Data Mining for Business

Linear Regression Model

Dr. Shipra Maurya

Department of Management Studies

IIT (ISM) Dhanbad

Email: shipra@iitism.ac.in

Linear Regression Model

- Difference between Correlation and Regression
- Difference between Simple Linear Regression and Multiple Linear Regression Models
- What does linear mean in Linear Regression Model –

Linear in Parameter	Linear in Variables	
	Yes	No
Yes	Linear Regression Model (LRM)	LRM
No	Non-Linear Regression Model (NLRM)	NLRM

- Difference between Population Regression Function (PRF) and Sample Regression Function (SRF)
- Difference between Parameter, Estimator and Estimate

Linear Regression Model…

- Difference between Point estimate and Interval estimate
- What is Residual term/error
- Significance of Error term/ Reasons for including error term
- Regression equation estimation methods:
    - Ordinary Least Square (OLS)
    - Maximum Likelihood Method (MLE)

Ordinary Least Square Method (OLS)

- OLS method is used to estimate the appropriate values of the coefficients α and β in the Linear Regression equation
- OLS entails taking each vertical distance from the point to the line, squaring it and then minimizing the total sum of the areas of squares or RSS or SSR (hence ‘least squares’)

$$y_t = \alpha + \beta x_t + u_t \quad \text{PRF}$$

$$\left.\begin{aligned} \hat{y}_t &= \hat{\alpha} + \hat{\beta} x_t \\ y_t &= \hat{\alpha} + \hat{\beta} x_t + \hat{u}_t \end{aligned}\right\} \text{SRF}$$

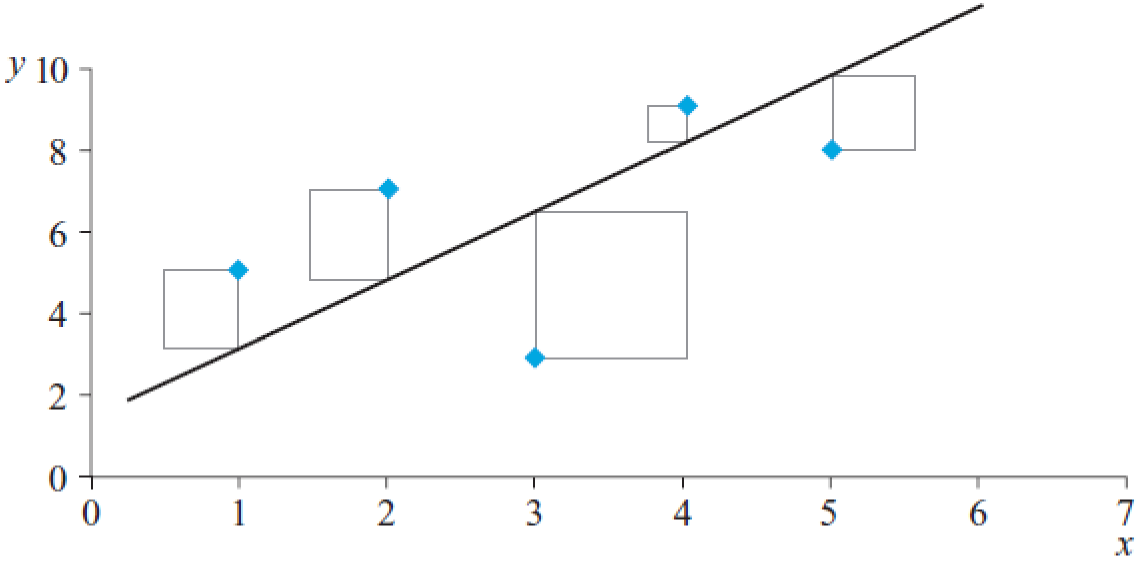


Figure 3.3 Method of OLS fitting a line to the data by minimising the sum of squared residuals

