BACHELOR OF ECONOMICS
THAMMASAT UNIVERSITY

FN211: Lecture Note 7
Regression Analyses

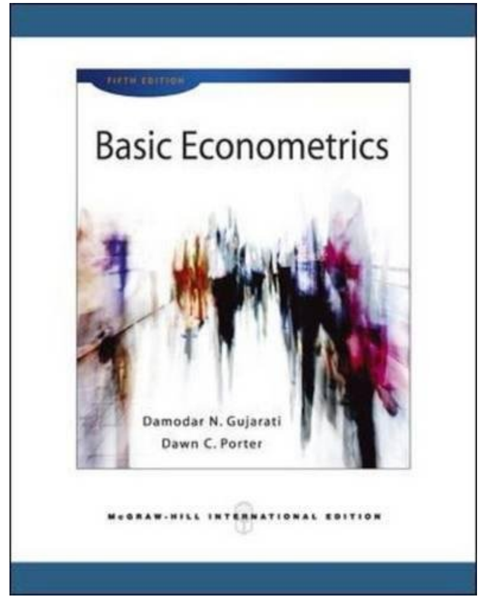
Dr. Winai Homsombat

Bachelor of Economics, International Program

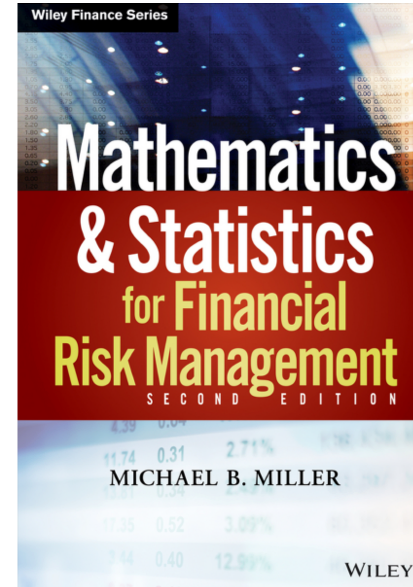
Thammasat University

Outline

- Simple Regression
- Multiple Regression
- Application



Reading: PART I



Reading: Chapter 10

Example in Excel File: bit.ly/211excel20

Recall: Basic statistics

- Arithmetic Average: $$E(r) = \sum_{s=1}^{n} p(s)r(s) = \frac{1}{n}\sum_{s=1}^{n} r(s)$$

- Variance: $$\hat{\sigma}^2 = \frac{1}{n}\sum_{s=1}^{n}\left[r(s) - \bar{r}\right]^2$$

- Covariance: $$\sigma_{XY} = \frac{\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)}{n}$$

Recall: Basic statistics

- Correlation: $$\rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

What is regression?

- **Correlation analysis VS. Regression analysis**

"CAUSE VS EFFECT"

