Introduction to Regression

Rob Hayward

February 15, 2014

Outline

Model for securities

Model security return. You have thought about this already as it is an important component of

- CAPM

Model for securities

Model security return. You have thought about this already as it is an important component of

- CAPM
  - Beta is the relationship between securities returns and the market

Model for securities

Model security return. You have thought about this already as it is an important component of

- CAPM
  - Beta is the relationship between securities returns and the market
- Diversification

Model for securities

Model security return. You have thought about this already as it is an important component of

- CAPM
  - Beta is the relationship between securities returns and the market
- Diversification
  - Distinguish market risk and ideosyncratic risk

Model for securities

Model security return. You have thought about this already as it is an important component of

- CAPM
  - Beta is the relationship between securities returns and the market
- Diversification
  - Distinguish market risk and ideosyncratic risk
- EMH

Model for securities

Model security return. You have thought about this already as it is an important component of

- CAPM
  - Beta is the relationship between securities returns and the market
- Diversification
  - Distinguish market risk and ideosyncratic risk
- EMH
  - What is the ideosyncratic or individual performance of the security?

Modelling

The model

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Modelling

The model

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Where

- y_t is the dependent variable

Modelling

The model

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Where

- y_t is the dependent variable
- α is an intercept or constant

Modelling

The model

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Where

- y_t is the dependent variable
- α is an intercept or constant
- x_t is the explanatory or independent variable(s)

Modelling

The model

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Where

- y_t is the dependent variable
- α is an intercept or constant
- x_t is the explanatory or independent variable(s)
- β is the key relationship

Modelling

The model

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Where

- y_t is the dependent variable
- α is an intercept or constant
- x_t is the explanatory or independent variable(s)
- β is the key relationship
- ε_t is the error that covers omitted variables, measurement error and other stochastic or random elements

Modelling

The model

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Modelling

The model

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Where

- y_t is the return of Bank of America

Modelling

The model

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Where

- y_t is the return of Bank of America
- α is an intercept or constant

Modelling

The model

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Where

- y_t is the return of Bank of America
- α is an intercept or constant
- x_t is the return of the market (S&P 500)

Modelling

The model

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Where

- y_t is the return of Bank of America
- α is an intercept or constant
- x_t is the return of the market (S&P 500)
- β is the relationship between BAC returns and the market returns

Modelling

The model

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Where

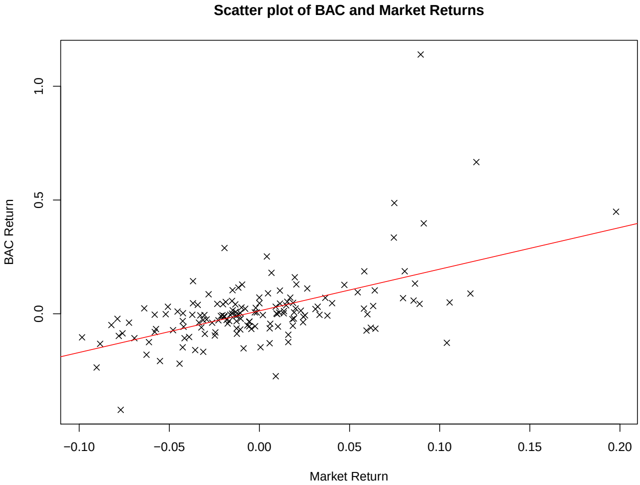
- y_t is the return of Bank of America
- α is an intercept or constant
- x_t is the return of the market (S&P 500)
- β is the relationship between BAC returns and the market returns
- ε_t is all the other factors that affect BAC returns

Caution!

> *"Essentially all models are wrong, but some are useful"*

(Box, 1987, p. 424)

S&P 500 and BAC



Solution 1

$y_t = a + bx_t + u_t$

Minimise the residuals

$$Min \sum_{t=1}^{t=T} u_t^2$$

$$Min \sum_{t=1}^{t=T} (y_t - a - bx_t)^2$$

Take, partial derivative to get the conditions.

