Quantitative Methods: Correlation and Regression

Master

Financial Economics

ANNÉE SCOLAIRE / ACADEMIC YEAR 2011-2012

Intervenant/Lecturer: Sergio FOCARDI

Correlation and Regression

- Correlation Analysis
- Linear Regression:models, estimation, diagnostic
- Multiple Linear Regression: model, estimation, diagnostic
- Dummy Variables

Textbooks and References

Textbook

- *Quantitative Investment Analysis*, CFA Institute Investment Series, 2nd edition.

References from the CFA Institute Research Foundation

- Fabozzi, F., S. Focardi, and C. Jonas, *Challenges in Quantitative Equity Management,* Research Foundation Publications (April 2008) available online at: http://www.cfapubs.org/toc/rf/2008/2008/2?cookieSet=1

- Fabozzi, F., S. Focardi, and P. Kolm, *Trends in Quantitative Finance*, Research Foundation of CFA Institute (April 21, 2006)

Articles on economics as a mathematical science

- Sergio Focardi and Frank J. Fabozzi, "The Reasonable Effectiveness of Mathematics in Economics" Spring 2009 *The American Economist*.

- Frank J. Fabozzi, Sergio M. Focardi, and Caroline L. Jonas, "Considerations on the Challenges in Quantitative Equity Management", *Quantitative Finance,* Volume 8, Issue 7, 2008, Pages 649 – 665.

- Sergio Focardi and Frank J. Fabozzi, "Black Swans and White Eagles: On Mathematics and Finance," *Mathematical Methods in Operations Research,* published online 5 September 2008.

Textbooks and References

Other Useful References

- The Professional Risk Managers' Handbook, *A Comprehensive Guide to Current Theory and Best Practices*, Edited by Carol Alexander and Elizabeth Sheedy. Volume II: Mathematical Foundations of Risk Measurement.
- Brooks, C. *Introductory Econometrics for Finance*, Cambridge University Press.
- Campbell, J. Y., Lo, A. W., and MacKinlay, C. A., *The Econometrics of Financial Markets*, Princeton University Press.
- DeGroot, M. H., *Probability and Statistics*, Addison Wesley.
- Fabozzi, F., S. Focardi, and P. Kolm, *Financial Modeling of the Equity Market: From CAPM to Cointegration*, Wiley
- Johnston, J. and J. Dinardo, *Econometric Methods*, 4th Edition, Mc Graw Hill.
- L'Habitant, Serge François, *Hedge Funds – Quantitative Insights*, John Wiley & Sons.
- Taylor, S., *Modelling Financial Time Series*, John Wiley & Sons.

Covariance and Correlation

Variance-covariance matrices

- Var-cov matrices are central to modern portfolio theory
- Estimation of the var-cov matrices is critical for portfolio management and asset allocation

Covariances

- Suppose returns are a multivariate random vector written as:

$$\mathbf{r}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t$$

- The random disturbances $\boldsymbol{\varepsilon}_t$ will be characterized by a covariance matrix $\boldsymbol{\Omega}$

- The entries $\sigma_{i,j}$ are the covariances between the returns of asset *i* and asset *j*.

Covariance

The *covariance* between the two variables is defined as:

$$\sigma_{X,Y} = Cov(X,Y) =$$
$$= E\left[\left(X - E[X]\right)\left(Y - E[Y]\right)\right] =$$
$$= E[XY] - E[X]E[Y]$$

Correlation coefficient

- The *correlation coefficient* is the covariance normalized with the product of the respective standard deviations:

$$\rho_{X,Y} = Corr(X,Y) =$$

$$= \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

- The correlation coefficient expresses a measure of linear dependence

Properties of correlation

- The correlation coefficient can assume values between -1 and +1 inclusive
- The variables *X,Y* are proportional without any noise term if and only if the correlation coefficient is +/-1
- If there is a noise term, then the correlation coefficient assumes a value intermediate between -1 and +1
- If variables are independent, the correlation coefficient is zero
- The converse is not true: two variables can exhibit non linear dependence with the correlation coefficient zero
- Uncorrelated variables are not necessarily independent. If the variables *X,Y* have a non linear dependence relationship, the correlation coefficient might become meaningless

Properties of correlation, cont'd..

- The correlation coefficient fully represents the dependence structure of multivariate normal distribution
- More in general, the correlation coefficient is unproblematic on elliptic distribution, i.e., distributions that are constants on ellipsoids
- In other cases, different measures of dependence are needed, e.g., copula functions
- "Modelling Dependence with Copulas and Applications to Risk Management", Paul Embrechts, Filip Lindskog, and Alexander McNeil

Empirical covariance

The empirical covariance between two variables is defined as:

$$\hat{\sigma}_{X,Y} = \frac{1}{n}\sum_{i=1}^{nn}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

where

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

are the empirical means of the variables

Empirical correlation coefficient

- The empirical correlation coefficient is the empirical covariance normalized with the product of the respective empirical standard deviations:

$$\hat{\rho}_{X,Y} = \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X \hat{\sigma}_Y}$$

- The empirical standard deviations are defined as

$$\hat{\sigma}_X = \sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}, \quad \hat{\sigma}_Y = \sqrt{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}$$

A sample of correlated data

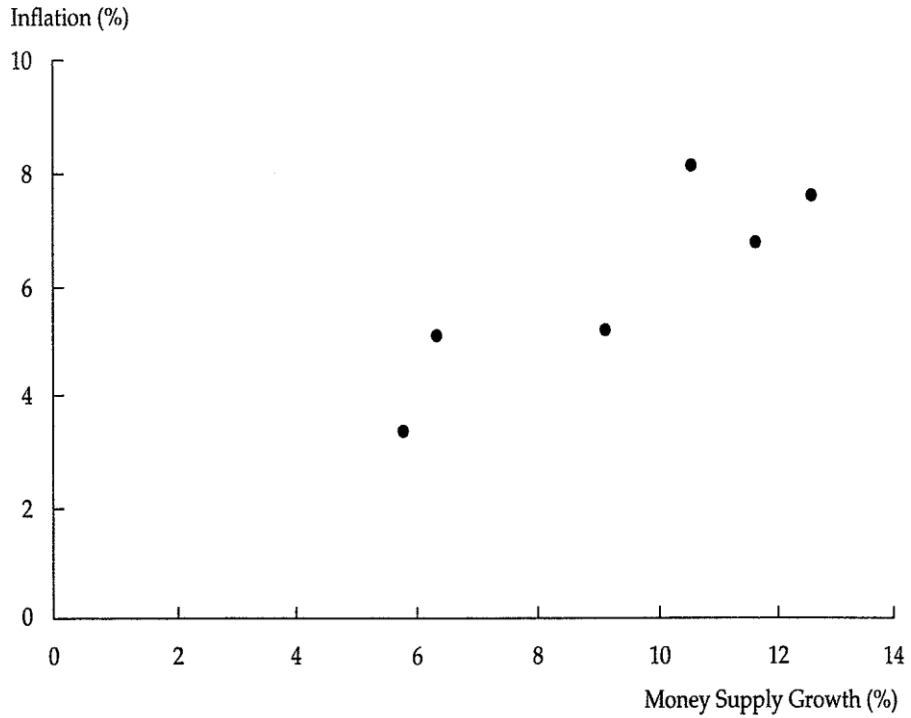
TABLE 8-1 Annual Money Supply Growth Rate and Inflation Rate by Country, 1970–2001

Country	Money Supply Growth Rate	Inflation Rate
Australia	11.66%	6.76%
Canada	9.15%	5.19%
New Zealand	10.60%	8.15%
Switzerland	5.75%	3.39%
United Kingdom	12.58%	7.58%
United States	6.34%	5.09%
Average	9.35%	6.03%

Source: International Monetary Fund.

Scatterplot of data

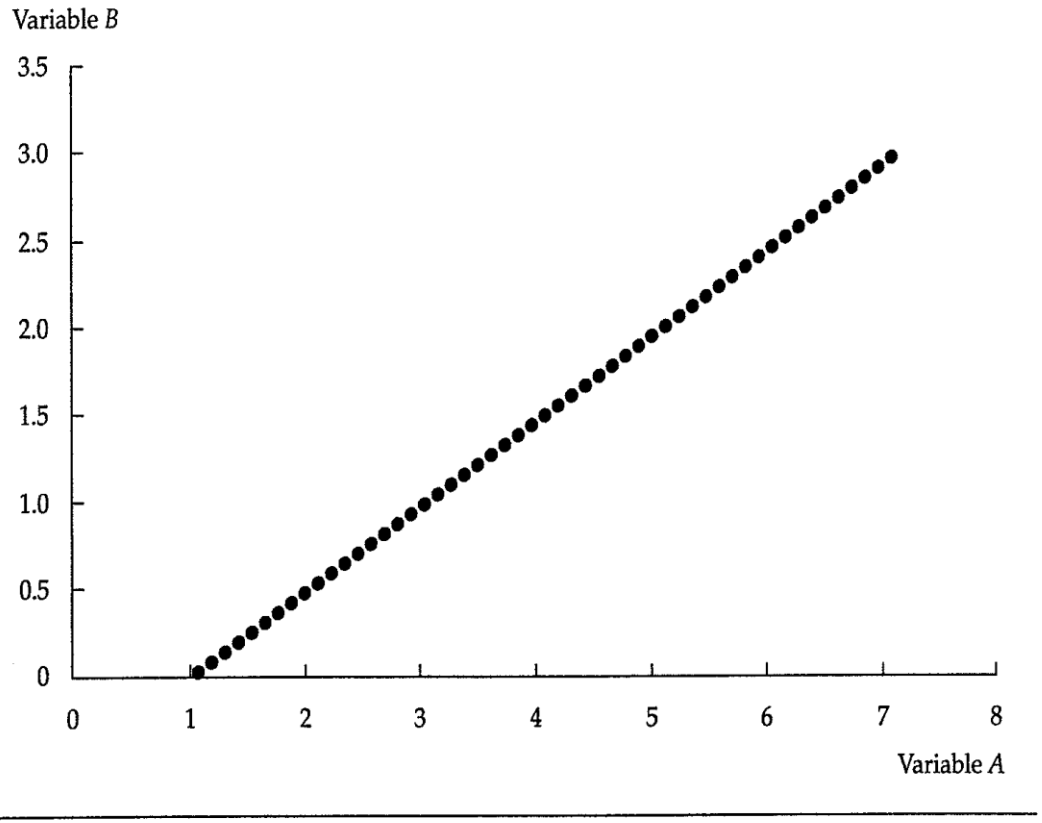
FIGURE 8-1 Scatter Plot of Annual Money Supply Growth Rate and Inflation Rate by Country: 1970–2001



Source: International Monetary Fund

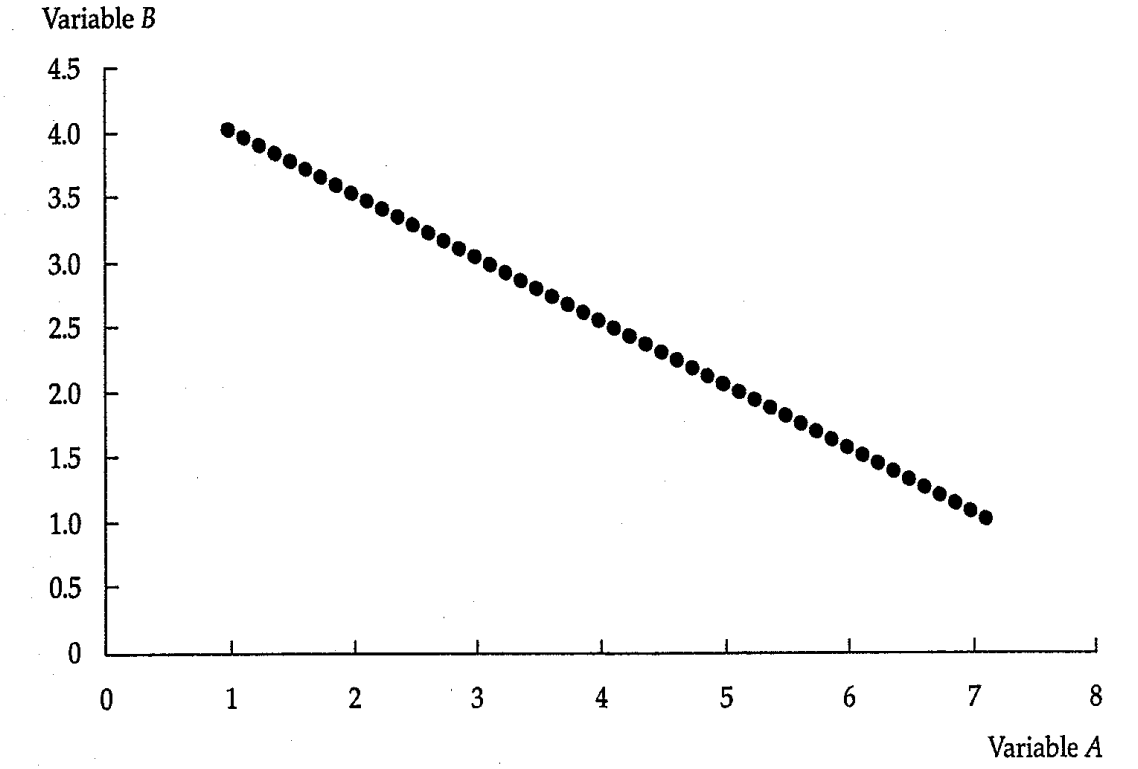
Perfect +1 correlation

FIGURE 8-2 Variables with a Correlation of 1 (positive)



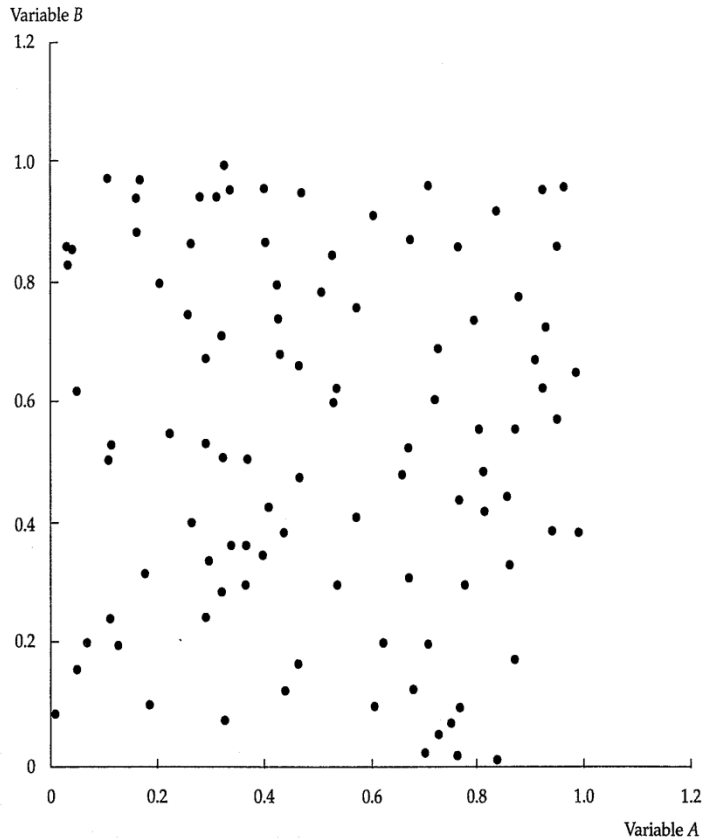
Perfect -1 correlation

FIGURE 8-3 Variables with a Correlation of −1 (negative)



