Statistics and Probability in equity analysis II: Linear Regressions

Class 3

FN 451

THAMMASAT
BUSINESS SCHOOL

Related topics

- Correlation vs regression coefficients: Why regressions?
- Estimation and interpretation
  - Assumptions in linear regressions
  - Estimating linear regressions
  - Hypothesis testing
  - Multiple regressions
- Regression applications in equity analysis
  - Beta estimation and hurdle rates in practice
  - Predictions
  - Market efficiency/ Event studies

Correlation vs regression coefficients

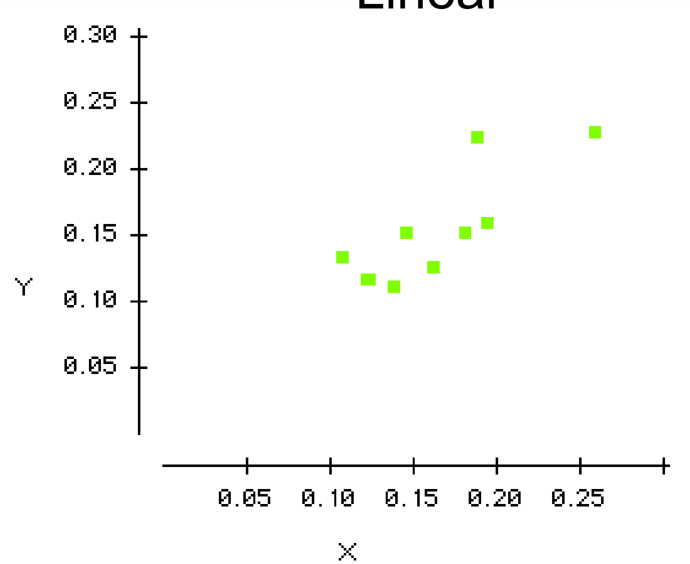
THAMMASAT
BUSINESS SCHOOL

Scatter plots

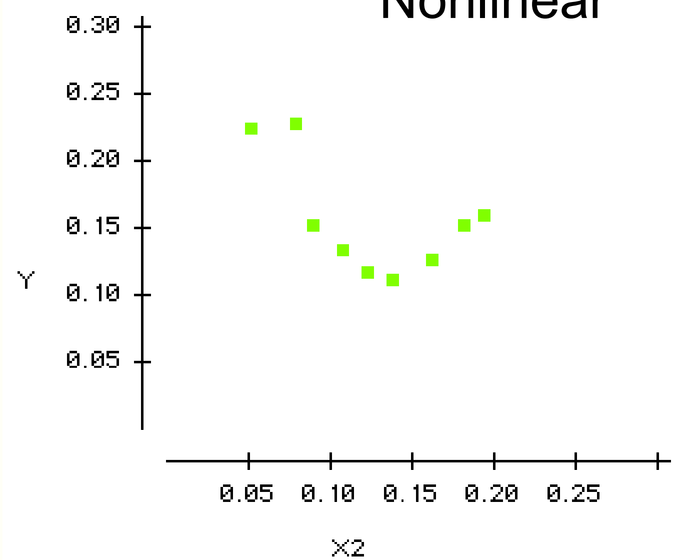
A scatter plot is a graph that shows the relationship between the observations for two data series in two dimensions.

- Scatter plots are formed by using the data from two different series to plot coordinates along the *x*- and *y*-axis, where one element of the data series forms the *x*-coordinate and the other the *y*-coordinate.

Linear



Nonlinear



THAMMASAT
BUSINESS SCHOOL

Sample covariance

Recall that covariance is the weighted average of the cross-product of each variable's departure from its mean.

- Sample covariance is calculated by using the same process as sample variance; however, rather than squaring the deviation of each observation from its mean, we take the product of two different variables' deviations from their respective means.

$$Cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

THAMMASAT
BUSINESS SCHOOL

Sample covariance

Focus On: Calculations

- Lending rates and current borrower burden are generally believed to be related. The following data cover the debt-to-income ratio for 10 borrowers and the interest rate they are being charged on five-year loans.
- What is the sample covariance between loan rate (*Y*) and debt-to-income ratio (*X*)?