Image Source: Introductory Econometrics for Finance by Chris Brook

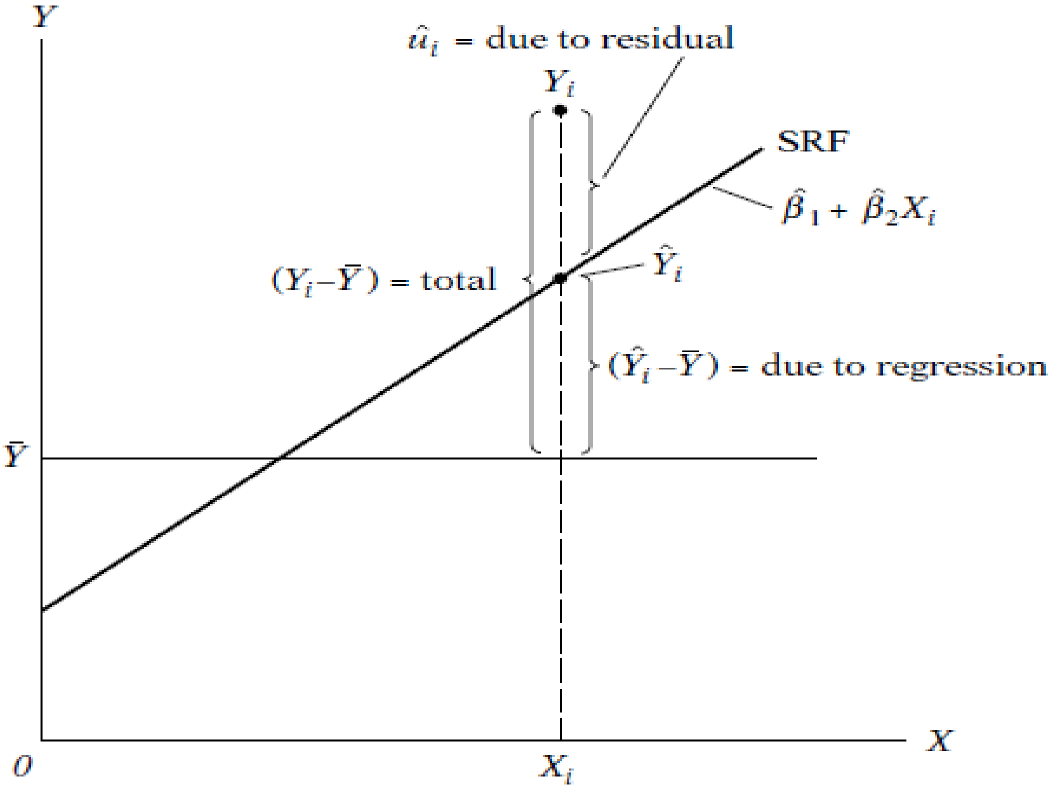
Ordinary Least Square Method (OLS)

- So minimizing the sum of squared distances is given by minimizing

$$\left(\sum_{t=1}^{T} \hat{u}_t^2\right)$$

- This sum is known as RSS. L represents RSS or the loss function that needs to be minimized w.r.t. α^ and β^