0 correlation

FIGURE 8-4 Variables with a Correlation of 0



Computing corr and cov

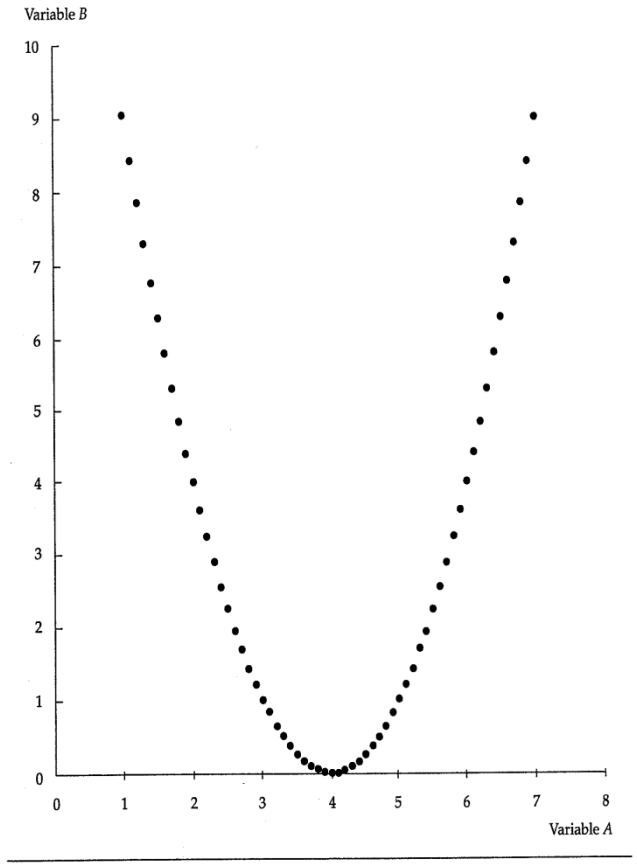
TABLE 8-2 Sample Covariance and Sample Standard Deviations: Annual Money Supply Growth Rate and Inflation Rate by Country, 1970–2001

Country	Money Supply Growth Rate X_i	Inflation Rate Y_i	Cross-Product $(X_i - \bar{X})(Y_i - \bar{Y})$	Squared Deviations $(X_i - \bar{X})^2$	Squared Deviations $(Y_i - \bar{Y})^2$
Australia	0.1166	0.0676	0.000169	0.000534	0.000053
Canada	0.0915	0.0519	0.000017	0.000004	0.000071
New Zealand	0.1060	0.0815	0.000265	0.000156	0.000449
Switzerland	0.0575	0.0339	0.000950	0.001296	0.000697
United Kingdom	0.1258	0.0758	0.000501	0.001043	0.000240
United States	0.0634	0.0509	0.000283	0.000906	0.000088
Sum	0.5608	0.3616	0.002185	0.003939	0.001598
Average	0.0935	0.0603			
Covariance			$= 0.000437$		
Variance				$= 0.000788$	0.000320
Standard deviation				$= 0.028071$	0.017889

Source: International Monetary Fund.
Notes:
1. Divide the cross-product sum by $n - 1$ (with $n = 6$) to obtain the covariance of X and Y.
2. Divide the squared deviations sums by $n - 1$ (with $n = 6$) to obtain the variances of X and Y.

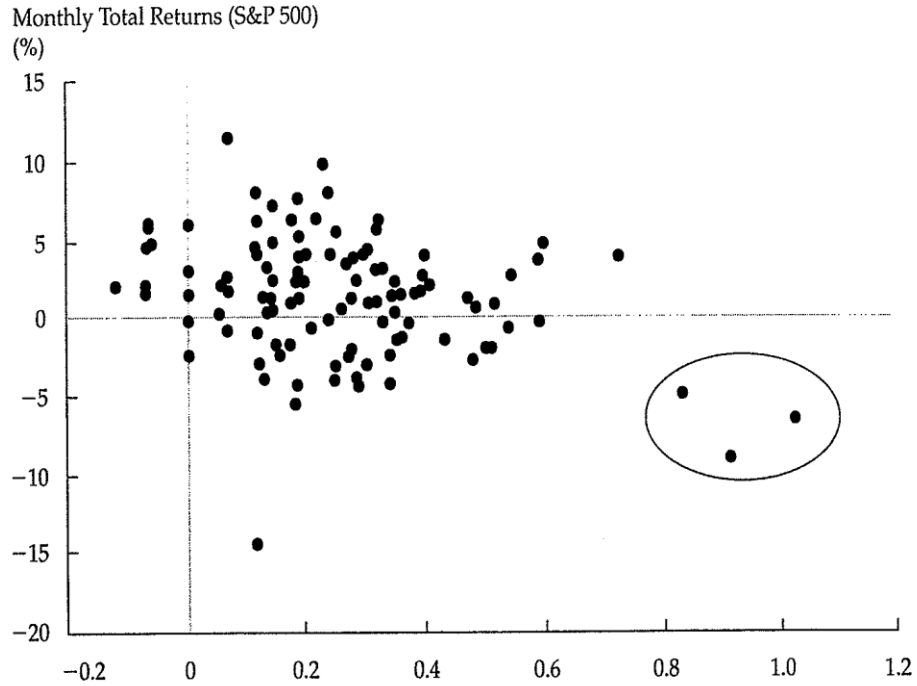
Deterministic functional link still zero correlation

FIGURE 8-5 Variables with a Strong Non-Linear Association



Sensitivity to outliers

FIGURE 8-6 U.S. Inflation and Stock Returns in the 1990s



Source: Ibbotson Associates

Testing the significance of correlation

- Given the empirical correlation coefficient r between two random variables with n samples
- We test the significance of correlation using the following test statistic:
- Test statistic: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
- Which has a Student's t distribution

EXAMPLE 8-7. Testing the Correlation between Money Supply Growth and Inflation.

Earlier in this chapter, we showed that the sample correlation between long-term money supply growth and long-term inflation in six industrialized countries was 0.8702 during the 1970–2001 period. Suppose we want to test the null hypothesis, H_0, that the true correlation in the population is 0 ($\rho = 0$) against the alternative hypothesis, H_a, that the correlation in the population is different from 0 ($\rho \neq 0$).

Recalling that this sample has six observations, we can compute the statistic for testing the null hypothesis as follows:

$$t = \frac{0.8702\sqrt{6-2}}{\sqrt{1-0.8702^2}} = 3.532$$

The value of the test statistic is 3.532. As the table of critical values of the t-distribution for a two-tailed test shows, for a t-distribution with $n - 2 = 6 - 2 = 4$ degrees of freedom at the 0.05 level of significance, we can reject the null hypothesis (that the population correlation is equal to 0) if the value of the test statistic is greater than 2.776 or less than -2.776. The fact that we can reject the null hypothesis of no correlation based on only six observations is quite unusual; it further demonstrates the strong relation between long-term money supply growth and long-term inflation in these six countries.

EXAMPLE 8-9. The Correlation Between Bond Returns and T-Bill Returns.

Table 8-5 showed that the sample correlation between monthly returns to U.S. government bonds and monthly returns to 30-day T-bills was 0.1119 from January 1926 through December 2002. Suppose we want to test whether the correlation coefficient is statistically significantly different from zero. There are 924 months during the period January 1926 to December 2002. Therefore, to test the null hypothesis, H_0 (that the true correlation in the population is 0), against the alternative hypothesis, H_a (that the correlation in the population is different from 0), we use the following test statistic:

$$t = \frac{0.1119\sqrt{924-2}}{\sqrt{1-0.1119^2}} = 3.4193$$

At the 0.05 significance level, the critical value for the test statistic is approximately 1.96. At the 0.01 significance level, the critical value for the test statistic is approximately 2.58. The test statistic is 3.4193, so we can reject the null hypothesis of no correlation in the population at both the 0.05 and 0.01 levels. This example shows that, in large samples, even relatively small correlation coefficients can be significantly different from zero.

is statistically significantly different from zero. There are 924 months during the period January 1926 to December 2002. Therefore, to test the null hypothesis, H_0 (that the true correlation in the population is 0), against the alternative hypothesis, H_a (that the correlation in the population is different from 0), we use the following test statistic:

$$t = \frac{0.1119\sqrt{924-2}}{\sqrt{1-0.1119^2}} = 3.4193$$

At the 0.05 significance level, the critical value for the test statistic is approximately 1.96. At the 0.01 significance level, the critical value for the test statistic is approximately 2.58. The test statistic is 3.4193, so we can reject the null hypothesis of no correlation in the population at both the 0.05 and 0.01 levels. This example shows that, in large samples, even relatively small correlation coefficients can be significantly different from zero.

Correlation of stock prices

- Unstable
- On average, stocks in the S&P 500 are correlated
- Average correlation in the range of 20%, subject to significant fluctuations
- Individual correlations are unstable

Correlation among stock returns

EXAMPLE 8-4. Correlations among Stock Return Series.

Table 8-4 shows the correlation matrix of monthly returns to three U.S. stock indexes during the period January 1971 to December 1999 and in three subperiods (the 1970s, 1980s, and 1990s).[10] The large-cap style is represented by the return to the S&P 500 Index, the small-cap style is represented by the return to the Dimensional Fund Advisors U.S. Small-Stock Index, and the broad-market returns are represented by the return to the Wilshire 5000 Index.

TABLE 8-4 Correlations of Monthly Returns to Various U.S. Stock Indexes

1971–1999	S&P 500	U.S. Small-Stock	Wilshire 5000
S&P 500	1.0000		
U.S. Small-Stock	0.7615	1.0000	
Wilshire 5000	0.9894	0.8298	1.0000
1971–1979	**S&P 500**	**U.S. Small-Stock**	**Wilshire 5000**
S&P 500	1.0000		
U.S. Small-Stock	0.7753	1.0000	
Wilshire 5000	0.9906	0.8375	1.0000
1980–1989	**S&P 500**	**U.S. Small-Stock**	**Wilshire 5000**
S&P 500	1.0000		
U.S. Small-Stock	0.8440	1.0000	
Wilshire 5000	0.9914	0.8951	1.0000
1990–1999	**S&P 500**	**U.S. Small-Stock**	**Wilshire 5000**
S&P 500	1.0000		
U.S. Small-Stock	0.6843	1.0000	
Wilshire 5000	0.9858	0.7768	1.0000

Source: Ibbotson Associates.

Correlation debt and equity returns

EXAMPLE 8-5. Correlations of Debt and Equity Returns.

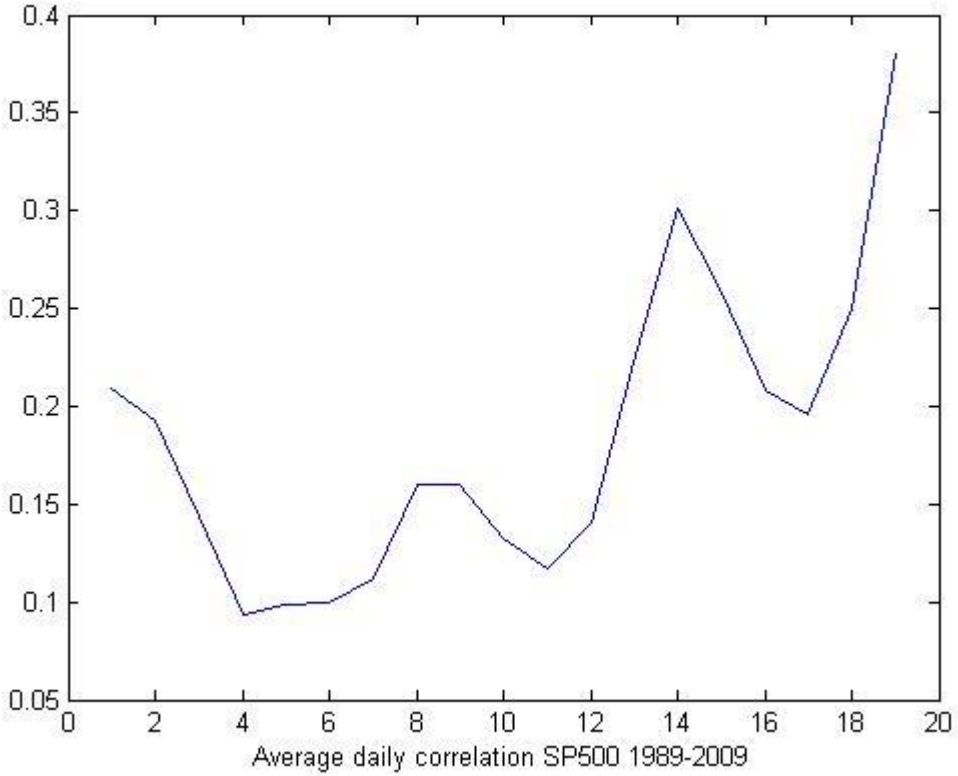
Table 8-5 shows the correlation matrix for various U.S. debt returns and S&P 500 returns using monthly data from January 1926 to December 2002.

TABLE 8-5 Correlations among U.S. Stock and Debt Returns, 1926–2002

All	S&P 500	U.S. Long-Term Corp.	U.S. Long-Term Govt.	U.S. 30-Day T-Bill	High-Yield Corp.
S&P 500	1.0000				
U.S. Long-Term Corp.	0.2143	1.0000			
U.S. Long-Term Govt.	0.1466	0.8480	1.0000		
U.S. 30-Day T-bill	−0.0174	0.0970	0.1119	1.0000	
High-Yield Corp.	0.6471	0.4274	0.3131	0.0174	1.0000

Source: Ibbotson Associates.

Average correlation coefficient SP500 1989-2009 1Y moving window



Instability of the covariance matrix

- The instability of the covariance matrix is well known to practitioners and academics
- "Estimating Covariance Matrices" Robert Litterman and Kurt Winkelmann, January 1998, Goldman Sachs, Risk Management Series
- "Estimation for Markowitz efficient portfolios.", Jobson, J. D. and Korkie, B. (1980), Journal of the American Statistical Association, 75:544

Garbage in, garbage out

- Estimates of the empirical covariance matrix are highly noisy
- Optimization becomes “error maximization” (Michaud)
- Optimization with noisy data produces risky corner portfolios

Random matrices

An advanced, modern presentation of the random nature of the cross-correlation matrix is based on the Theory of Random Matrices (RMT).