Solution 2

$$\frac{\delta u}{\delta a} = \sum_{t=1}^{t=T} 2(y_t - a - bx_t) = 0$$

$$\frac{\delta u}{\delta b} = \sum_{t=1}^{t=T} 2x_t(y_t - a - bx_t) = 0$$

The following can be skipped. It comes from the appendix to chapter 4 of Econometrics (Stock and Watson)

Solution 3

collecting terms and dividing by n. Remember that summing and dividing by T is equal to the mean.

$$\bar{y} = a + b\bar{x} = 0 \tag{1}$$

and

$$\frac{1}{T}\sum_{t=1}^{T} x_t y_t - a\bar{x} - b\frac{1}{T}\sum_{t-1}^{T} x_t^2 = 0 \tag{2}$$

Substitute Equation 1 into 2

$$\frac{1}{T}\sum x_t y_t - (\bar{y} - b\bar{x})\,\bar{x} - b\frac{1}{T}\sum_{t-1}^{T} x_t^2$$

Solution 4

$$\frac{1}{T}\sum x_t y_t - \bar{y}\bar{x} + b\bar{x}^2 - b\frac{1}{T}\sum x^2$$

$$\frac{1}{T}\sum x_t y_t - \bar{y}\bar{x} + b\left(\bar{x}^2 - \frac{1}{T}\sum x^2\right)$$

$$b = \frac{\frac{1}{T}\sum x_t y_t - \bar{y}\bar{x}}{\frac{1}{T}\sum x^2 - \bar{x}^2}$$

$$b = \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sum(x_t - \bar{x})^2}$$

Solution: matrix form

In matrix form

$$\begin{aligned}\mathbf{y} =& \mathbf{X}\beta + \mathbf{u}\\ \mathbf{u} =& \mathbf{y} - \mathbf{X}\beta\\ \mathbf{u}'\mathbf{u} =& (\mathbf{X}\beta + \mathbf{u})'(\mathbf{X}\beta + \mathbf{u})\end{aligned}$$

Taking derivative and re-arranging (see textbook for proof)

$$\beta = (\mathbf{X}`\mathbf{X})^{-1}\mathbf{X}`\mathbf{y}$$

Regression Table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0130	0.0105	1.23	0.2203
SPY.R	1.8303	0.2240	8.17	0.0000

The Adjusted R^2 is 0.29, therefore nearly 30% of the BAC returns are explained by the returns of the market. 95% confidence intervals for the β are 1.39 to 2.27.

Ordinary Least Squares (OLS)

There are two important qualities that are desiable in an *estimator*

- Unbiased: the expected value of the estimator is equal to the population value

Ordinary Least Squares (OLS)

There are two important qualities that are desiable in an *estimator*

- Unbiased: the expected value of the estimator is equal to the population value
- Efficient: Estimates should be as close as possible to the true value

Ordinary Least Squares (OLS)

There are two important qualities that are desiable in an *estimator*

- Unbiased: the expected value of the estimator is equal to the population value
- Efficient: Estimates should be as close as possible to the true value

Ordinary Least Squares (OLS)

There are two important qualities that are desiable in an *estimator*

- Unbiased: the expected value of the estimator is equal to the population value
- Efficient: Estimates should be as close as possible to the true value

Given a number of assumptions OLS is the BLUE **B**est, **L**inear, **U**nbiased, **E**stimator.

OLS Assumptions

The assumptions for BLUE OLS qualities

OLS Assumptions

The assumptions for BLUE OLS qualities

- The errors have a zero mean

OLS Assumptions

The assumptions for BLUE OLS qualities

- The errors have a zero mean
- The errors are *independent and identically distributed* (iid)

OLS Assumptions

The assumptions for BLUE OLS qualities

- The errors have a zero mean
- The errors are *independent and identically distributed* (iid)
  - No serial correlation (errors related to each other)

OLS Assumptions

The assumptions for BLUE OLS qualities

- The errors have a zero mean
- The errors are *independent and identically distributed* (iid)
  - No serial correlation (errors related to each other)
  - Hetroskedasticity (some errors are systematically larger than others)

OLS Assumptions

The assumptions for BLUE OLS qualities

- The errors have a zero mean
- The errors are *independent and identically distributed* (iid)
  - No serial correlation (errors related to each other)
  - Hetroskedasticity (some errors are systematically larger than others)
- Explanatory variables are not related to the error

OLS Assumptions

The assumptions for BLUE OLS qualities

- The errors have a zero mean
- The errors are *independent and identically distributed* (iid)
  - No serial correlation (errors related to each other)
  - Hetroskedasticity (some errors are systematically larger than others)
- Explanatory variables are not related to the error
- Additionally, assume *normal errors* if we want to use normal assumption to compute *t-tests* of coefficients

Issues

Therefore, there are a number of potential problems

- Functional form

Issues

Therefore, there are a number of potential problems

- Functional form
    - Linear form

Issues

Therefore, there are a number of potential problems

- Functional form
    - Linear form
    - Missing or superfluous variables

Issues

Therefore, there are a number of potential problems

- Functional form
    - Linear form
    - Missing or superfluous variables
    - Structural breaks

Issues

Therefore, there are a number of potential problems

- Functional form
  - Linear form
  - Missing or superfluous variables
  - Structural breaks
- Evidence of problems

Issues

Therefore, there are a number of potential problems

- Functional form
  - Linear form
  - Missing or superfluous variables
  - Structural breaks
- Evidence of problems
  - Serial correlation in the residuals

Issues

Therefore, there are a number of potential problems

- Functional form
  - Linear form
  - Missing or superfluous variables
  - Structural breaks
- Evidence of problems
  - Serial correlation in the residuals
  - Hetroscedasticity in the residuals

Model Problems: Missing Variables

If explanatory variables are missing

Model Problems: Missing Variables

If explanatory variables are missing

- Does theory suggest other variables are important?