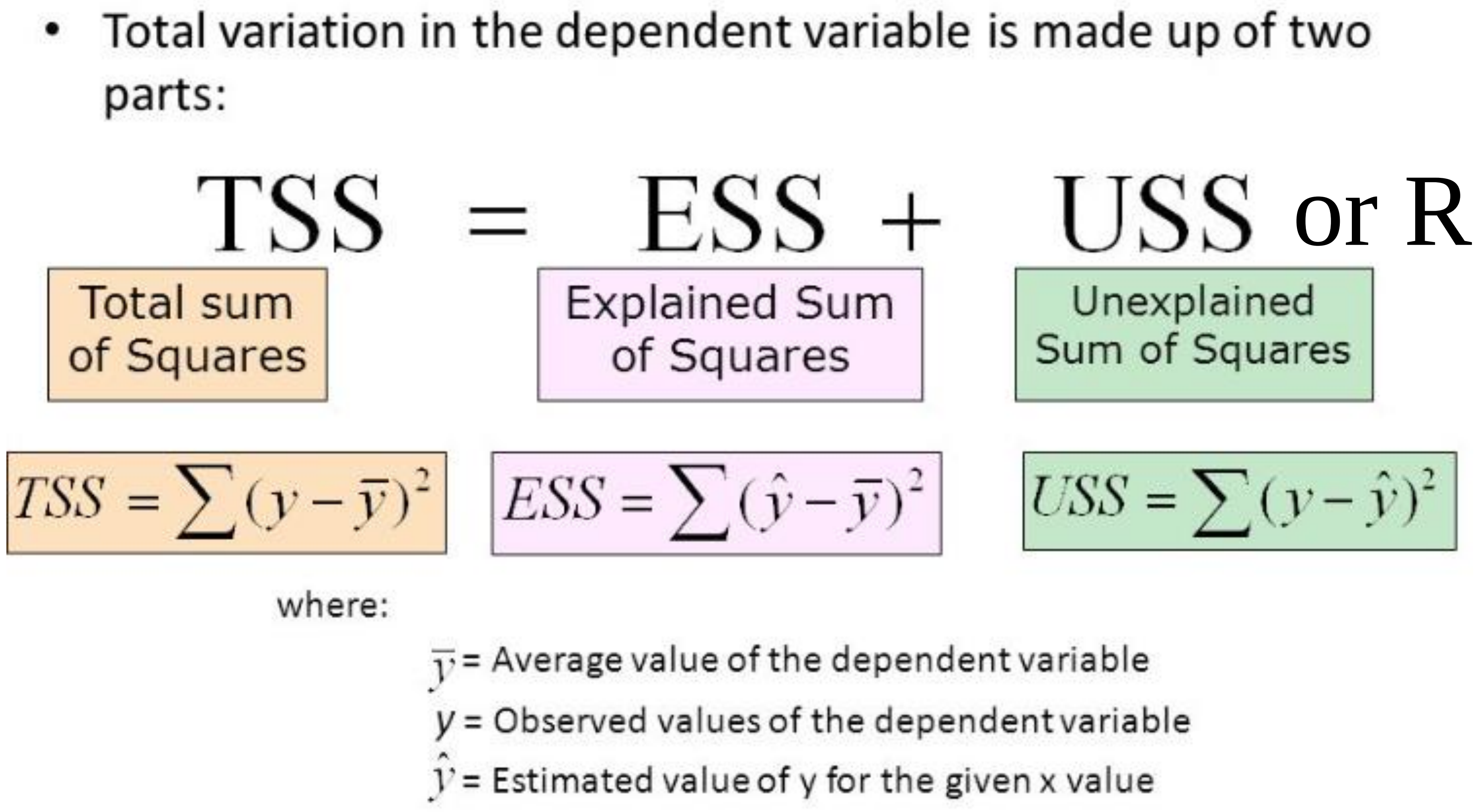
$$L = \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 = \sum_{t=1}^{T} (y_t - \hat{\alpha} - \hat{\beta} x_t)^2$$



Y
$\hat{u}_i$ = due to residual
Y_i
SRF
$\hat{\beta}_1 + \hat{\beta}_2 X_i$
$(Y_i - \bar{Y})$ = total
$\hat{Y}_i$
$(\hat{Y}_i - \bar{Y})$ = due to regression
$\bar{Y}$
0
X_i
X

Image Source: Introductory Econometrics for Finance by Chris Brook

Explained and Unexplained Variation



- Total variation in the dependent variable is made up of two parts:

$$\text{TSS} = \text{ESS} + \text{USS} \text{ or RSS}$$

Total sum of Squares

Explained Sum of Squares

Unexplained Sum of Squares

$$TSS = \sum (y - \bar{y})^2$$

$$ESS = \sum (\hat{y} - \bar{y})^2$$

$$USS = \sum (y - \hat{y})^2$$

where:

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

y = Observed values of the dependent variable

$\hat{y}$ = Estimated value of y for the given x value

Image Source: Internet

Ordinary Least Square Method (OLS)

- The coefficients are estimated using following equations:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

