Queen Mary
University of London

Maths & Stats Pre-Sessional

Basics of Regression Analysis

Lecturer: Claudio Vallar
School of Economics and Finance

Basics of Regression Analysis

In this session:

- Regression is a statistical method that attempts to fit a model to data to quantify the relationship between the dependent (outcome) variables and the predictor (independent) variable(s).
- Two-variable linear regression.
- Multiple-variable linear regression.

For more extensive reading, refer to Chapter 11 and 12 of Newbold, P., Carlson, W., and Thorne, B. (2010). Statistics for Business and Economics, Pearson, 7th Edition

Two-Variable Linear Regression

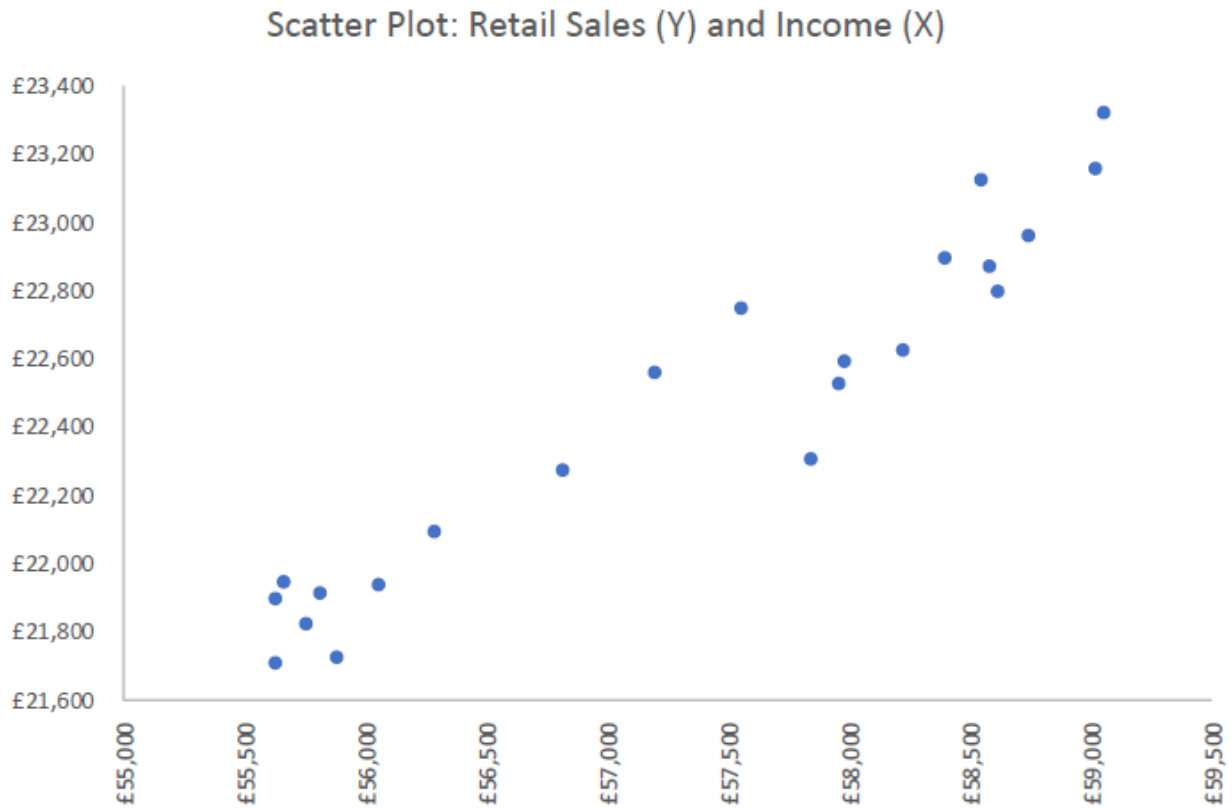
Example:

Suppose we manage 22 retail stores in 22 different locations. We have data on the disposable income per household (X) and retail sales per household (Y) by store/location in the following table.