Model Problems: Missing Variables

If explanatory variables are missing

- Does theory suggest other variables are important?
- Tests of structural form and residuals will indicate problems

Model Problems: Missing Variables

If explanatory variables are missing

- Does theory suggest other variables are important?
- Tests of structural form and residuals will indicate problems
- Estimates of coefficients will be biased if there is a relationship between estimated and missing variables

Model Problems: Missing Variables

If explanatory variables are missing

- Does theory suggest other variables are important?
- Tests of structural form and residuals will indicate problems
- Estimates of coefficients will be biased if there is a relationship between estimated and missing variables
- Estimated errors will be too large or too small

Model Problems: Missing Variables

If explanatory variables are missing

- Does theory suggest other variables are important?
- Tests of structural form and residuals will indicate problems
- Estimates of coefficients will be biased if there is a relationship between estimated and missing variables
- Estimated errors will be too large or too small

Model Problems: Missing Variables

If explanatory variables are missing

- Does theory suggest other variables are important?
- Tests of structural form and residuals will indicate problems
- Estimates of coefficients will be biased if there is a relationship between estimated and missing variables
- Estimated errors will be too large or too small

Solution: Add missing variable or a proxy

Model Problems: Superfluous Variables

Unnecessary variables

Model Problems: Superfluous Variables

Unnecessary variables

- Theory and the t-statistic should guide

Model Problems: Superfluous Variables

Unnecessary variables

- Theory and the t-statistic should guide
- Estimates are unbiased but inefficient.

Model Problems: Superfluous Variables

Unnecessary variables

- Theory and the t-statistic should guide
- Estimates are unbiased but inefficient.
- Multicolinearity is a problem

Model Problems: Superfluous Variables

Unnecessary variables

- Theory and the t-statistic should guide
- Estimates are unbiased but inefficient.
- Multicolinearity is a problem
    - Measuring the same thing twice

Model Problems: Superfluous Variables

Unnecessary variables

- Theory and the t-statistic should guide
- Estimates are unbiased but inefficient.
- Multicolinearity is a problem
  - Measuring the same thing twice
  - Singular matrix

Model Problems: Superfluous Variables

Unnecessary variables

- Theory and the t-statistic should guide
- Estimates are unbiased but inefficient.
- Multicolinearity is a problem
  - Measuring the same thing twice
  - Singular matrix

Model Problems: Superfluous Variables

Unnecessary variables

- Theory and the t-statistic should guide
- Estimates are unbiased but inefficient.
- Multicolinearity is a problem
  - Measuring the same thing twice
  - Singular matrix

Solution: Remove superfluous variable. Be careful of *the dummy variable problem*

Appropriate model

There are two additional issues to be aware of

Appropriate model

There are two additional issues to be aware of

- Linear model

Appropriate model

There are two additional issues to be aware of

- Linear model
  - Non-linear relationship

Appropriate model

There are two additional issues to be aware of

- Linear model
    - Non-linear relationship
    - Can variables be transformed (logs)

Appropriate model

There are two additional issues to be aware of

- Linear model
    - Non-linear relationship
    - Can variables be transformed (logs)
- Structural breaks

Appropriate model

There are two additional issues to be aware of

- Linear model
  - Non-linear relationship
  - Can variables be transformed (logs)
- Structural breaks
  - Shifts in parameters

Appropriate model

There are two additional issues to be aware of

- Linear model
  - Non-linear relationship
  - Can variables be transformed (logs)
- Structural breaks
  - Shifts in parameters
  - Use dummy variables

Textbooks

All in the library

Textbooks

All in the library

- C. Dougherty, "Introduction to Econometrics", OUP

Textbooks

All in the library

- C. Dougherty, "Introduction to Econometrics", OUP
- JH Stock and M Watson "Introduction to Econometrics", Pearson

Textbooks

All in the library

- C. Dougherty, "Introduction to Econometrics", OUP
- JH Stock and M Watson "Introduction to Econometrics", Pearson
- D Gujarati, "Basic Econometrics", McGraw-Hill

Bibliography

Box, G. E. (1987), *Empirical Model Building and Response Surfaces*, John Wiley and sons.