Independent variable VS Dependent variable

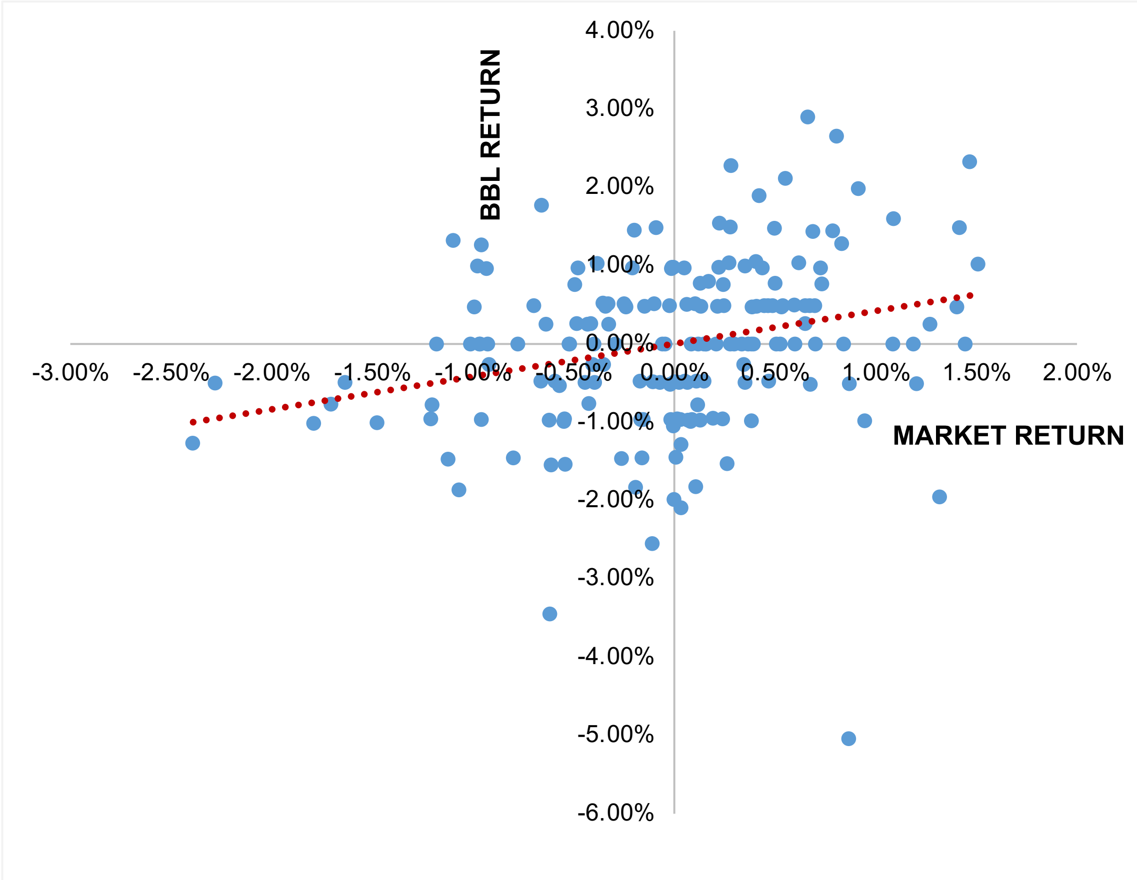
Example

1. Movement between returns of the 2 stocks during 1st quarter of 2019
2. The impact of SET index on ADVANCE stock return
3. Relation between Finance and Statistics scores of BE students

Intuition – Index Model

Estimation Model:

$$R_{BBL} = \alpha + \beta R_m + \varepsilon_t$$



Classical Linear Regression Analysis

Simple Linear Regression: $y = \beta_0 + \beta_1 x_1 + \epsilon$

Multiple Linear Regression: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$

(General form)

Simple Regression

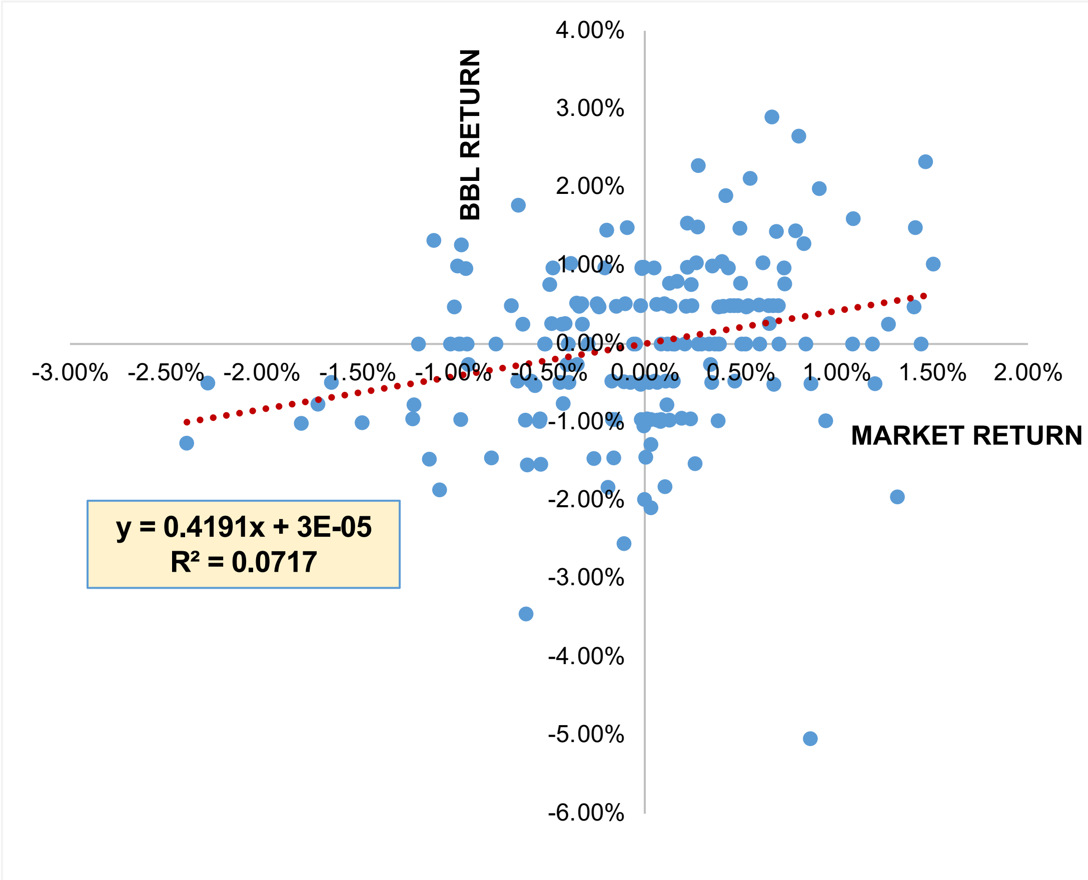
The important parameter of estimation

- Coefficient
- Standard deviation
- R-square and Adjust R-square

  (Test of Goodness of Testing)
- Z, t, F statistics

  (Hypothesis Testing)