“Random matrix approach to cross correlation in financial data”, Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luìs A. Nunes Amaral, Thomas Guhr, and H. Eugene Stanley, *PHYSICAL REVIEW E*, VOLUME 65, 066126

“Collective Origin of the Coexistence of Apparent RMT Noise and Factors in Large Sample Correlation Matrices”, Y. Malevergne, and D. Sornette, Cond-Mat 02/ 0115, 1, no. 4 (October 2002)

L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, “Noise Dressing of Financial Correlation Matrices,” *Physics Review Letter* 83 (1999), pp. 1467–1470

Random matrices

- The theory of random matrices was developed in the 1950s in the domain of quantum physics
- M.L. Mehta, *Random Matrix Theory* (New York: Academic Press, 1995).
- A random cross-correlation matrix can be thought as the cross-correlation matrix of a set of independent random walks
- As such, its entries are a set of zero-mean IID variables
- The mean of the random correlation coefficients is zero as these coefficients have a symmetrical distribution in the range [-1,+1]

Limit distribution of the eigenvalues of random matrices

- Consider *N* sample random walks of length *M*
- Suppose that both the number of sample points *M* and the number *N* of time series tend to infinity
- Suppose that both *M* and *N* tend to infinity with a fixed ratio *Q*

Distribution of eigenvalues

- The density of eigenvalues of the random cross-correlation matrix tends to the following distribution:

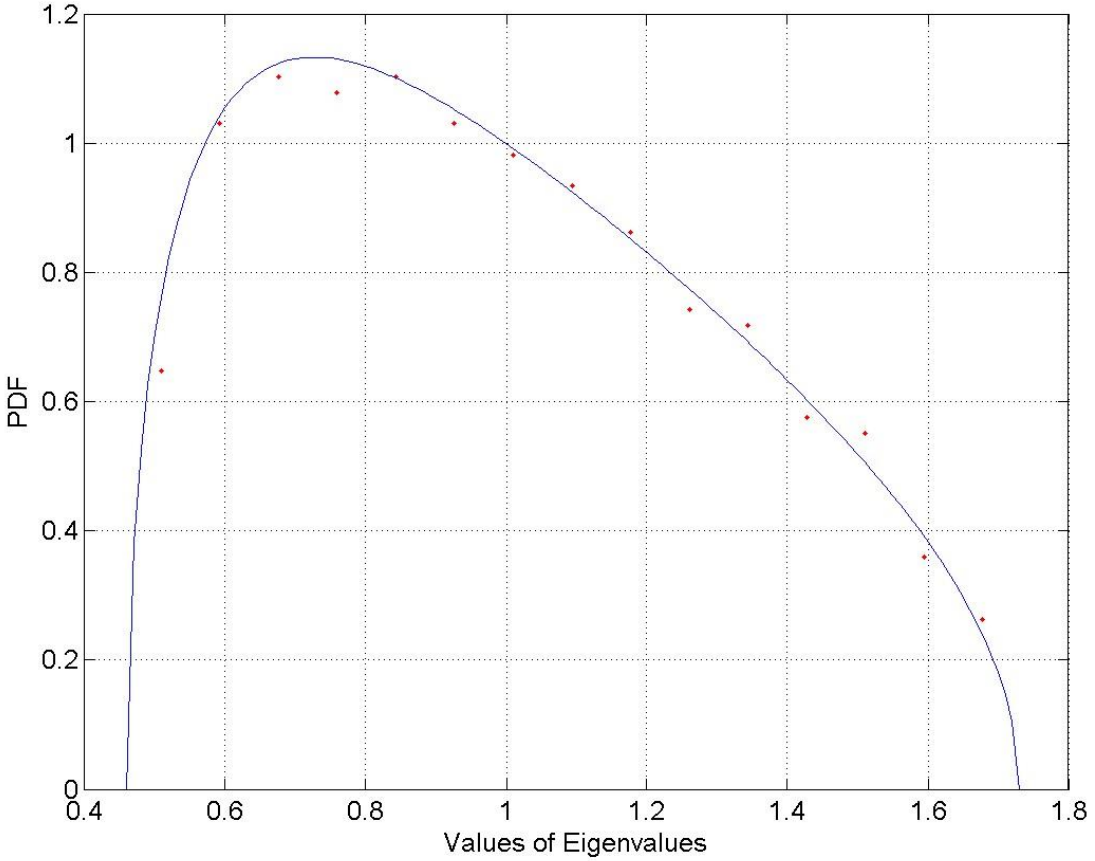
$$\rho(\lambda) = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_{\max} - \lambda)(\lambda_{\min} - \lambda)}}{\lambda}$$

$$N, M \to \infty,\ Q = {}^{N}\!/_{M} \geq 1$$

$$\lambda_{\max,\min} = \sigma^2 \left(1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}\right)$$

- where σ^2 is the average eigenvalue of the matrix (average correlation is zero)

Theoretical distribution of eigenvalues and simulation on 500 generated random walks



Load SP500

```
clear all
close all
clc

load('C:\Documents and Settings\s_focardi\MATLAB_PROGRAMS\CRSP\SP500_30Y_SERIES')
load('C:\Documents and Settings\s_focardi\MATLAB_PROGRAMS\CRSP\SP500_30Y_FLAGS')
load('C:\Documents and Settings\s_focardi\MATLAB_PROGRAMS\CRSP\SP500_30Y_INDEX_EW')
load('C:\Documents and Settings\s_focardi\MATLAB_PROGRAMS\CRSP\SP500_30Y_INDEX_VW')

DATES=PRICES(2:end,1);
SRN=PRICES(1,2:end);
PRICES=PRICES(2:end,2:end);
FLAGS=FLAGS(2:end,2:end);

[T N]=size(PRICES)

[I J] =find(FLAGS==0);

for k=1:length(I);
   PRICES(I(k),J(k))=NaN;
end
clear I
clear J

[I J] =find((PRICES)<=0);

for k=1:length(I);
   PRICES(I(k),J(k))=NaN;
end
RET=price2ret(PRICES);

[I J] =find(abs(RET)>=1);

for k=1:length(I);
   PRICES(I(k),J(k))=NaN;
end
clear I
clear J
```

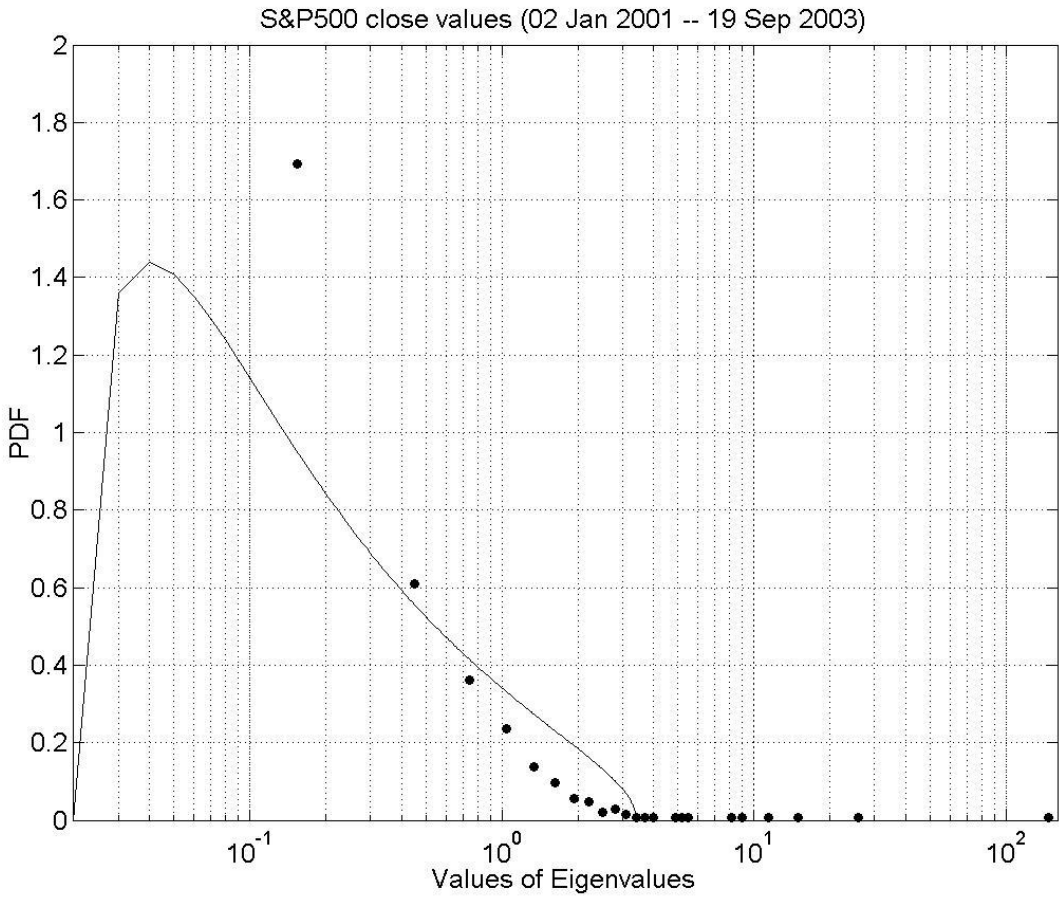
Compute average correlation

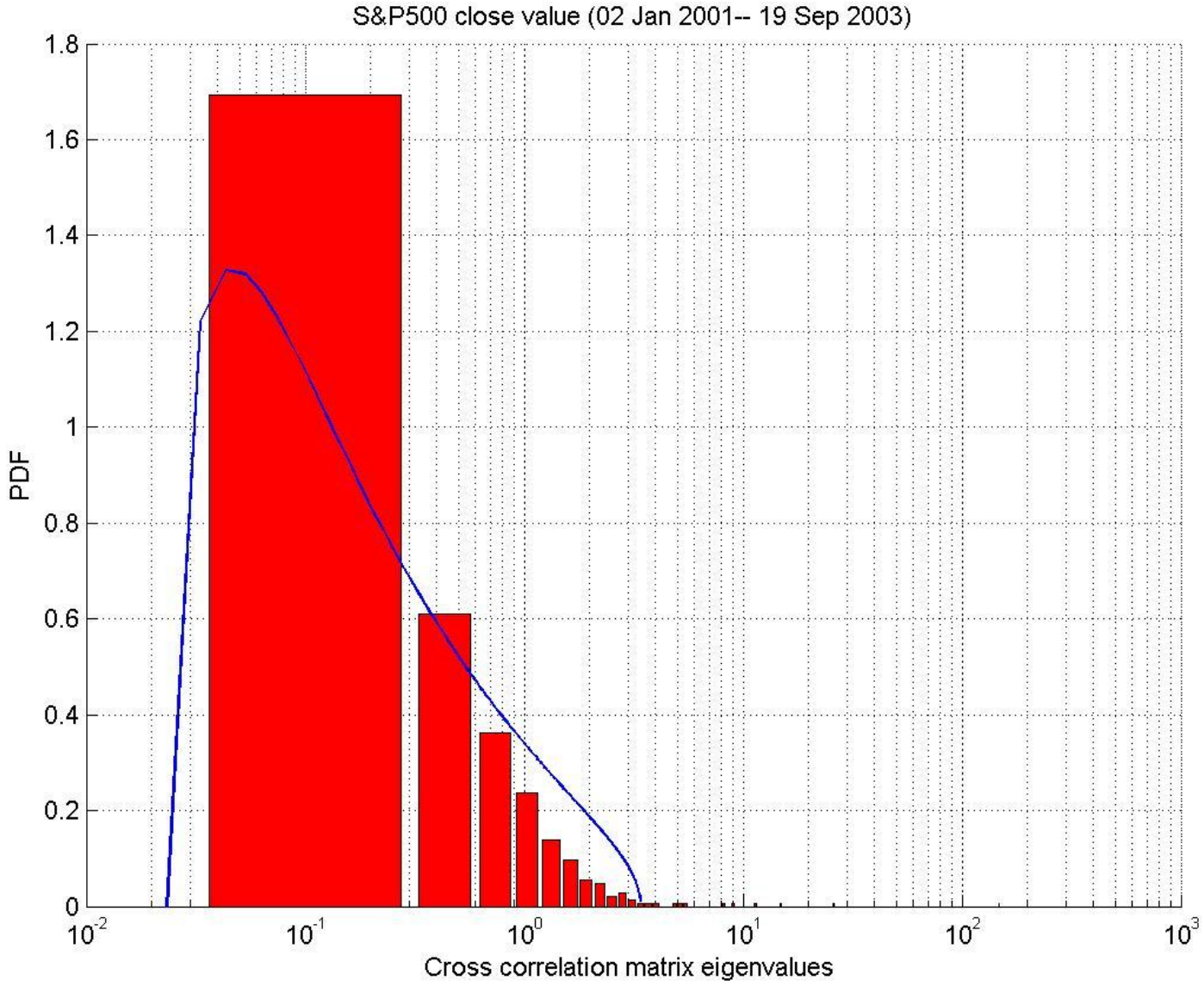
```
in_window=500;
out_window=250;
Parameters.in_window=in_window;
Parameters.out_window=out_window;
Parameters.t0 = in_window;

Pace=out_window
t0=Parameters.t0;

for t=t0:Pace:T-out_window

    d=(t-t0)/Pace+1;
    Parameters.t=t;
    Parameters.d=d;

    [IDX,LPrices,DPrices,LRet, DRet]=DFA_CreateLocalPricesReturns(PRICES,Parameters);

    DPrices=log(DPrices);
    LPrices=log(LPrices);

    [T N]=size(DPrices);

    X=[1:T];

    CC=corrcoef(DRet);
    VCC=find(vech(CC)<0.99);
    CC=CC(VCC);
    MCC(d)=mean(vech(CC));
end

   figure
   plot(MCC)
```

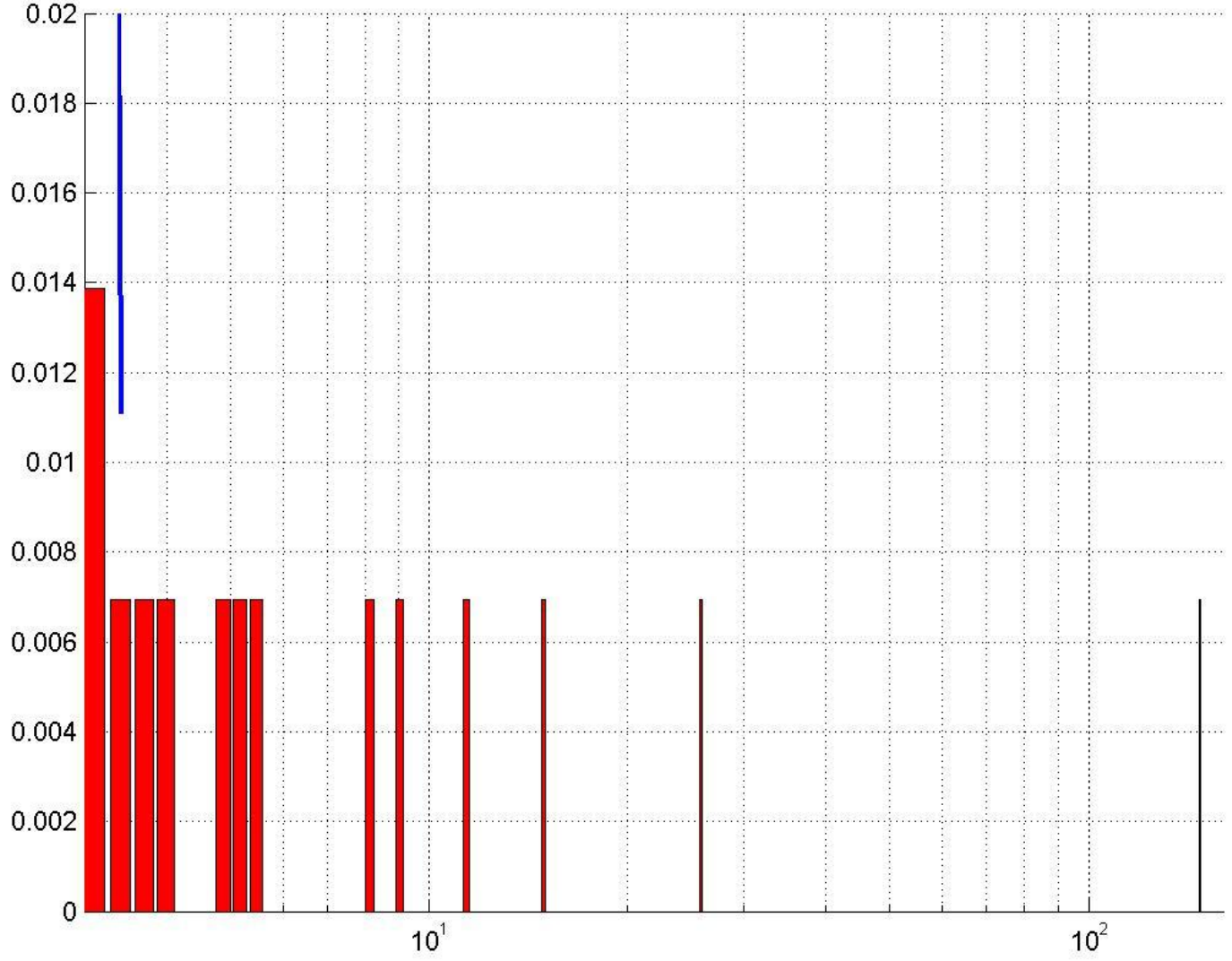
Compute M-P law

```
beta=L/T
  clear effe
  clear s
  clear effe2
  clear s2
  a=(1-sqrt(beta))^2
  b=(1+sqrt(beta))^2
  s=zeros(1,500);
  s2=zeros(1,500);
  for ss=1:500
    s(ss)=ss/100;


    if s<=a
      effe(ss)=0;
    elseif s<b
       effe(ss)=sqrt((s(ss)-a)*(b-s(ss)))/(2*pi*s(ss)*beta);

    else
       effe(ss)=0;
    end
  end


  figure
  plot(s,effe)
  title('Plot of Marcenko-Pastur law for LxL correlation matrix')
  xlabel('beta=L/T')

  E=flipud(eig(corrcoef(X)))
  LL=(0:0.1:5)
  n=hist(E,LL)
  figure
  plot(LL,n/L)
  title('Plot of empirical distribution for LxL correlation matrix')
  xlabel('beta=L/T')
```

