2012

Analyzing Relations between dataset-IV

Author: Naseha Sameen

About: Multiple Regression what it is and how and when to use

Multiple Regression

LAB NOTEBOOK

NASEHA SAMEEN

Contents

Multiple Regression

Multiple Regression Meaning and Perspective:

Multiple Regression tells us the exact kind of **linear association that exists between a single dependent variable and several independent variable**. It tells us the exact kind of linear association that exists between those variables. It is the simultaneous combination of multiple factors to assess how and to what extent they affect a certain outcome.

The **technique breaks down when the nature of the factors** themselves is of an unmeasurable or **pure-chance nature**. It is extremely powerful when you are trying to develop a "model" for predicting a wide variety of outcomes.

Regression equation allows us to express the relationship between two (or more) variables algebraically. Multiple Regression is used to study the relationship. Using multiple regression we can test models about precisely which set of variables is influencing the outcome.

Multiple Regression tells you the following characteristics about the relationship between the dependent and independent or predictor variable

- **Direction** - Whether the relationship is positive (+) or negative (-).
- **Magnitude** - The size of the correlation coefficient dependent and the set of predictor variables indicates the ***relative*** importance of each predictor.
- **Nature** - If the relationship is linear or other types relationship such curvilinear.
- **Indicate Redundancy** - The relationship may make some predictive variable redundant in their predictive effort and are not needed to produce the optimal prediction. Individually, a particular independent variable may be correlated to the dependent variable but together with other independent variable, the variable may not be needed as other variables are explaining the variance.

Type of Multiple Regression:

There are three types of Multiple Regression – Standard, Hierarchical, Stepwise Regression. The application depends on the type of scenario or problem we are trying to solve for.

Line Item	Standard	Hierarchical	Stepwise
Where is it used	To evaluates relationship between dependent and independent variable	To evaluates the relationship between dependent and independent variable after controlling the effects of some independent variable on the dependent variable	To identify the subset of independent variables that has strongest relation to the dependent variable and most effective in predicting the dependent variable
Variable Analyzed	All at once. All variable are entered in regression equation together, at the same time	Variables are entered in two stages – 1st where all independent variables which are to be controlled are entered. 2nd – all variables whose relationship is to be	Variables are added to the regression equation one at a time, using the statistical criterion of maximizing the R^2 of the included variables.

		examined after controls are entered	
When to use	Multiple R and R^2 to measure Strength of the relationship between independent & dependent variables	A statistical test of the change in R^2 from the first stage is used to evaluate the importance of the variables entered in the second stage	When none of the possible addition can make a statistically significant improvement in R^2, the analysis stops.
			Semipartial correlations are used. In a forward stepwise regression, the variable which would add the largest increment to R^2 (i.e. the variable which would have the largest semipartial correlation) is added next (provided it is statistically significant). In a backwards stepwise regression, the variable which would produce the smallest decrease in R^2 (i.e. the variable with the smallest semipartial correlation) is dropped next (provided it is not statistically significant.)

Multiple Regression is not Multivariate Regression

Regression analysis is used to predict the value of one or more responses from a set of predictors. It can also be used to estimate the linear association between the predictors and responses. Predictors can be continuous or categorical or a mixture of both.
When there are several (p >1) criterion variables, we could just fit p separate models

$Y_1 = X\beta_1$
$Y_2 = X\beta_2$
$Y_3 = X\beta_3$
…
…
$Y_p = X\beta_p$

Multivariate tests provide a way to understand the structure of relations across separate response measures. In particular: how many “dimensions” of responses are important? And how do the predictors contribute to these? However, they do not give simultaneous tests for all regressions and does not take correlations among the y’s into account

Difference between Multiple & Multivariate Regression

Multiple Regression	Multivariate Regression
Multiple Regression applies to the number of predictors that enter the model (or equivalently the design matrix) with a single outcome (Y response),	Multivariate refers to a matrix of response vectors. Here we have multiple dependent variables and multiple independent variables.
Example - Suppose that a university wishes to refine its admission criteria so that they admit 'better' students.	
Dependent Variable = Student Grade (G)	Dependent Variable = Student Grade for say 5 yrs (G1, G2, G3, G4, G5)
Multiple independent variables = Attendance, Gender, Term End Marks, Extra Curricular	Multiple independent variables = Attendance, Gender, Term End Marks, Extra Curricular
Result from Analysis - which one of the independent variables are good predictors for the dependent variable.	Result from the Analysis - track student performance across time and which one of the independent variables predict G scores better performance across time.
Which of the independent variables should be considered and which one to be ignored while student's admission	*Same* independent variables predict performance across time so that their choice of admissions criteria ensures that student performance is consistently high across all four years.
Independent variables 1 Regression Y = XB	Independent variables 2+ Regression Y = X B
Categorical – ANOVA	Categorical – MANOVA

Objective of using Multiple Linear Regression:

There are two general applications for multiple regression: prediction and explanation and also to test hypothesis in model. However, the last point is a subset of multiple regression for explanation.