$$E(R_{Kbank} \mid R_m) = \alpha + \beta R_m$$
$$E(R_{Kbank} \mid R_m = 1) = \ldots$$



Estimation & Interpretation – Ordinary Least Square (OLS)

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Ordinary Least Square: Begin with **Population Regression Function (PRF)**

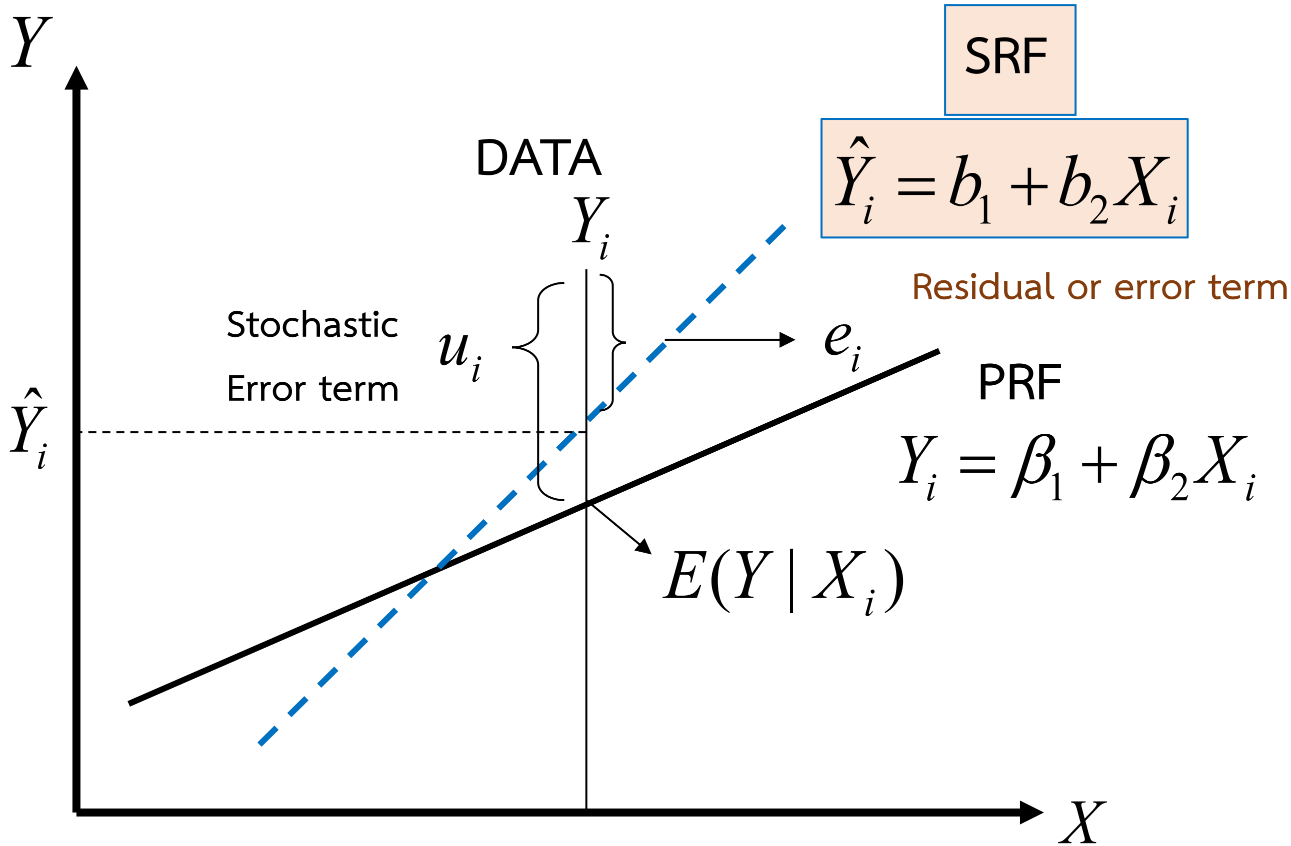
$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