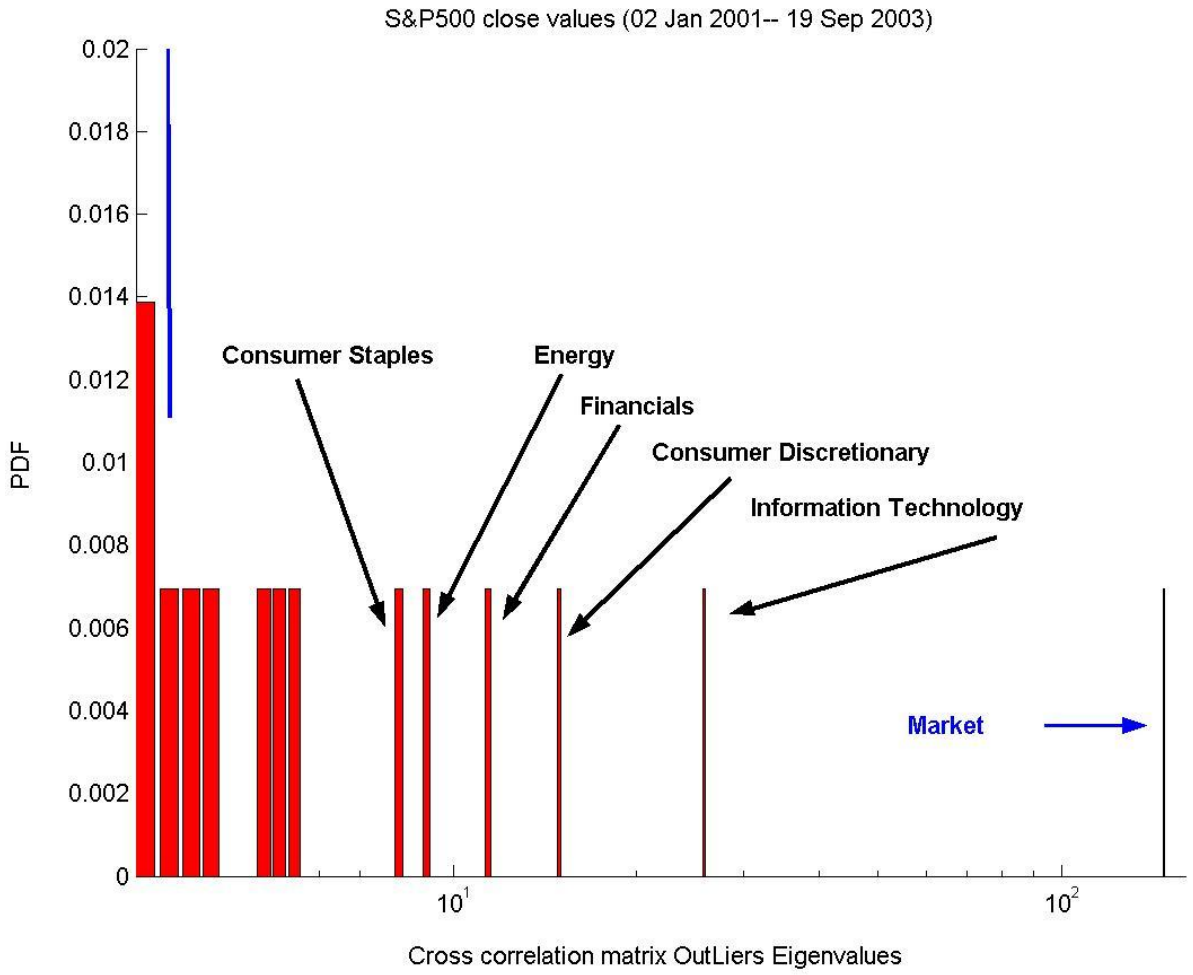
PDF Eigenvalues




S&P500 close value (02 Jan 2001-- 19 Sep 2003)
PDF
Cross correlation matrix eigenvalues
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
10^-2
10^-1
10^0
10^1
10^2
10^3


0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
10^1
10^2

Sectors


S&P500 close values (02 Jan 2001-- 19 Sep 2003)
PDF
0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
Consumer Staples
Energy
Financials
Consumer Discretionary
Information Technology
Market
10^1
10^2
Cross correlation matrix OutLiers Eigenvalues

487	Information Technology	"Xilinx, Inc" National Semiconductor KLA-Tencor Corp. Cisco Systems QLogic Corp.	Altera Corp. PMC-Sierra Inc. Linear Technology Corp. Applied Micro Circuits Broadcom Corporation
486	Consumer Discretionary	Reebok International Nordstrom "Liz Claiborne, Inc." V.F. Corp. Federated Dept. Stores	Leggett & Platt May Dept. Stores Kohl's Corp. Target Corp.
485	Financials	North Fork Bancorporation Wells Fargo Federal Home Loan Mtg. Golden West Financial Fannie Mae	Marsh & McLennan Progressive Corp. SLM Corporation Washington Mutual
484	Energy	EOG Resources BJ Services Rowan Cos. Anadarko Petroleum	Burlington Resources Nabors Industries Ltd. Noble Corporation Apache Corp.
483	Consumer Staples	"ConAgra Foods, Inc.“ Clorox Co. Colgate-Palmolive Procter & Gamble Wrigley (Wm) Jr.	Gillette Co. Heinz (H.J.) Campbell Soup Avon Products

Consequences of the random nature of correlation matrices

- The entries of the var-cov or correlation matrix are nearly random
- Information is concentrated in some aggregates
- Aggregates are the eigenvectors corresponding to the largest eigenvalues

Dimensionality Reduction

- We need to reduce the dimensionality of the problem
- i.e., we need to reduce the number of correlation parameters to estimate
- Otherwise we fit the empirical covariance matrix to noise

Linear Regression I

Concept of regression

- Regression represents a functional relationship between a dependent variable and one or more independent variables plus noise
- The regression function is the locus of the conditional expectation of the dependent variable given the independent variables
- Properties of regression critically depend on noise (residuals)

Linear Regression with One Independent Variable

- Linear regression quantifies the strength of the linear relationship between two variables…
- Tests hypotheses about the relation between two variables…
- Uses one variable to predict another variable.
- It assumes a linear relationship between the dependent and the independent variable

$$Y = b_0 + b_1 X + \varepsilon$$

- Y is the dependent variable and X the independent variable.
- b0, the intercept, and b1, the slope coefficient, are called the regression coefficients.
- The error term represents the portion of the dependent variable that cannot be explained by the independent variable.
- Cross-sectional and time series regressions.

Regression model definition

- Simple regression: $Y = b_0 + b_1 X + \varepsilon$
- Design matrix:

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_1 \end{bmatrix}, b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- Regression in matrix form:

$$Y = Xb + \varepsilon$$

Interpretation of the Error Terms

- For a same Xi we can have a different Yi. The relationship is an approximate one.
- Even though we could use a lot of variables (not recommended based on a principle of model parsimony) a model is always an approximation.
- Sometimes (survey data) the error term can capture some measurement error.
- It could also be an approximation error introduced by the choice of a linear form while the true relationship is nonlinear.
- Contrary to the Yi these errors are unobservable. Once we estimate the parameters we will have estimated values for these residual errors.

Estimation of the Linear Regression Model

The estimation computes a line that best fits the observations by choosing values for the b0 and b1 coefficients in order to minimize the sum of squared vertical distances between the observations and the regression lin

Estimation simple regression (one independent variable)

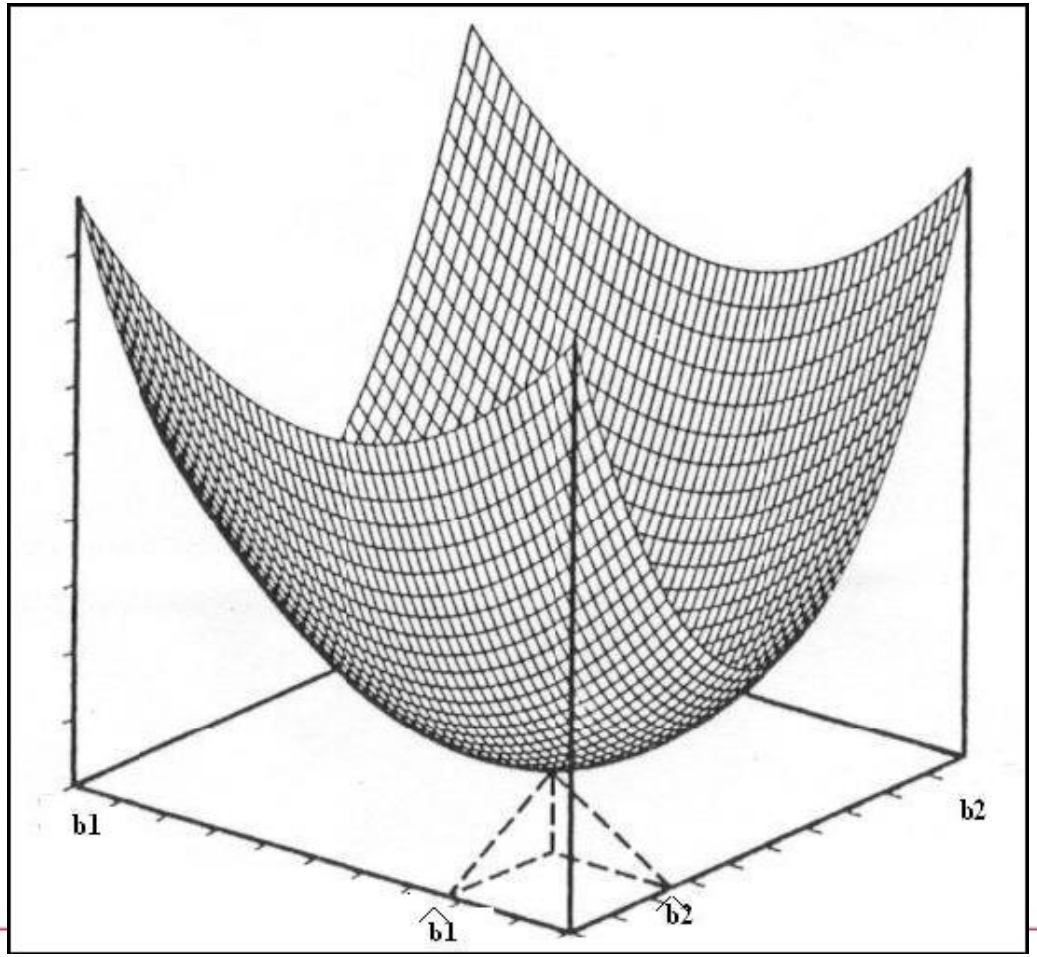
- Method of the Least Squares (LS)

- Given a sample $Y_i, X_i, i = 1,2,\ldots,n$

- Minimize the sum of the squares of the differences $(Y_i - b_0 - b_1 X_i)^2$

$$\min_{b_0,b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 = \min_{b_0,b_1} \sum_{i=1}^{n} \varepsilon_i^{\,2}$$

Estimation: minimize a function of b0 and b1



Estimators of the regression coefficients

- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ sample means
- $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2, s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$ sample variances
- $s_{XY} = \mathrm{cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$ sample covariance
- $\hat{b}_0 = \bar{Y} - \bar{X}\hat{b}_1$
- $\hat{b}_1 = \frac{s_{XY}}{s_X^2}$

estimators of the regression coefficients

Estimation regression

- Least square estimators: minimize the sum of squared errors

$$\min_b\left(\varepsilon'\varepsilon\right)=\min_b\left(\left(Y-Xb\right)'\left(Y-Xb\right)\right)$$

- Applying First Order Conditions (partial derivatives equal to 0) we obtain

$$\frac{\partial\left(\left(Y-Xb\right)'\left(Y-Xb\right)\right)}{\partial b}=X'\left(Y-Xb\right)$$

$$X'Y=X'Xb$$

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

If only one variable:

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & n\bar{X} \\ n\bar{X} & \sum_{i=1}^{n} X_i^2 \end{bmatrix},$$

$$\begin{bmatrix} n & n\bar{X} \\ n\bar{X} & \sum_{i=1}^{n} X_i^2 \end{bmatrix}^{-1} = \frac{1}{n\sum_{i=1}^{n} X_i^2 - n^2\bar{X}^2} \begin{bmatrix} \sum_{i=1}^{n} X_i^2 & -n\bar{X} \\ -n\bar{X} & n \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \bar{Y} \\ \sum_{i=1}^{n} X_i Y_i \end{bmatrix}$$

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \frac{1}{n\sum_{i=1}^{n} X_i^2 - n^2\bar{X}^2} \begin{bmatrix} \sum_{i=1}^{n} X_i^2 & -n\bar{X} \\ -n\bar{X} & n \end{bmatrix} \begin{bmatrix} \bar{Y} \\ \sum_{i=1}^{n} X_i Y_i \end{bmatrix} = \frac{1}{n\sum_{i=1}^{n} X_i^2 - n^2\bar{X}^2} \begin{bmatrix} \sum_{i=1}^{n} X_i^2 \bar{Y} - n\bar{X}\sum_{i=1}^{n} X_i Y_i \\ -n^2\bar{X}\bar{Y} + n\sum_{i=1}^{n} X_i Y_i \end{bmatrix}$$

Remarks

- The regression line goes through the center of gravity
- The slope coefficient is slightly different from the coefficient of correlation and not bounded

$$\hat{b}_1 = \rho \frac{s_Y}{s_X}$$

- The estimators b0 and b1 are random variables.

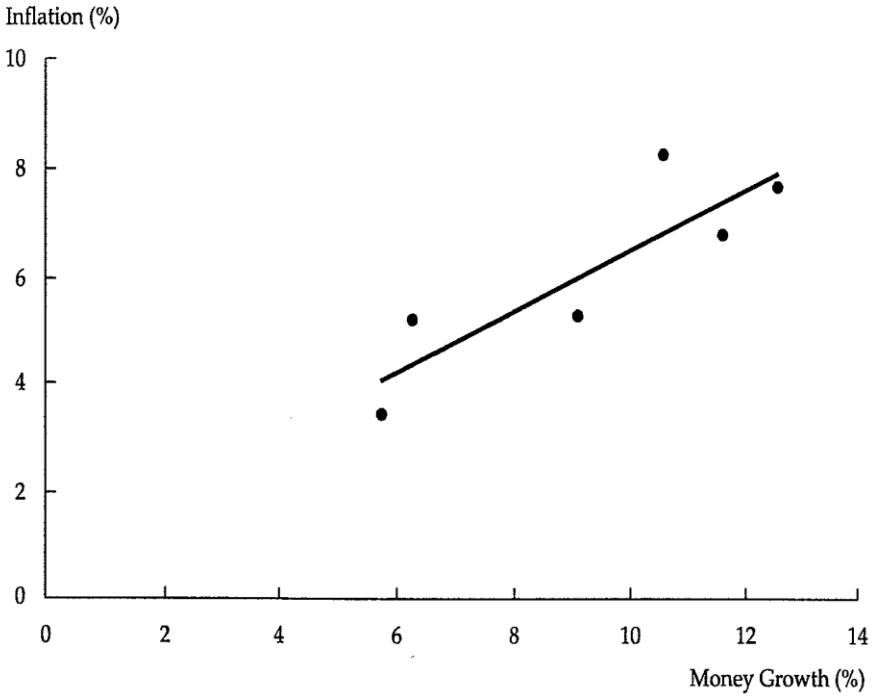
Assumptions of the Linear Regression Model

- The relationship between the dependent variable Y and the independent variable Y is linear in the parameters b0 and b1. It means that the coefficients cannot be raised at a power or multiplied or divided by another coefficient.
- It does not prevent to raise the variable X to be raised at a power.
- The independent variable, X, is not random.
- The expected value of the error term is equal to 0.
- The variance of the error term is the same for all observations
- The error terms are uncorrelated across observations. Consequently, E(eiej) = 0 for all i not equal to j.
- The error term, e, is normally distributed.

