Introduction to Linear Regression

Often, engineers interested in relationship among the variables or they want to develop a method of prediction

Example:

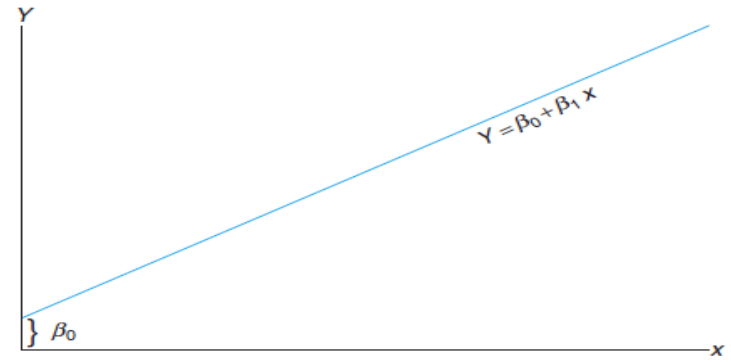
- Several automobiles with the same engine volume and they will not all have the same gas mileage
- Houses in the same part of country that have the same square footage of living space will not all be sold for the same price
- Engine volume and square meter of living space are, respectively, natural independent variables, or cause

A reasonable form of a relationship between the effect y and the cause x is the linear relationship.

$$Y = \beta_0 + \beta_1 x_0$$

Where, β_0 is the intercept and β_1 is the slope.

If the relationship is exact, then it is a deterministic relationship between two scientific variables.



Simple Linear Regression Model

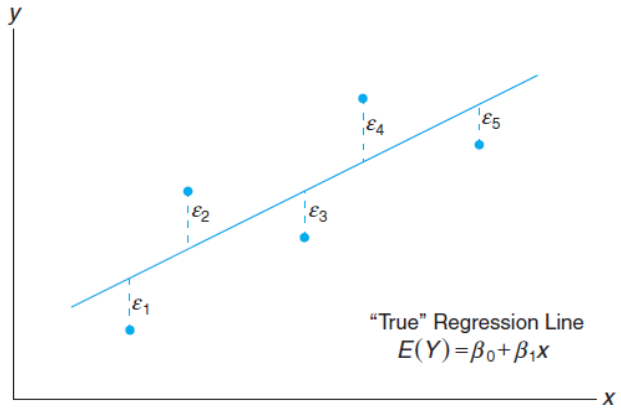
$$\boldsymbol{Y = \beta_0 + \beta_1 x_0 + \varepsilon}$$

Where, β_0 *and* $\beta_1$are unknown intercept and slope parameters, respectively, and $\boldsymbol{\varepsilon}$ is a random variable that is assumed to be distributed with $E(\varepsilon) = 0$ and

$Var(\varepsilon) = \sigma^2$. The quantity σ^2 is often called the error variance or residual variance.

The quantity $\boldsymbol{Y}$ is a random variable since $\boldsymbol{\varepsilon}$ is random

The quantity $\boldsymbol{\varepsilon}$ is often called **random error** or **random disturbance,** has constant variance.



The Fitted Regression Line (Least Squares)

An important aspect of regression analysis is to estimate the parameters $\beta_0 \; and \; \beta_1$ (regression coefficients). Suppose we denote the estimates b_0 for β_0 and β_1 for b_1. Then the estimated or fitted regression line is given by

$$\widehat{\boldsymbol{Y}} = \boldsymbol{b_0} + \boldsymbol{b_1 x_0}$$

Where, $\widehat{\boldsymbol{Y}}$ is the predicted or fitted value

Residual: Error in Fit

Given a set of regression data $\{(x_i, y_i), i = 1,2, \dots, n\}$ and a fitted model, $\widehat{\boldsymbol{y}}_{\boldsymbol{i}} = \boldsymbol{b_0} + \boldsymbol{b_1 x_0}$ the i^{th}residual $e_i$is given by $e_i = y_i - \hat{y}_i$, i=1, 2,…, n. obviously, if a set of n residuals is large, then the fit of the model is not good. Small residuals are a sign of good fit. We shall find$b_0 and \; b_1$, so that the sum of the squares of the residuals is a minimum.