However, PRF is not directly observable, we estimate it from **Sample Regression Function (SRF)**

$$Y_i = b_1 + b_2 X_i + e_i$$

$$Y_i = \hat{Y}_i + e_i$$

Estimation & Interpretation – Ordinary Least Square (OLS)



Estimation & Interpretation – Ordinary Least Square (OLS)

- This method creates the high probability that the *estimators* are the best estimation to the *parameters*
- It tries to minimize the residuals squares of estimation.

$$\min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} (Y_i - b_1 - b_2 X_i)^2$$

Estimation & Interpretation – Ordinary Least Square (OLS)

OLS Estimators

$$b_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

$$= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$= \frac{\sum x_i y_i}{\sum x_i^2}$$

$$= \frac{\sum x_i Y_i}{\sum X_i^2 - n\bar{X}^2} = \frac{\sum X_i y_i}{\sum X_i^2 - n\bar{X}^2}$$

$$b_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

$$= \bar{Y} - b_2 \bar{X}$$

Estimation & Interpretation – Ordinary Least Square (OLS)

Precision or Standard Errors of Least-Squares Estimates

Estimation of σ^2 :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} [Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)]^2}{n-2}$$

Number of parameter

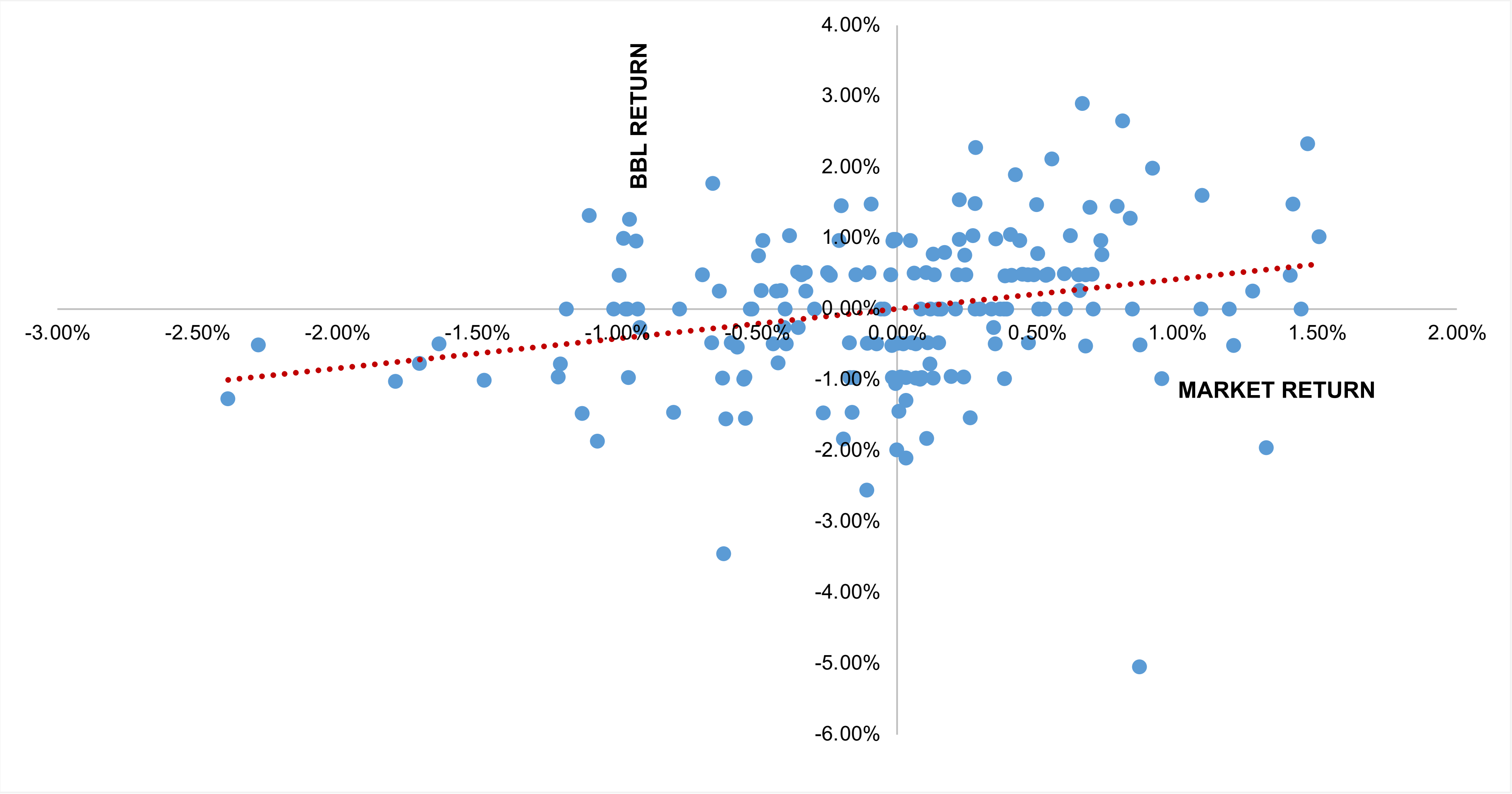
$$\text{var}(b_2) = \frac{\sigma^2}{\sum x_i^2}$$

