4)(7)b_1 - (10)(7)b_2$$
$$(13)(-10) = (3)(-10)b_1 + (7)(-10)b_2$$
$$or$$
$$-14 = 28b_1 - 70b_2$$
$$-130 = -30b_1 - 70b_2$$

- So $$b_1 = \frac{116}{58} = 2$$

Multiple Regression
Classical Assumptions (Continued)

Key Classical Assumptions

Normality	$e_i \approx N \quad \forall i$
Zero Mean	$E(e_i) = 0 \quad \forall i$
Homoskedasticity	$s_e^2 = s^2 \quad \forall i$
Zero Autocorrelation	$COV(e_i, e_j) = 0 \quad \forall i, j \quad i \neq j$
Independence of X_i, e_i	$COV(e_i, X_i) = 0 \quad \forall i$
Linearity	Linear in Parameters
No Exact Linear Relationship	No Multicollinearity
Non-stochastic X_i	Fixed X

Multiple Regression
Estimators

- Derive estimators in same manner as before
  - Minimize the error sum of squares
  - Solve normal equations for unknown parameter estimators
    - Have *p+1* normal equations
- Estimator written in matrix notation

$$\hat{b} = (\mathbf{X'}\,\mathbf{X})^{-1}\,\mathbf{X'}\,\mathbf{Y}$$

Multiple Regression
Estimators (Continued)

- Note that estimator form is no different than previously derived estimator for simple OLS
  - Multiple regression

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

  - Simple regression

$$\hat{b}_1 = \frac{S_{XY}}{S_{XX}}$$
$$= S_{XX}^{-1} S_{XY}$$

Multiple Regression
Estimators (Continued)

- Other estimator formulas

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

$$\hat{\mathbf{Y}} = \mathrm{X}\hat{b} = \mathbf{HY} \qquad \mathrm{H} = (\mathrm{X'X})^{-1}\mathrm{X'}$$

$$s_{\hat{\mathbf{Y}}}^2 = s^2\mathbf{H}$$

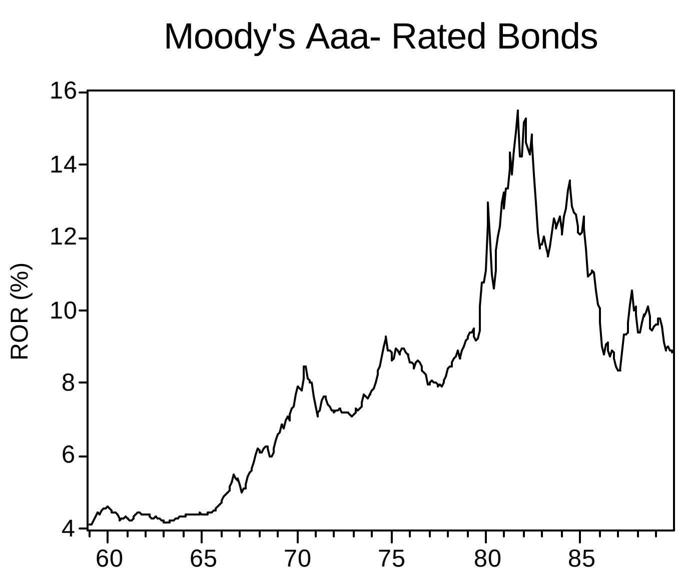
$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