Discussion of the assumptions

- Assumption 1 is crucial for the validity of the linear regression model.
- Assumptions 2 and 3 ensure that the estimators are unbiased.
- Assumptions 4, 5 and 6 allow us to determine the distribution of the estimators bˆ0 and bˆ1 of the coefficients b1 and b2 and test whether these coefficients have a particular value.
- Assumption 4 states that the variance of the error term is the same for all observations (homoscedasticity assumption). It will be released later.
- Assumption 5 states that the error terms are uncorrelated across observations. It is necessary for correctly estimating the variances of the estimators bˆ0 and bˆ1 of the coefficients b1 and b2. It will be released later.
- Assumption 6 states that the error term is normally distributed and allows to test easily a particular hypothesis about a linear regression model.

FIGURE 8-8 Fitted Regression Line Explaining the Inflation Rate Using Growth in the Money Supply by Country: 1970–2001



Source: International Monetary Fund

TABLE 8-2 (excerpted)

	Money Supply Growth Rate X_i	Inflation Rate Y_i	Cross-Product $(X_i - \bar{X})(Y_i - \bar{Y})$	Squared Deviations $(X_i - \bar{X})^2$	Squared Deviations $(Y_i - \bar{Y})^2$
Sum	0.5608	0.3616	0.002185	0.003939	0.001598
Average	0.0935	0.0603			
Covariance			**0.000437**		
Variance				**0.000788**	0.000320
Standard deviation				0.028071	0.017889

Usefulness of the simple regression : Standard Error of Estimates

- Does not tell if statistically significant
- Standard Error of Estimate SEE

$$SEE = \left(\frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \hat{b}_0 - \hat{b}_1 X_i \right)^2 \right)^{\frac{1}{2}} = \left(\frac{1}{n-2} \sum_{i=1}^{n} \varepsilon_i^2 \right)^{\frac{1}{2}}$$

- Requires: estimates of the regression coefficients plus observations
- Similar to std of residuals but uses n-2 instead of n-1

EXAMPLE 8-12. Computing the Standard Error of Estimate.

Recall that the estimated regression equation for the inflation and money supply growth data shown in Figure 8-8 was $Y_i = 0.0084 + 0.5545X_i$. Table 8-7 uses this estimated equation to compute the data needed for the standard error of estimate.

TABLE 8-7 Computing the Standard Error of Estimate

Country	Money Supply Growth Rate X_i	Inflation Rate Y_i	Predicted Inflation Rate $\hat{Y}_i$	Regression Residual $Y_i - \hat{Y}_i$	Squared Residual $(Y_i - \hat{Y}_i)^2$
			0.0084 + 0.5545(0.1166)		
Australia	0.1166	0.0676	0.0731	−0.0055	0.000030
Canada	0.0915	0.0519	0.0591	−0.0072	0.000052
New Zealand	0.1060	0.0815	0.0672	0.0143	0.000204
Switzerland	0.0575	0.0339	0.0403	−0.0064	0.000041
United Kingdom	0.1258	0.0758	0.0782	−0.0024	0.000006
United States	0.0634	0.0509	0.0436	0.0073	0.000053
Sum					0.000386

Source: International Monetary Fund.

$$\hat{\varepsilon}_i > \left(\frac{0.000386}{6-2}\right)^{1/2} = 0.009823$$

Usefulness of the simple regression : Coefficient of Determination

- In the simple regression the Coefficient of Determination can be determined as the square of the correlation coefficient between the independent and the dependent variable
- Alternative interpretation (needed in multiple regression)
- If no regression is known the best point estimate of the variable Y is its mean $\overline{Y}$

Coefficient of Determination

- A measure of forecast accuracy is the sample variance of Y
- Call Total variation the sum of squared deviations

$$\text{Total Variation}(Y) = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$$

- To compute we need the samples and the sample mean

Coefficient of Determination

- If we know regression the forecast of Y conditional on X offers better estimate of Y:
- The unexplained variation is the sum of the squared residuals

$$\sum_{i=1}^{n}\left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2 = \sum_{i=1}^{n}\varepsilon_i^2$$

- To compute we need the samples plus the regression coefficients

Coefficient of Determination

- The explained variation is the total variation minus the unexplained variation

- Total variation=
- Unexplained variation+Explained variation

- The Coefficient of Determination is the ratio of the explained variation to total variation

Coefficient of Determination

- Coefficient of determination=
- Explained variation/Total variation=

- 1-Unexplained variation/Total variation

- To compute we need samples, estimated regression coefficients, residuals

EXAMPLE 8-13. Inflation Rate and Growth in the Money Supply.

Using the data in Table 8-7, we can see that the unexplained variation from the regression, which is the sum of the squared residuals, equals 0.000386. Table 8-8 shows the computation of total variation in the dependent variable, the long-term rate of inflation.

TABLE 8-8 Computing Total Variation

Country	Money Supply Growth Rate X_i	Inflation Rate Y_i	Deviation from Mean $Y_i - \bar{Y}$	Squared Deviation $(Y_i - \bar{Y})^2$
Australia	0.1166	0.0676	0.0073	0.000053
Canada	0.0915	0.0519	−0.0084	0.000071
New Zealand	0.1060	0.0815	0.0212	0.000449
Switzerland	0.0575	0.0339	−0.0264	0.000697
United Kingdom	0.1258	0.0758	0.0155	0.000240
United States	0.0634	0.0509	−0.0094	0.000088
	Average:	0.0603	Sum:	0.001598

Source: International Monetary Fund.

The average inflation rate for this period is 6.03 percent. The next-to-last column shows the amount each country's long-term inflation rate deviates from that average; the last column shows the square of that deviation. The sum of those squared deviations is the total variation in Y for the sample (0.001598), shown in Table 8-8.

The coefficient of determination for the regression is

$$\frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}} = \frac{0.001598 - 0.000386}{0.001598} = 0.7584$$

Note that this method gives the same result rounded to two decimal places, 0.76, that we obtained earlier (the difference at greater decimal places results from rounding). We will use this method again in the chapter on multiple regression; when we have more than one independent variable, this method is the only way to compute the coefficient of determination.

Regression diagnostic

- Assume the regression model is correct
- Regression coefficients and estimates are random variables that depend on the sample
- Compute tests and confidence intervals
- But do not tell if the regression model is well specified

Estimators of the regression coefficients as random variables

- Unbiased estimators: the expectation of the estimators coincide with their true values
- Theoretical concept cannot be checked in practice, the expectation cannot be computed
- But useful for inference

Estimators of the regression coefficients as random variables

- Variance of the estimators of the regression coefficients:

$$\text{var}(b_1) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad \text{var}(b_0) = \frac{\sigma_\varepsilon^2 \sum_{i=1}^{n}(X_i)^2}{n\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\text{cov}(b_0, b_1) = -\frac{\sigma_\varepsilon^2 \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- To compute we need samples, sample mean and an estimate of the variance of residuals

Variance of the residuals

- An unbiased estimator of the variance of the residuals is the following:

$$\hat{\sigma}_{\varepsilon}^{2} = \frac{1}{n-2}\sum_{i=1}^{n}(\varepsilon_i)^2$$

- To compute we need the residuals, hence the full regression model

```
clear all;
close all;
clc;
X=2*randn(1000,1);
for i=1:1000
%    X=2*randn(1000,1);
   E=0.2*randn(1000,1);
   Y=1+0.6*X+E;
   X2=[ones(1000,1) X];
   [B,BINT,R,RINT,STATS] =regress(Y,X2);
%    STATS=regstats(W,V,'linear')
   beta(:,i)=B;
%    STATS.beta
   bci(:,:,i)=BINT;
   e2=R'*R/998;
   e2=0.04
k(:,:,i)=e2*(X2'*X2)^(-1);
% STATS.covb
   clear V
   clear W
end
  mean(k,3)
  cov(beta')
```

Covariance matrix coefficients

ans =

1.0e-004 *

0.4004 0.0001
0.0001 0.1003

ans =

1.0e-004 *

0.3814 0.0055
0.0055 0.1069

Confidence intervals for the regression coefficients

- Under the assumption of normality, the confidence interval for the regression coefficients is determined as follows
- We determine the critical value of the distribution corresponding to the level we desire and we construct the confidence interval as:

$$\hat{b}_i \pm t_C s_{\hat{b}_i}$$

- The degree of freedom is the number of samples minus the number of parameters to estimate

Testing hypothesis about the regression coefficients

- Suppose we want to test the hypothesis that a regression coefficient assume a value b_i
- We use the test statistic : $\dfrac{\hat{b}_i - b_i}{s_{\hat{b}_i}}$
- Which follows a *t*-distribution

Suppose we regress a stock's returns on a stock market index's returns and find that the slope coefficient ($\hat{b}_1$) is 1.5 with a standard error ($s_{\hat{b}_1}$) of 0.200. Assume we used 62 monthly observations in our regression analysis. The hypothesized value of the parameter (b_1) is 1.0, the market average slope coefficient. The estimated and the population slope coefficients are often called beta, because the population coefficient is often represented by the Greek symbol beta (β) rather than the b_1 we use in this text. Our null hypothesis is that $b_1 = 1.0$ and $\hat{b}_1$ is the estimate for b_1. We will use a 95 percent confidence interval for our test, or we could say that the test has a significance level of 0.05.

Our confidence interval will span the range $\hat{b}_1 - t_c s_{\hat{b}_1}$ to $\hat{b}_1 + t_c s_{\hat{b}_1}$, or

$$\hat{b}_1 \pm t_c s_{\hat{b}_1} \qquad \textbf{(8-9)}$$

where t_c is the critical t value.[31] The critical value for the test depends on the number of degrees of freedom for the t-distribution under the null hypothesis. The number of degrees

of freedom equals the number of observations minus the number of parameters estimated. In a regression with one independent variable, there are two estimated parameters, the intercept term and the coefficient on the independent variable. For 62 observations and two parameters estimated in this example, we have 60 degrees of freedom (62 − 2). For 60 degrees of freedom, the table of critical values in the back of the book shows that the critical t-value at the 0.05 significance level is 2.00. Substituting the values from our example into Equation 8-9 gives us the interval

$$\begin{aligned}\hat{b}_1 \pm t_c s_{\hat{b}_1} &= 1.5 \pm 2.00(0.200) \\ &= 1.5 \pm 0.400 \\ &= 1.10 \text{ to } 1.90\end{aligned}$$

Under the null hypothesis, the probability that the confidence interval includes b_1 is 95 percent. Because we are testing $b_1 = 1.0$ and because our confidence interval does not include 1.0, we can reject the null hypothesis. Therefore, we can be 95 percent confident that the stock's beta is different from 1.0.

In practice, the most common way to test a hypothesis using a regression model is with a t-test of significance. To test the hypothesis, we can compute the statistic

$$t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \tag{8-10}$$

This test statistic has a t-distribution with $n - 2$ degrees of freedom because two parameters were estimated in the regression. We compare the absolute value of the t-statistic to t_c. If the absolute value of t is greater than t_c, then we can reject the null hypothesis. Substituting the values from the above example into this relationship gives the t-statistic associated with the probability that the stock's beta equals 1.0 ($b_1 = 1.0$).

$$\begin{aligned} t &= \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \\ &= (1.5 - 1.0)/0.200 \\ &= 2.50 \end{aligned}$$

Because $t > t_c$, we reject the null hypothesis that $b_1 = 1.0$.

The t-statistic in the example above is 2.50, and at the 0.05 significance level, $t_c = 2.00$; thus we reject the null hypothesis because $t > t_c$. This statement is equivalent to saying that we are 95 percent confident that the interval for the slope coefficient does not contain the value 1.0. If we were performing this test at the 0.01 level, however, t_c would be 2.66 and we would not reject the hypothesis because t would not be greater than t_c at this significance level. A 99 percent confidence interval for the slope coefficient does contain the value 1.0.

EXAMPLE 8-14. Estimating Beta for General Motors Stock.

You are an investor in General Motors stock and want an estimate of its beta. As in the text example, you hypothesize that GM has an average level of market risk and that its required return in excess of the risk-free rate is the same as the market's required excess return. One regression that summarizes these statements is

CAPM

$$(R - R_F) = \alpha + \beta\,(R_M - R_F) + \varepsilon \qquad \textbf{(8-11)}$$

where R_F is the periodic risk-free rate of return (known at the beginning of the period), R_M is the periodic return on the market, R is the periodic return to the stock of the company, and β is the covariance of stock and market return divided by the variance of the market return, $\text{Cov}(R, R_M)/\sigma_M^2$. Estimating this equation with linear regression provides an estimate of β, $\hat{\beta}$, which tells us the size of the required return premium for the security, given expectations about market returns.[33]

Suppose we want to test the null hypothesis, H_0, that $\beta = 1$ for GM stock to see whether GM stock has the same required return premium as the market as a whole. We need data on returns to GM stock, a risk-free interest rate, and the returns to the market index. For this example, we use data from January 1998 through December 2002

($n = 60$). The return to GM stock is R. The monthly return to 30-day Treasury bills is R_F. The return to the S&P 500 is R_M.[34] We are estimating two parameters, so the number of degrees of freedom is $n - 2 = 60 - 2 = 58$. Table 8-9 shows the results from the regression $(R - R_F) = \alpha + \beta(R_M - R_F) + \varepsilon$.

TABLE 8-9 Estimating Beta for GM Stock

Regression Statistics	
Multiple R	0.5549
R-squared	0.3079
Standard error of estimate	0.0985
Observations	60

	Coefficients	Standard Error	t-Statistic
Alpha	0.0036	0.0127	0.2840
Beta	1.1958 ($\hat{\beta}$)	0.2354	5.0795

We are testing the null hypothesis, H_0, that β for GM equals 1 ($\beta = 1$) against the alternative hypothesis that β does not equal 1 ($\beta \neq 1$). The estimated $\hat{\beta}$ from the regression is 1.1958. The estimated standard error for that coefficient in the regression, $s_{\hat{\beta}}$ is 0.2354. The regression equation has 58 degrees of freedom (60 − 2), so the critical value for the test statistic is approximately $t_c = 2.00$ at the 0.05 significance level. Therefore, the 95 percent confidence interval for the data for any particular hypothesized value of β is shown by the range

$$\hat{\beta} \pm t_c s_{\hat{\beta}}$$
$$1.1958 \pm 2.00(0.2354)$$
$$0.7250 \text{ to } 1.6666$$

In this case, the hypothesized parameter value is $\beta = 1$, and the value 1 falls inside this confidence interval, so we cannot reject the hypothesis at the 0.05 significance level. This means that we cannot reject the hypothesis that GM stock has the same systematic risk as the market as a whole.

Another way of looking at this issue is to compute the *t*-statistic for the GM beta hypothesized parameter using Equation 8-10:

$$t = \frac{\hat{\beta} - \beta}{s_{\hat{\beta}}} = \frac{1.1958 - 1.0}{0.2354} = 0.8318$$

This *t*-statistic is less than the critical *t*-value of 2.00. Therefore, neither approach allows us to reject the null hypothesis. Note that the *t*-statistic associated with $\hat{\beta}$ in the regression results in Table 8-9 is 5.0795. Given the significance level we are using, we cannot reject the null hypothesis that $\beta = 1$, but we can reject the hypothesis that $\beta = 0$.[35]

EXAMPLE 8-15. Explaining Company Value Based on Returns to Invested Capital.