$$se(b_2) = \frac{\sigma}{\sqrt{\sum x_i^2}}$$

$$\text{var}(b_1) = \frac{\sum X_i^2}{n \sum x_i^2} \sigma^2$$

$$se(b_1) = \sqrt{\frac{\sum X_i^2}{n \sum x_i^2} \sigma^2}$$

What is regression output?



Regression Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.2677
R Square	0.0717
Adjusted R Square	0.0665
Standard Error	0.0104
Observations	180

ANOVA

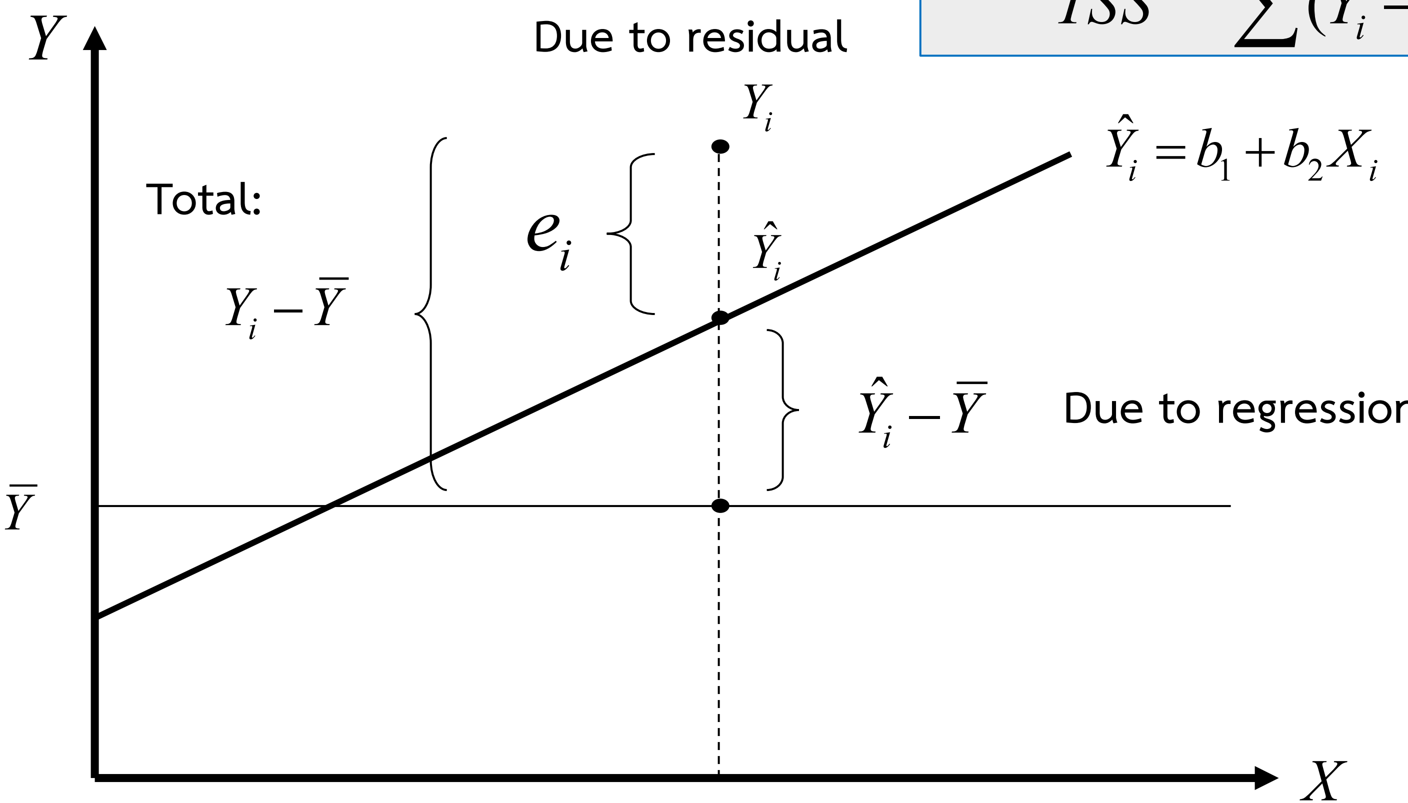
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.0015	0.0015	13.7448	0.0003
Residual	178	0.0194	0.0001		
Total	179	0.0209			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.9327E-05	0.0008	0.0377	0.9700	-0.0015	0.0016
X Variable 1	0.4191	0.1130	3.7074	0.0003	0.1960	0.6421