$$s_{\mathbf{e}}^2 = s^2(\mathbf{I} - \mathbf{H}) = s^2\mathbf{M}$$

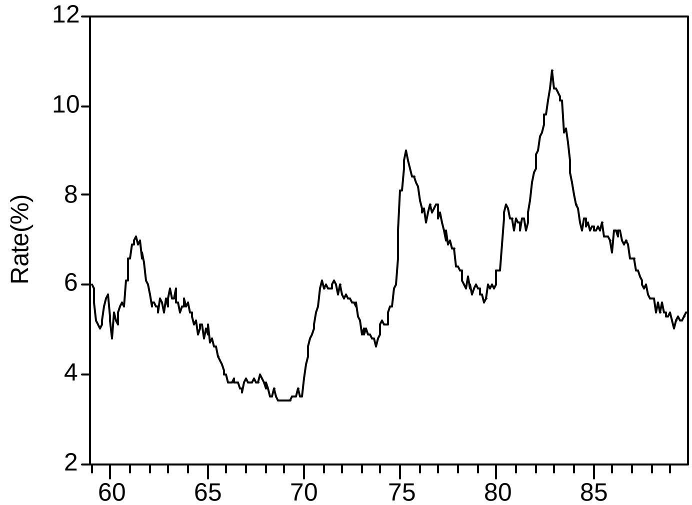
$$\mathrm{s}^2 = \frac{\mathbf{e'e}}{\mathrm{n - p - 1}}$$

Multiple Regression
Estimators (Continued)

- Properties
  - Desirable small sample properties still hold for general case of multiple regression
    - Linearity
    - Unbiasedness
    - Minimum variance
  - Under expanded classical assumptions, OLS estimators are still BLUE

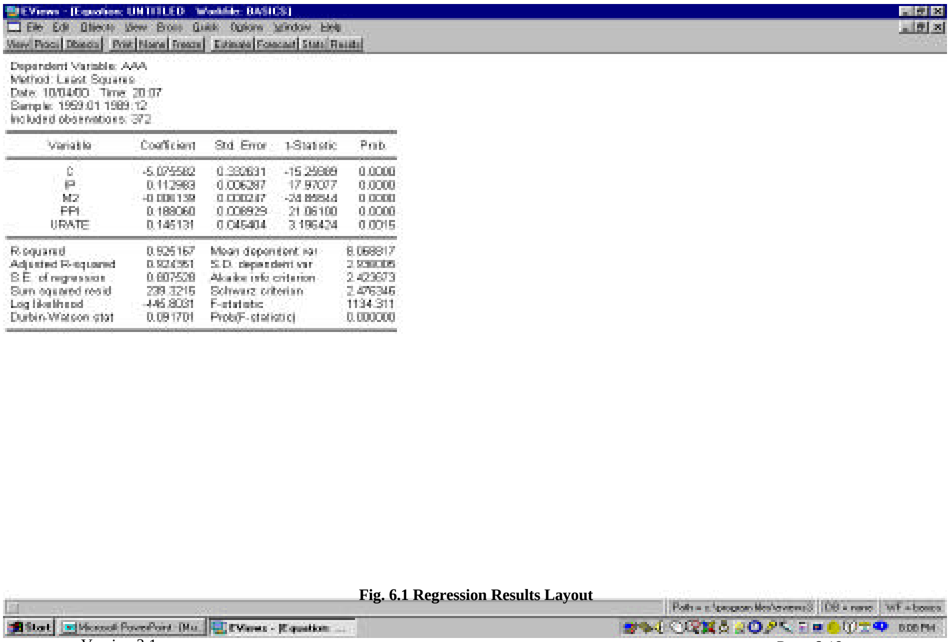


Unemployment Rate



Variable Coding Sheet

Variable	Mnemonic	Source	Possible Values
Moody's AA Rated Corporate Bonds	AAA	Moody's	% per Annum
Industrial Production Index	IP	Federal Reserve	Index, 1987=100, SA
Money Supply	M2	Federal Reserve	$Billion
Producer Price Index, Finished Goods	PPI	BLS	Index, 1982=100, NSA
Unemploy. Rate	URATE	BLS	% for 16+ older



Dependent Variable: AAA
Method: Least Squares
Date: 10/04/00 Time: 20:07
Sample: 1959:01 1989:12
Included observations: 372

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-5.075582	0.332631	-15.25889	0.0000
IP	0.112983	0.006287	17.97077	0.0000
M2	-0.006139	0.000247	-24.85844	0.0000
PPI	0.188060	0.008929	21.06100	0.0000
URATE	0.145131	0.045404	3.196424	0.0015