Retail Store	Income (X)	Retail Sales (Y)	Retail Store	Income (X)	Retail Sales (Y)
1	£ 55,641	£ 21,886	12	£ 57,850	£ 22,301
2	£ 55,681	£ 21,934	13	£ 57,975	£ 22,518
3	£ 55,637	£ 21,699	14	£ 57,992	£ 22,580
4	£ 55,825	£ 21,901	15	£ 58,240	£ 22,618
5	£ 55,772	£ 21,812	16	£ 58,414	£ 22,890
6	£ 55,890	£ 21,714	17	£ 58,561	£ 23,112
7	£ 56,068	£ 21,932	18	£ 59,066	£ 23,315
8	£ 56,299	£ 22,086	19	£ 58,596	£ 22,865
9	£ 56,825	£ 22,265	20	£ 58,631	£ 22,788
10	£ 57,205	£ 22,551	21	£ 58,758	£ 22,949
11	£ 57,562	£ 22,736	22	£ 59,037	£ 23,149

Two-Variable Linear Regression

What is the relationship between income and retail sales? One way to visualize the relationship is to do a scatterplot.



Regression Model

- The classical linear regression model is a way of examining the nature and form of the relationship among two or more variables.

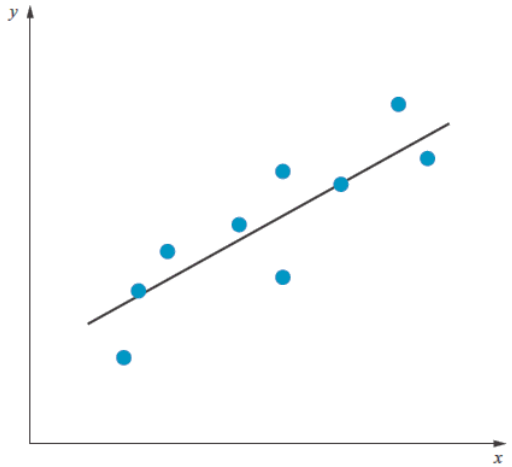
$$Y_t = \beta_1 + \beta_2 X_{1t} + u_t$$

Where:

- Y_t is the dependent variable
- X_{1t} is the independent variable (called also regressor)
- β_1, β_2 are called (population) coefficients or parameters. They are unknown need to be estimated by running a model.
- u_t is the error term or disturbance term. We assume is normally distributed with mean 0 and variance σ^2.

How are the values of coefficients determined?

- β_0 and β_1 are chosen so that the vertical distances (errors) from the data points to the fitted line are minimised.



- The most common method used to fit a line to the data is known as **Ordinary Least Squares (OLS)**
- The method of OLS can be viewed as equivalent to minimising the sum of the squared error terms.

$$\text{OLS: } \min \sum_{t=1}^{T} \hat{u}_t^2$$

How are the values of coefficients determined?

The OLS estimators are the solution to the problem

$$\min_{\hat{\beta}_1,\hat{\beta}_2} \sum_{i=1}^{n} \hat{u}_i^2 = \min_{\hat{\beta}_1,\hat{\beta}_2} \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

First order conditions (FOC):

$$\frac{\partial(\sum_{i=1}^{n} \hat{u}_i^2)}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

Re-arranging yields:

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i \Rightarrow$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i}{n} - \hat{\beta}_2 \frac{\sum_{i=1}^{n} X_i}{n} \Rightarrow$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

How are the values of coefficients determined?

First order conditions (FOC):

$$\frac{\partial(\sum_{i=1}^{n}\hat{u}_i^2)}{\partial\hat{\beta}_2} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)X_i = 0$$

Re-arranging yields:

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)X_i = 0 \Rightarrow \sum_{i=1}^{n}(Y_i - \bar{Y} + \hat{\beta}_2\bar{X} - \hat{\beta}_2 X_i)X_i = 0$$

$$\sum_{i=1}^{n} X_iY_i - \bar{Y}\sum_{i=1}^{n} X_i + \hat{\beta}_2\bar{X}\sum_{i=1}^{n} X_i - \hat{\beta}_2\sum_{i=1}^{n} X_i^2 = 0 \Rightarrow$$

$$\sum_{i=1}^{n} X_iY_i - n\bar{Y}\bar{X} + n\hat{\beta}_2(\bar{X})^2 - \hat{\beta}_2\sum_{i=1}^{n} X_i^2 = 0 \Rightarrow$$

$$\hat{\beta}_2(n\bar{X}^2 - \sum_{i=1}^{n} X_i^2) = n\bar{Y}\bar{X} - \sum_{i=1}^{n} X_iY_i \Rightarrow$$

$$\hat{\beta}_2 = \frac{n\bar{Y}\bar{X} - \sum_{i=1}^{n} X_iY_i}{n\bar{X}^2 - \sum_{i=1}^{n} X_i^2} = \frac{\sum_{i=1}^{n} X_iY_i - n\bar{Y}\bar{X}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

Two-Variable Linear Regression

- The **slope coefficient estimator** is

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{cov(X,Y)}{s_X^2} = \frac{cov(X,Y)s_Y}{s_X s_Y s_X} = r\frac{s_Y}{s_X}$$

Where r is the correlation between Y and X and s_Y (s_X) is the sample standard deviation of Y (X).