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \qquad b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}$$

$$E(\varepsilon) = \sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

The coefficient of determination is defined as

$$r^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

A coefficient of determination of **0.85** is considered quite well.

Random errors

In regression model, ε_t, random error or residual is normally distributed with mean value 0 and standard deviation σ_i

$$\varepsilon_i \sim N(0, \sigma_i)$$

Example: In a study of pollution in a water stream, the concentration of pollution is measured at 5 different locations. The locations are at different distances to the pollution source. In the table below, these distances and the average pollution are given:

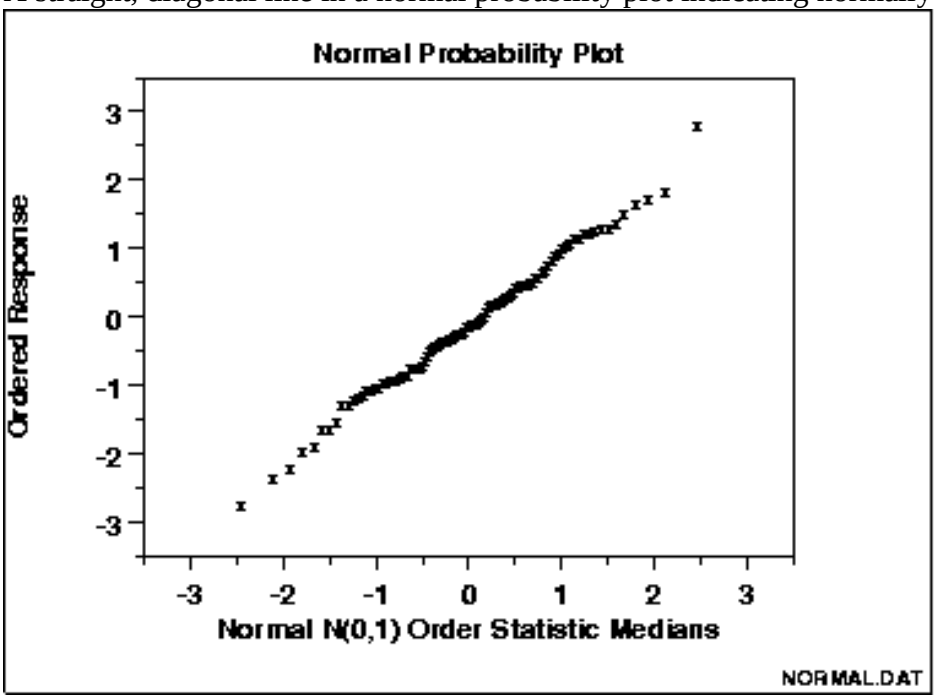
a- A researcher claims that there is linear relation between Distance to the pollution source (in km) and Average concentration, check it and find relation
b- Plot residuals and show that mean value of errors is zero or very close to zero
c- Assume that there is one location with 10.5 distance from pollution source, estimate the value of average concentration
d- Check normality of residuals

Distance to the pollution source (in km)	2	4	6	8	10
Average concentration	11.5	10.2	10.3	9.68	9.32

