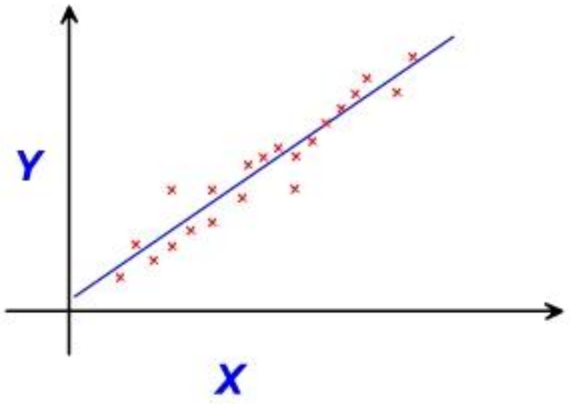
Regression

Day 1

Regression Analysis

A statistical procedure for analyzing associative relationships between metric-dependent variable and one or more independent variables



Regression Analysis

Ways in which Regression analysis can be used

- To determine whether the independent variables explain a significant variation in the dependent variable: whether a relationship exists.
- To determine how much of the variation in the dependent variable can be explained by the independent variables: strength of the relationship.
- To determine the structure or form of the relationship: the mathematical equation relating the independent and dependent variables.
- To predict the values of the dependent variable.
- To control for other independent variables when evaluating the contributions of a specific variable or set of variables.

Bivariate Regression

Bivariate Regression

- A procedure for deriving a mathematical relationship, in the form of an equation, between a single metric dependent variable and a single metric-independent variable.

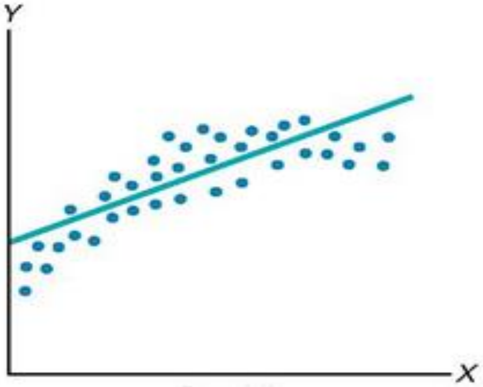
- It can be used to answer the following questions:
  - Can variation in **sales** be explained in terms of variation in **advertising expenditures**? What is the structure and form of this relationship, and can it be modelled mathematically by an equation describing a straight line?
  - Can the variation in **market share** be accounted for by the size of the **sales force**?
  - Are consumers' perceptions of **quality** determined by their perceptions of **price**?

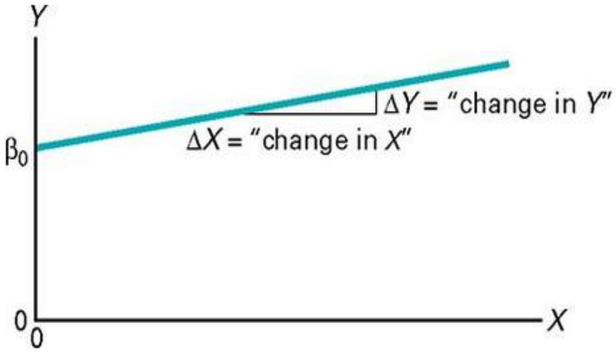
Bivariate Regression

- Bivariate Regression is the simplest form of regression analysis, and is also known as Ordinary Least-Squares regression or linear regression.

- The basic simple linear regression model equation is,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$





- Y_i – Dependent variable or criterion variable for observation i
- β_0 – Y intercept for the population
- β_1 – Slope for the population
- X_i – Independent variable for observation I
- ε_i – Random error in Y for observation i

Bivariate Regression

- The simple linear regression equation: The prediction line

$$\hat{Y}_i = b_0 + b_1 X_i$$

  - $\hat{Y}_i$ - predicted value of Y for observation i
  - b_0 – Sample Y intercept
  - b_1 – Sample slope
  - X_i – value of X for observation i

- This equation requires to determine two regression coefficients b_0 and b_1 The most common approach to finding b_0 and b_1 is using Least Square Method.

Least-Squares Method

Least-Squares Method

- Minimizes the sum of squared differences between the actual values (Y_i) and the predicted values ($\hat{Y}_i$)
- Using the $\hat{Y}_i = b_0 + b_1X_i$, the sum of squared differences is equal to

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}[Y_i - (b_0 + b_1X_i)]^2$$

Least-Squares Method

- This method determines the values of b_0 & b_1 that minimize the sum of squared differences around the prediction line.

- Any values for b_0 & b_1 other than those determined by the least-squares method result in a greater sum of squared differences between the actual values (Y_i) and the predicted values ($\hat{Y}_i$)

Example for interpreting the Y intercept, b0 and the slope b1

- A statistics professor wants to use the number of hours a student studies for a statistics final exam (X) to predict the final score (Y). A regression model is fit based on data collected from a class during a previous semester, with the following results, $\hat{Y}_i = 35.0 + 3X_i$

- *What is the interpretation of the Y intercept,* b_o *and the slope* b_1*?*

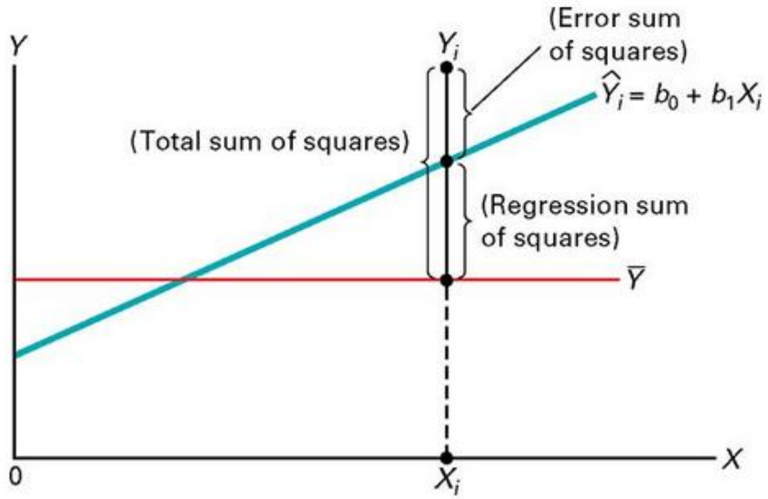
Example for interpreting the Y intercept, b0 and the slope b1

- The Y intercept b_0 = 35.0 indicates that when a student does not study for the final exam, the predicted mean final exam score is 35.0
- The slope b_1 = 3 indicates that for each increase of one hour in studying time, the predicted change in the mean final score is +3.0
- In other words, the final exam score is predicted to increase by a mean of 3 points for each one hour increase in studying time.

Measures of Variation

Measures of Variation

- When using Least square methods to determine regression coefficients, 3 measures of variation needs to be computed.



- SST = SSR + SSE
- SST – Total sum of squares
- SSR – Regression sum of squares (Explained variation)
- SSE – Error sum of squares (Unexplained variation)

Measures of Variation

- Total sum of squares

$$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

- Regression sum of squares

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

$$= b_0\sum_{i=1}^{n} Y_i + b_1\sum_{i=1}^{n} X_iY_i - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

- Error sum of squares

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$= \sum_{i=1}^{n} Y_i^2 - b_0\sum_{i=1}^{n} Y_i - b_1\sum_{i=1}^{n} X_iY_i$$

Coefficient of Determination

&

Standard Error of the Estimate

Correlation Coefficient (r)

- Correlation Coefficient is the statistic summarizing the strength of association between two metric (interval of ratio scaled) variables.
- It is an index used to determine whether a linear or a straight line relationship exists between X and Y

Note:

- *It was originally proposed by Karl Pearson, hence (r) is also known as Pearson Correlation Coefficient.*

- *It is also referred to as Simple Correlation, Bivariate Correlation or merely the Correlation Coefficient.*

Correlation Coefficient (r)

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Division of the numerator and denominator by *n* – 1 gives

$$r = \frac{\frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}}{\sqrt{\sum_{i=1}^{n}\frac{(X_i - \overline{X})^2}{n-1} \sum_{i=1}^{n}\frac{(Y_i - \overline{Y})^2}{n-1}}}$$

$$= \frac{COV_{xy}}{S_x S_y}$$

- X_{bar} and Y_{bar} denote sample means

- s_x and s_y are the standard deviations

- COV_{xy} is the covariance between X and Y

Correlation Coefficient (r)

- Covariance is a systematic relationship between two variables in which change in one implies a corresponding change in the other (COV_{xy}).

- The covariance between X and Y, measures the extent to which X and Y are related.

- The covariance may be either positive or negative

Correlation Coefficient (r)

- Division of COV$_{xy}$ by s$_x$s$_y$ achieves standardization, so r varies between -1.0 and 1.0
- It is absolute number and is not expressed in any unit of measurement.
- The correlation coefficient will be the same regardless of their underlying units of measurement.
- r can also be expressed in terms of the decompositionof total variation. (i.e. r^2)

Coefficient of Determination (r^2)

- r^2 = Explained variation / Total variation

- r^2 = SSR / SST

- The ratio of the regression sum of squares (SSR) to the total sum of squares (SST) measures the proportion of variation in Y that is explained by the linear relationship of the independent variable X with the dependent variable Y in the regression model

- r^2 must be a value between 0 and 1. It cannot be negative.

- Larger r^2 indicates a strong linear relationship between two variables.

Coefficient of Determination (r^2)

- r^2 = Explained variation / Total variation
- r^2 = SSR / SST
- The ratio of the regression sum of squares (SSR) to the total sum of squares (SST) measures the proportion of variation in Y that is explained by the linear relationship of the independent variable X with the dependent variable Y in the regression model
- r^2 must be a value between 0 and 1. It cannot be negative.
- Larger r^2 indicates a strong linear relationship between two variables.