Multiple Regression for Prediction:

When one uses MR for prediction, one is using a sample to create a regression equation that would optimally predict a particular phenomenon within a particular population. Here the goal is to use the equation to predict outcomes for individuals not in the sample used in the analysis.

Also, to improve the accuracy in predicting values

Multiple Regression for Explanation:

When one uses is for explanatory purposes to explore relationships between multiple variables in a sample to shed light on a phenomenon, with a goal of generalizing this new understanding to a population.

A second use of multiple regression is to try to understand the functional relationships between the dependent and independent variables, to try to see what might be causing the variation in the dependent variable.

Multiple regression is a statistical way to try to answer questions like "If all the other measured variable x2 to xn remained same, would regression variable y on variable x1 be significant?"

Multiple Regression for Hypothesis testing in Model:

It helps to test the main three hypothesis explaining the relationship in a model

- The variation explained by the model is not due to chance. This is determined by F test.
- That the slope of the regression line is significantly different than zero. This is determined by t test of the β parameter.
- That the y intercept is significantly different than zero (t test of the constant parameter). This test result can be ignored unless there is some reason to believe that the y intercept should be zero

Factors that can affect the result of Multiple Regression

- These factors can provide incorrect result and inference, hence these are to be checked and removed before modelling data

Nature of Dependent Variable

- **Multiple regression** analysis is used when **one is interested in predicting a continuous dependent variable from a number of independent variables**. If **dependent variable is dichotomous, then logistic regression** should be used.

- If the split between the two levels of the dependent variable is close to 50-50, then both logistic and linear regression will end up giving you similar results.

- The independent variables used in regression can be either continuous or dichotomous. Independent variables with more than two levels can also be used in regression analyses, but they first must be converted into variables that have only two levels, called Dummy Coding

- One point to keep in mind with regression analysis is that causal relationships among the variables cannot be determined. While the terminology is such that we say that X "predicts" Y, we cannot say that X "causes" Y.

- Error in status of Independence state of the variable leads to theoretical conclusions, inflated standard errors

Multicollinearity

- In practice, the problem of multicollinearity occurs when some of the x variables are highly correlated. It can have significant impact on the quality and stability of the model by leading to

i. **Unstable partial regression coefficients** which won't hold up when applied to a new sample of cases.

ii. Their **shared variance with the dependent or criterion variable may be redundant**. Regression weights that don't really reflect the independent contribution to prediction of each of the predictors

iii. **Independent Variable that has more than a .3 correlation with the dependent variable and less than .7 with any other Independent variable can be possible multicollinear predictor**.

- There are two ways to detect multicollinearity
  - i. The presence of multicollinearity can be detected by examining the correlation matrix (say r= ± 0.9 and above).
  - ii. Check - the correlation matrix, if there is are large correlations between pairs of explanatory variables.
  - iii. 'Tolerance' associated with a predictor. The tolerance of Xi is defined as 1 – R^2 correlation between that x_i and the remaining x variables.
  - iv. Tolerance = 1– R^2
  - v. The inverse of the tolerance is called the **variance inflation factor (VIF).** 1/tolerance
  - vi. The **higher VIF**, the **greater the multicollinearity. When there is no multicollinearity the value of VIF equals 1.**
  - vii. Multicollinearity problems have to be dealt with (by getting rid of redundant predictor variables or other means) if VIF approaches 10 (means 10% of the variance in the predictor is not explained by the combination of the other predictors)
  - viii. **Lesser the VIF, better is the model.**

    Error in Collinearity to inflated standard errors

Singularity

- Singularity exists when there is perfect correlation between explanatory variables. The presence of either affect the interpretation of the explanatory variables effect on the response variable.

Endogeneity

- Regression measures the **effect of changes in the independent variable on the dependent variable**. Endogeneity **occurs when that relationship is either backwards or circular**, **meaning that changes in the dependent variable cause changes in the independent variable**. This circular relationship, if it is strong, can bias the results of the regression. Try to remove endogenous variables. There are strategies for reducing the bias if removing the endogenous variable is not an option.