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

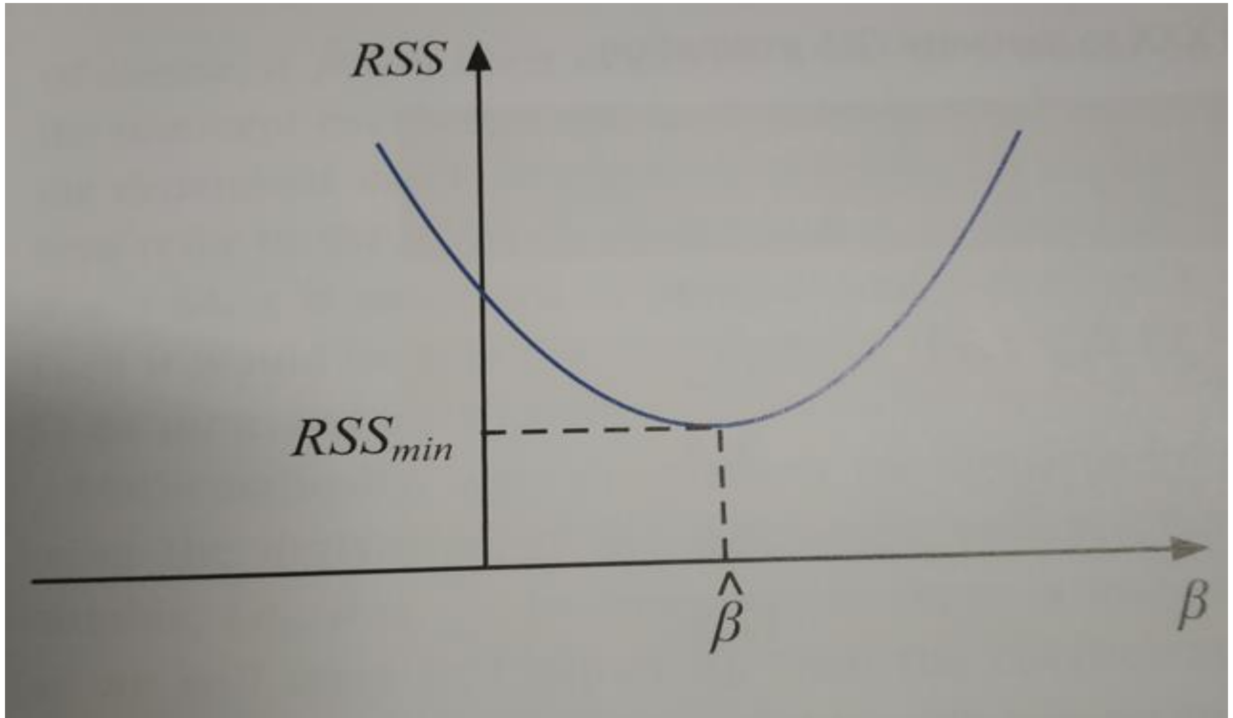


Figure 3.5 How RSS varies with different values of β

Image Source: Introductory Econometrics for Finance by Chris Brook

BLUE Properties of OLS Estimator

- **Best** - means that the OLS estimator β^ has minimum variance among the class of linear unbiased estimators
- **Linear** - α^ and β^ are linear estimators – that means that the formulae for α^ and β^ are linear combinations of the random variables
- **Unbiased** - on average, the actual values of α^ and β^ will be equal to their true values
- **Estimator** - α^ and β^ are estimators of the true value of α and β

Image Source: Introductory Econometrics for Finance by Chris Brook

Standard errors (1/2)

- It is a measure of precision for regression estimates α^ and β^.
- Standard error helps understand the impact of sample variability on regression estimates.
- It gives only a general indication of the likely accuracy of the regression parameters. They do not show how accurate a particular set of coefficient estimates is.
- If the standard errors are small, it shows that the coefficients are likely to be precise on average, not how precise they are for this particular sample.

Source: Introductory Econometrics for Finance by Chris Brook

Standard Errors (2/2)

- Everything else being equal, the smaller this quantity is, the closer is the fit of the line to the actual data
- T is the sample size, s is the estimated standard deviation of error term, x is the independent variable/feature

$$SE(\hat{\alpha}) = s\sqrt{\frac{\sum x_t^2}{T\sum (x_t - \bar{x})^2}} = s\sqrt{\frac{\sum x_t^2}{T\left(\left(\sum x_t^2\right) - T\bar{x}^2\right)}} \qquad s = \sqrt{\frac{\sum \hat{u}_t^2}{T-2}}$$

$$SE(\hat{\beta}) = s\sqrt{\frac{1}{\sum (x_t - \bar{x})^2}} = s\sqrt{\frac{1}{\sum x_t^2 - T\bar{x}^2}}$$