Standard Error of the Estimate

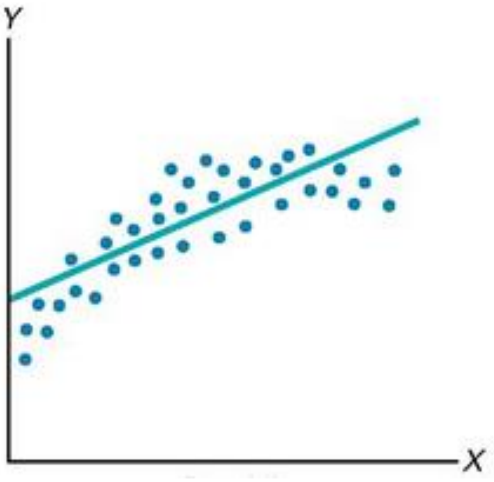
- It measures the variability of the observed Y values to the predicted Y values.

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

- Although the least-square method produces the line that fits the data with the minimum amount of prediction error, unless all the observed data points fall on a straight line, the prediction line is not a perfect predictor.

- The figure shows the variability around the prediction line.

- Note: Many observed values of Y fall near the prediction line, but most of the values are exactly not on the line.



Standard Error of the Estimate

- The interpretation of the standard error of the estimate is similar to that of the standard deviation.

- The standard deviation measures variability around the mean.

- The standard error of the estimate measures variability around the prediction line.

Assumptions of Regression

Regression Assumptions

- Linearity
- Independence of errors
- Normality of error
- Equal variance or Homoscedasticity

Residual Analysis

- Residual Analysis visually evaluates the above assumptions and helps to determine whether the regression model that has been selected is appropriate.
- The residual or estimated error value e_i, is the difference between the observed (Y_i) and the predicted ($\hat{Y}_i$) values of the dependent variable for a given value X_i
- A residual appears on a scatter plot as the vertical distance between an observed value of *Y* and the prediction line.
- $e_i = Y_i - \hat{Y}_i$

Linearity

- First, To check whether the relationship between the input and output variable is linear.
- Here Scatter plot is used, it reveals quadratic relationship between the input in X axis and output in Y axis

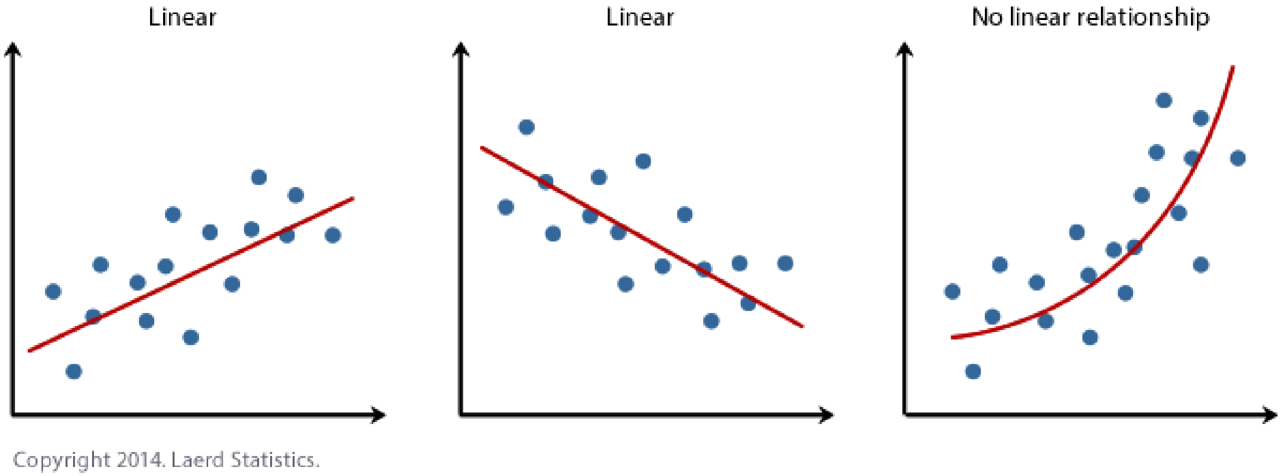
- Second, To evaluate linearity, residuals are plotted on the vertical axis against the corresponding Xi values of the independent variable on the horizontal axis.
- Here Residual plot is used, it reveals the quadratic relationship between the residuals and the input presented by the x-axis.

Linearity

- Relationship between the input variables and the output variable is assumed to be linear.

- The linearity can be checked using the scatter plots.

- If the relationship is not linear, create derived variables that transform the inputs so that the relationship to the target becomes linear.



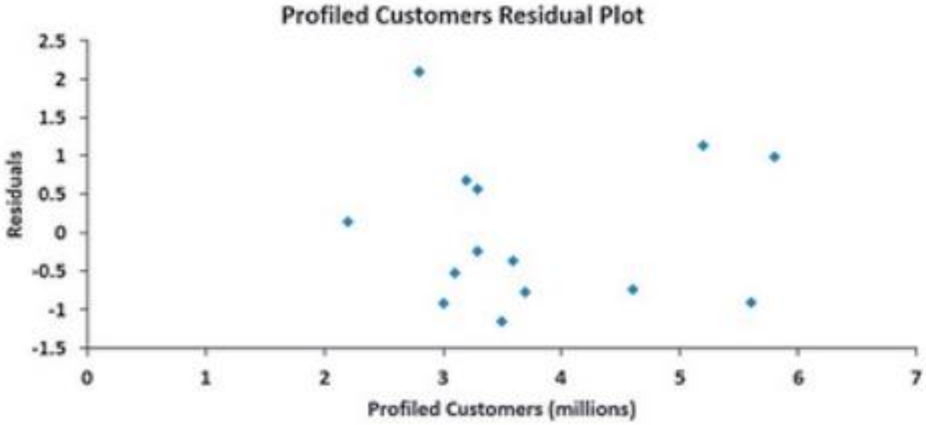
Linearity

- Using Residual plot, If the linear model is appropriate for the data, you will not see any apparent pattern in the plot.

- If the linear model is not appropriate, in the residual plot, there will be a relationship between the X_i values and the residuals e_i

- To assess linearity, residuals are plotted against the independent variable.

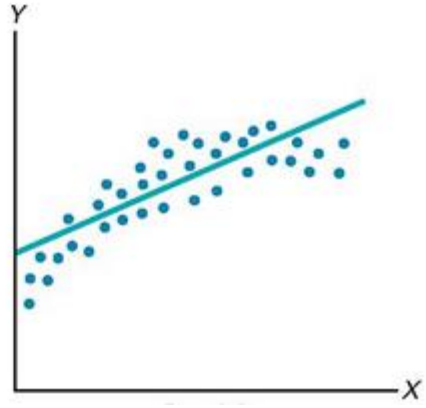
- Example for a residual plot with no clear pattern is shown below,

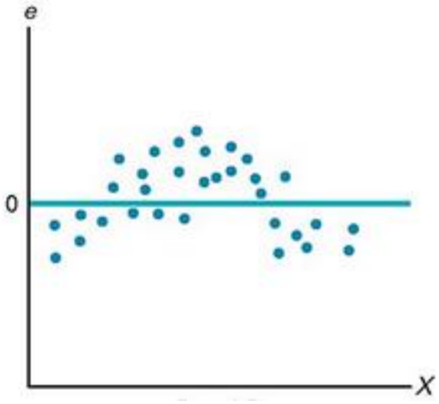


Linearity

- Plot – 1

Plot - 2





- In the above plots, plot-1 shows that relationship between X and Y are curvilinear.

- Plot-2 shows that there is a clear relationship between X_i and e_i

- For this case, a quadratic or curvilinear model is a better fit and should be used instead of the simple linear model.

Linearity

Linear Rainbow Test

- The basic idea of the Rainbow-Test is that even if the true relationship is nonlinear, over a subsample of data given a good linear fit can be achieved.

- The null hypothesis is rejected whenever the overall fit is significantly inferious to the fit of the subsample.

- The test statistic under H0 follows a F distribution with df1 and df2 degree of freedom.

- This particular procedure compares a subsample consisting of all data points without the upper and lower quartile (of time index), thus 50 data points.

- If the true relationship is concave or convex, the null hypothesis should be rejected.

Independence of errors

- This requires the errors (ε_i) be independent of one another.

- This assumption is particularly important when data are collected over a period of time.

- In such cases, the errors in a specific time period are sometimes correlated with those of the previous time period.

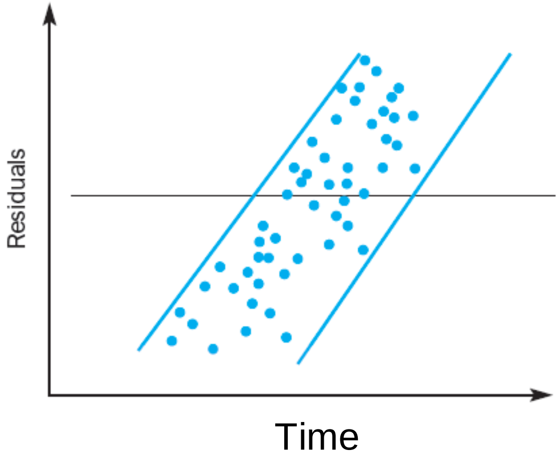
- This assumption can be evaluated by plotting the residuals in the order or sequence in which the data were collected.

- If the values of Y are part of Time Series, a residual may sometimes be related to the residual that precedes it.

- If the relationship exists between consecutive residuals, the plot will often show a cyclical pattern.