Some financial analysts have argued that one good way to measure a company's ability to create wealth is to compare the company's return on invested capital (ROIC) to its weighted-average cost of capital (WACC). If a company has an ROIC greater than its cost of capital, the company is creating wealth; if its ROIC is less than its cost of capital, it is destroying wealth.[36]

Enterprise value (EV) is a market-price-based measure of company value defined as the market value of equity and debt minus the value of cash and investments. Invested capital (IC) is an accounting measure of company value defined as the sum of the book values of equity and debt. Higher ratios of EV to IC should reflect greater success at wealth creation in general. Mauboussin (1996) argued that the spread between ROIC and WACC helps explains the ratio of EV to IC. Using data on companies in the food-processing industry, we can test the relationship between EV/IC and (ROIC–WACC) using the regression model given in Equation 8-12.

$$EV_i/IC_i = b_0 + b_1(ROIC_i - WACC_i) + \varepsilon_i \quad \textbf{(8-12)}$$

where the subscript i is an index to identify the company. Our null hypothesis is H_0: $b_1 \leq 0$, and we specify a significance level of 0.05. If we reject the null hypothesis, we have evidence of a statistically significant relationship between EV/IC and (ROIC–WACC). We estimate Equation 8-12 using data from nine food-processing companies for 2001.[37] The results of this regression are displayed in Table 8-10 and Figure 8-11.

TABLE 8-10 Explaining Enterprise Value/Invested Capital by the ROIC–WACC Spread

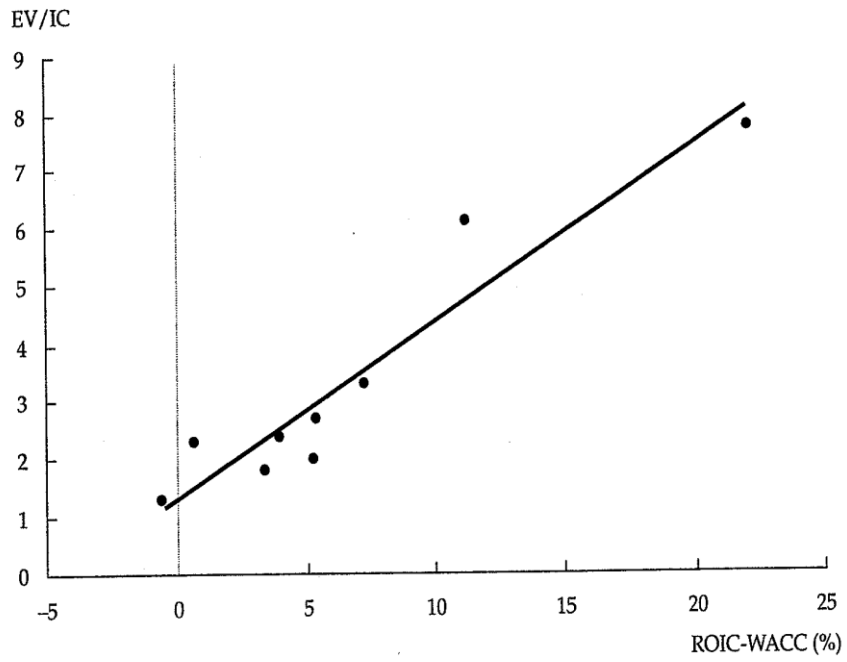
Regression Statistics

Multiple *R*	0.9469
R-squared	0.8966
Standard error of estimate	0.7422
Observations	9

	Coefficients	Standard Error	*t*-Statistic
Intercept	1.3478	0.3511	3.8391
Spread	30.0169	3.8519	7.7928

Source: Nelson (2003).

FIGURE 8-11 Fitted Regression Line Explaining Enterprise Value/ Invested Capital Using ROIC–WACC Spread for the Food Industry



Source: CSFB Food Investors Handbook 2003

ANalysis OF Variance
ANOVA

- ANOVA is a statistical procedure to apportion variability to different sources
- In regression analysis we use ANOVA to determine the usefulness of individual variables
- An important test conducted in ANOVA is the F test
- Which tests the null that all coefficients are zero

To correctly determine the test statistic for the null hypothesis that the slope coefficient equals 0, we need to know the following:

- the total number of observations (n);
- the total number of parameters to be estimated (in a one-independent-variable regression, this number is two: the intercept and the slope coefficient);
- the sum of squared errors or residuals, $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$, abbreviated SSE. This value is also known as the residual sum of squares; and
- the regression sum of squares, $\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$, abbreviated RSS. This value is the amount of total variation in Y that is explained in the regression equation. Total variation (TSS) is the sum of SSE and RSS.

F test

- The F test (to determine whether the slope coefficient is zero) is based on an F statistic constructed using the previous values
- The F statistic is the ratio of the average regression sum of squares to the average sum of squared errors
- the average regression sum of squares is computed by dividing the regression sum of squares by the number of slope parameters
- the average sum of squared errors is computed by dividing the sum of squared errors by the numbers of observations minus the number of parameters

Analysis of variance: The F-test

- The F statistic tests the null that all regression coefficients are zero. To compute we need:
- The total number of parameters and of observations
- The residuals sum of squares, i.e., the sum of squarerd residuals SSE

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{b}_0 - \hat{b}_1 X_i \right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

- The regression sum of squares RSS

$$RSS = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y} \right)^2$$

- The following holds: TSS=RSS+SEE

F statistic

- The F statistic is:

$$F = \frac{RSS/1}{SSE/n-2}$$

- We test the null that the regression has no explanatory power

Often, mutual fund performance is evaluated based on whether the fund has positive alpha—significantly positive excess risk-adjusted returns.[40] One commonly used method of risk adjustment is based on the capital asset pricing model. Consider the regression

$$(R_i - R_F) = \alpha_i + \beta_i(R_M - R_F) + \varepsilon_i \quad \textbf{(8-14)}$$

where R_F is the periodic risk-free rate of return (known at the beginning of the period), R_M is the periodic return on the market, R_i is the periodic return to Mutual Fund i, and β_i is the fund's beta. A fund has zero risk-adjusted excess return if $\alpha_i = 0$. If $\alpha_i = 0$, then $(R_i - R_F) = \beta_i(R_M - R_F) + \varepsilon_i$ and taking expectations, $E(R_i) = R_F + \beta_i(R_M - R_F)$, implying that β_i completely explains the fund's mean excess returns. If, for example, $\alpha_i > 0$, the fund is earning higher returns than expected given its beta.

In summary, to test whether a fund has a positive alpha, we must test the null hypothesis that the fund has no risk-adjusted excess returns (H_0: $\alpha = 0$) against the alternative hypothesis of nonzero risk-adjusted returns (H_a: $\alpha \neq 0$).

EXAMPLE 8-17. Performance Evaluation: The Dreyfus Appreciation Fund.

Table 8-12 presents results evaluating the excess return to the Dreyfus Appreciation Fund from January 1998 through December 2002. Note that the estimated beta in this regression, $\hat{\beta}_i$, is 0.7902. The Dreyfus Appreciation Fund was estimated to be about 0.8 times as risky as the market as a whole.

TABLE 8-12 Performance Evaluation of Dreyfus Appreciation Fund, January 1998 to December 2002

Regression Statistics	
Multiple *R*	0.9280
R-squared	0.8611
Standard error of estimate	0.0174
Observations	60

ANOVA	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MSS)	*F*
Regression	1	0.1093 RSS	0.1093	359.64
Residual	58	0.0176 SSE	0.0003	
Total	59	0.1269		

	Coefficients	Standard Error	*t*-Statistic
Alpha	0.0009	0.0023	0.4036
Beta	0.7902	0.0417	18.9655

Source: Center for Research in Security Prices, University of Chicago.

Note also that the estimated alpha ($\hat{\alpha}$) in this regression is positive (0.0009). The value of the coefficient is only a little more than one-third the size of the standard error for that coefficient (0.0023), so the *t*-statistic for the coefficient is only 0.4036. Therefore, we cannot reject the null hypothesis ($\alpha = 0$) that the fund did not have a significant excess return beyond the return associated with the market risk of the fund. This result means that the returns to the fund were explained by the market risk of the fund and there was no additional statistical significance to the excess returns to the fund during this period.[41]

Because the *t*-statistic for the slope coefficient in this regression is 18.9655, the *p*-value for that coefficient is less than 0.0001 and is approximately zero. Therefore, the probability that the true value of this coefficient is actually 0 is microscopic.

How can we use an *F*-test to determine whether the slope coefficient in this regression is equal to 0? The ANOVA portion of Table 8-12 provides the data we need. In this case,

- the total number of observations (n) is 60;
- the total number of parameters to be estimated is 2 (intercept and slope);
- the sum of squared errors or residuals, SSE, is 0.0176; and
- the regression sum of squares, RSS, is 0.1093.

Therefore, the *F*-statistic to test whether the slope coefficient is equal to 0 is

$$\frac{0.1093/1}{0.0176/(60-2)} = 360.19$$

(The slight difference from the *F*-statistic in Table 8-12 is due to rounding.) The ANOVA output would show that the *p*-value for this *F*-statistic is less than 0.0001 and is exactly the same as the *p*-value for the *t*-statistic for the slope coefficient. Therefore, the *F*-test tells us nothing more than we already knew from the *t*-test. Note also that the *F*-statistic (359.64) is the square of the *t*-statistic (18.9655).

Confidence interval for the prediction

- Regression gives a point forecast
- The variance of the forecast error is

$$\hat{s}_f^2 = \hat{\sigma}_\varepsilon^2 \sum_{i=1}^{n} \left[1 + \frac{1}{n} + \frac{(X_i - X)^2}{(n-1)s_X^2} \right]^2$$

- The confidence interval for the prediction error is: $\hat{Y} \pm t_C s_f$

EXAMPLE 8-18. Predicting the Ratio of Enterprise Value to Invested Capital.

We continue with the example of explaining the ratio of enterprise value to invested capital among food-processing companies by the spread between the return to invested capital and the weighted-average cost of capital (ROIC–WAAC). In Example 8-15, we estimated the regression given in Table 8-10.

TABLE 8-10 (repeated) Explaining Enterprise Value/Invested Capital by the ROIC–WACC Spread

Regression Statistics

Multiple *R*	0.9469
R-squared	0.8966
Standard error of estimate	0.7422
Observations	9

	Coefficients		Standard Error	*t*-Statistic
Intercept	1.3478	b_0	0.3511	3.8391
Spread	30.0169	b_1	3.8519	7.7928

You are interested in predicting the ratio of enterprise value to invested capital for a company if the return spread between ROIC and WACC is 10 percentage points. What is the 95 percent confidence interval for the ratio of enterprise value to invested capital for that company?

Using the data provided in Table 8-10, take the following steps:

$\hat{Y} = b_0 + b_1 X_i$

1. Make the prediction: Expected EV/IC = 1.3478 + 30.0169(0.10) = 4.3495. This regression suggests that if the return spread between ROIC and WACC (X_i) is 10 percent, the EV/IC ratio will be 4.3495.
2. Compute the variance of the prediction error. To compute the variance of the forecast error, we must know
   - the standard error of the estimate of the equation, $s = 0.7422$ (as shown in Table 8-10);
   - the mean return spread, $\overline{X} = 0.0647$ (this computation is not shown in the table); and
   - the variance of the mean return spread in the sample, $s_x^2 = 0.004641$ (this computation is not shown in the table).

Using these data, you can compute the variance of the forecast error (s_f^2) for predicting EV/IC for a company with a 10 percent spread between ROIC and WACC.

$$s_f^2 = 0.7422^2 \left[1 + \frac{1}{9} + \frac{(0.10 - 0.0647)^2}{(9 - 1)0.004641}\right]$$
$$= 0.630556$$

In this example, the variance of the forecast error is 0.630556, and the standard deviation of the forecast error is $s_f = (0.630556)^{1/2} = 0.7941$.

3. Determine the critical value of the *t*-statistic. Given a 95 percent confidence interval and $9 - 2 = 7$ degrees of freedom, the critical value of the *t*-statistic, t_c, is 2.365 using the tables in the back of the book.
4. Compute the prediction interval. The 95 percent confidence interval for EV/IC extends from $4.3495 - 2.365(0.7941)$ to $4.3495 + 2.365(0.7941)$, or 2.4715 to 6.2275.

In summary, if the spread between the ROIC and the WACC is 10 percent, the 95 percent prediction interval for EV/IC will extend from 2.4715 to 6.2275. The small sample size is reflected in the relatively large prediction interval.

Linear Regression II

Multiple Linear Regression and Issues

- We extend the simple linear model with one independent variable to a multiple regression model with several independent variables
- We question several assumptions of the regression model that may be violated in practice and invalidate the results we put forward

Multiple regression model

- We often hypothesize that a variable may be explained by more than one independent variable: Model for explaining equity returns more general than the market model, consumer expenditures explained by both income and price, etc.
- The general form of a multiple regression model is:

$$Y = b_0 + b_1 X_1 + \cdots + b_K X_K + \varepsilon$$

- A slope coefficient bj measure how much the dependent variable Y changes when the independent variable Xj changes by one unit, holding all the other independent variables constant.
- We refer to the intercept b0 and b1; : : : ;bk as regression coefficients.

Multiple Linear Regression

- Assumptions of the Multiple Linear Model
- The relationship between the dependent variable Y and the independent variables X1;X2; : : : ;Xk is linear as described by the previous equation.
- The independent variables X1;X2; : : : ;Xk are not random
- No exact linear relation exists between two or more of the independent variables.
- The expected value of the error term, conditioned on the independent variables, is 0:

- E(ejX1;X2; : : : ;Xk) = 0:

- The variance of the error term is the same for all observations: E(e2i) = s2e .
- The error term is uncorrelated across observations: E(eiej) = 0;
- The error term is normally distributed.

Remarks

- These assumptions are almost exactly the same as those for the single-variable model except for 2.;
- even a strong linear relation may cause problems as we will see.
- We will provide later in this section tests to check if the assumptions are fulfilled and propose some solutions.