Estimation & Interpretation – Ordinary Least Square (OLS)

Coefficient of Determination

$$r^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$



Estimation & Interpretation – Ordinary Least Square (OLS)

Coefficient of Determination: R^2 and Adjusted R^2

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2}$$

$$\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2 / (n - \boxed{k})}{\sum y_i^2 / (n-1)}$$

→ k: number of parameters in the model including the intercept term

Adjust: Adjusted for the degree of freedom associated with the sum of square

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

Estimation & Interpretation – Ordinary Least Square (OLS)

Analysis of Variance (ANOVA)

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2$$
$$= 2b_2^2 \sum X_i^2 + \sum e_i^2$$

Source of Variation	SS	Df	MSS
Due to regression (ESS)	$\sum \hat{y}_i^2$	1	$b_2^2 \sum x_i^2$
Due to residual (RSS)	$\sum e_i^2$	$n-2$	$\frac{\sum e_i^2}{n-2} = \hat{\sigma}^2$
TSS	$\sum y_i^2$	$n-1$	

$$F = \frac{MSS \quad of \quad ESS}{MSS \quad of \quad RSS}$$

Hypothesis Testing and Interval Estimation

Test individual coefficient

test statistic to verify or falsify the null hypothesis

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Testing of the significance of regression coefficient: t - test

$$t = \frac{\overline{X} - \mu_0}{S_X / \sqrt{n}}$$

Hypothesis Testing and Interval Estimation

Confidence interval derivation

$$\Pr(-t_{\alpha/2} \le t \le t_{\alpha/2}) = 1 - \alpha$$

$100(1-\alpha)\%$ confidence interval for β_1 and β_2

$$\hat{\beta}_1 \pm t_{\alpha/2} S_{\hat{\beta}_1} \qquad \hat{\beta}_2 \pm t_{\alpha/2} S_{\hat{\beta}_2}$$

Example

The relationship between nominal exchange rate and relative prices. From the annual observations from 1980 to 1994, the following regression results were obtained, where Y = exchange rate of the German mark to the U.S. dollar (GM/$) and X = ratio of the U.S. consumer price index to the German consumer price index; that is, X represents the relative prices in the two countries:

$$\hat{Y}_t = 6.682 - 4.318 X_t \qquad r^2 = 0.528$$
$$\text{se} = (1.22)(1.333)$$

a. Interpret this regression. How would you interpret r^2?
b. Does the negative value of X_t make economic sense? What is the underlying economic theory?

Key Assumptions

1. Regressors are NOT correlated
2. Error variance is constant

$$\mathrm{var}(u_i) = \sigma_i^2$$

3. Error terms are NOT correlated

$$E(\varepsilon_i, \varepsilon_j) \neq 0, \quad \mathrm{cov}(\varepsilon_i, \varepsilon_j) \neq 0$$

4. Model is correct, i.e. liner vs. log form
5. Regressor and error term are correlated

$$E(\mathrm{X}_i, \varepsilon_i) \neq 0, \quad \mathrm{cov}(\mathrm{X}_i, \varepsilon_i) \neq 0$$

Example

From a sample of 10 observations, the following results were obtained:

$$\sum Y_i = 1110 \qquad \sum X_i = 1700 \qquad \sum X_i Y_i = 205{,}500$$

$$\sum X_i^2 = 322{,}000 \qquad \sum Y_i^2 = 132{,}100$$

with coefficient of correlation $r = 0.9758$. But on rechecking these calculations it was found that two pairs of observations were recorded:

Y	X		Y	X
90	120	instead of	80	110
140	220		150	210

What will be the effect of this error on r? Obtain the correct r.