Independence of errors

- A plot of residuals against time will give us an understanding of whether the error terms are correlated or uncorrelated
- If the plot shows a random pattern, then the assumption lies true.
- Durbin Watson Test is used for examining the correlations between the error terms



Plot indicating a Linear Relationship between residuals and Time

Independence of errors

Durbin-Watson Test

- Durbin-Watson statistic is used to measure autocorrelation.
- This statistic measures the correlation between each residual and the residual over the previous time period.
- Durbin-Watson statistic will approach 0, if successive residuals are positively autocorrelated.
- If the residuals are not correltated, It will be close to 2.
- if the residuals are negatively autocorrelated, It will be greater than 2 and could even approach its maximum value of 4.
- The Null Hypothesis of the test is that there is no serial correlation.

Normality

- This assumption requires the errors (*ε_i*) be normally distributed at each value of X.

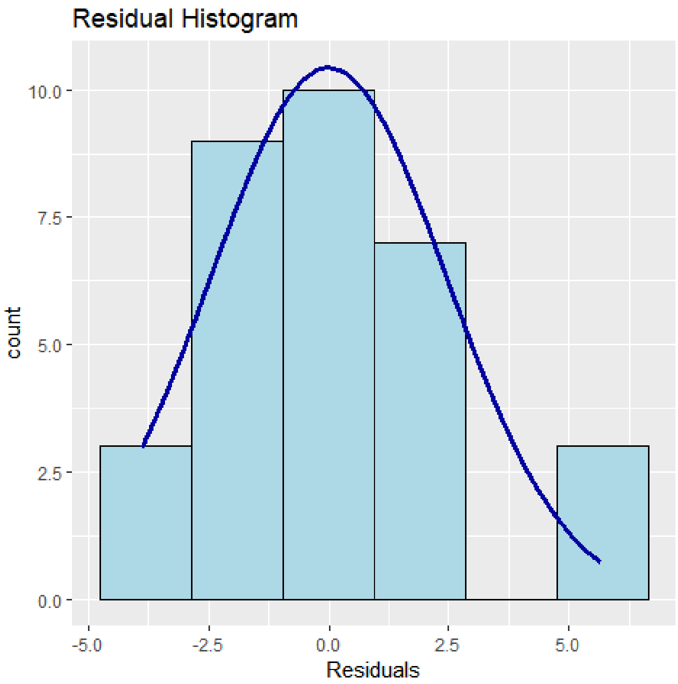
- As long as the distribution of the errors at each level of X is not extremely different from a normal distribution, inferences about β_0 and β_1 are not seriously affected.

- Normality can be evaluated by constructing a histogram of standardized residuals or a normal probability plot of standardized residuals, Q-Q Plots.

Normality

Histogram of standardized residuals

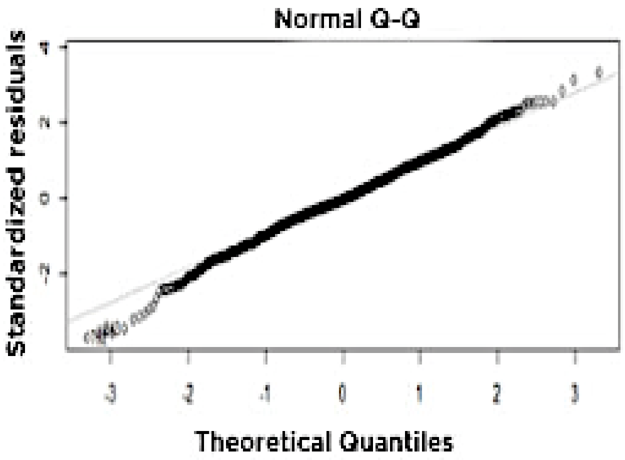
- For histogram, the residuals are organized into a frequency distribution.
- The histogram is plotted with residuals in x axis and frequency in Y axis.

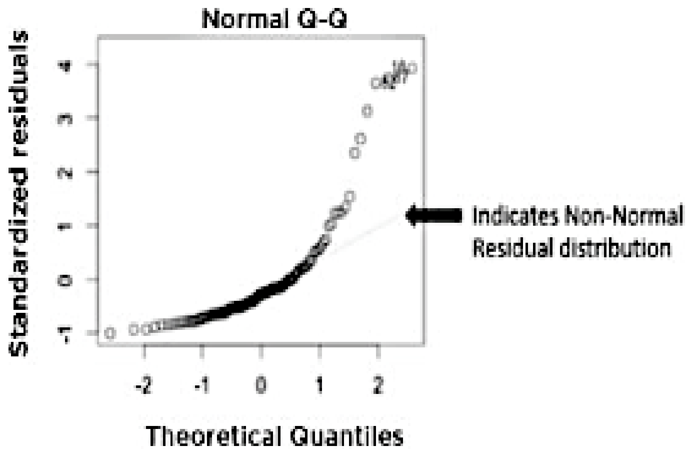


Normality

Q-Q plot

- This plot compares the observed data against the quantiles (Q) of the assumed distribution.
- This plots the standardized (z-score) residuals against the theoretical normal quantiles.
- Below are the Q-Q plot’s showing Linearity vs Non Linearity

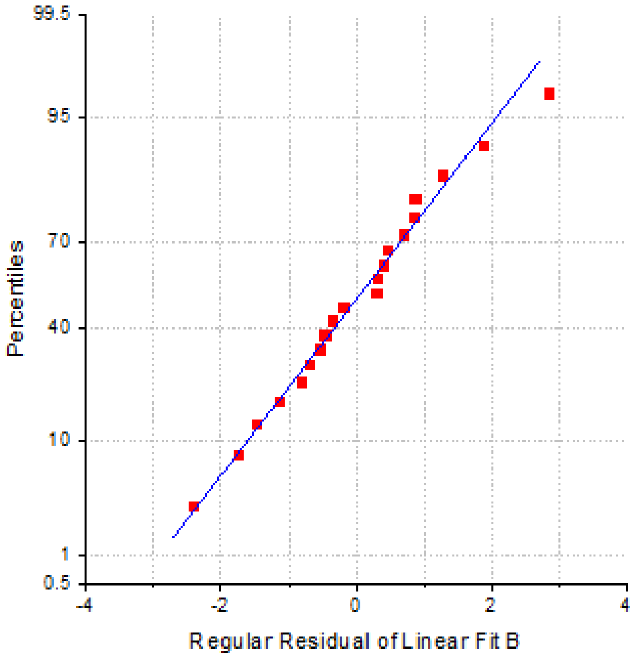




Indicates Non-Normal Residual distribution

Normality

- Normal Probability plot is a special case of Q-Q probability plot for a normal distribution.
- It can be used to check whether the variance is normally distributed as well.
- The plot is based on the percentiles versus ordered residuals, the percentiles are estimated by (i – 3/8) / (n + ¼)
- Where n is the total number of data and the i is the ith data.



Normality

Jarque Bera Test

- The Jarque–Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution.

- Note that this test generally works good for large enough number of data samples(>2000) as the test statistics asymptotically has a chi squared distribution with degrees 2 of freedom.

- If the Data is not normal a non linear transformation (e.g. Log Transformation) can fix the issue.

Equal Variance or Homoscedasticity

- This assumption requires that the variance of the errors be constant for all values of X.

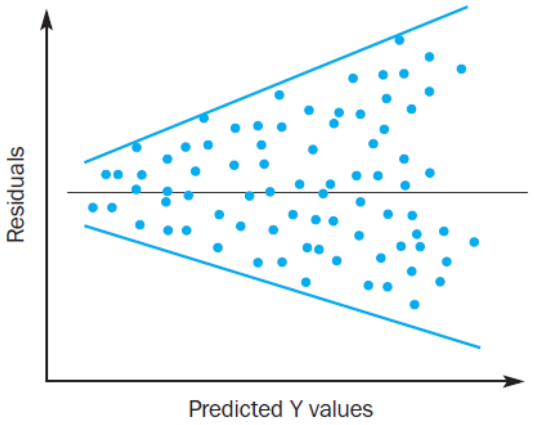
- In other words, the variability of Y values is the same when X is a low value as when X is a high value.

- This can be examined by plotting the standardized residu$\hat{Y}_i$s against the standardized predicted values of the dependent variable

- If the pattern is not random, the variance of the error term is not constant

Equal Variance or Homoscedasticity

- Below is the residual plot indicating that variance is not constant.

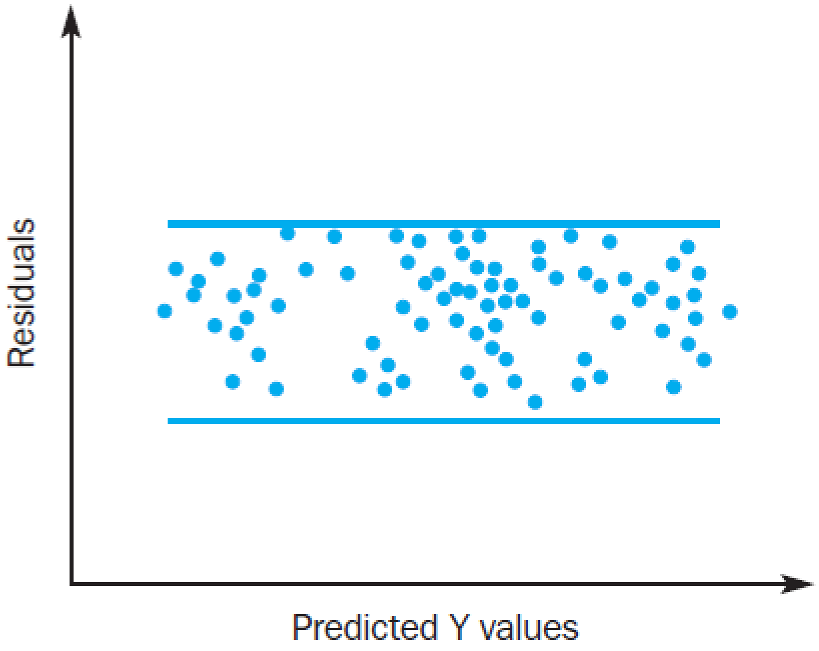


- This plot is fan or funnel shaped because the variability of the residuals increases dramatically as X increases.