Multiple linear regression
Regression Model

- Let us consider one dependent variable Y and K explanatory factors:

$$Y = b_0 + b_1 X_1 + \cdots + b_K X_K + \varepsilon$$

- And in matrix representation:

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{k,1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1,n} & \cdots & X_{k,n} \end{bmatrix}, b = \begin{bmatrix} b_0 \\ \vdots \\ b_k \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y = Xb + \varepsilon$$

Estimation multiple regression

- Least square estimators: minimize the sum of squared errors

$$\min_b\left(\varepsilon'\varepsilon\right)=\min_b\left(\left(Y-Xb\right)'\left(Y-Xb\right)\right)$$

- Applying First Order Conditions (partial derivatives equal to 0) we obtain

$$\frac{\partial\left(\left(Y-Xb\right)'\left(Y-Xb\right)\right)}{\partial b}=X'\left(Y-Xb\right)=0$$

$$X'Y=X'Xb$$

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

Orthogonality conditions

- Observe that:

$$\frac{\partial\left(\left(Y - Xb\right)'\left(Y - Xb\right)\right)}{\partial b} = 0$$

$$X'\left(Y - Xb\right) = X'\varepsilon = 0$$

- The principle of least squares holds iff residuals are orthogonal to the independent variables

Mean and variance of coefficients

- The estimate of b is unbiased

$$E\left(\hat{b}\right)=E\left(\left(X'X\right)^{-1}X'Y\right)=E\left(\left(X'X\right)^{-1}X'\left(bX+\varepsilon\right)\right)=$$
$$bE\left(\left(X'X\right)^{-1}X'X\right)+E\left(\left(X'X\right)^{-1}X'\varepsilon\right)=b$$

- Lets compute the var-cov of regression coefficients:

$$E\left(\left(\hat{b}-b\right)\left(\hat{b}-b\right)'\right)=E\left(\left(X'X\right)^{-1}X'\varepsilon\varepsilon'X\left(X'X\right)^{-1}\right)=\left(\left(X'X\right)^{-1}X'E\left(\varepsilon\varepsilon'\right)X\left(X'X\right)^{-1}\right)=$$
$$=\left(\left(X'X\right)^{-1}X'\sigma_{\varepsilon}^{2}IX\left(X'X\right)^{-1}\right)=\sigma_{\varepsilon}^{2}\left(X'X\right)^{-1}$$

$$\hat{\sigma}_{\varepsilon}^{2}=\frac{\varepsilon'\varepsilon}{n-\left(k+1\right)}$$

Var cov for simple regression

$$DM = \begin{bmatrix} 1 & X \end{bmatrix}$$

$$DM'DM = \begin{bmatrix} n & n\bar{X} \\ n\bar{X} & X'X \end{bmatrix}$$

$$|DM| = nX'X - n^2\bar{X}^2 = n(X-\bar{X})'(X-\bar{X}) = n\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$(DM'DM)^{-1} = \frac{\begin{bmatrix} X'X & -n\bar{X} \\ -n\bar{X} & n \end{bmatrix}'}{n\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{\begin{bmatrix} X'X & -n\bar{X} \\ -n\bar{X} & n \end{bmatrix}}{n\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

$$\text{var}(b_0) = \frac{\sigma_\varepsilon^2 X'X}{n\sum_{i=1}^{n}(X_i-\bar{X})^2}, \quad \text{var}(b_1) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}, \quad \text{cov}(b_0,b_1) = -\frac{\sigma_\varepsilon^2 n\bar{X}}{n\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

Tests to be conducted

- Tests on the regression coefficients: unilateral (sign tests) or bilateral (tests of a particular value) Student-t tests.
- Construction of confidence intervals for the regression coefficients.
- Test on the overall usefulness of the regression model: test whether all slope coefficients are equal to zero. This is the F-test

Multiple Linear Regression

- Adjusted Coefficient of Determination
- The coefficient of determination R2 is increasing with the number of independent variables included in the regression.
- Therefore one can artificially increase this statistic.
- To control for this, we can use an alternative measure of goodness of fit called adjusted $\bar{R}^2$, denoted $\bar{R}^2$

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1} \right) \left(1 - R^2 \right)$$

- Note that this measure of fit does not automatically increase when another variable is added to a regression. Note also that this adjusted measure can take negative values, while R2 is always positive.
- A high R2 is not a guarantee that the regression is well specified in the sense of including the correct set of variables.

Predicting the Dependent Variable in a Multiple Regression Model

- Collect the $\hat{b}_0, \ldots, \hat{b}_k$ coming out of the regression output.
- Determine the assumed values of the independent variables
- Compute the predicted value of the dependent variable Yˆi using the
- equation: $$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_1 + \cdots + \hat{b}_k X_k$$
- Two practical points:
- 1) Make sure the assumptions of the regression are met;
- 2) Be cautious about values of the independent variable that are outside the range of the data on which the model was estimated; such predictions are often unreliable.

Point estimate and prediction

- Prediction
- In vector form $\hat{Y}_i = X_i \hat{b}$
- This point estimator is unbiased:

$$E\left(\hat{Y} - Y_i\right) = E\left(X_i \hat{b} - Y_i\right) = E\left(X_i \hat{b} - X_i b - \varepsilon_i\right) =$$
$$X_i E\left(\hat{b} - b\right) - E\left(\varepsilon_i\right) = 0$$

- And its standard error is given by:

$$\sigma_i = \sqrt{E\left(\hat{Y} - Y_i\right) E\left(\hat{Y} - Y_i\right)'} =$$
$$\sqrt{X_i E\left(\hat{b} - b\right) E\left(\hat{b} - b\right) X_i} =$$
$$\sqrt{X_i \left(X' X\right)^{-1} X_i}$$

Using Dummy Variables in Regressions

- Often we use qualitative variables as independent variables in a regression.
- A qualitative variable takes discrete values to indicate that a particular firm or individual belongs to a certain group.
- For example, instead of putting the exact level of income of a household (which is rarely asked to people in surveys) in a regression, we put the income group (1 to 5) if income is divided in five classes (often the question asked is about a range of income).
- A dummy variable takes on a value of 1 if a particular condition is true and 0 if the condition is false.
- Example: test if the mean return is the same say in January than during the remaining months of the year.

$$Y_t = b_0 + b_1 X_{1t} + \varepsilon_t$$

- The dummy variable X_{1t} has a value of 1 for each January and a value of zero for every other month of the year.
- WARNING: Exercise care in choosing the number of dummy variables in a regression
- If you want to distinguish the mean between each of the four quarters of the year you would include dummy variables for just three quarters. Otherwise assumption 2 is violated.

Heteroscedasticity and autocorrelation

- The estimators we obtained for the regression coefficients and their properties rest on several assumptions. We will look at two key assumptions that are often not fulfilled in linear regression models.
- 1 The variance of the error term is the same for all observations:

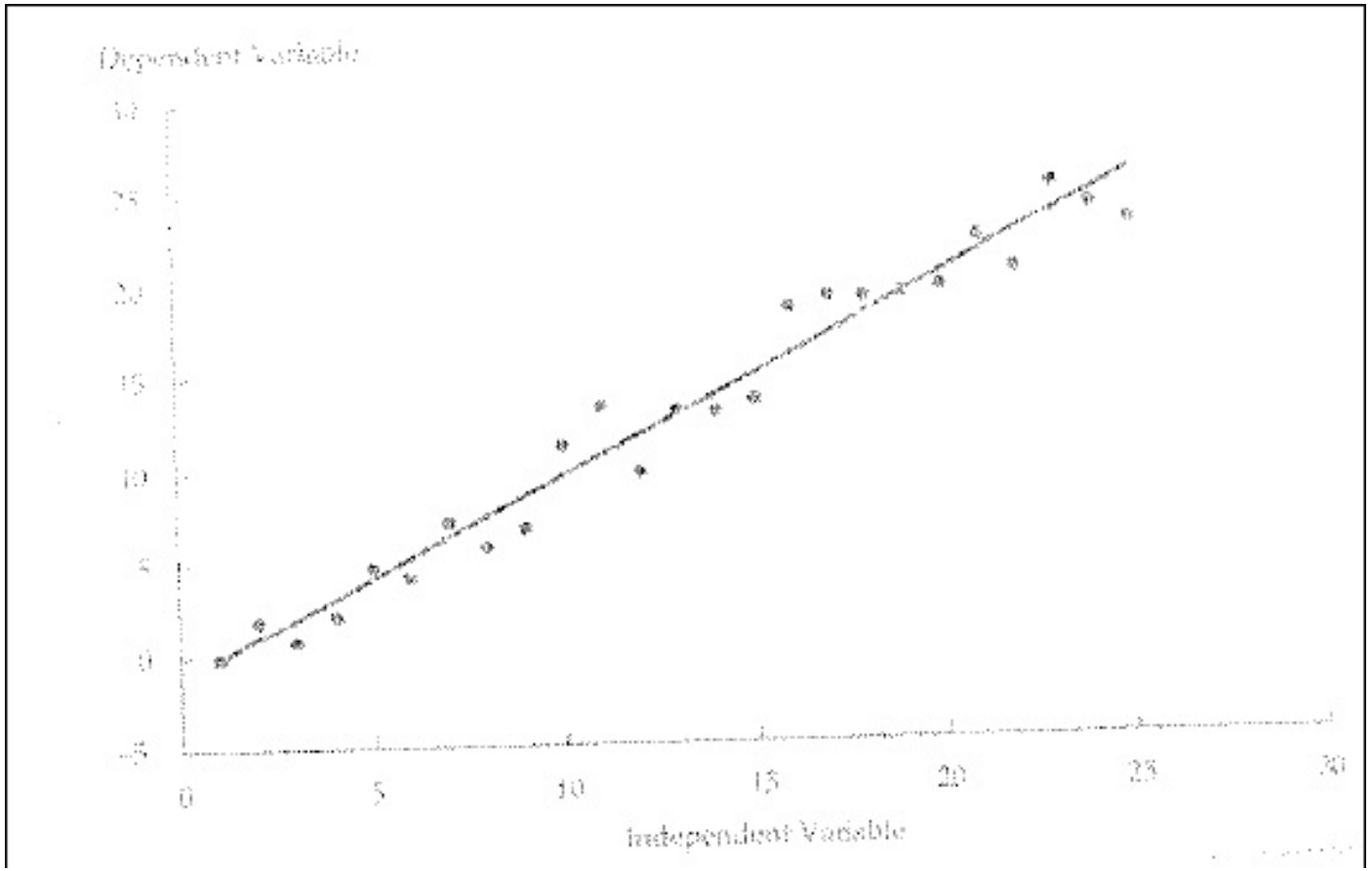
$$E\left(\varepsilon_i^2\right) = s^2$$

- 2 The error term is uncorrelated across observations:

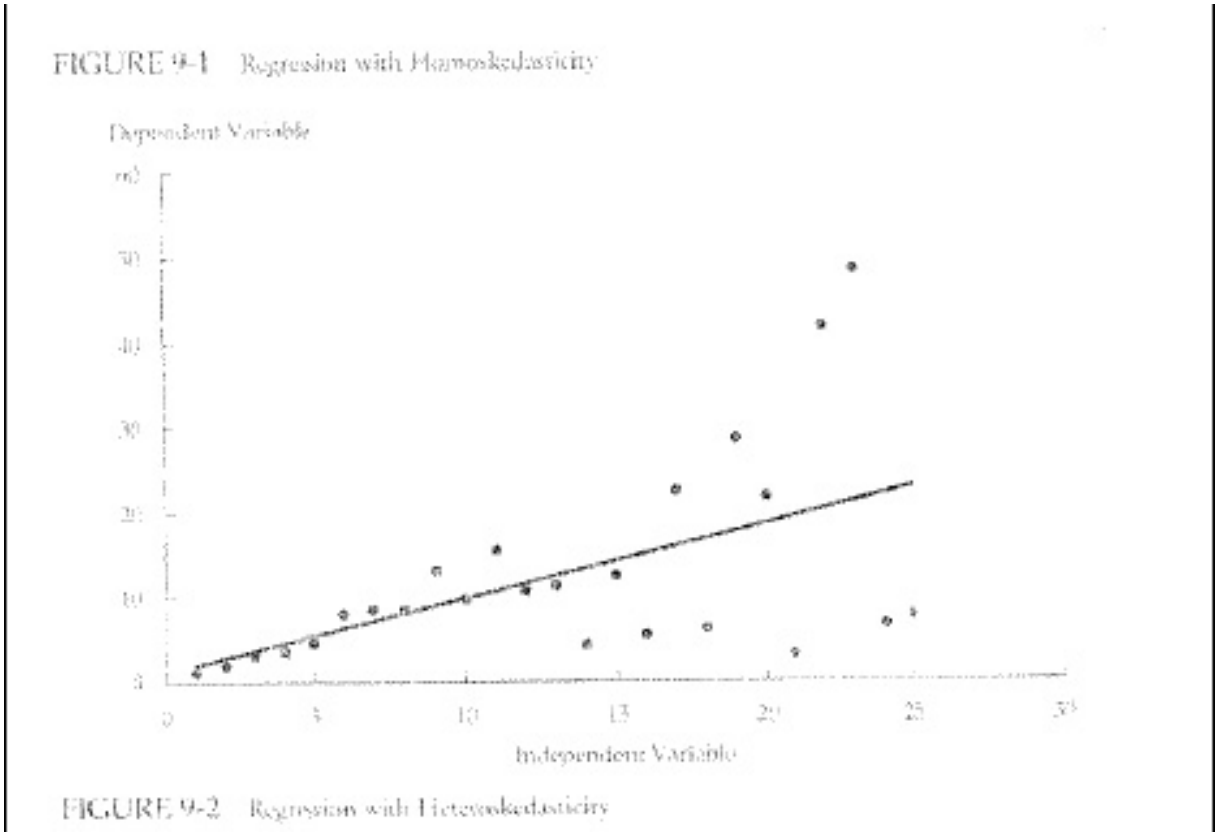
$$E\left(\varepsilon_i \varepsilon_j\right) = 0, \; i \neq j$$

- The non-fulfilment of the first is called heteroscedasticity, as opposed to homoscedasticity. It can happen in both cross-sectional and time-series regressions.
- The non-fulfilment of the second is called autocorrelation or serial correlation and is a time series phenomenon.
- For each problem, we will see the consequences of the problem for the regression properties, learn about the tests to detect the problem, and finally look at ways to correct the problem

Homoscedastic errors



Heteroscedastic errors



Heteroscedasticity
Consequences of heteroscedasticity

- Heteroscedasticity does not affect the consistency of the estimator.
- Heteroscedasticity can lead to mistakes in inference.
- The estimators of the standard errors of the regression will be biased unless they are corrected for heteroscedasticity.
- The F-test and the t-tests are unreliable.
- In regressions with financial data, often standard errors will be underestimated and t-stats inflated.

Heteroscedasticity
Correcting for Heteroscedasticity

- Two different methods to correct for heteroscedasticity.
  - Computing Robust standard errors.
  - Generalized least squares (GLS)
- The first method corrects the standard errors of the regression coefficients of the linear model, to account for the conditional heteroscedasticity (heteroscedasticity in the error variance that is correlated with the values of the independent variables in the regression). Software packages can produce corrected standard errors.
- The second method (GLS) modifies the original equation in an attempt to eliminate the heteroscedasticity. The new modified regression equation is then estimated under the assumption that the errors are homoscedastic.

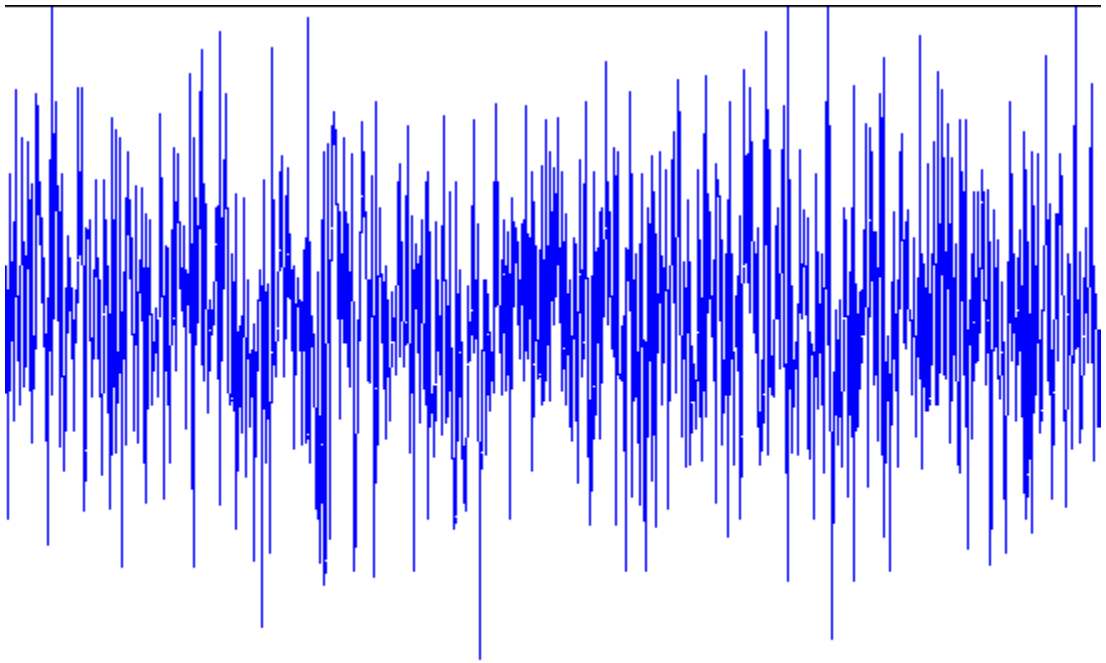
Generalized Least Squares GLS

- OLS-efficiency is subject to the following assumption:

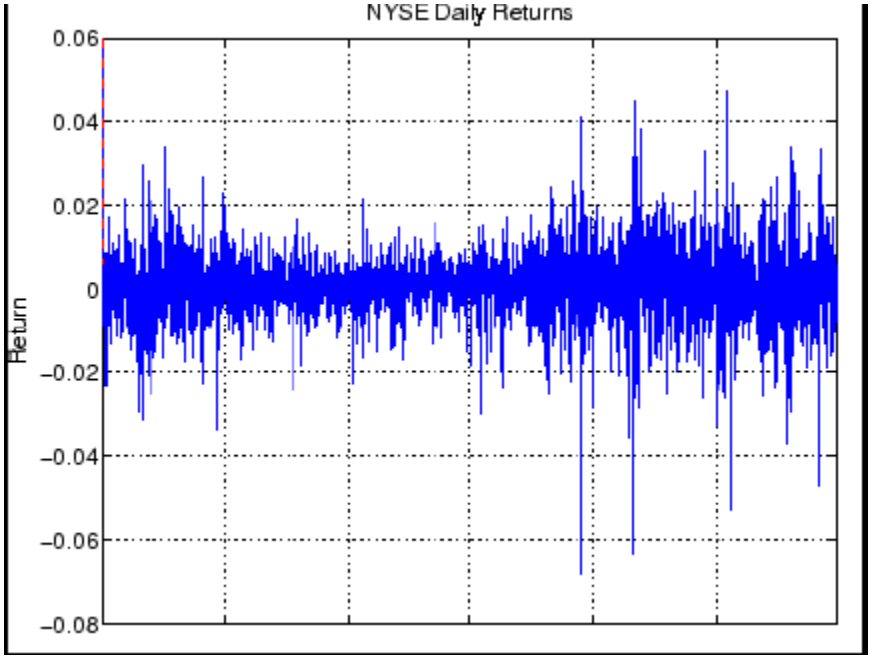
$$\varepsilon \approx N\left(0, \sigma^2 I\right)$$

- This translates to:
- 1Homoscedastic variances
- Absence of autocorrelation

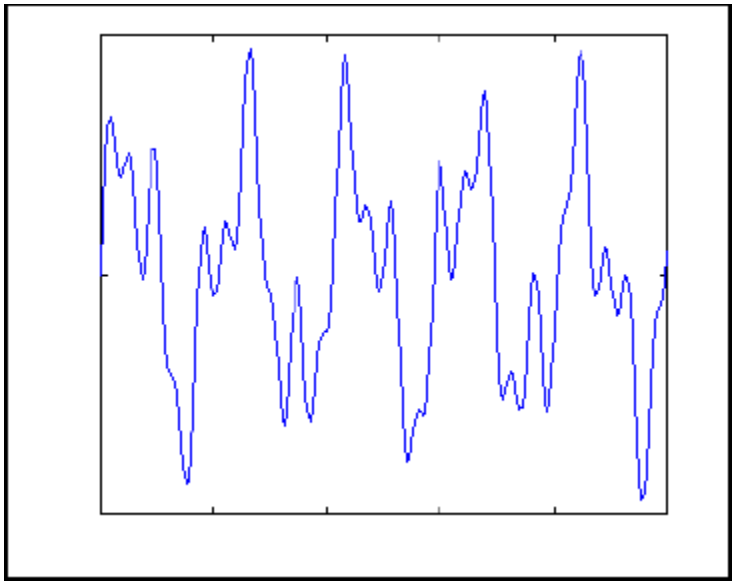
White noise



Heteroscedastic errors



Autocorrelated errors



Generalized Least Squares

- The more general assumption for the error process is given by:

$$\varepsilon \approx N\left(0, \sigma^2 \Psi\right)$$

- Remarks
- OLS is still unbiased
- OLS is not efficient
- An efficient estimator is given by:

$$b_{GLS} = \left(X' \Psi^{-1} X\right)^{-1} X' \Psi^{-1} Y$$

- This estimator is called the Generalized Least Squares estimator (GLS)

Explanations

- Since Y is positive definite (being a variance-covariance matrix), its inverse is positive definite.
- Then we have seen that it is possible to find a nonsingular matrix P such that:

$$\Psi^{-1} = P'P$$

- Therefore:

$$b_{GLS} = \left(X'P'PX\right)^{-1} X'P'PY = \left(\left(PX\right)'PX\right)^{-1}\left(PX\right)'PY$$

- This is exactly the vector of estimated coefficients that would be obtained from the OLS regression of the vector PY on the matrix PX.

Explanations

- The error terms of the transformed regression model obey the OLS assumptions. Starting with the original model:

$$Y = XB + \varepsilon$$

- Premultiplying by the matrix P:

$$PY = PXB + P\varepsilon$$

- Define $v_t = P\varepsilon_t$
- The variance of the transformed model is:

$$Var(v_t) = E(P\varepsilon\varepsilon' P') = PE(\varepsilon\varepsilon')P' = \sigma_\varepsilon^2 P\Psi P' = \sigma_\varepsilon^2 PP'(P')^{-1}P' = \sigma_\varepsilon^2 I$$

Cross sectional heteroscedasticity

- Definition
- A cross-section is heteroscedastic if its variance is not constant over groups

$$\begin{bmatrix} \sigma_1^2 I_1 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma_2^2 I_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \sigma_g^2 I_{ng} \end{bmatrix}$$

Goldfeld-Quandt Test

- This test is applicable if there is a single variable (say X_i) that is thought to be an indicator of the heteroscedasticity.
- Procedure
- Reorder the observations by the value of X_i
- Omit c central observations.
- Fit separate regressions by OLS to the first and last $(n-c)/2$ observations.
- Let RSS1 and RSS2 be the residual sums of squares from the two regressions, the subscript 1 indicating that from the smaller Xi and 2 that from the larger Xi values.
- Then

$$R = \frac{RSS1}{RSS2} \approx F\left((n-c-2k)/2, (n-c-2k)/2 \right)$$

- If R > F0:95, one would reject the assumption of homoscedasticity at the 5
- percent level.
- The power will depend on how many central observations are excluded.

Heteroscedasticity of time series

- A time-series is heteroscedastic if its variance is not constant over time

$$\begin{bmatrix} \sigma_1^2 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma_2^2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \sigma_T^2 \end{bmatrix}$$

- Test
- White test
- Breusch-Pagan test
- ARCH test

Heteroscedasticity
White Test

- Idea: This asymptotic test does not require to specify a model to determine the heteroscedasticity.
- Procedure
- One simply computes an auxiliary regression of the squared OLS residuals on a constant and all variables, their squares and their cross-product.
- For two variables include

$$\left\{1, X_{2t}, X_{3t}, X_{2t}^2, X_{3t}^2, X_{2t}, X_{3t}\right\}$$

- On the null hypothesis of homoscedasticity, nR^2 is asymptotically distributed as

$$\chi^2(5)$$

- If homoscedasticity is rejected, there is no indication of the form of heteroscedasticity.

Heteroscedasticity
Breusch-Pagan Test

- Idea: It is assumed that heteroscedasticity takes the form

$$E(\varepsilon_t) = 0, \ \forall t$$
$$\sigma_t^2 = E(\varepsilon_t^2) = h(z_t'\alpha)$$
$$z'_t \left[1, z_{2t}, \ldots, z_{pt}\right]$$

- Where $z'_t \left[1, z_{2t}, \ldots, z_{pt}\right]$ is a vector of known variables and alpha a vector of *p* unknown parameters.
- Procedure
- Estimate the original model by OLS. Obtain the OLS residuals e_t
- Perform an auxiliary regression of e_t^2 on z_t
- Then TR^2 from this regression is asymptotically distributed as $\chi^2(p-1)$ under the null of homoscedasticity.

Heteroscedasticity
ARCH Test

- Idea: In speculative markets such as exchange rates and stock market returns, large and small errors tend to occur in clusters.
- We postulate a relation of the form:

$$\sigma_t^2 = \hat{\alpha}_0 + \hat{\alpha}_1 \varepsilon_{t-1}^2 + \cdots + \hat{\alpha}_p \varepsilon_{t-p}^2$$

- Procedure
- Fit Y to X by OLS and obtain the residuals $\{e_t\}$
- Compute the OLS regression:

$$e_t^2 = \alpha_0 + \alpha_1 e_{t-1}^2 + \cdots + \alpha_p e_{t-p}^2 + error$$

- Test the null hypothesis: H0 : a1 = a2 = = ap = 0 (F-test)
- F =(RestrictedRSS-UnrestrictedRSS)/p/UnrestrictedRSS=(T –p-1) is distributed as F(p;T –p-1)

Autocorrelation

- A time-series is autocorrelated if disturbances depend on past disturbances, for example an AR(1) $\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$
- We can write: $\text{var}(\varepsilon_t) = \rho^2 \text{var}(\varepsilon_{t-1}^2) + \text{var}(\nu_t)$

$$\sigma_\varepsilon^2 = \frac{\sigma_\nu^2}{1-\rho^2}$$

$$\Psi = \begin{bmatrix} \varepsilon_1^2 & \cdots & \cdots & \varepsilon_1\varepsilon_T \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \varepsilon_T\varepsilon_1 & \cdots & \cdots & \varepsilon_T^2 \end{bmatrix} = \frac{\sigma_\nu^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^{T-1} \\ \rho & 1 & \\ & & \\ \rho^{T-1} & & 1 \end{bmatrix}$$

Autocorrelation

- Consequences of autocorrelation
- Autocorrelation can lead to mistakes in inference.
- The estimators of the standard errors of the regression will be biased unless they
- are corrected for autocorrelation.
- The F-test and the t-tests are unreliable.
- Tests
- Durbin-Watson test
- Box-Pierce-Ljung Statistic

Autocorrelation
Durbin-Watson test

- Given the AR(1) process, the null hypothesis of zero autocorrelation will be: H0 : r = 0
- The e are unobservable, so need to rely on the estimated residuals *e = Y -Xb.*
- These might exhibit some autocorrelation even if the null is true.
- No exact finite sample test valid for any X matrix
- The procedure:
- The DW statistic is given by:

$$DW = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2}$$

- The Durbin-Watson statistic is closely related to the first-order autocorrelation
- coefficient. Expanding the statistic, for large T: $DW \cong 2(1-\hat{\rho})$

- Heuristically, the range of d is from 0 to 4.
- DW < 2 for positive autocorrelation of the e
- DW > 2 for negative autocorrelation of the e
- DW=2 for zero autocorrelation of the e

Autocorrelation
Box-Pierce-Ljung Statistic

- The Box-Pierce Q statistic is based on the squares of the first p autocorrelation coefficients of the OLS residuals. The statistic is defined as:

$$Q = T\sum_{j=1}^{p} r_j^2,\ r_j = \frac{\sum_{t=j+1}^{T} e_t e_{t-j}}{\sum_{t=1}^{T} e_t^2}$$

- The test
- Under the null of zero autocorrelation for the residuals, Q will have an asymptotic χ^2 distribution.
- Remark:
- An improved small-sample performance is expected from the revised Ljung-Box statistic:

$$Q' = T(T+2)T\sum_{j=1}^{p} \frac{r_j^2}{n-j}$$

Autocorrelation
Correcting for Autocorrelation

- Two different methods to correct for autocorrelation.
- Computing Robust standard errors.
- Generalized least squares (GLS).
- The first method corrects the standard errors of the regression coefficients of the linear model, to account for the serial correlation. Software packages can produce corrected standard errors.
- The second method (GLS) modifies the original equation in an attempt to eliminate the serial correlation. The new modified regression equation is then estimated under the assumption that the errors are uncorrelated.
- The first method uses a method developed by Hansen (1982) that simultaneously corrects for serial correlation and heteroscedasticity. Often it is called the Newey-West method of correcting standard errors.
- The correction will have the effect of increasing the standard errors of the regression coefficients (lowering the t-statistics) and making the statistical inference more robust.

Multicollinearity

- When some or all of the independent variables in a multiple regression are correlated, it is difficult or impossible to disentangle their separate explanatory effects on Y.
- When there is multicollinearity, the regression coefficients are unstable in the degree of statistical significance magnitude and sign. The R2 may be high but the standard errors are also high and consequently, the *t*- statistics are small indicating an apparent lack of significance.
- The problem is particularly prevalent in multi-factor models that use time series to model equity returns. Style factors, industry indices and even major market factors can be highly correlated.

Potential Solutions to Multicollinearity

- Principal Components
- Add further sample data.
- Drop variables that are highly correlated with each other.
- Pooling of cross-section and time-series data.

Conditional mean and conditional variance

- Consider a stationary time series $(X_1, \ldots, X_t, \ldots)$
- The unconditional mean and variance are constant $\mu = E(X_t)$

$$\sigma^2 = E\left[(X_t - \mu)^2\right]$$

- If a stationary time series is generated by an AR process the conditional mean depends on past values of the X_t

$$X_t = \alpha_0 + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p}$$

Conditional mean and conditional variance

- We can also define stationary processes with a non constant conditional variance or conditional volatility (std)
- The ARCH/GARCH is a family of stationary models with time varying conditional variance
- Stationary models have constant unconditional meana nd variance but can have time varying conditional mean and variance

Empirical Regularities (Stylized Facts)

- Thick tails
- Volatility clustering
- Thick tails and volatility clustering intimately related: link between dynamic (conditional) volatility behavior and (unconditional) heavy tails.
- Leverage effects: changes in stock prices negatively correlated with changes in stock volatility
- Long-memory and persistence: volatility is highly persistent
- Co-movements in volatilities
- Volatility and information arrival: links between volatility and trading volume, dividend announcements or macroeconomic data releases.

Conditional Variance

- The conditional variance σ_t^2 of an innovation ε_t process is by definition:

$$E_{t-1}(\varepsilon_t) = 0, \ \mathrm{t} = 1,2,\ldots$$

$$\sigma_t^2 = \mathrm{var}_{t-1}(\varepsilon_t) = E_{t-1}(\varepsilon_t^2), \ \mathrm{t} = 1,2,\ldots$$

- In autoregressive conditional heteroscedastic models, this conditional
- variance depends on the past of the innovations $[\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots]$

- Standardized process $z_t = \varepsilon_t \left(\sigma_t^2\right)^{-\frac{1}{2}}$
  so that $\varepsilon_t = \left(\sigma_t^2\right)^{\frac{1}{2}} z_t$
- Mean-zero, time invariant and variance of unity
- If conditional distribution of z_t is assumed to be time invariant and Gaussian, the unconditional distribution for ε_t is leptokurtic.
- The key insight of GARCH lies in the distinction between conditional and unconditional variances of the innovations process

Conditional Variance
Linear ARCH (q) model: Engle (1982)

$$\sigma_t^2 = w + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2$$

- Parameters must satisfy $w > 0,\ \alpha_i \geq 0,\ i = 1, \ldots, q$
- Defining: $v_t = \varepsilon_t^2 - \sigma_t^2$ the model can be rewritten as

$$\varepsilon_t^2 = w + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + v_t$$

- Since $E_{t-1}(v_t) = 0$, the model corresponds directly to an AR(q) for the ε_t^2 innovations squared,
- The process is covariance stationary if and only if $\sum_{i=1}^{q} \alpha_i < 1$ then:

$$\text{var}(\varepsilon_t) = \sigma^2 = \frac{w}{1 - \alpha_1 - \cdots - \alpha_q}$$

- In empirical applications of ARCH(q) models, a long lag length and a large number of parameters are often called for.

GARCH (p,q) model

- To circumvent this problem, Bollerslev (1986) proposed a generalized ARCH, GARCH (p,q)

$$\sigma_t^2 = w + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$$

- For the conditional variance to be well-defined, all coefficients in infinite ARCH representation must be positive.
- For GARCH (1,1), positivity of σ_t^2 requires $w \geq 0,\ \alpha_1 \geq 0,\ \beta_1 \geq 0,$

$$\sigma_t^2 = w + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$