The slope coefficient $\hat{\beta}_1$ is an estimate of the change in Y when X changes by one unit.

- The **constant or intercept estimator** is

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}$$

Where $\bar{y}$ is the sample mean of Y and $\bar{x}$ is the sample mean of X .

Under maintained assumption on the error terms, it can be shown that the least squares coefficient estimators are unbiased and have minimum variance (**BLUE estimator**)

Variability in the Regression Analysis

The total variability in a regression analysis, TSS, can be partitioned into a component explained by the regression, ESS, and a component due to unexplained error, RSS:

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

with the components defined as:

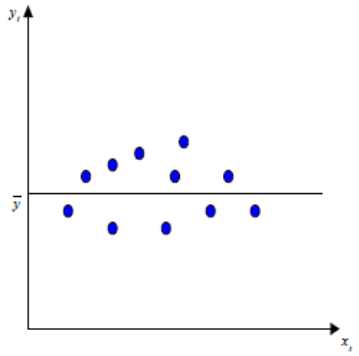
1. Sum of squares total: $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$
2. Residual sum of squares error: $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(u_i)^2$
3. Explained sum of squares regression: $ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

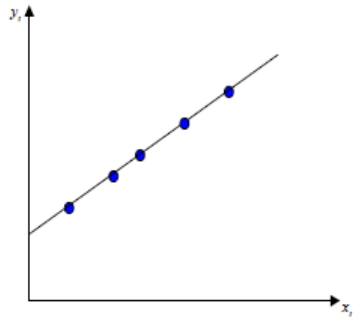
Variability in the Regression Analysis

The coefficient of determination, R^2, for a regression equation is defined as :

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

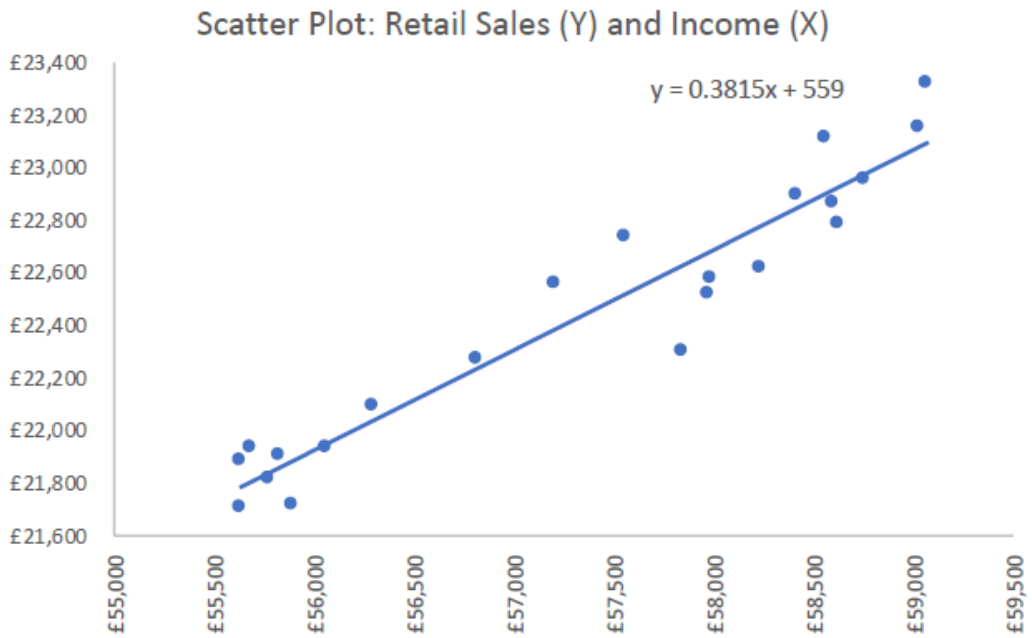
- R^2 tells us how well the estimated regression explains the data. In other words, R^2 measures the proportion of the total variation in Y explained by the regression model.
- The value of R^2 varies from 0 to 1 and higher values indicate a better regression