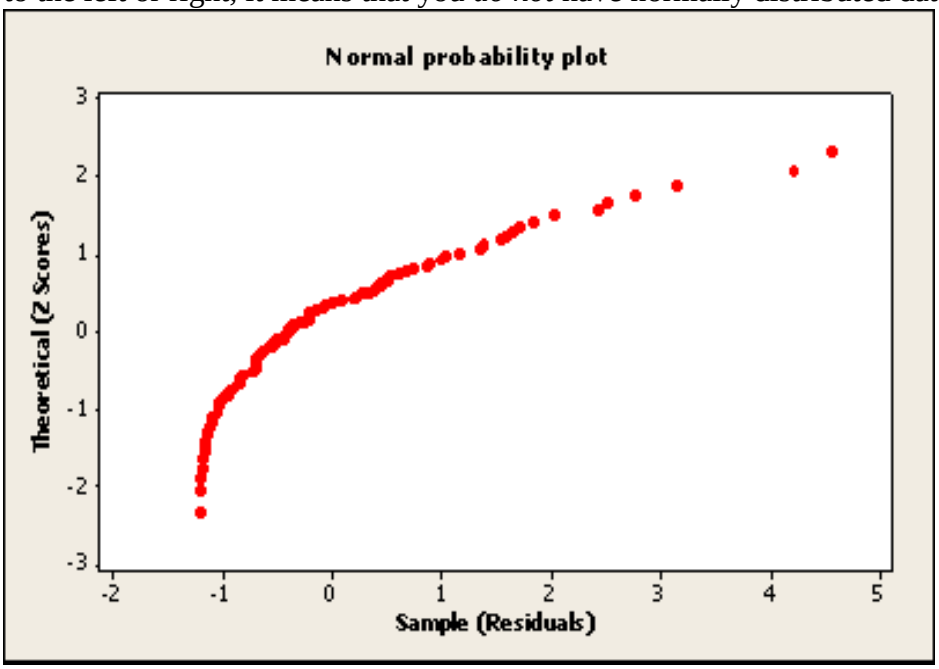
Test of normality

When you have a set of data that you think might have a normal distribution, a graph of your data can help you decide whether or not your data is normal. There is a specialized type of plot you can create, called a **normal probability plot**. A normal probability plot graphs z-scores (normal scores) against your data set.

A straight, diagonal line in a normal probability plot indicating normally distributed data.



A straight, diagonal line means that you have normally distributed data. If the line is skewed to the left or right, it means that you *do not* have normally distributed data.



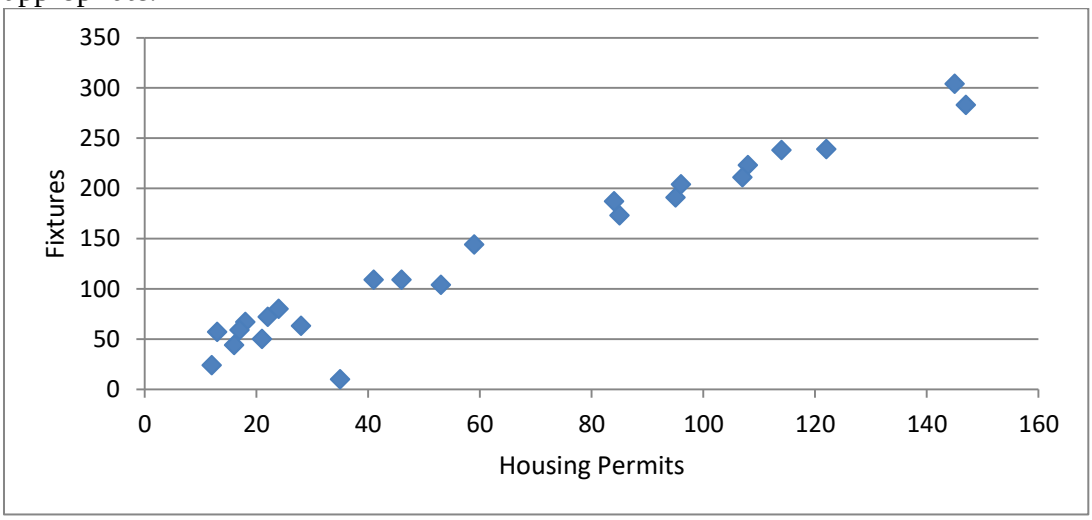
How to Draw a Normal Probability Plot By Hand

1. Arrange your x-values (errors) in ascending order.
2. Calculate $f_i = \frac{i-0.375}{n+0.25}$, where i, is the position of the data value in the ordered list and n is the number of observations.
3. Find the z-score for each f_i
4. Plot your x-values on the horizontal axis and the corresponding z-score on the vertical axis.