Image Source: Introductory Econometrics for Finance by Chris Brook

Statistical Inference Approaches

Test of significance approach

Step 1 – Find **t-statistic** = $\frac{\beta^{\wedge} - \beta}{SE(\beta^{\wedge})}$

H0 : β = 0

H1 : β ≠ 0

Step 2 – Find **t-critical value** for the subjectively defined level of significance (usually denoted as α – not intercept) and degree of freedom

Step 3 – if t-statistic lies in the rejection region then reject the null hypothesis (H0), else do not reject H0

Confidence interval approach

Step 1 – Find **t-critical value** for the subjectively defined level of significance (usually denoted as α – not intercept) and degree of freedom

Step 2 – Find the confidence interval for β using below equation:

$$(\hat{\beta} - t_{crit} \cdot SE(\hat{\beta}), \hat{\beta} + t_{crit} \cdot SE(\hat{\beta}))$$

Step 3 – if beta^ values lies within the confidence interval then do not reject the null hypothesis (H0)

Source: Introductory Econometrics for Finance by Chris Brook

Test of significance

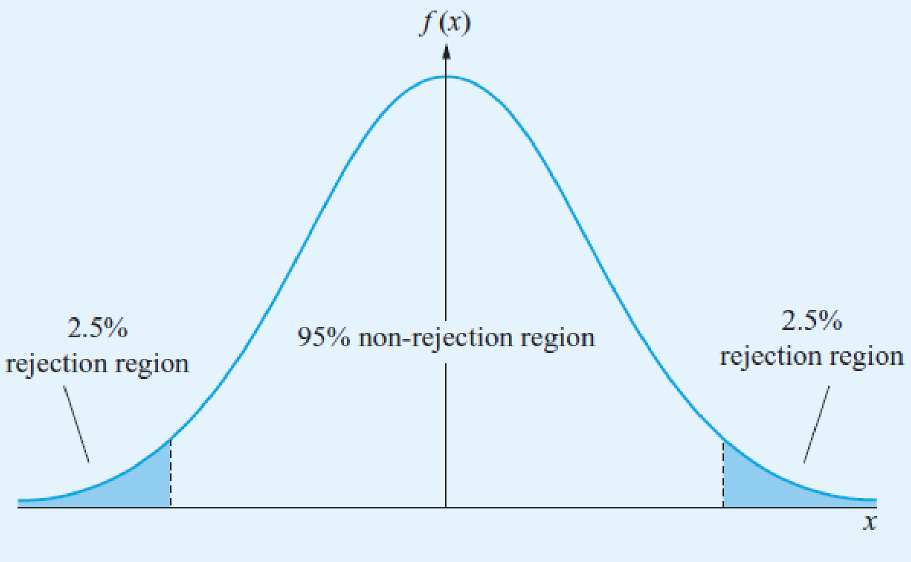


Figure 3.12 Rejection regions for a two-sided 5% hypothesis test

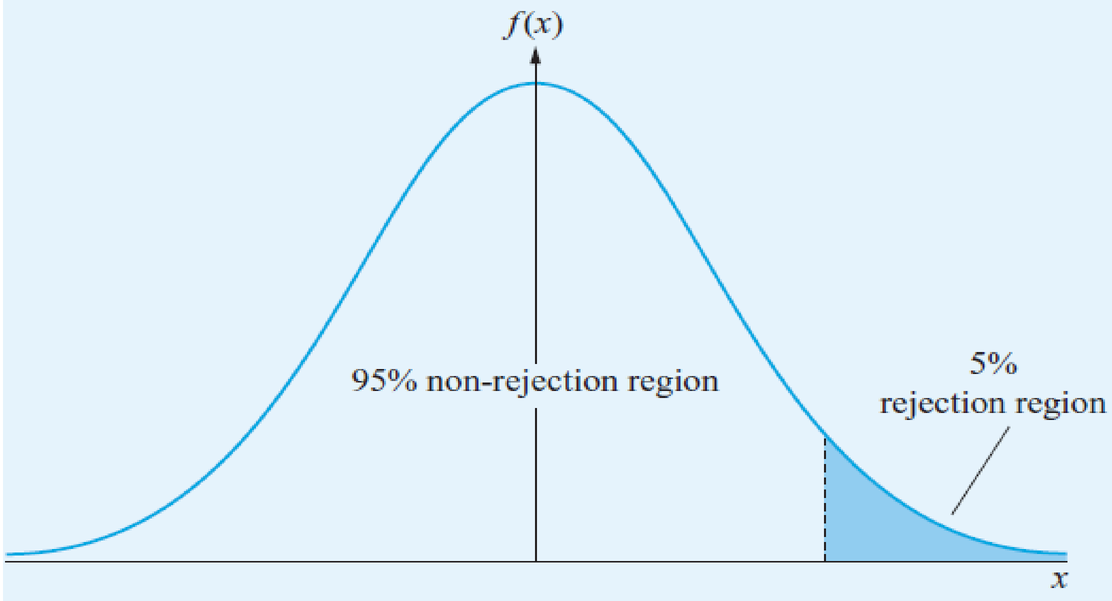


Figure 3.14 Rejection region for a one-sided hypothesis test of the form $H_0 : \beta = \beta^*$, $H_1 : \beta > \beta^*$

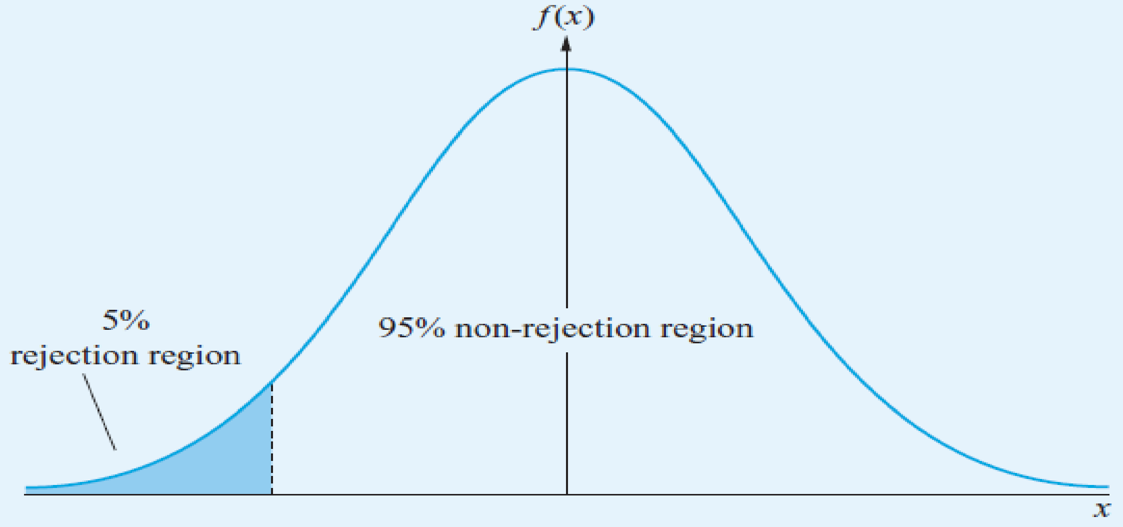


Figure 3.13 Rejection region for a one-sided hypothesis test of the form $H_0 : \beta = \beta^*$, $H_1 : \beta < \beta^*$

P-value

- P-value is also known as the exact level of significance
- It is different from level of significance (α) as α is determined outside the model and is dependent on the researcher. The most chosen values of α are 1%, 5% and 10%
- 1 – level of significance (α) = Confidence interval
- **How to calculate the p-value for t-test in excel**: =TDIST(T-statistic value, number of degrees of freedom, one/two tail test)
- If p-value > level of significance then do not reject the null hypothesis and vice-versa

Goodness of Fit measures

- **T-test statistic** – for measuring significance of individual features in the model
- **R^2 and Adjusted R^2** – indicates how well the model explains variations in the target variable – details have already been discussed in earlier topic. The value ranges between 0 and 1
- **F-test statistic** – for measuring the overall model performance

Note - 1. single hypotheses involving one coefficient can be tested using a t- or an F-test, but multiple hypotheses can be tested only using an F-test

2. R^2 is also known as coefficient of determination

F-test

$\boldsymbol{H_0}$: $\beta_1 = \beta_2 = \ldots. = \beta_k = 0$ (no linear relationship – none of the features impact target variable)