- As this plot shows unequal variances of the residuals at different levels of X, the equal variance assumption becomes invalid.

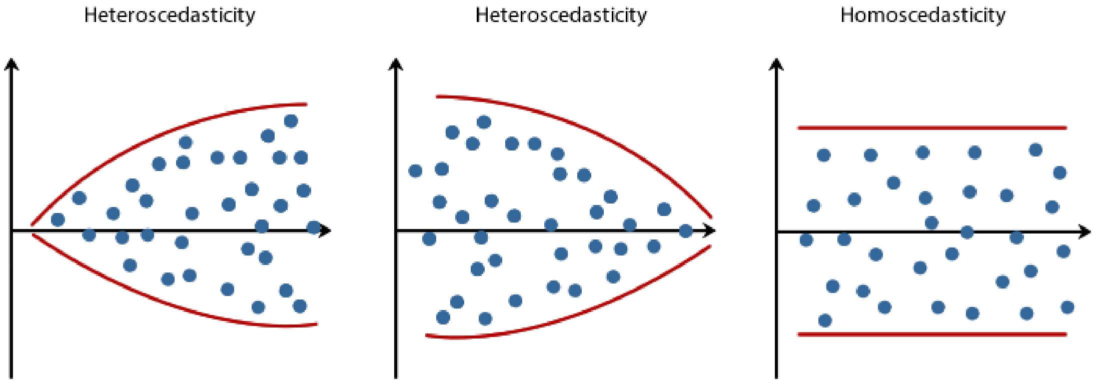
Equal Variance or Homoscedasticity

- Plot of residuals indicating that a fitted model is Appropriate



Equal Variance or Homoscedasticity

- Plots showing the difference between homoscedasticity and heteroscedasticity.



Copyright 2014. Laerd Statistics.

- A classic example of heteroscedasticity is If you model household consumption based on income, you’ll find that the variability in consumption increases as income increases.

Equal Variance or Homoscedasticity

Tests for heteroscedasticity of errors in regression

Goldfeld-Quandt Test:

- Given a known time T, the Goldfeld-Quandt tests the null-hypothesis: variances at time 1..T and T..n are equal.

Breusch-Pagan-Test:

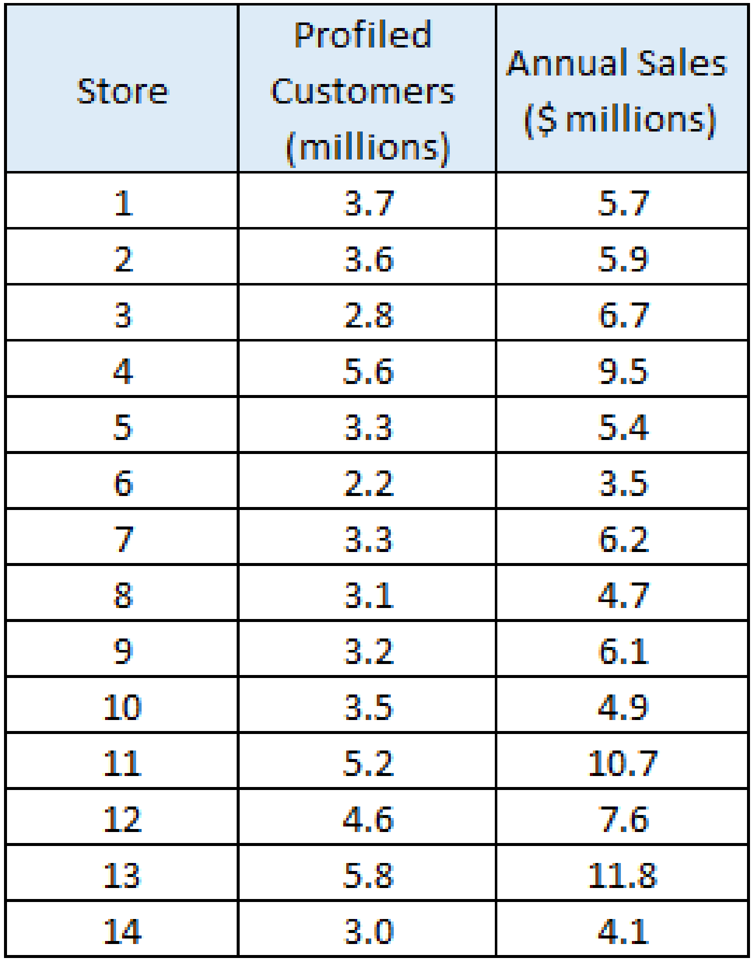
- It is a chi-squared test: the test statistic is distributed nχ2 with k degrees of freedom.
- If the test statistic has a p-value below an appropriate threshold (e.g. $p < 0.05$) then the null hypothesis of homoskedasticity is rejected and heteroskedasticity assumed.

Inferences about the Slope and Correlation Coefficient

Inferences about the Slope and Correlation Coefficient

- t Test for the slope
- F Test for the slope
- Confidence Interval Estimate for the slope
- t Test for the Correlation Coefficient

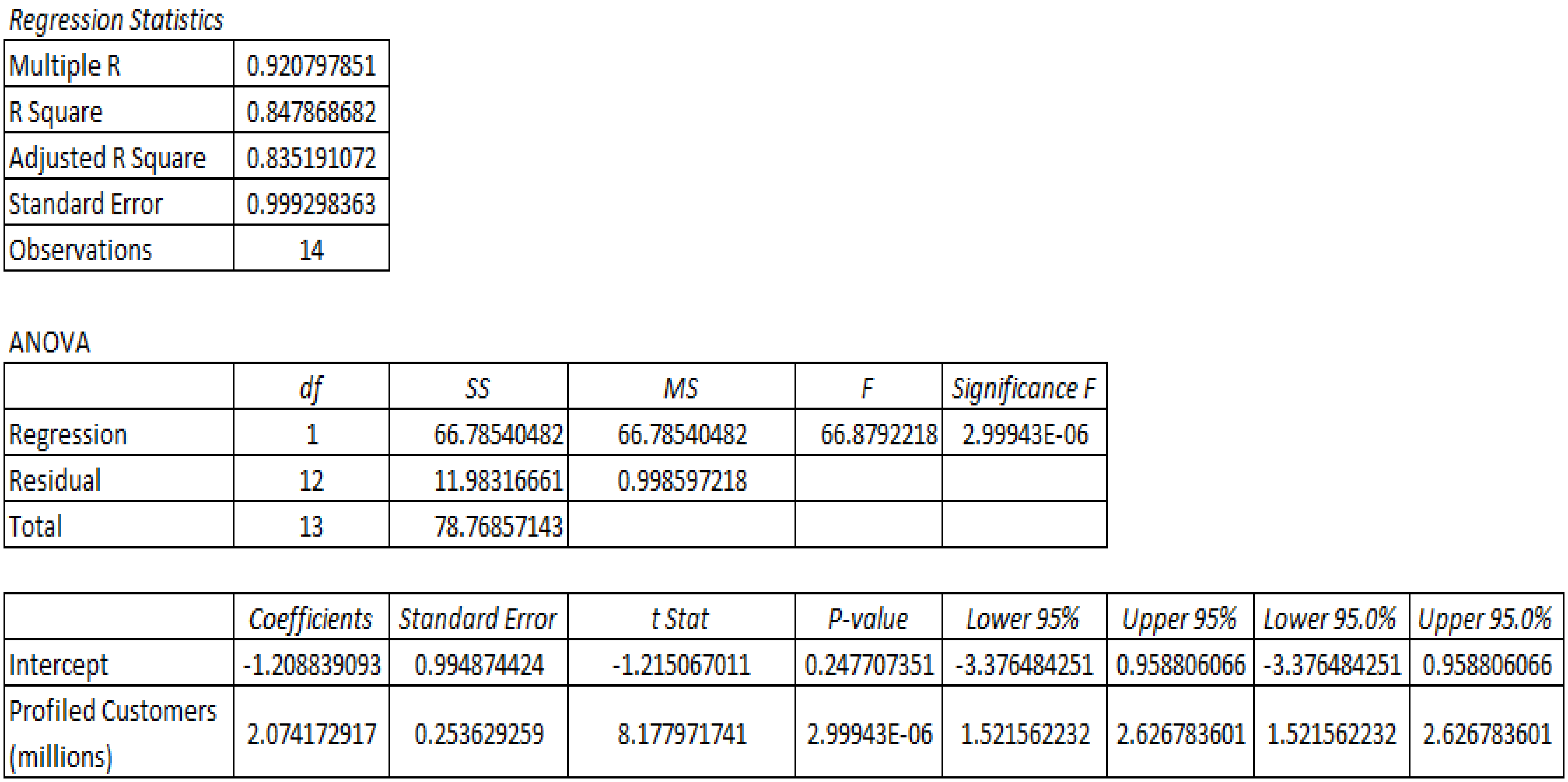
Sample Data



Store	Profiled Customers (millions)	Annual Sales ($ millions)
1	3.7	5.7
2	3.6	5.9
3	2.8	6.7
4	5.6	9.5
5	3.3	5.4
6	2.2	3.5
7	3.3	6.2
8	3.1	4.7
9	3.2	6.1
10	3.5	4.9
11	5.2	10.7
12	4.6	7.6
13	5.8	11.8
14	3.0	4.1

Number of Profiled Customers (in millions) and Annual Sales (in $ millions) for a sample of 14 Sunflower Apparel Stores

Regression summary output for sample data



Regression Statistics

Multiple R	0.920797851
R Square	0.847868682
Adjusted R Square	0.835191072
Standard Error	0.999298363
Observations	14

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	66.78540482	66.78540482	66.8792218	2.99943E-06
Residual	12	11.98316661	0.998597218		
Total	13	78.76857143			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.208839093	0.994874424	-1.215067011	0.247707351	-3.376484251	0.958806066	-3.376484251	0.958806066
Profiled Customers (millions)	2.074172917	0.253629259	8.177971741	2.99943E-06	1.521562232	2.626783601	1.521562232	2.626783601

t Test for the slope

- To determine the existence of a significant linear relationship between X and Y variables, must test whether the Beta1 (the population slope) is equal to 0.