Client	*Y*	*X*	Y-Yhat	X-Xhat	Product
1	0.1595	0.1952	0.0070	0.0323	0.0002
2	0.1171	0.1239	–0.0354	–0.0390	0.0014
3	0.1171	0.1229	–0.0354	–0.0400	0.0014
4	0.1269	0.1625	–0.0256	–0.0004	0.0000
5	0.1343	0.1078	–0.0182	–0.0551	0.0010
6	0.1523	0.1470	–0.0002	–0.0159	0.0000
7	0.1523	0.1823	–0.0002	0.0194	0.0000
8	0.2295	0.2599	0.0770	0.0970	0.0075
9	0.1112	0.1384	–0.0413	–0.0245	0.0010
10	0.2247	0.1890	0.0722	0.0261	0.0019
Mean	0.1525	0.1629		Sum =	0.0144
StDev	0.0427	0.0454		Cov =	0.0016

THAMMASAT BUSINESS SCHOOL

Correlation Coefficient

The correlation coefficient measures the extent and direction of a linear association between two variables.

- If the sample covariance is denoted as $s_{x,y}$, then the sample correlation coefficient is the sample covariance divided by each sample standard deviation or

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y}$$

- Continuing with our example, the sample correlation coefficient is then

$$r_{x,y} = \frac{0.0016}{0.0427(0.0454)} = 0.8253$$

- From this result, we can conclude that there is a strong linear relationship between the debt-to-income ratio of the borrowers and the loan rate they are charged. Furthermore, we can conclude that the relationship has a positive sign, indicating that an increase in the debt-to-income ratio is associated with a higher loan rate.

Spurious correlation

Spurious correlation is estimated correlation that arises because of the estimating process, not because of a fundamental underlying linear association.

Potential sources of spurious correlation:

1. Correlation between two variables that reflects chance relationships in a particular dataset.
2. Correlation induced by a calculation that mixes each of two variables with a third.
3. Correlation between two variables arising not from a direct relationship between them but from their relationship to a third variable.

See website

http://www.tylervigen.com/spurious-correlations

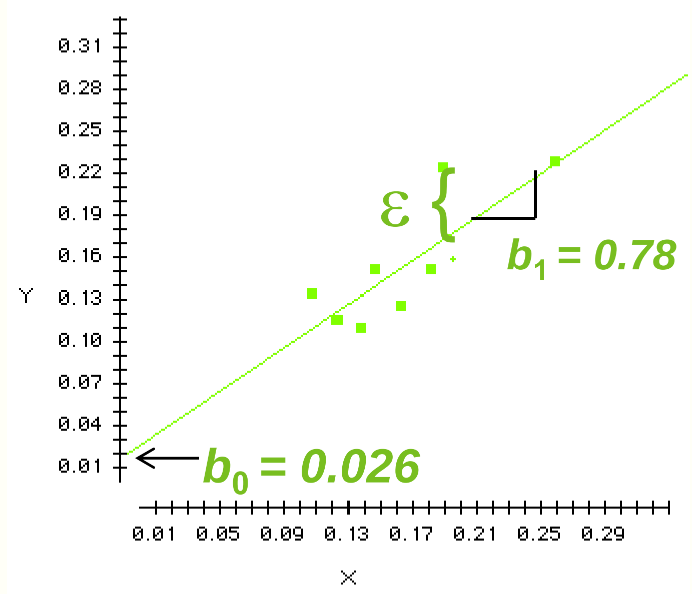
THAMMASAT
BUSINESS SCHOOL

The Basics of Linear regression

Linear regression allows us to describe one variable as a linear function of another variable.

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

- The **independent variable (X_i)** is the variable you are using to explain changes in the **dependent variable (Y_i),** the variable you are attempting to explain.
- The linear regression estimation process chooses parameter estimates to minimize the sum of the squared departures of the predicted values from the observed values.
  - b_0 is known as the intercept and b_1 is known as the slope coefficient.
  - If the value of the independent variable increases by one unit, the value of the dependent variable changes by b_1 units.