R-squared

Example (cont'd). If we graph the estimated linear relation in the same figure of the scatterplot, we have



We can also compute the R^2 for this example:

$$R^2 = \frac{b_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 91.9\%.$$

Multiple Linear Regression Model

We can generalize the two variable linear regression to multiple variable linear regression when an outcome variable is determined by several independent variables.

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + u_t$$

Where:

- Y_t is the dependent variable
- $X_{1t}, \ldots, X_{kt}$ arethe independent variables (called also regressors)
- $\beta_0, \beta_1, \ldots, \beta_k$ are called coefficients or parameters. They are unknown need to be estimated by running a model.
- u_t is the error term or disturbance term.

The Assumptions Underlying the CLRM

In order to use OLS, a model must satisfy the following assumptions:

1. **Linearity**: the model must be linear in parameters (e.g., $Y_t = \alpha + \beta X_t + u_t$)
2. **No multicollinearity** – independent variables are not highly correlated between them
3. $\mathbf{E(u_t) = 0}$ – the expected value of error terms is equal to zero
4. **Homoskedasticity** – the variance of the error terms is constant
5. **No Autocorrelation** - no correlation between error terms
6. $\mathbf{Cov(X_t, u_t) = 0}$
7. **Error terms are normally distributed**

Example - Taylor Rule

The Taylor rule is a simple monetary policy rule linking mechanically the level of the policy rate to deviations of inflation from its target and of output from its potential (the output gap).

The Taylor Rule (1993) takes the following form:

$$i_t = \beta_0 + \beta_1(\pi_t - \pi_t^*) + \beta_2(y_t - y_t^*) + u_t$$

- where α is the stabilizing interest rate of an economy (when $\pi_t = \pi_t^*$ and $y_t = y_t^*$),
- $(\pi_t - \pi_t^*)$ is the inflation gap between the actual values of the inflation (π_t) and a desired level (π_t^*),
- $(y_t - y_t^*)$ is the output gap that is the difference between the real GDP and the potential real GDP

Taylor Rule – Estimation Results

- We obtain the following estimation output in Eviews:

Dependent Variable: INT
Method: Least Squares
Sample (adjusted): 1976Q1 2020Q3
Included observations: 179 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.576754	0.123247	4.679655	0.0000
INFLATION	0.165749	0.020664	8.020974	0.0000
OUTPUT_GAP	0.103249	0.040825	2.529047	0.0123

R-squared	0.291099	Mean dependent var	1.300049
Adjusted R-squared	0.283043	S.D. dependent var	1.331186
S.E. of regression	1.127159	Akaike info criterion	3.093896
Sum squared resid	223.6059	Schwarz criterion	3.147316
Log likelihood	-273.9037	Hannan-Quinn criter.	3.115557
F-statistic	36.13584	Durbin-Watson stat	0.055367
Prob(F-statistic)	0.000000		

Estimation Results

Orange Box: here the estimated value of the coefficients ($\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$). .

Therefore the estimated value of the regression is:

$$Int_t = 0.557 + 0.166 * inflation_t + 0.103 * output\_gap_t$$

- $\hat{\beta}_1$ and $\hat{\beta}_2$ measure the change in the mean value of the dependent variable for each unitary change of the associated regressors, holding all the other regressors as constant.
- For example: $\hat{\beta}_1 = 0.1657$ means that a 1% increases of inflation causes a 0.166% increase of interest rate.
- $\hat{\beta}_0$ is the constant and it is positive. It informs us that if the economy was moving along the stable path (i.e., inflation = target inflation and real GDP = potential GDP), then the stabilising interest rate would be equal to 0.577.

Estimation Results

Gray Box: these are the **standard errors.**

Standard error is a measure of reliability or precision of the estimate

Standard Errors (SE) depend on the sample variance.

The greater the sample variance is, then the more dispersed the errors are about their mean value and therefore the more dispersed y will be about its mean value.

Estimation Results

Red Box: **R-squared (R^2)** is a measure of goodness of fit. How well the sample regression line fits the dataset. It measures the proportion of the variation of the dependent variable that has been explained by the regression.