- H0: Beta1 = 0 [There is a no linear relationship (the slope is zero).]
- H1: Beta1 ≠ 0 [There is a linear relationship (the slope is not zero).]

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}}$$

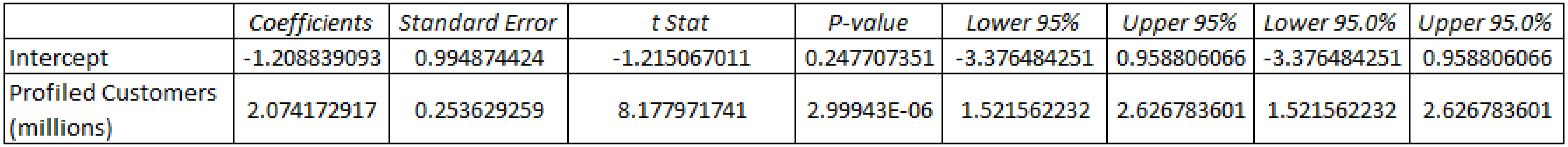
Where,

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}}$$

$$SSX = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

t Test for the slope

- t test result for the sample data to test whether there is a significant linear relationship between the number of profiled customers and the annual sales at the 0.05 level of significance,



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.208839093	0.994874424	-1.215067011	0.247707351	-3.376484251	0.958806066	-3.376484251	0.958806066
Profiled Customers (millions)	2.074172917	0.253629259	8.177971741	2.99943E-06	1.521562232	2.626783601	1.521562232	2.626783601

- At 0.05 level of significance, the critical value of t with n-2 = 12 degrees of freedom is 2.1788

- Here,
  - t-stat = 8.17797 > 2.1788
  - P-value is 0.000, which is less than alpha 0.05

  H0 is rejected

- Hence, we can conclude that there is a significant linear relationship between mean annual sales and the number of profiled customers

F Test for the slope

- As an alternative to t-test, F-test can be used to determine whether the slope is statistically significant

$$F_{STAT} = \frac{MSR}{MSE}$$

- Where,

$$MSR = \frac{SSR}{1} = SSR$$

$$MSE = \frac{SSE}{n-2}$$

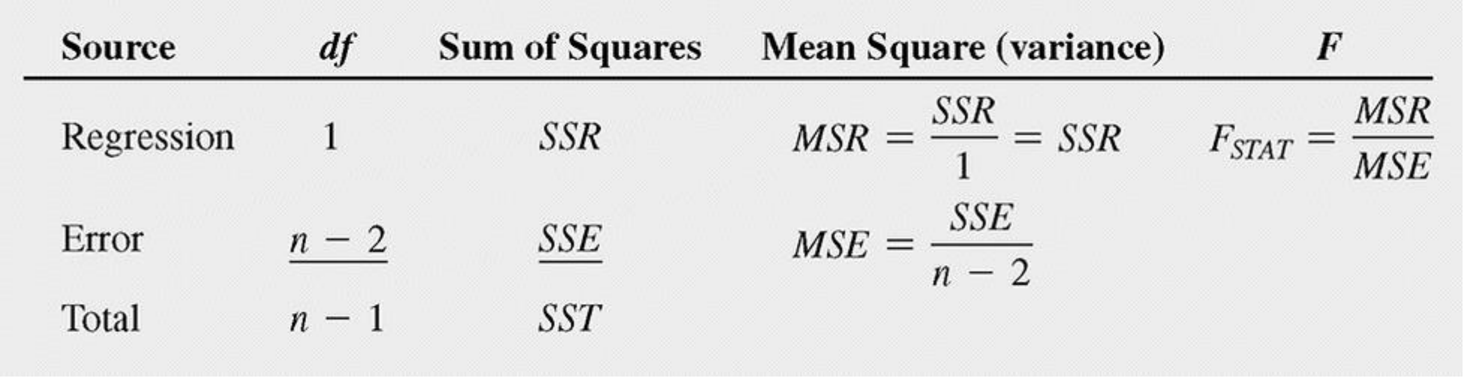
- Here the decision rule is,

  Reject H_0 if $F_{stat} > F\alpha$; otherwise do not reject H_0

- The F_{stat} test statistic follows an F distribution with 1 and n-2 degrees of freedom.

F Test for the slope

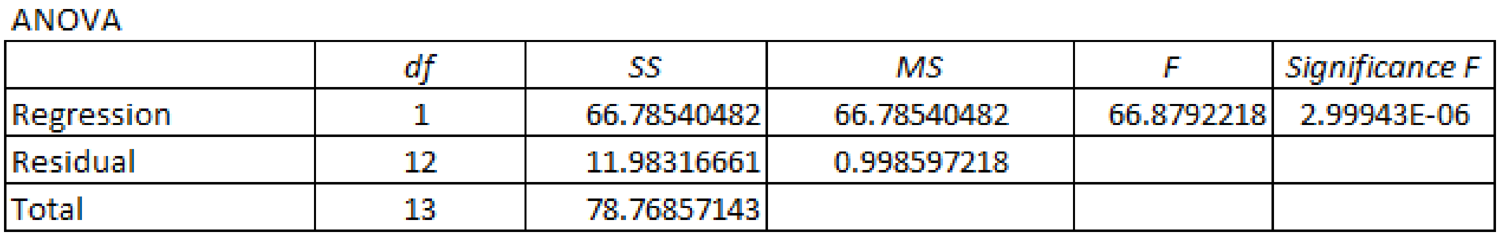
- ANOVA table for testing the significance of a regression coefficient.



Source	*df*	Sum of Squares	Mean Square (variance)	*F*
Regression	1	SSR	$MSR = \frac{SSR}{1} = SSR$	$F_{STAT} = \frac{MSR}{MSE}$
Error	$n - 2$	SSE	$MSE = \frac{SSE}{n - 2}$	
Total	$n - 1$	SST		

F Test for the slope

- F-test result for the sample data



ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	66.78540482	66.78540482	66.8792218	2.99943E-06
Residual	12	11.98316661	0.998597218		
Total	13	78.76857143			

- Using level of significance 0.05, the critical value of F distribution with 1 and 12 degrees of freedom is 4.75

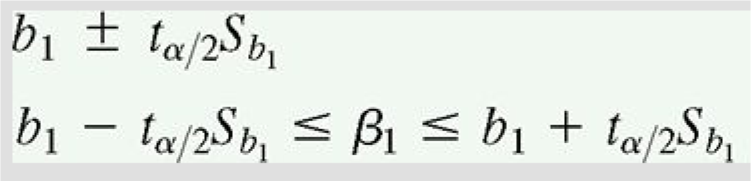
- Here,
  - F-stat = 66.88 > 4.75
  - P-value is 0.000 < 0.05

  H0 is rejected

- Hence, we can conclude that there is a significant linear relationship between mean annual sales and the number of profiled customers

Confidence Interval Estimate for the Slope

- This can be constructed by taking the sample slope b1, and adding and subtracting the critical t value multiplied by the standard error of the slope

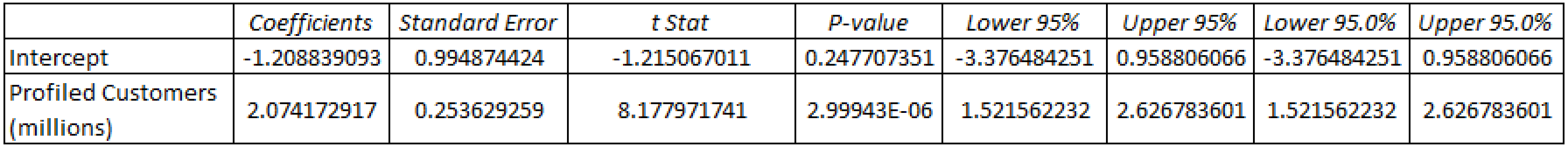


$$b_1 \pm t_{\alpha/2} S_{b_1}$$

$$b_1 - t_{\alpha/2} S_{b_1} \leq \beta_1 \leq b_1 + t_{\alpha/2} S_{b_1}$$

- T $_{\alpha/2}$ → critical value corresponding to an upper tail probability of $\alpha/2$ from the t distribution with n-2 degrees of freedom

Confidence Interval Estimate for the Slope



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.208839093	0.994874424	-1.215067011	0.247707351	-3.376484251	0.958806066	-3.376484251	0.958806066
Profiled Customers (millions)	2.074172917	0.253629259	8.177971741	2.99943E-06	1.521562232	2.626783601	1.521562232	2.626783601

- $b_1 = 2.0742$; $n = 14$; $S_{b1} = 0.2536$

- For 95% confidence interval estimate, $\alpha/2 = 0.025$, from critical values of t table $T_{\alpha/2} = 2.1788$

- By applying the values in the equation, we get
- $1.52 \leq \beta_1 \leq 2.63$

Confidence Interval Estimate for the Slope

$$1.52 \le \beta_1 \le 2.63$$

- This shows, 95% confidence that the estimated population slope is between 1.52 and 2.63