Example (with solution): Mary is manager of the plumbing department of a big store. Each month, she must place an order for bathroom plumbing fixtures. If she orders more than the number sold, the excess fixtures represent money the company cannot use elsewhere. If she orders too few, sales are lost to the store's competitors.
Mary has been thinking about how she might anticipate the demand for fixtures.
We wish to estimate a dependent (effect) variable –sales of plumbing fixtures in our example- and the value of the independent (cause) variable is related to an observable value of one or more independent variables –housing starts in the example. We call this process causal forecasting, because the value of the dependent variable is often caused by, or at least highly correlated with, the value of the independent variable.
The first thing Mary should do is make a scatter plot of her data. Table below gives the number of housing permits issued in an area and the number of plumbing fixtures sold in store for each month of last two years.
Housing permits and plumbing fixture sales

Data Point	Number of Permits	Number of Fixtures
1	22	72
2	16	44
3	24	80
4	95	191
5	84	187
6	13	57
7	114	238
8	147	283
9	96	204
10	59	144
11	35	10
12	41	109
13	28	63
14	21	50
15	18	67
16	46	109
17	145	304
18	122	239
19	108	223
20	85	173
21	107	211
22	53	104
23	17	59
24	12	24

A scatter plot is given in figure. This plot has housing permits, sorted in ascending order, as its x-axis and the corresponding fixture sales as its y-axis. The scatter plot seems to show a linear relationship between housing starts and fixture demand, so simple linear regression is appropriate.



Scatter plot of permits and sales

The underlying method is

$$y_i = B_0 + B_1 x_i + \varepsilon_i \qquad i = 1,2,\dots,n$$

Where; y_i = the value of dependent variable at time *i*

x_i = the value of independent variable at time *i*

ε_i = random error in the model

B_0 = intercept of the straight line relating y_i and x_i

B_1 = slope of the line

n = the total number of periods of the available data

We wish to choose estımate of B_0 and B_1 say b_0 and b_1, so that a straight line fits the data as closely as possible. To do so, we minimize the sum of the squared differences between the actual sales and the sales indicated by the model. The difference is the "error" of estimation. By squaring the difference we ensure that the value is nonnegative, penalizing both underestimates and overestimates. Squaring the difference also causes more weight to be placed on larger differences than on smaller differences. Thus we are minimizing the error in our prediction (estimation).

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \qquad \text{(Slope)}$$

$$b_0 = \bar{y} - b_1\bar{x} \text{ (Intercept)}$$

We calculated these values for Mary's problem; they are given in table, from these we compute

$$b_1 = \frac{(24 * 294095) - (1508 * 3337)}{(24 * 140928) - (1508)^2} = 1.828252$$

$$b_0 = \frac{3337}{24} - 1.83\left(\frac{1508}{24}\right) = 24.16647$$

This results in the regression equation

$$\hat{y}_i \cong 24.17 + 1.83x_i$$

Where $\hat{y}$ is the estimate for the number of plumbing fixtures sold in a month given that there were *x* housing permits. If there are 23 housing permits in January, we would expect to sell about 24.17+1.83*23=66 fixtures in that month.

For Mary's data, $r^2 = 0.98$, indicates an excellent fit, because the regression equation explains 98% of the variance.

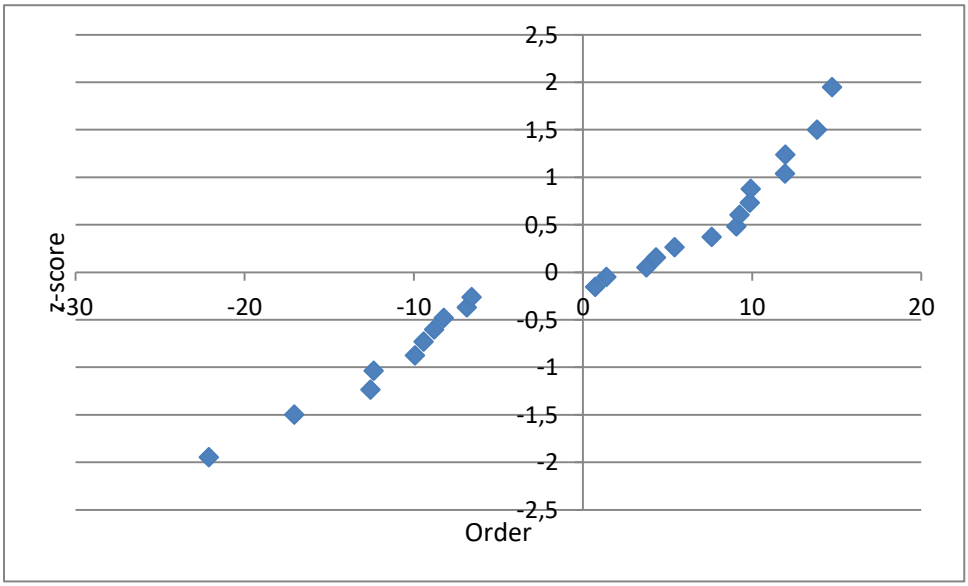
In general, b_1 can be positive or negative. A positive value implies that the dependent variable increase as the independent variable increases or those they are positively correlated. A negative $\hat{b}$ implies the opposite.