Estimation and interpretation

Assumptions underlying linear regression

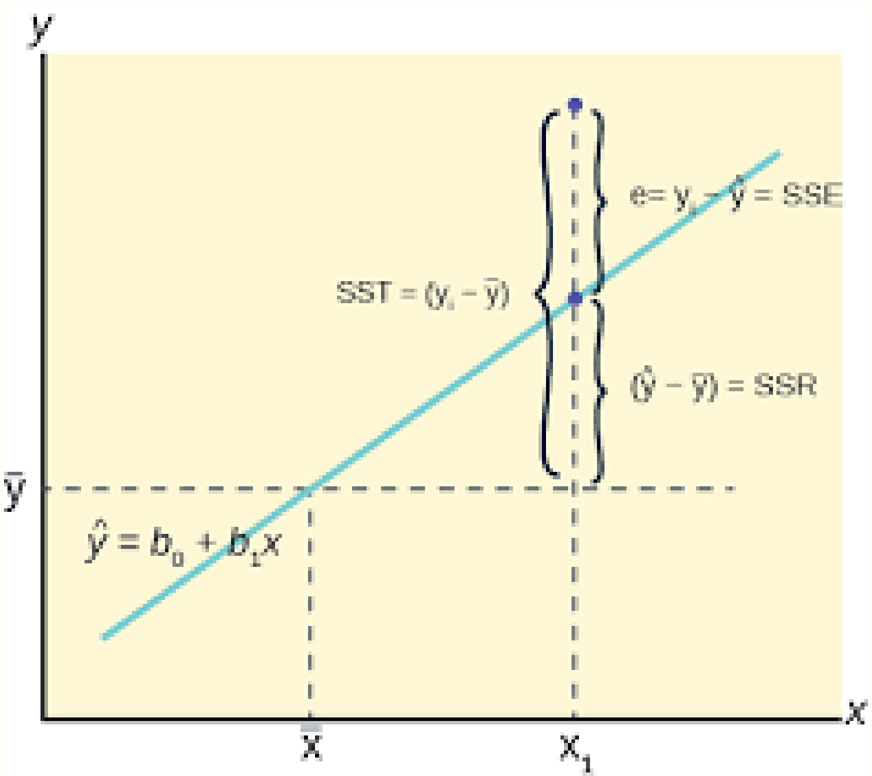
$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

1. The relationship between the dependent variable, *Y*, and the independent variable, *X*, is linear in the parameters b_0 and b_1.
2. The independent variable, *X*, is not random.
3. The expected value of the error term is 0 → $E(\varepsilon) = 0$.
4. The variance of the error term is the same for all observations.
5. The error term, ε, is uncorrelated across observations.

   Consequently, $E(\varepsilon_i, \varepsilon_j) = 0$ for all *i* not equal to *j*.
6. The error term, ε, is normally distributed.

THAMMASAT BUSINESS SCHOOL

Visualizing variation



Measures of Variation

- Total variation is made up of two parts:

$$SST \quad = \quad SSR \quad + \quad SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares

$SST = \sum(Y_i - \bar{Y})^2$ $\quad SSR = \sum(\hat{Y}_i - \bar{Y})^2$ $\quad SSE = \sum(Y_i - \hat{Y}_i)^2$

where:

$\bar{Y}$ = Average value of the dependent variable

Y_i = Observed values of the dependent variable

$\hat{Y}_i$ = Predicted value of Y for the given X_i value

Department of Statistics, ITS Surabaya

Slide-25

THAMMASAT BUSINESS SCHOOL

The Basics of Linear regression

Focus On: Regression Output

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

Regression Output	Coefficient Estimates	Standard Error	*t*-Statistic
b_0	0.0258	0.0315	0.8197
b_1	0.7774	0.1872	4.1534

$$Y_i = 0.0258 + 0.7774 X_i$$

THAMMASAT
BUSINESS SCHOOL

Standard error of the estimate

The standard error of the estimate gives us a measure of the goodness of fit for the relationship.

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

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

Client	Y	Predicted Y	Residuals²
1	0.1595	0.1776	0.0003
2	0.1171	0.1222	0.0000
3	0.1171	0.1214	0.0000
4	0.1269	0.1522	0.0006
5	0.1343	0.1096	0.0006
6	0.1523	0.1401	0.0001
7	0.1523	0.1676	0.0002
8	0.2295	0.2279	0.0000
9	0.1112	0.1334	0.0005
10	0.2247	0.1728	0.0027
		SEE =	0.0006

Coefficient of determination

The coefficient of determination is the portion of variation in the dependent variable explained by variation in the independent variable(s).

- Total variation = Unexplained variation + Explained variation; therefore, we can calculate it two ways.
  1. Square the correlation coefficient when we have one dependent and one independent variable.
  2. We can use the above relationship to determine the unexplained portion of the total variation as the sum of the squared prediction errors divided by the total variation in the dependent variable when we have more than one independent variable.
     – Because we have one independent and one dependent variable in our regression, the coefficient of determination is **$0.8253^2 = 0.6811$**.
       - The debt-to-income ratio explains 68.11% of the variation in loan rate.