R-squared	0.925167	Mean dependent var	8.068817
Adjusted R-squared	0.924351	S.D. dependent var	2.938006
S.E. of regression	0.807528	Akaike info criterion	2.423673
Sum squared resid	239.3215	Schwarz criterion	2.476346
Log likelihood	-445.8031	F-statistic	1134.311
Durbin-Watson stat	0.091701	Prob(F-statistic)	0.000000

Fig. 6.1 Regression Results Layout

Multiple Regression
Goodness of Fit

ANOVA Table

Source of Variation	DF	Sum of Squares	Mean Square	F-Ratio
Regression	p	SSR	MSR= SSR/p	MSR/MSE
Residual	n - p - 1	SSE	MSE= SSE/n-p-1	
Total	n - 1	SST	MST= SST/n-1	

Multiple Regression
New R^2

- Have to adjust conventional R^2
  - R^2 can be made arbitrarily large simple by adding independent variables to model
  - New R^2 is called ***Adjusted R²*** or $\overline{R}^2$

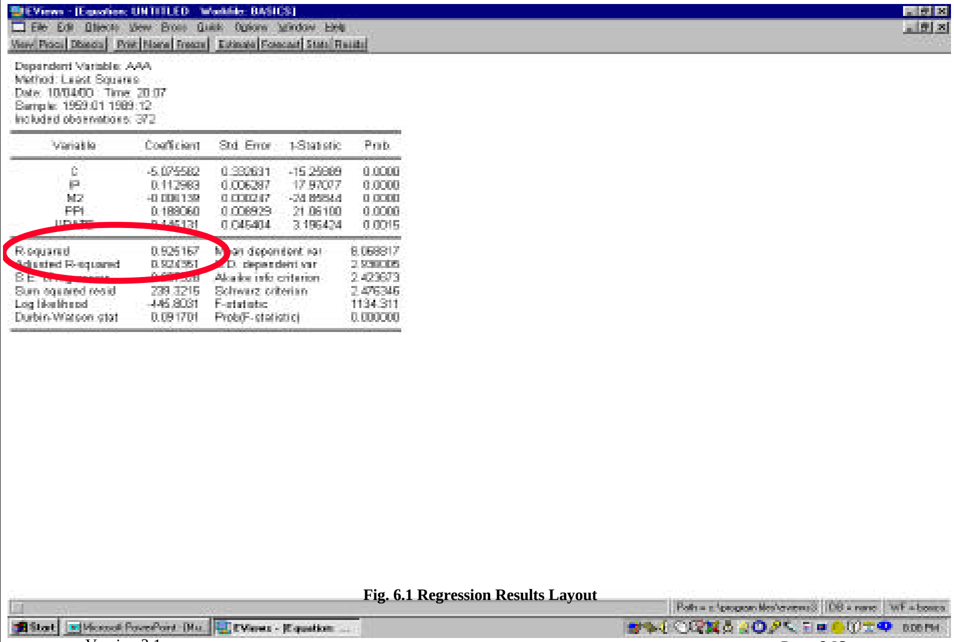
$$\overline{R}^2 = 1 - \left(\frac{n-1}{n-p-1}\right)(1-R^2)$$
$$= 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$
$$= 1 - \frac{s^2}{s_Y^2}$$

Multiple Regression
New R^2 (Continued)

- Notice that

$$\overline{R}^2 \le R^2$$

One modeling objective
  - Maximize $\overline{R}^2$
- Property of $\overline{R}^2$
  - Can decline if added independent variable has no or little explanatory power
  - Reason
    - Increase n - p - 1



Dependent Variable: AAA
Method: Least Squares
Date: 10/04/00 Time: 20:07
Sample: 1959:01 1989:12
Included observations: 372