- This indicates that for each increase of 1 million profiled customers, predicted annual sales are estimated to increase by **at least** $1,521,600 **but no more than** $2,626,800

- Since both the values are above 0, there is an evidence that significant linear relationship exists between the variables

- If the interval had included 0, conclusion would have been that there is no evidence of a significant linear relationship between the variables.

t Test for the Correlation Coefficient

- Here the null and alternative hypothesis are

$$H_0: \rho = 0 \text{ (no correlation)}$$
$$H_1: \rho \neq 0 \text{ (correlation)}$$

- Test statistic for determining the existence of a significant correlation,

$$t_{STAT} = \frac{r - \rho}{\sqrt{\dfrac{1 - r^2}{n - 2}}}$$

- Where,

$$\begin{aligned} r &= +\sqrt{r^2} \quad \text{if} \quad b_1 > 0 \\ r &= -\sqrt{r^2} \quad \text{if} \quad b_1 < 0 \end{aligned}$$

- The tSTAT test statistic foloows a t distribution with n-2 degrees of freedom

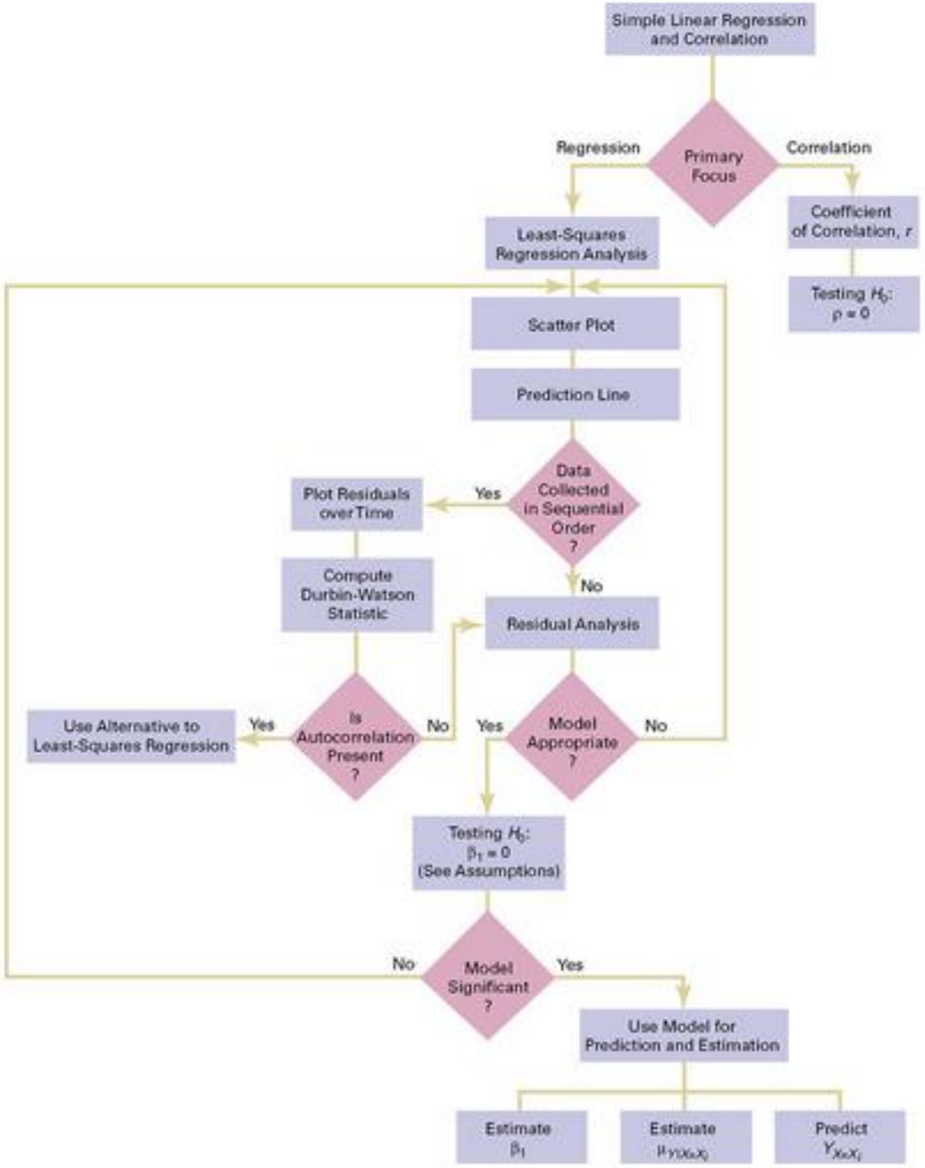
t Test for the Correlation Coefficient

- From the summary statstics of the sample data
- r^2 is 0.8479 and b1 is +2.0742
- b1>0, so the correlation coefficient is the positive root of r^2
- So, r = +0.9208
- Substituting r in the tSTAT, we get tSTAT as 8.178

- tSTAT = 8.178 > 2.1788
- At 0.05 level of significance, null hypothesis is rejected. Concluding that there is a significant association between annual sales and the number of profiled customers.

- Note: This tSTAT is equivalent to tSTAT test statistic found when testing whether the population slope, beta1 is equal to zero.

Road map for Simple Linear Regression


Simple Linear Regression and Correlation
Primary Focus
Regression
Correlation
Least-Squares Regression Analysis
Coefficient of Correlation, r
Testing H_0: $\rho = 0$
Scatter Plot
Prediction Line
Data Collected in Sequential Order ?
Yes
No
Plot Residuals over Time
Compute Durbin-Watson Statistic
Residual Analysis
Is Autocorrelation Present ?
Yes
No
Use Alternative to Least-Squares Regression
Model Appropriate ?
Yes
No
Testing H_0: $\beta_1 = 0$ (See Assumptions)
Model Significant ?
No
Yes
Use Model for Prediction and Estimation
Estimate β_1
Estimate $\mu_{Y|X=X_i}$
Predict $Y_{X=X_i}$

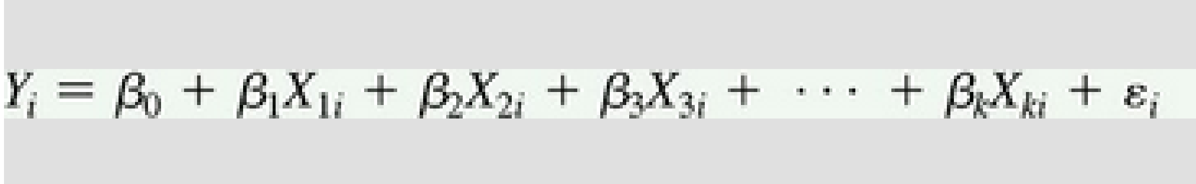
Multiple Regression

- A statistical technique that simultaneously develops a mathematical relationship between two or more independent variables and an interval-scaled dependent variable. (**Multiple regression** involves a single dependent variable and two or more independent variables)

- The questions raised in the context of bivariate regression can also be answered via multiple regression by considering additional independent variables:
  - Can variation in sales be explained in terms of variation in advertising expenditures, prices and level of distribution?
  - Can variation in market shares be accounted for by the size of the sales force, advertising expenditures and sales promotion budgets?
  - Are consumers' perceptions of quality determined by their perceptions of prices, brand image and brand attributes?

Multiple Regression

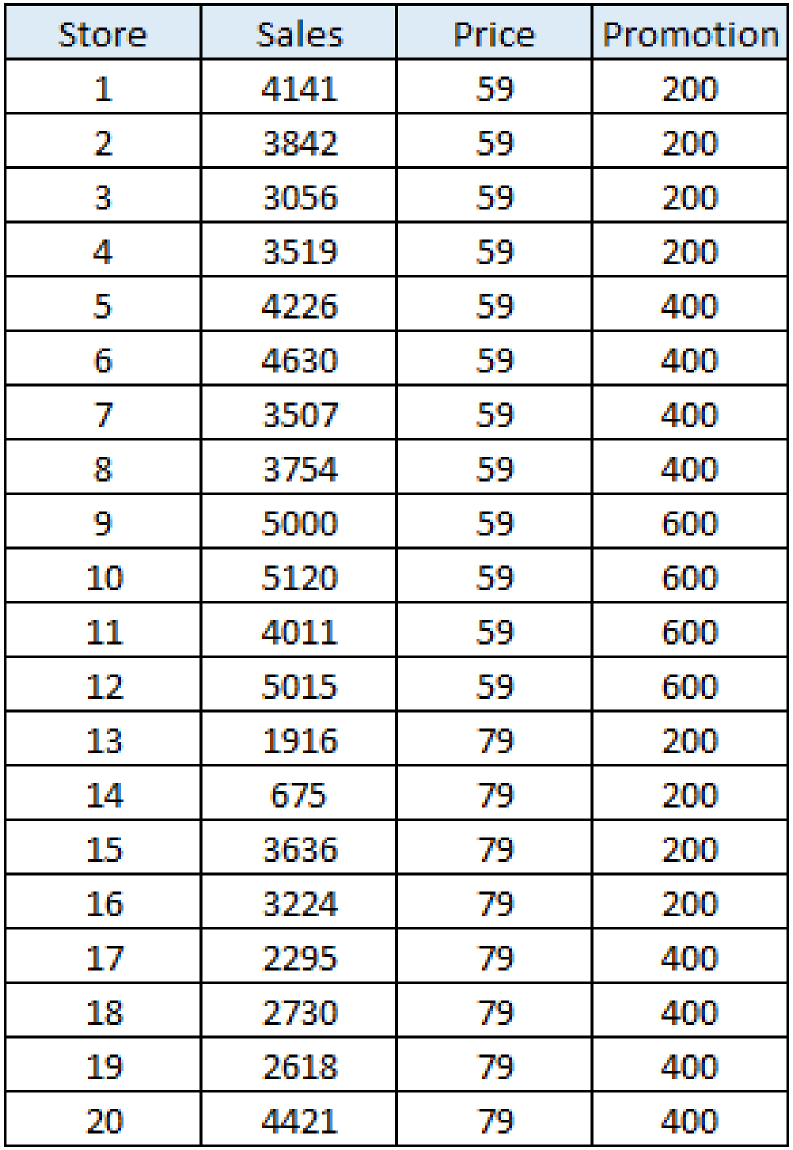
- The general form of multiple regression model is,