THAMMASAT
BUSINESS SCHOOL

Hypothesis Testing

- Test the significance of the constant term in the same way as the slope parameter
- Although the conventional test for significance is at the 5% level, we also test at the 1% and 10% levels
- Use of the t-distribution tells us what to expect ‘by chance’
- For finite samples, when applying the t-test, we need to allow for degrees of freedom
- The t-test can be applied to either one or two tailed tests
- The t-test is an absolute value, so we can ignore the sign

The T-test

- If the t-test statistic exceeds the critical value, reject the null hypothesis, if we are testing if the coefficient equals zero, this means it is significant
- If the test statistic is below the critical value accept the null hypothesis.
- To find the critical value, you need to know the degrees of freedom, which equal n-k-1.

THAMMASAT
BUSINESS SCHOOL

Regression Hypothesis Tests

- If assumptions are met, the sampling distribution of the slope (b) approximates a T-distribution
- Standard deviation of the sampling distribution is called the standard error of the slope (σ_b)
- Population formula of standard error:

$$\sigma_b = \sqrt{\frac{\sigma_e^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}}$$

- Where σ_e^2 is the variance of the regression error

THAMMASAT BUSINESS SCHOOL

Regression Hypothesis Tests

- Estimating σ_e^2 lets us estimate the standard error:

$$\hat{\sigma}_e = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N-2}} = \sqrt{\frac{SS_{ERROR}}{N-2}} = \sqrt{MS_{ERROR}}$$

- Now we can estimate the S.E. of the slope:

$$\hat{\sigma}_b = \sqrt{\frac{MS_{ERROR}}{\sum_{i=1}^{N} (X_i - \overline{X})^2}}$$

THAMMASAT
BUSINESS SCHOOL

Regression Hypothesis Tests

- Finally: A t-value can be calculated:
  - It is the slope divided by the standard error

$$t_{N-2} = \frac{b_{YX}}{s_b}$$

- Where s_b is the sample point estimate of the standard error
- The t-value is based on N-2 degrees of freedom

THAMMASAT
BUSINESS SCHOOL

Hypotheses Test

$$H_0 : \hat{\beta} = 0$$

$$H_1 : \hat{\beta} \neq 0$$

$$T = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})} = \frac{0.4745 - 0}{0.055} = 8.6$$

Critical value is 2.04 (at df approx 30)

$3 > 2.04$, reject H_0

THAMMASAT BUSINESS SCHOOL

Significance Testing

$$\hat{y}_t = 5.54 + 0.47 x_t$$

$$(0.97) \quad (0.05)$$

$$[5.7] \qquad [8.6]$$

(Standard errors in parentheses)

[T-statistics in parentheses]

THAMMASAT
BUSINESS SCHOOL

Prediction and Linear regression

Focus On: Calculating Predicted Values

- Continuing with our example, we can calculate predicted values for our dependent variable given our estimated regression model and values for our independent variable.

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

$$Y_i = 0.0258 + 0.7774(0.18) = 0.1657$$

- If we want to predict the value of a loan rate for a borrower with a debt-to-income ratio of 18%, we substitute our estimated coefficients and a value of X = 0.18 to get

- For our estimated relationship, a borrower with an **18% debt-to-income ratio** would be expected to have a **16.57% loan rate**.

THAMMASAT
BUSINESS SCHOOL

Multiple regression

With multiple regression, we can analyze the association between more than one independent variable and our dependent variable.

Returning to our analysis of the determinants of loan rates, we also believe that the number of lines of credit the client currently employs is related to the loan rate charged. Accordingly, we model a multiple linear regression of the relationship:

- General form → $Y_i = b_0 + b_1 X_{1,i} + b_2 X_{2,i} \dots + b_k X_{k,i} + \varepsilon_i$

- Specific form → $\text{Loan rate}_i = b_0 + b_1 \text{DTI}_i + b_2 \text{Open lines}_i + \varepsilon_i$

  where DTI is the debt-to-income ratio and Open lines is the number of existing lines of credit the client already possesses.

THAMMASAT
BUSINESS SCHOOL

Multiple regression

- Coefficient estimates output

	Coefficients	Standard Error	*t*-Stat	*p*-Value
Intercept	0.0066	0.0352	0.1879	0.8563
DTI	0.7576	0.1845	4.1068	0.0045
Open lines	0.0059	0.0052	1.1429	0.2906

- The coefficient estimates are both **positive**, indicating that increases in the DTI and open lines are associated with increases in the loan rate. But only the DTI is significant at the 95% or better level as indicated by its *t*-stat of **4.1068**.
- **A 1% increase in the debt-to-income ratio leads to a 75.76 bp increase in loan rate, *holding the number of open lines constant.***

Multiple regression

Focus On: Hypothesis Testing

We can test the hypothesis that the true population slope coefficient for the association between open lines and loan rate is zero.