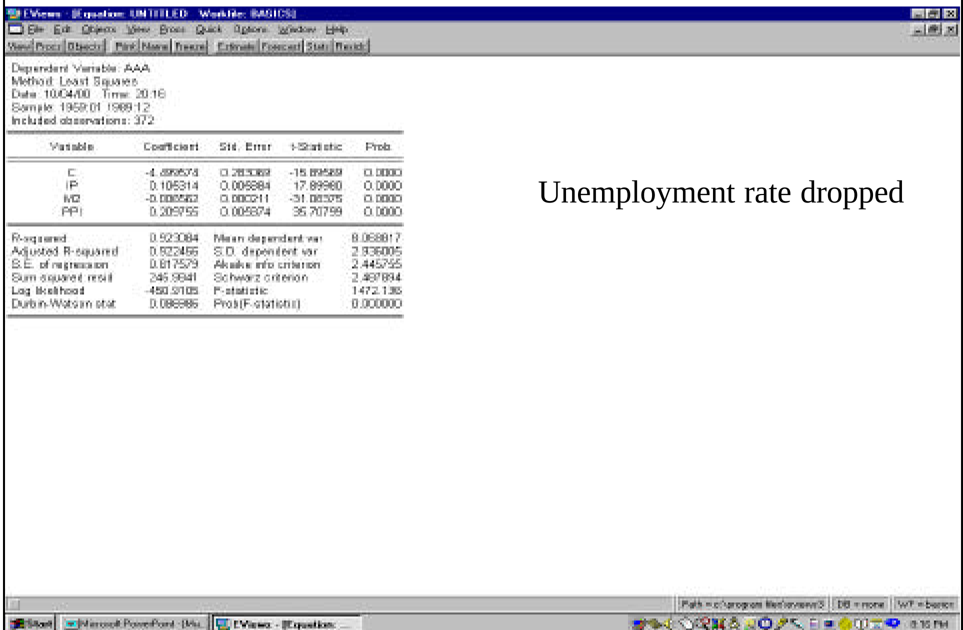
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-5.075582	0.332631	-15.25889	0.0000
IP	0.112983	0.006287	17.97077	0.0000
M2	-0.006139	0.000247	-24.85844	0.0000
PPI	0.188060	0.008929	21.06100	0.0000
URATE	0.145131	0.045404	3.196424	0.0015

R-squared	0.925167	Mean dependent var	8.068817
Adjusted R-squared	0.924351	S.D. dependent var	2.938006
S.E. of regression	[illegible]	Akaike info criterion	2.423673
Sum squared resid	239.3215	Schwarz criterion	2.476346
Log likelihood	-445.8031	F-statistic	1134.311
Durbin-Watson stat	0.091701	Prob(F-statistic)	0.000000

Fig. 6.1 Regression Results Layout

Multiple Regression
Alternative Models

- Can now develop alternative models by adding or deleting variables
  - Each model is a new regression run
  - Criteria for selecting variables will be discussed later