Here is 0.29 which means that our model can explain 29% of the total variability of the dependent variable.

Estimation Results

Blue Box: the t-statistic for statistical significance. The t-statistic is calculated as a ratio between the estimated coefficient and the associated standard error.

Purple Box: the associated p-values (probability values).

P-Value rule:

- If $p_{value} \leq \alpha$, then we reject H_0
- If $p_{value} > \alpha$, then we fail to reject H_0

In this case we are testing whether the coefficient is statistically significant. (e.g., $H_0: \beta_1 = 0$). When the null hypothesis is rejected, then the coefficient is significant and statistically different from zero.

Estimation Results

Green Box: the F-test measures of the overall significance of the regression.

In other words, it tests the model which we estimated against a model with the constant as only regressor. It tells us whether all the regressors (a part from the constant) are jointly significant or not.

Whenever the probability value associated to the F-test is less or equal to 0.10, then we may safely say that the overall significance of our model is ok.

Estimation Results in R

Simple Output

```
Call:
lm(formula = Sales ~ Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
 -3385  -2097    258   1726   3034

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1383.4714  1255.2404   1.102    0.296
Spend         10.6222     0.1625  65.378 1.71e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2313 on 10 degrees of freedom
Multiple R-squared:  0.9977, Adjusted R-squared:  0.9974
F-statistic:  4274 on 1 and 10 DF,  p-value: 1.707e-14
```

Multiple Regression Output

```
Call:
lm(formula = Sales ~ Spend + Month, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-1793.73 -1558.33    -1.73  1374.19  1911.58

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -567.6098  1041.8836  -0.545  0.59913
Spend         10.3825     0.1328  78.159 4.65e-14 ***
Month        541.3736   158.1660   3.423  0.00759 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1607 on 9 degrees of freedom
Multiple R-squared:  0.999, Adjusted R-squared:  0.9988
F-statistic:  4433 on 2 and 9 DF,  p-value: 3.368e-14
```

Hypothesis Testing

- Test of significance is a procedure, which allows us to use the sample results to verify whether a null hypothesis is true or false.

- The decision about the rejection or the not rejection of the null hypothesis is based on the value of the test statistic, which we obtain from our data.

- We may test several single or joint hypotheses.

Hypothesis Testing – t test

- We want to test the following hypothesis:

$$H_0: \beta = \beta^* \text{ vs } \quad H_1: \beta \neq \beta^*$$

- If we want to test the significance of a coefficient, then we want to test: H_0: β= 0 vs H_1: β≠ 0

Note that the alternative hypothesis is a composite one (we do not know whether the true value is larger or smaller than 0. But if we reject the null, we know for sure that it is not equal to 0).

- The test statistic will be constructed as follows:

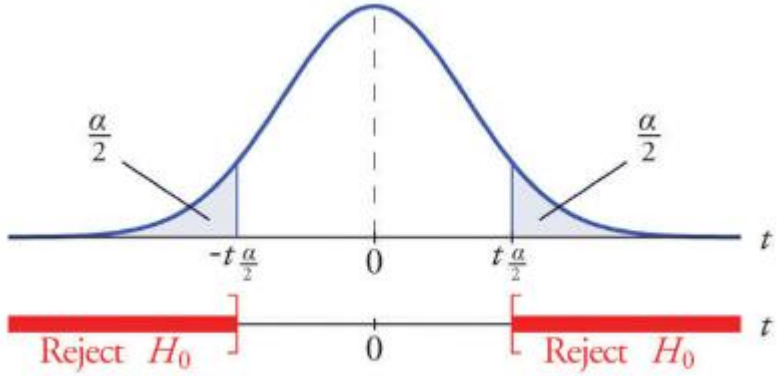
$$t = \frac{\hat{\beta} - \beta_{Ho}}{SE(\hat{\beta})}$$

- Conclusion of the test (for a two-sided test):

If the test statistic lies in the rejection area (i.e. $|t| \geq t\_crit$), then we reject the null hypothesis.

If the test statistic does not lie in the rejection area (i.e. $|t| < t\_crit$), we fail to reject the null hypothesis.

Hypothesis Testing – t test



- The t-distribution is defined by the degrees of freedom. These are related to the sample size.