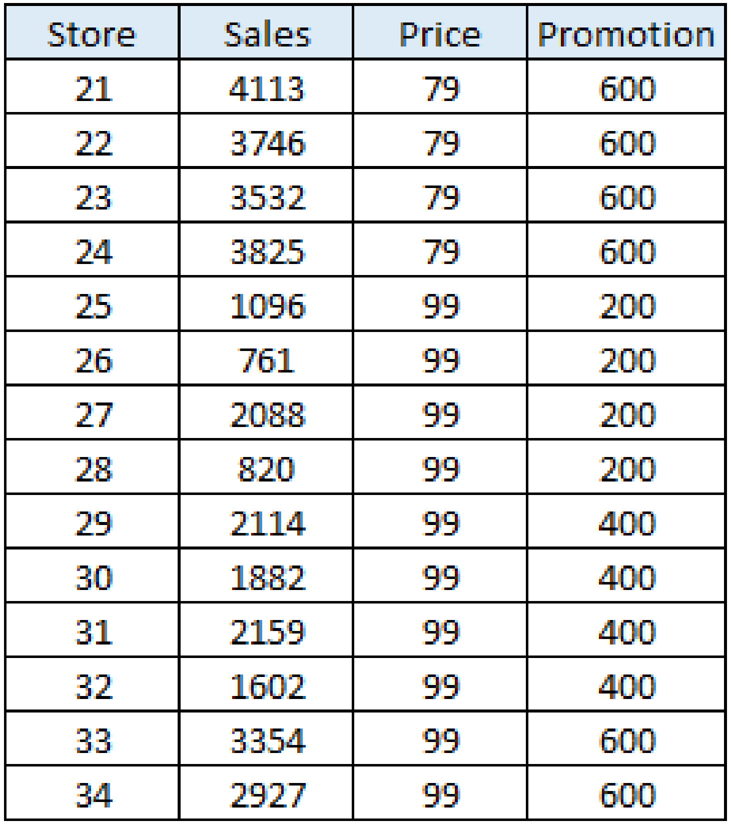


$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

- β_0 – Y Intercept
- β_1 – Slope of Y variable with variable X1
- β_2 – Slope of Y variable with variable X2
- β_k – Slope of Y variable with variable X_k
- e – Random error in Y for observation i

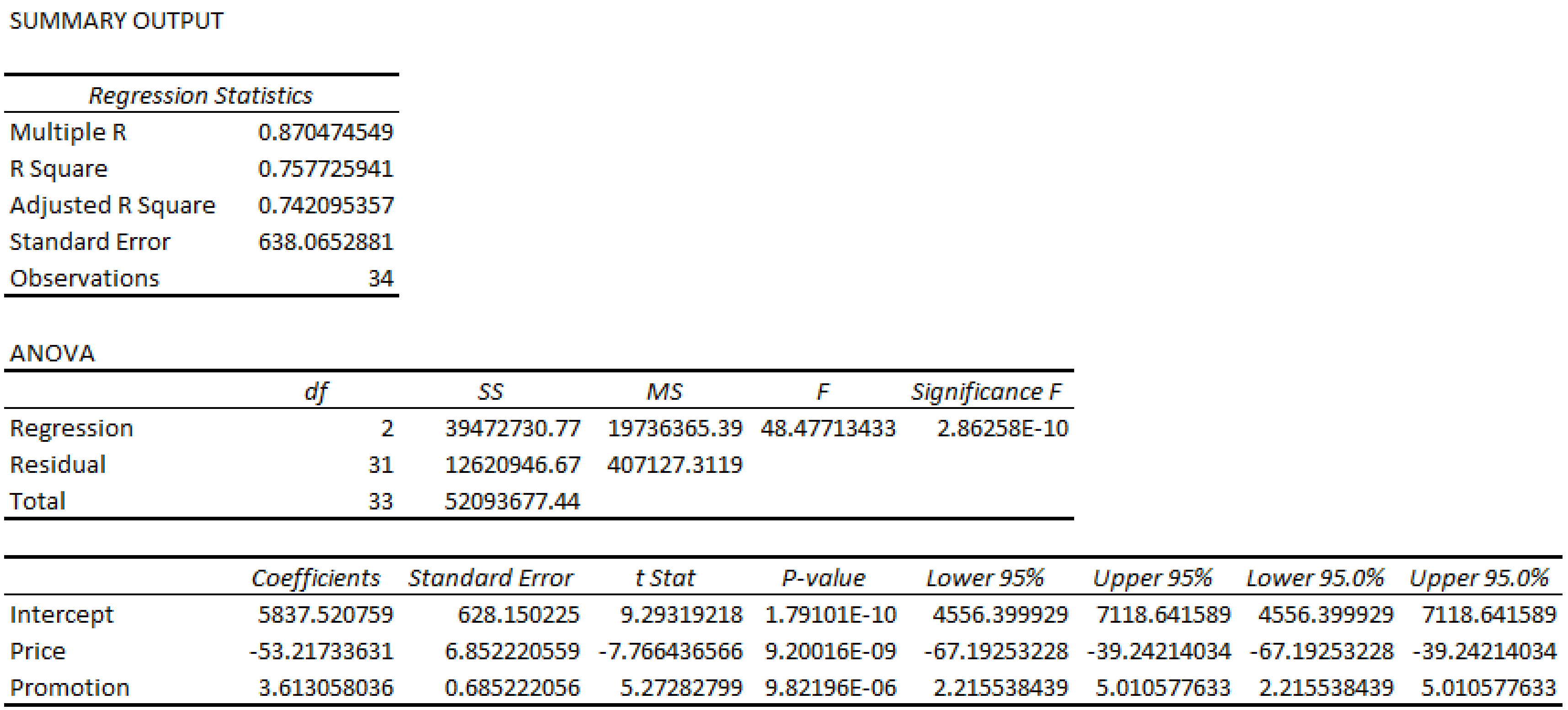
Sample Data





Store	Sales	Price	Promotion
1	4141	59	200
2	3842	59	200
3	3056	59	200
4	3519	59	200
5	4226	59	400
6	4630	59	400
7	3507	59	400
8	3754	59	400
9	5000	59	600
10	5120	59	600
11	4011	59	600
12	5015	59	600
13	1916	79	200
14	675	79	200
15	3636	79	200
16	3224	79	200
17	2295	79	400
18	2730	79	400
19	2618	79	400
20	4421	79	400
21	4113	79	600
22	3746	79	600
23	3532	79	600
24	3825	79	600
25	1096	99	200
26	761	99	200
27	2088	99	200
28	820	99	200
29	2114	99	400
30	1882	99	400
31	2159	99	400
32	1602	99	400
33	3354	99	600
34	2927	99	600

Regression output for sample data



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.870474549
R Square	0.757725941
Adjusted R Square	0.742095357
Standard Error	638.0652881
Observations	34

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	39472730.77	19736365.39	48.47713433	2.86258E-10
Residual	31	12620946.67	407127.3119		
Total	33	52093677.44			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	5837.520759	628.150225	9.29319218	1.79101E-10	4556.399929	7118.641589	4556.399929	7118.641589
Price	-53.21733631	6.852220559	-7.766436566	9.20016E-09	-67.19253228	-39.24214034	-67.19253228	-39.24214034
Promotion	3.613058036	0.685222056	5.27282799	9.82196E-06	2.215538439	5.010577633	2.215538439	5.010577633

$$\hat{Y}_i = 5{,}837.5208 - 53.2173X_{1i} + 3.6131X_{2i}$$

Predicting the Dependent Variable Y

- Multiple regression equation to predict values of the dependent variable.
- For example, what are the predicted means sales for a store charging 79 cents during a month in which promotional expenditures are $400?

$$\hat{Y}_i = 5{,}837.5208 - 53.2173X_{1i} + 3.6131X_{2i}$$

$$\begin{aligned}\hat{Y}_i &= 5{,}837.5208 - 53.2173(79) + 3.6131(400)\\ &= 3{,}078.57\end{aligned}$$

- For the stores charging 79 cents and $400 in promotional expenditures the mean sales will be 3,078.57

Methods used to evaluate the overall multiple regression model

Methods used to evaluate the overall multiple regression model

- Rsquare
- adjusted Rsquare
- RMSE
- the Overall F test

Coefficient of Multiple Determination

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{SSR}{SST}$$

- Using the sample store data

- $r^2 = 0.7577$

- This indicates that 75.77% of the variation of sales is explained by the variation in the price and in the promotional expenditures.

Adjusted r^2

- In case of multiple linear regression, SSE will decrease as the number of explanatory variables increases, and SST remains constant.

- So, it is possible that r^2 will increase even when there is no statistically significant relationship between the explanatory variable and the response variable.

- To compensate this, r^2 is adjusted by normalizing both SSE and SST with the corresponding degrees of freedom.

Adjusted r^2

- The adjusted r^2 takes into account the number of independent variables as well as the sample size.

$$r^2_{\text{adj}} = 1 - \left[(1 - r^2)\frac{n - 1}{n - k - 1}\right]$$

- $r^2 = 0.7577$; $n=34$; $k=2$

$$\begin{aligned} r^2_{\text{adj}} &= 1 - \left[(1 - 0.7577)\frac{34 - 1}{34 - 2 - 1}\right] \\ &= 1 - \left[(0.2423)\frac{33}{31}\right] \\ &= 1 - 0.2579 \\ &= 0.7421 \end{aligned}$$

- 74.21% of the variation in sales is explained by the multiple regression model adjusted for the number of independent variables and sample size.