1. Formulate hypothesis → ***H_0: b_2 = 0 versus H_a : b_2 ≠ 0 (a two-tailed test)***

2. Identify appropriate test statistic → $t = \dfrac{\hat{b}_2 - b_2}{s_{\hat{b}_2}}$

3. Specify the significance level → **0.05 leading to a critical value of 2.4469**

4. Collect data and calculate test statistic → $t = \dfrac{0.0059 - 0}{0.0052} = 1.1429^{*}$

5. Make the statistical decision → **Fail to reject the null because 1.1429 < 2.4469**

THAMMASAT
BUSINESS SCHOOL

Multiple regression predicted values

Focus On: Calculations

Returning to our multiple linear regression, what loan rate would we expect for a borrower with an **18%** DTI and **3** open lines of credit?

$$\widehat{\text{Loan rate}}_i = 0.0066 + 0.7576(0.18) + 0.0059(3)$$

$$= 0.1607$$

THAMMASAT
BUSINESS SCHOOL

Uncertainty in linear regression

There are two sources of uncertainty in linear regression models:

1. Uncertainty associated with the **random error term**.
   - The random error term itself contains uncertainty, which can be estimated from the standard error of the estimate for the regression equation.
2. Uncertainty associated with the **parameter estimates**.
   - The estimated parameters also contain uncertainty because they are only estimates of the true underlying population parameters.
   - For a single independent variable, as covered in the prior chapter, estimates of this uncertainty can be obtained.
   - For multiple independent variables, the matrix algebra necessary to obtain such estimates is beyond the scope of this text.

Multiple regression: anova

Focus On: Regression Output

	df	SS	MSS	*F*	Significance *F*
Regression	2	0.0120	0.0060	9.6104	0.0098
Residual	7	0.0044	0.0006		
Total	9	0.0164			

- The analysis of variance section of the output provides the *F*-test for the hypothesis that all the coefficient estimates are jointly zero. The high value of this *F*-test leads us to reject the null that all the coefficients are jointly zero, concluding that at least one coefficient estimate is nonzero.
- Combined with the coefficient estimates, this model suggests that the loan rate is fairly well described by the level of the debt-to-income ratio for the client, but that the number of outstanding open lines does not make a strong contribution to that understanding.

F-test

Focus On: Calculations

- The *F*-test for a multiple regression determines whether the slope coefficients, taken together simultaneously as a group, are all zero. The test statistic is (idea is we compare 2 models 1, 2, where there are more parameters in model 2.

$$F = \frac{\text{SSR1} - \text{SSR2} / k2 - k1}{SSR2 / [n - (k2)]}$$

- From our regression output, this is 9.6104

which is greater than the critical value for an **F(0.05,2,8)=4.4590** leading us to **reject** the null hypothesis of all coefficient estimates being equal to zero.

THAMMASAT
BUSINESS SCHOOL

R^2 and Adjusted R^2

Focus On: Regression Output

- Regression specification output from our example regression provides
  - Multiple R is the correlation coefficient for the degree of association between the independent variables and the dependent variable.
  - R^2 is our familiar correlation estimate → the independent variables explain **73.3%** of the variation in the dependent variable.
    - Adjusted R^2 is a more appropriate measure for a correlation estimate that accounts for the presence of multiple independent variables and it is **65.68%**.

Regression Statistics	
Multiple R	0.8562
R^2	0.7330
Adjusted R^2	0.6568
Standard Error	0.0250
Observations	10

THAMMASAT
BUSINESS SCHOOL

Limitations of regression analysis

1. Parameter instability occurs when regression relationships change over time.
   - This instability generally occurs when the underlying population from which the sample is drawn has changed fundamentally in some way.
   - Example: regime shifts in regulatory or monetary policy
2. Public knowledge of the relationships may decrease or eliminate their usefulness.
3. Violation of the underlying assumptions makes hypothesis tests and prediction intervals invalid, and we may not be certain as to whether the assumptions have been violated.

THAMMASAT
BUSINESS SCHOOL

Summary

- We are often interested in knowing the extent of the relationship between two or more financial variables.
- We can assess this relationship in several ways, including
    - correlation, which measures the degree to which two variables move together, and
    - linear regression, which describes at a more fundamental level the nature of any linear relationship between two variables.
- We can combine hypothesis testing with linear regression and correlation to test beliefs about the nature and extent of relationships between two or more variables

THAMMASAT
BUSINESS SCHOOL