Regression computation for Mary's data

Data Point	x_i	y_i	x_i^2	$x_i y_i$
1	22	72	484	1584
2	16	44	256	704
3	24	80	576	1920
4	95	191	9025	18145
5	84	187	7056	15708
6	13	57	169	741
7	114	238	12996	27132
8	147	283	21609	41601
9	96	204	9216	19584
10	59	144	3481	8496
11	35	102	1225	3570
12	41	109	1681	4469
13	28	63	784	1764
14	21	50	441	1050
15	18	67	324	1206
16	46	109	2116	5014
17	145	304	21025	44080
18	122	239	14884	29158
19	108	223	11664	24084
20	85	173	7225	14705
21	107	211	11449	22577
22	53	104	2809	5512
23	17	59	289	1003
24	12	24	144	288
Total	1508	3337	140928	294095

For example of Mary's store we try to check the normality of residuals.

y_i	$\hat{y}_i$	ε_i	Order	i	$f_i = \frac{i - 0.375}{n + 0.25}$	z-score
72	64,38801	7,611986	-22,1055	1	0,025773196	-1,95
44	53,4185	-9,4185	-17,0638	2	0,067010309	-1,50
80	68,04452	11,95548	-12,5598	3	0,108247423	-1,24
191	197,8504	-6,85041	-12,3575	4	0,149484536	-1,04
187	177,7396	9,260362	-9,91951	5	0,190721649	-0,88
57	47,93375	9,066254	-9,4185	6	0,231958763	-0,738
238	232,5872	5,412802	-8,78943	7	0,273195876	-0,60
283	292,9195	-9,91951	-8,21321	8	0,31443299	-0,48
204	199,6787	4,321338	-6,85041	9	0,355670103	-0,37
144	132,0333	11,96666	-6,56789	10	0,396907216	-0,26
102	88,15529	13,84471	0,733938	11	0,43814433	-0,16
109	99,1248	9,875198	1,382314	12	0,479381443	-0,05
63	75,35753	-12,3575	3,753246	13	0,520618557	0,05
50	62,55976	-12,5598	4,321338	14	0,56185567	0,16
67	57,07501	9,924994	5,412802	15	0,603092784	0,26
109	108,2661	0,733938	7,611986	16	0,644329897	0,37
304	289,263	14,73699	9,066254	17	0,68556701	0,48
239	247,2132	-8,21321	9,260362	18	0,726804124	0,60
223	221,6177	1,382314	9,875198	19	0,768041237	0,738
173	179,5679	-6,56789	9,924994	20	0,809278351	0,88
211	219,7894	-8,78943	11,95548	21	0,850515464	1,04
104	121,0638	-17,0638	11,96666	22	0,891752577	1,24
59	55,24675	3,753246	13,84471	23	0,932989691	1,50
24	46,10549	-22,1055	14,73699	24	0,974226804	1,95



Normal probability plot of residuals

Discussion about mean value of errors

In simple regression model, there is one dependent variable and one independent variable and we wish to estimate dependent variable by independent variable.

It is obvious that, there is difference between exact value and estimated value. This difference refers to random error.

Mean value of random errors must be 0. Elsewhere there is at least another independent variable, or we made mistake/mistakes in our calculations.

For example of Mary's store we try to check this property.