RMSE – Root Mean Square Error

- The RMSE is the square root of the variance of the residuals.

- It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values.

- Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit.

- As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable.

- Lower values of RMSE indicate better fit.

RMSE – Root Mean Square Error

- Formula for RMSE

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

- It can also be constructed using below,

$$RMSError = \sqrt{1 - r^2} SD_y$$

Test for the significance of the overall multiple regression model

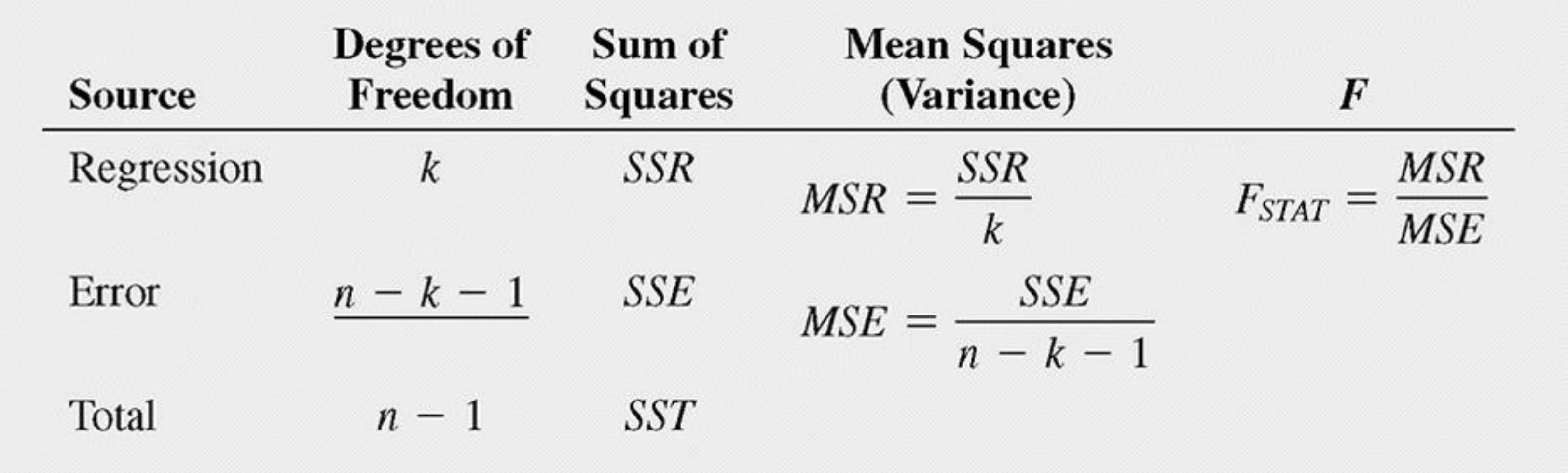
- Overall F test is used to determine whether there is a significant relationship between the dependent variable and the entire set of independent variables.

- H0: β1 = β2 = … = βk = 0
- (There is no linear relationship between the dependent variable and the independent variables)

- H1: At least one βj ≠ 0, j = 1,2,..,k
- (There is a linear relationship between the dependent variable and the independent variables)

Overall F test

- F_{STAT} = MSR / MSE



Source	Degrees of Freedom	Sum of Squares	Mean Squares (Variance)	*F*
Regression	k	SSR	$MSR = \frac{SSR}{k}$	$F_{STAT} = \frac{MSR}{MSE}$
Error	$n - k - 1$	SSE	$MSE = \frac{SSE}{n - k - 1}$	
Total	$n - 1$	SST		

- The decision rule is
- Reject H0 at the α level of significance if $F_{STAT} > F\alpha$;
- Otherwise do not reject H0

Overall F test

- Using the sample store data, the critical value of the F distribution with 2 and 31 degrees of freedom is apprx 3.32
- Using the regression summary output, F_{STAT} test statistic is 48.4771

- F-stat = 48.4771 > 3.32
- P-value is 0.000 < 0.05

- H0 is rejected and we conclude that at least one of the independent variables is related to the sales.

Limitations of Multiple Regression Model

- Multicollinearity – If the independent variables are correlated with each other. If they are, then the regression coefficients cannot be estimated

- Wild fluctuations in one or more of the independent variables crumbles the model and will be highly unreliable

- In forecasting problems, multiple regression, at best, can work for short and medium term only. It cannot be successfully used for long term forecasting.

Multicollinearity

- Multicollinearity arises when intercorrelations among the predictors are very high.

- Multiple regression and Stepwise regression are complicated by the presence of multicollinearity

Multi-collinearity and Interactions

Multicollinearity

Multicollinearity can result in several problems, including the following:

i. The partial regression coefficients may not be estimated precisely. The standard errors are likely to be high.

ii. The magnitudes as well as the signs of the partial regression coefficients may change from sample to sample.

iii. It becomes difficult to assess the relative importance of the independent variables in explaining the variation in the dependent variable.

iv. Predictor variables may be incorrectly included or removed in stepwise regression.

Multicollinearity

- Variance Inflation Factor

- It is a measure used for identifying the existence of multi-collinearity

- VIF = $1/(1-R_i^2)$, where i is the predictor.

- the square root of a given variable’s VIF shows how much larger the standard error is, compared with what it would be if that predictor were uncorrelated with the other features in the model.

- If no features are correlated, then all values for VIF will be 1.

Interaction (cross-product term)

- An interaction occurs if the effect of an independent variable on the dependent variable changes according to the value of a second independent variable.

- Example:
- When the price of a product is low, It is possible that advertising will have a large effect on sales of a product
- However, if the price of the product is too high, increases in advertising will not dramatically change sales.
- Here, price and advertising are said to interact.

Note:

- Interactions does not imply collinearity and collinearity does not imply there are interactions

Dummy Variables

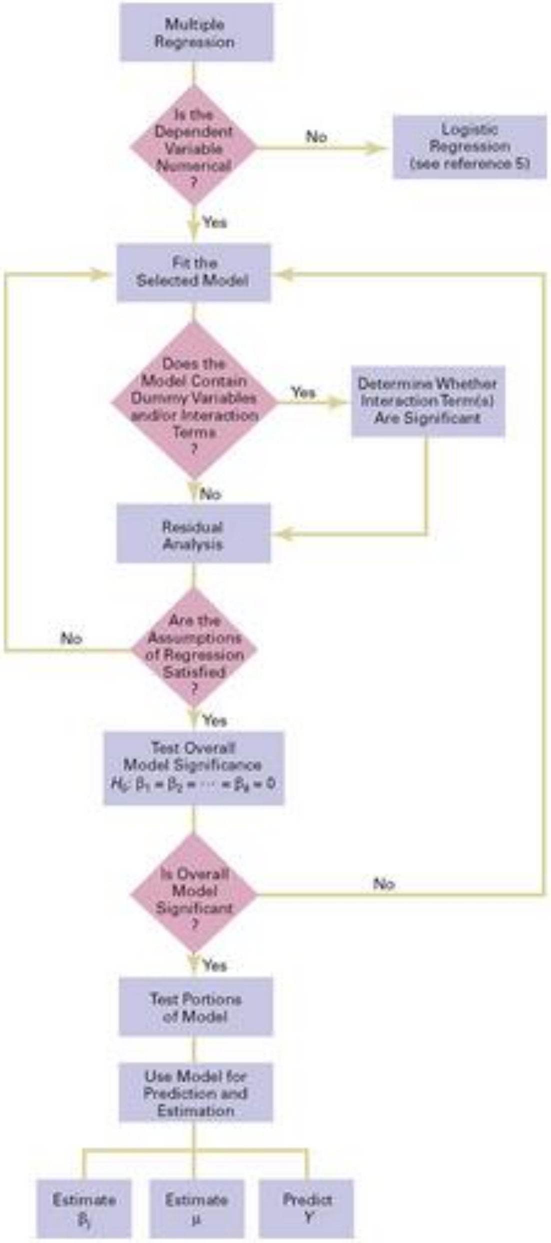
Dummy Variables

- Dummy variable is used to include a categorical independent variable in a regression model

- A variable Xd recodes the categories of a categorical variable using the numerical values 0 and 1

- In the special case of a categorical independent variable that has only two categories, only one dummy variable is defined Xd, and use the values 0 and 1 to represent the two categories.

Roadmap for Multiple Regression


Multiple Regression
Is the Dependent Variable Numerical ?
No
Logistic Regression (see reference 5)
Yes
Fit the Selected Model
Does the Model Contain Dummy Variables and/or Interaction Terms ?
Yes
Determine Whether Interaction Term(s) Are Significant
No
Residual Analysis
Are the Assumptions of Regression Satisfied ?
No
Yes
Test Overall Model Significance
$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$
Is Overall Model Significant ?
No
Yes
Test Portions of Model
Use Model for Prediction and Estimation
Estimate β_j
Estimate μ
Predict Y

References

Books:

- Marketing Research An Applied Orientation – Naresh K.Malhotra and Satyabhushan Dash
- Business Statistics A First Course – David M Levine, Kathryn A Szabat, David F Stephan and P.K. Viswanathan
- Applied Predictive Analytics – Dean Abbott

Others:

- https://cran.r-project.org/web/packages/olsrr/vignettes/residual_diagnostics.html
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6387894/
- ftp://ess.r-project.org/Software/CRAN-attic/src/contrib/lmtest.pdf
- https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/
- http://statweb.stanford.edu/~susan/courses/s60/split/node60.html

Thank You