$\boldsymbol{H_1}$: $atleast\ one\ \beta_i \neq 0$ (atleast one feature impacts target variable)

F-statistic =$\dfrac{Explained\ variance}{Unexplained\ variance}$ OR $test\ statistic = \dfrac{RRSS - URSS}{URSS} \times \dfrac{T - k}{m}$

- URSS = residual sum of squares from unrestricted regression
- RRSS = residual sum of squares from restricted regression
- m = number of restrictions
- T = number of observations
- k = number of regressors in unrestricted regression including the constant

Assumptions of CLRM

- Since we have u_t (error term) as well in the model, there are certain assumptions that are made w.r.t. u_t

1. E(ut) = 0 : The errors have zero mean
2. var(ut) = σ2 < ∞ : The variance of the errors is constant and finite over all values of xt - Homoscedasticity
3. cov(ui, uj) = 0 : The errors are linearly independent of one another – No autocorrelation
4. cov(ut, xt) = 0 : There is no relationship between the error and corresponding x variate – No Endogeneity
5. ut ~ N(0, σ2) : means that ut is normally distributed

Heteroscedasticity

- If the variance of errors is not constant, it is known as Heteroscedasticity
- **Consequence of Heteroscedasticity:** OLS estimators will still give unbiased coefficient estimates, but they are no longer BLUE and will not have minimum variance
- **Detecting Heteroscedasticity:** Residual Graph, Goldfeld-Quandt Test and White's Test
- **Solutions for Heteroscedasticity:**
  - Transforming the variables into logs
  - To correct the standard error for heteroscedasticity, use "robust" option in econometrics packages. Robust option uses White's test to correct the standard errors

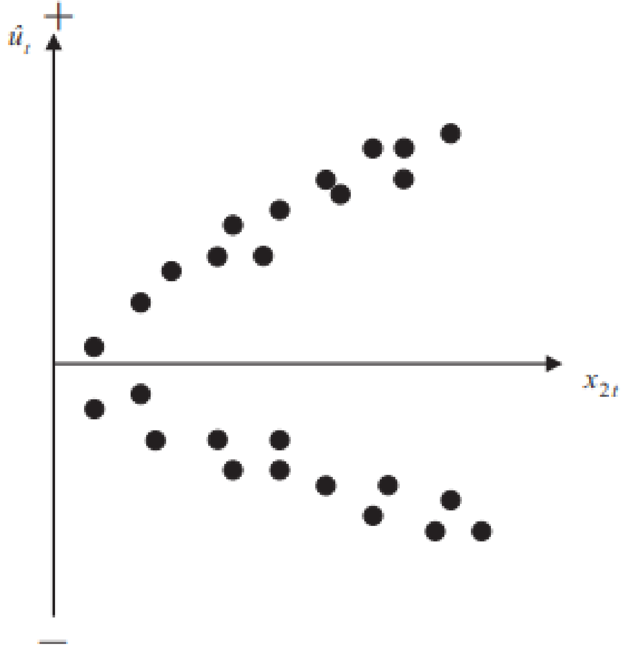
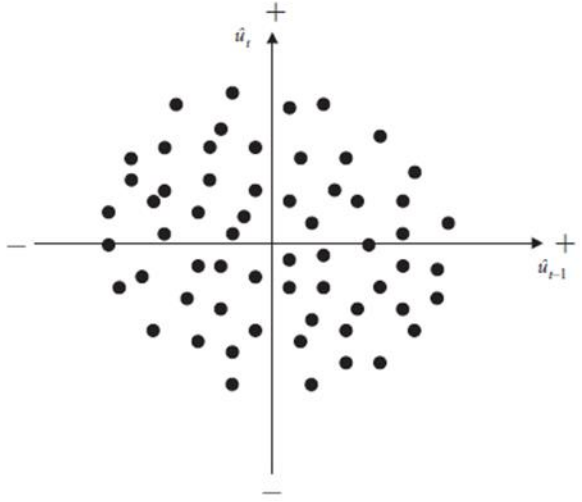