Data Point	x_i	y_i	$\hat{y}_i$	$\varepsilon_i = y_i - \hat{y}_i$
1	22	72	64,38801	7,611986
2	16	44	53,4185	-9,418502
3	24	80	68,04452	11,955482
4	95	191	197,8504	-6,85041
5	84	187	177,7396	9,260362
6	13	57	47,93375	9,066254
7	114	238	232,5872	5,412802
8	147	283	292,9195	-9,919514
9	96	204	199,6787	4,321338
10	59	144	132,0333	11,966662
11	35	102	88,15529	13,84471
12	41	109	99,1248	9,875198
13	28	63	75,35753	-12,357526
14	21	50	62,55976	-12,559762
15	18	67	57,07501	9,924994
16	46	109	108,2661	0,733938
17	145	304	289,263	14,73699
18	122	239	247,2132	-8,213214
19	108	223	221,6177	1,382314
20	85	173	179,5679	-6,56789
21	107	211	219,7894	-8,789434
22	53	104	121,0638	-17,063826
23	17	59	55,24675	3,753246
24	12	24	46,10549	-22,105494
Total	1508	3337		0,000704

$$E(\varepsilon_i) = \frac{\sum_{i=1}^{n} \varepsilon_i}{n} = \frac{0.000704}{24} = 0.00003 \cong 0$$

Example: In a study, The table below shows the number of absences,x, in an IENG/MANE385 course and the final exam grade,y, for 7 students.

a- A researcher claims that there is linear relation between Absence sessions and Final grade, find relation and check accuracy of relation
b- Plot residuals and show that mean value of errors is zero or very close to zero
c- Assume that absence sessions of one student is 5, estimate Finl grade of this student
d- Check normality of residuals

Number of absences, x	1	0	2	6	4	3	3
Final grade, y	95	90	90	55	70	80	85

Analysis of Variance Technique

In the estimation and hypothesis testing, we were restricted in each case to considering no more than two population parameters. Such was the case, for example, in testing for the equality of two population means using independent samples from normal population with common but unknown variance, where it was necessary to obtain a pooled estimate ofσ^2.

In this case, the problem would involve one factor with two levels.

In the $k > 2$ sample problem, it will be assumed that there are k samples from k populations. One very common procedure used to deal with testing population means is called the analysis of variance, or ANOVA.

Hypothesis testing in one-way ANOVA

$$\begin{cases} H_0: \mu_0 = \mu_1 = \cdots = \mu_k \\ H_1: At\ least\ two\ of\ means\ are\ not\ equal \end{cases}$$

Treatment:	1	2	$\cdots$	i	$\cdots$	k	
	y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{k1}	
	y_{12}	y_{22}	$\cdots$	y_{i2}	$\cdots$	y_{k2}	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	y_{1n}	y_{2n}	$\cdots$	y_{in}	$\cdots$	y_{kn}	
Total	$Y_{1.}$	$Y_{2.}$	$\cdots$	$Y_{i.}$	$\cdots$	$Y_{k.}$	$Y_{..}$
Mean	$\bar{y}_{1.}$	$\bar{y}_{2.}$	$\cdots$	$\bar{y}_{i.}$	$\cdots$	$\bar{y}_{k.}$	$\bar{y}_{..}$

$$SST\ (total\ sum\ of\ squares) = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} {y_{ij}}^2 - \frac{y_{..}^2}{N}$$

$$SSA\ (treatment\ sum\ of\ squares) = n\sum_{i=1}^{k}(y_{ij} - \bar{y}_{..})^2 = \frac{1}{n}\sum_{i=1}^{k} y_{i.}^2 - \frac{y_{..}^2}{N}$$

$$SSE(error\ sum\ of\ squares) = SST - SSA$$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed f
Treatments	SSA	$k-1$	$s_1^2 = \frac{SSA}{k-1}$	$\frac{s_1^2}{s^2}$
Error	SSE	$k(n-1)$	$s^2 = \frac{SSE}{k(n-1)}$	
Total	SST	$kn-1$		

Example: Test the hypothesis $\mu_0 = \mu_1 = \cdots = \mu_5$ at the 0.05 level of significance for the data of the following table on absorption of moisture by various types of cement aggregates.

Aggregate:	1	2	3	4	5	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854
Mean	553.33	569.33	610.50	465.17	610.67	561.80