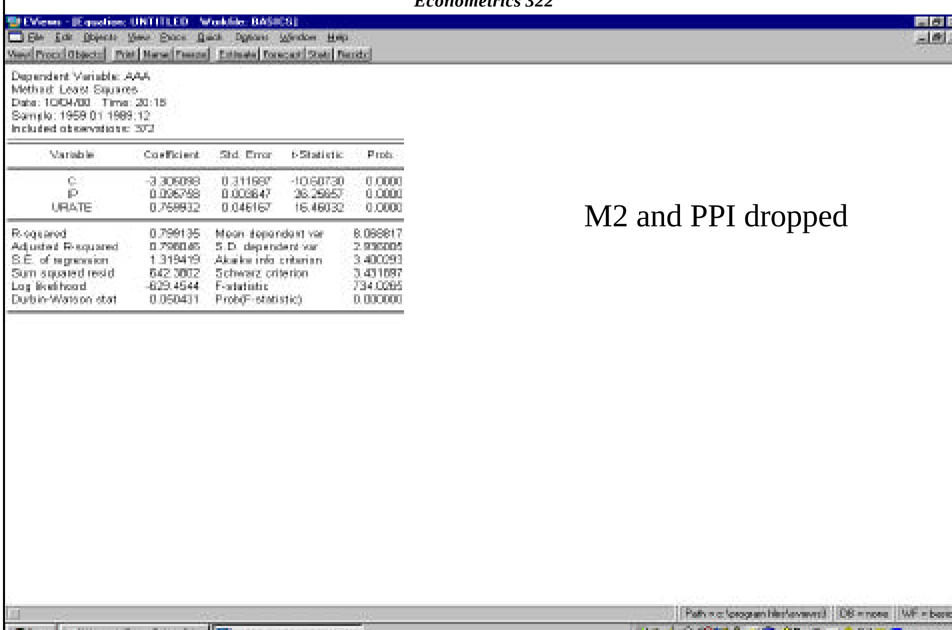


Dependent Variable: AAA
Method: Least Squares
Date: 10/04/00 Time: 20:16
Sample: 1959:01 1989:12
Included observations: 372

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-4.499574	0.283089	-15.89569	0.0000
IP	0.105314	0.005884	17.89980	0.0000
M2	-0.006562	0.000211	-31.08375	0.0000
PPI	0.209755	0.005874	35.70759	0.0000

R-squared	0.923084	Mean dependent var	8.068817
Adjusted R-squared	0.922456	S.D. dependent var	2.936005
S.E. of regression	0.817579	Akaike info criterion	2.445755
Sum squared resid	245.9841	Schwarz criterion	2.487894
Log likelihood	-450.9105	F-statistic	1472.136
Durbin-Watson stat	0.088986	Prob(F-statistic)	0.000000

Unemployment rate dropped



Dependent Variable: AAA
Method: Least Squares
Date: 10/04/00 Time: 20:15
Sample: 1959:01 1989:12
Included observations: 372

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.306098	0.311697	-10.60730	0.0000
IP	0.096798	0.003647	26.25657	0.0000
URATE	0.759932	0.046167	16.46032	0.0000

R-squared	0.799135	Mean dependent var	8.068817
Adjusted R-squared	0.798046	S.D. dependent var	2.936005
S.E. of regression	1.319419	Akaike info criterion	3.400293
Sum squared resid	642.3802	Schwarz criterion	3.431897
Log likelihood	-629.4544	F-statistic	734.0265
Durbin-Watson stat	0.050431	Prob(F-statistic)	0.000000

M2 and PPI dropped

Model Portfolio

	Models		
Independent Variables	1	2	3
Constant	-5.075* (0.0000)	-4.499* (0.0000)	-3.305* (0.0000)
IP	0.112* (0.0000)	0.105* (0.0000)	0.095* (0.0000)
M2	-0.006* (0.0000)	-0.006* (0.0000)	
PPI	0.188* (0.0000)	0.209* (0.0000)	
URATE	0.145* (0.0000)		0.759* (0.0000)
R^2	0.925	0.923	0.799
R^2 Adjusted	0.924	0.922	0.798
F-stat	1134.311* (0.0000)	1472.136* (0.0000)	734.026* (0.0000)
N	372	372	372

Notes: p-values in parentheses; * = significant

Multiple Regression
Model Selection

- Develop ***Model Portfolio***
- Very important topic
  - Will discuss shortly

Multiple Regression
Knowledge Checks

- Outline the procedure for deriving the estimator vector for the parameters. How does this compare to the procedure used in Topic 5?
- What is the final form in matrix notation for the parameter vector estimator? For the standard linear model, what is the first element of the estimator?

Econometrics 322

Multiple Regression
Knowledge Checks (Continued)

- What is the problem with R^2?
- How does $\bar{R}^2$ correct for this problem?
- What is an upper bound on $\bar{R}^2$? Can it ever decrease or be negative?