Multiple Regression

Estimation & Interpretation – OLS

The Generic Multiple Regression Model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + \varepsilon_i \quad i = 1,\ldots,n$$

Estimation of regression parameters:

The use of the matrix notation allows a view of how the data are housed in software programs.

$$\begin{aligned}
Y_1 &= \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \cdots + \beta_k X_{k1} + u_1 \\
Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \cdots + \beta_k X_{k2} + u_2 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
Y_n &= \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \cdots + \beta_k X_{kn} + u_n
\end{aligned}$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{nx1} \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1k} \\ 1 & X_{21} & X_{22} & \ldots & X_{2k} \\ \vdots & \vdots & \vdots & & \\ 1 & X_{n1} & X_{n2} & \ldots & X_{nk} \end{bmatrix}_{nxk} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}_{kx1} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{nx1}$$

Estimation & Interpretation – OLS

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\underset{n \times 1}{\mathbf{y}} = \underset{n \times k}{\mathbf{X}} \; \underset{k \times 1}{\boldsymbol{\beta}} + \underset{n \times 1}{\mathbf{u}}$$

$$\underset{n \times 1}{\mathbf{y}} = \underset{n \times k}{\mathbf{X}} \; \underset{k \times 1}{\boldsymbol{\beta}} + \underset{n \times 1}{\mathbf{u}}$$

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki})^2$$

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = [\hat{u}_1 \quad \hat{u}_2 \quad \cdots \quad \hat{u}_n] \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix} = \hat{u}_1^2 + \hat{u}_2^2 + \cdots + \hat{u}_n^2 = \sum \hat{u}_i^2$$

Estimation & Interpretation – OLS

From:

$$\hat{u}'\hat{u} = (\mathrm{y} - \mathrm{X}\hat{\beta})'(\mathrm{y} - \mathrm{X}\hat{\beta})$$

$$= \mathrm{y}'\mathrm{y} - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$$

FOC:

$$\frac{\partial(\hat{\mathrm{u}}'\hat{\mathrm{u}})}{\partial\hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

$$\left(X'X\right)\hat{\beta} = X'y$$

Hence, $\hat{\beta} = \left(X'X\right)^{-1} X'y$

Estimation & Interpretation – OLS

OLS Assumptions with matrix notation

Scalar notation			Matrix notation
1. $E(u_i) = 0,$	for each i	(3.2.1)	**1.** $E(\mathbf{u}) = \mathbf{0}$ where **u** and **0** are $n \times 1$ column vectors, **0** being a null vector
2. $E(u_i u_j) = 0$	$i \neq j$	(3.2.5)	**2.** $E(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I}$ where **I** is an $n \times n$ identity matrix
$= \sigma^2$	$i = j$	(3.2.2)	
3. $X_2, X_3, \ldots, X_k$ are nonstochastic or fixed			**3.** The $n \times k$ matrix **X** is nonstochastic, that is, it consists of a set of fixed numbers
4. There is no exact linear relationship among the X variables, that is, no multicollinearity		(7.1.7)	**4.** The rank of **X** is $\rho(\mathbf{X}) = k$, where k is the number of columns in **X** and k is less than the number of observations, n
5. For hypothesis testing, $u_i \sim N(0, \sigma^2)$		(4.2.4)	**5.** The **u** vector has a multivariate normal distribution, i.e., $\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$

Estimation & Interpretation – OLS

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

Substituting $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ into the preceding expression gives

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{(\mathbf{X}\boldsymbol{\beta} + \mathbf{u})} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\end{aligned}$$

Therefore, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$

$$\begin{aligned}\text{var-cov}(\hat{\boldsymbol{\beta}}) &= E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\ &= E\{[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}]'\} \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\underline{\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}}]\end{aligned}$$

$$\begin{aligned}\text{var-cov}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}\mathbf{u}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

$$\boxed{E(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I}}$$

Estimation & Interpretation – OLS

Coefficient of determination R-square:

$$R^2 = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{Y}^2}{\mathbf{y}'\mathbf{y} - n\bar{Y}^2}$$

Testing the overall significance of regression:

$$F = \frac{(\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{Y}^2)/(k-1)}{(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y})/(n-k)}$$

ANOVA table for *k*-variable linear regression model:

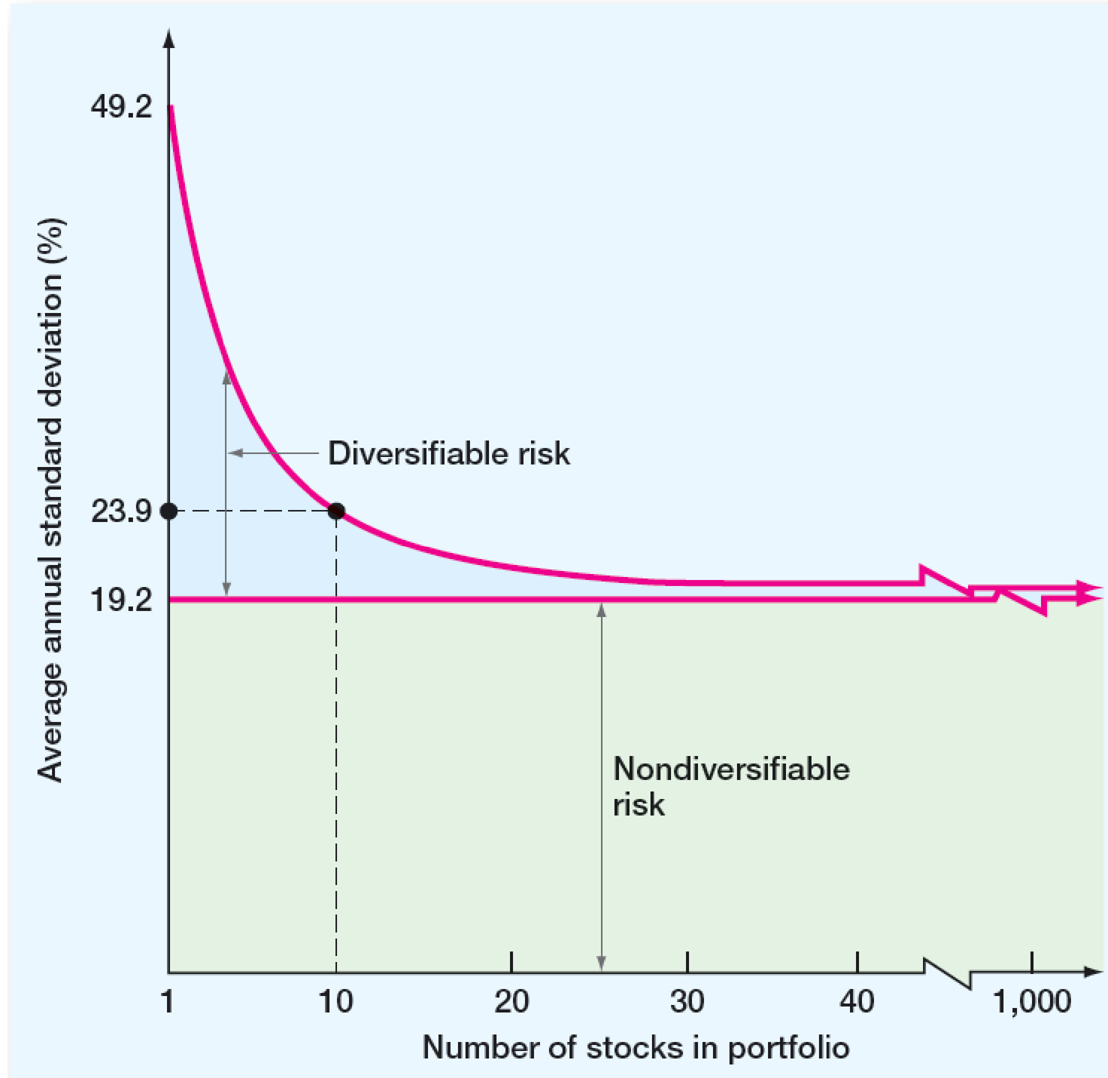
Source of variation	SS	df	MSS
Due to regression (that is, due to $X_2, X_3, \ldots, X_k$)	$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{Y}^2$	$k-1$	$\frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{Y}^2}{k-1}$
Due to residuals	$\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$	$n-k$	$\frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}}{n-k}$
Total	$\mathbf{y}'\mathbf{y} - n\bar{Y}^2$	$n-1$	

Application in Finance

Recall: Total Risk

- Total risk = systematic risk + unsystematic risk
- The standard deviation of returns is a measure of total risk.
- For well-diversified portfolios, unsystematic risk is very small.
- Consequently, the total risk for a diversified portfolio is essentially equivalent to the systematic risk.

Recall: Portfolios



Index Model Regression Equation

Excess return of security i

$$R_i(t) = \alpha_i + \beta_i R_{S\&P500}(t) + e_i(t)$$

Expected excess return when the market excess return is zero

Sensitivity of security i's return to changes in the return of the market

Expected excess return of the market

Zero-mean, firm-specific surprise in security i's return in month t.
(the *residual*)

Regression and Risk Management

- Simple Linear Regression:

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

OR

$$Y = \underbrace{\alpha + \beta X}_{\text{systematic}} + \underbrace{\varepsilon}_{\text{idiosyncratic}}$$

1) A part that <u>**CAN**</u> be explained by the model
2) A part that <u>**CANNOT**</u> be explained by the model

Regression and Risk Management

- Variance = Systematic risk + Firm-specific risk:

$$\sigma_i^2 = \beta_i^2 \sigma_M^2 + \sigma^2(e_i)$$

Why?

BACHELOR OF ECONOMICS
THAMMASAT UNIVERSITY

Question?