Image Source: Introductory Econometrics for Finance by Chris Brook

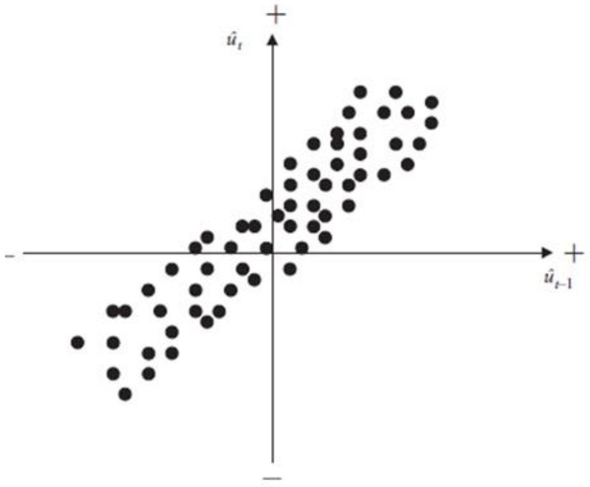
Autocorrelation (1/2)

- If the errors are correlated with each other, it is an indication of presence of autocorrelation
- Concept of lagged value
- **Detection of Autocorrelation** – Residual graphs, Durbin-Watson test, Breusch-Godfrey Test
- **Consequences of Autocorrelation** - OLS estimators will still give unbiased coefficient estimates, but they are no longer BLUE and will not have minimum variance
- **Solution for Autocorrelation** – use Cochrane-Orcutt procedure

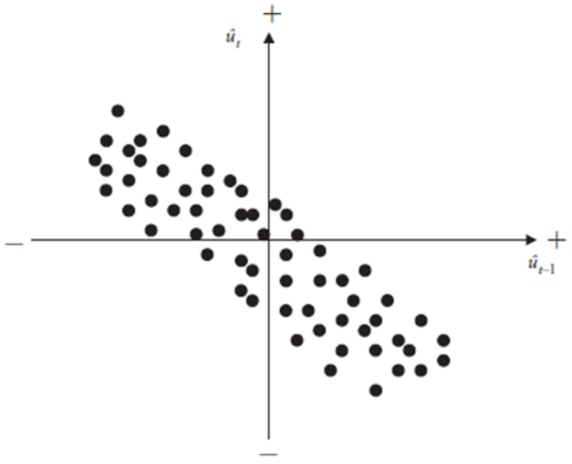
Autocorrelation (2/2)



No autocorrelation



Positive autocorrelation



Negative autocorrelation

Endogeneity

- If the independent variables are correlated with the error term, this is known as endogeneity issue
- **Consequences of Endogeneity**: Results in biased and inconsistent parameter estimates and a fitted line that appears to capture the features of the data much better than it does in reality
- **Causes of Endogeneity**: omission of features and simultaneous causality
- **How to detect Endogeneity**: Hausman Test for Endogeneity
- **Solution of Endogeneity:** introducing instrumental variables, using methods which model endogenous variables

Normality Assumption

- ut ~ N(0, σ2) : means that ut is normally distributed
- **Testing for normality** – Jarque-Berra test

$$W = T\left[\frac{b_1^2}{6} + \frac{(b_2-3)^2}{24}\right] \qquad b_1 = \frac{E[u^3]}{(\sigma^2)^{3/2}} \quad \text{and} \quad b_2 = \frac{E[u^4]}{(\sigma^2)^2}$$

- b1 represents Skewness coefficient, b2 represents Kurtosis coefficient, T is sample size
- H0: The data is normally distributed
- H1: The data is not normally distributed

Multicollinearity (1/2)

- An implicit assumption when using OLS estimation method
- If independent variables are highly correlated with each other, it is known as Multicollinearity
- **Detecting Multicollinearity-** Correlation and VIF (Variance Inflation Factor). Any value of VIF exceeding 10 (some researchers also use threshold of 5) indicates a problem of multicollinearity
- **Consequences**- R^2 will be high but the individual coefficients will have high standard errors, so that the regression ‘looks good’ as a whole, but the individual variables are not significant.

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

Source: Introductory Econometrics for Finance by Chris Brook

Multicollinearity (2/2)

- **Solution to Multicollinearity:**
  - Drop one of the collinear variables,
  - Transform the highly correlated variables into a ratio and include only the ratio,
  - Use PCA (Principal Component Analysis),
  - Ridge regression

Source: Introductory Econometrics for Finance by Chris Brook

Thank you!