  Degrees of freedom: n – k where K is the number of parameters (in the model)

- The 100(1 − α)% confidence interval for the population regression slope β_1 is given by

$$\hat{\beta}_1 - t_{n-k,\alpha/2} SE(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + t_{n-k,\alpha/2} SE(\hat{\beta}_1)$$

Hypothesis Testing – F Test

- The interesting point is that we may also test several hypotheses at the same time. We have already encounter such a possibility, when we discussed the F - test in the estimation output.
- Taylor claims that a conducive monetary policy should imply that $\beta_1 = \beta_2 = 0.5$. We may interpret this hypothesis by claiming that the two components, which determines the monetary policy, should have the same impact, i.e. they have an equal weight.
- More specifically, we want to test the joint hypothesis

$$H0 : \beta_1 = 0.5 \text{ and } \beta_2 = 0.5$$

$$H1 : \beta_1 \neq 0.5 \text{ and/or } \beta_2 \neq 0.5$$

How to Perform a F-test

How to run an hypothesis test:

- Define H_0 and H_1
- Calculate the test statistic. This is given by the formula

$$test\ statistic = \frac{RRSS - URSS}{URSS} \text{x} \frac{T - k}{m}$$

where RRSS = RSS from restricted regression

URSS = RSS from unrestricted regression

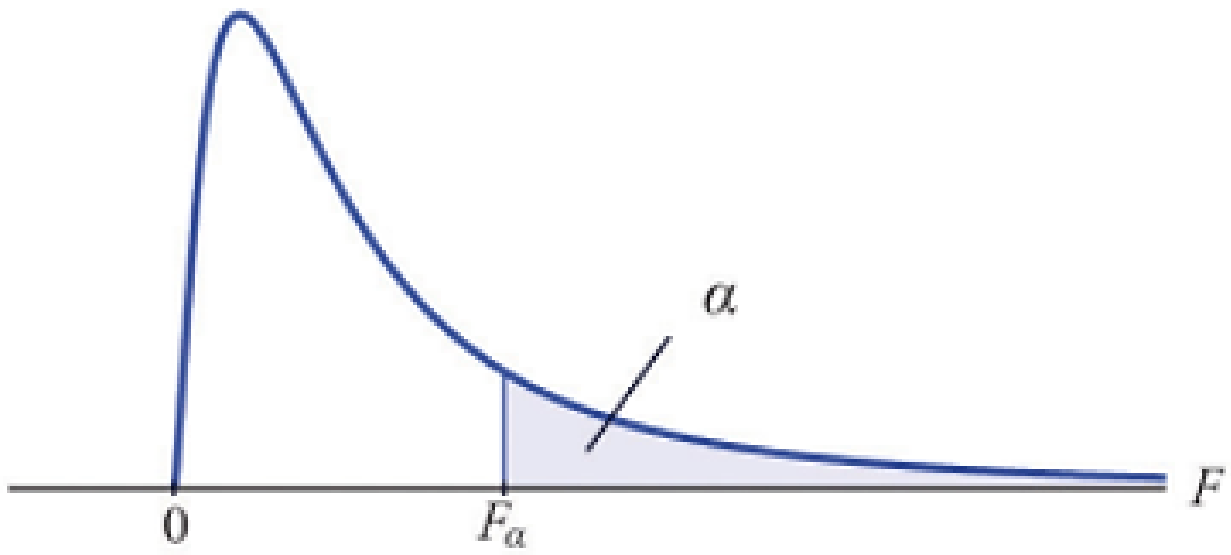
T = number of observations

K = number of parameters (in the unrestricted regression)

m = number of restrictions

How to Perform a F-test

- Choose a “significance level” denoted as α. It determines the rejection area.



- Obtain the critical values ($F_{crit} or F_a$) using distribution tables.

Critical values denotes the rejection area.

How to Perform a F-test

- Conclusion of the test:

If the test statistic lies in the rejection area (i.e. $f \geq f_{crit}$), then we reject the null hypothesis.

If the test statistic does not lie in the rejection area (i.e. $f < f_{crit}$), we fail to reject the null hypothesis.

Alternatively:

P-Value Approach

- This approach is widely used. Statistical software provides the p-value anytime a test is performed.
- P-Value rule:
  - If $p_{value} \leq \alpha$, then we reject H_0
  - If $p_{value} > \alpha$, then we fail to reject H_0