- In the home value example, perceived quality of the local schools might affect home values. But the perceived quality is likely also related to the actual quality, and the actual quality is at least partially a result of funding levels. Funding levels are often related to the property tax base, or the value of local homes. So… **good schools increase home values, but high home values also improve schools. This circular relationship**

Semi partial (Part) Correlation

- **Partial and semi partial correlations** provide another means of assessing the **relative "importance" of independent variables in determining Y**. Basically, they show how much each variable uniquely contributes to R^2 over and above that which can be accounted for by the other IVs.

- **Semi partial correlations (also called part correlations) indicate the "unique" contribution of an independent variable**. Specifically, the squared semi partial correlation for a variable tells us how much **R^2 will decrease if that variable is removed** from the regression equation.

- To get Xk's unique contribution to R^2, first regress Y on all the X's. Then regress Y on all the X's except Xk. The difference between the R^2 values is the squared semi partial. Alternatively, the standardized coefficients and the Tolerances can be used to compute the semi partials and squared semi partials. The more "tolerant" a variable is (i.e. the less highly correlated it is with the other IVs), the greater its unique contribution to R2 will be. This is generally used in Step Regression

Partial Correlation

- The partial correlation coefficient can be viewed **as an adjustment of the simple correlation taking into account the effect of a control variable**: r(X ; Y / Z) i.e. correlation between X and Y controlled for Z.
- **Partial Correlation gives the influence of one Independent Variable over Dependent Variable when some variables are held as constant while examining the relations between X and Y.** With assignment we can do this by design. If we regress variable X on variable Z, then subtract X' from X, we have a residual e. This e will be uncorrelated with Z, so any correlation X shares with another variable Y cannot be due to Z.
- **Partial correlation analysis is aimed at finding correlation between two variables after removing the effects of other variables.** This type of analysis **helps spot spurious correlations** (i.e. correlations explained by the effect of other variables) as well as to **reveal hidden correlations** - i.e correlations masked by the effect of other variables.
- The central concept in partial correlation analysis is the partial correlation coefficient rxy.z between variables x and y , adjusted for a third variable z . Both x and y are presumed to be linearly related to z :

$$x = Az + B + d_x;$$

$$y = Cz + D + d_y;$$

- The partial correlation coefficient rxy.z is defined as the correlation coefficient between residuals dx and dy in this model.
- The partial correlation coefficient rxy.z between x and y adjusted for z may be computed from the pairwise values of the correlation between variables x , y , and z (r_{xy}, r_{yz}, r_{xz}) :

- $r_{xy.z}$ = • r_{xy} □ r_{xz} r_{yz}/Square Root((1□r_{xz}^2)(1□r_{yz}^2))

- The rxy.z takes on values between -1 and 1.
- Another way of denoting Partial correlation is For example $r_{12.34}$ is the correlation of variables 1 and 2, controlling for variables 3 and 4. Partial correlation $r_{12.34}$

How to Measure degree of Relationship:

The computations are more complex, however, because the interrelationships among all the variables must be taken into account in the weights assigned to the variables.

Things get much more complicated when your multiple independent variables are related to with each other. In other words, when the independent variables “interact” with each other as well as with the dependent variable. In this case, in order to be able to make predictions you need to break all of the correlations down so that you can figure out the value of multiple R.

In multiple regression analysis, the end goal is to find the nature of the relationship itself between the dependent variable and the independent variable called Predictor Variables

The analysis consists of choosing and fitting an appropriate model, done by the method of **least squares**, with a view to exploiting the relationship between the variables to help estimate the expected response for a given value of the independent variable.

For example, if we are interested in the effect of age on height, then by fitting a regression line, we can predict the height for a given age.

A linear regression equation is usually written
Y = m1*x1 + m2*x2 + m3*x3 + m4*x4 + mn*xn + b + e

where
Y is the dependent variable
b is the intercept
m1 is the slope or regression coefficient of variable x1
x1 is the independent variable
M2 is the slope or regression coefficient of variable x2
x2 is the independent variable
……
……
mn is the slope or regression coefficient of variable xn
xn is the independent variable
e is the error term

The equation will specify the average magnitude of the expected change in Y given a change in x_1 to x_n. The regression equation is often represented on a scatterplot by a regression line.

Linear regression quantifies goodness of fit with r^2, sometimes shown in uppercase as R^2. If you put the same data into correlation, the square of r from correlation will equal r^2 from regression.

Word of Caution: Observations are independent and the depended variable Y should be random. The depended variables (response) should be normally distributed.

Multiple regression is "dependence technique', it is required to specify the dependent and independent variables. Extra care in selecting the depended and independent variable. The dependent Variable should be a metric or continuous variable for linear regression. If your Dependent Variable is categorical such 1 = low, 2 = average and 3 = high, then a different regression method called Logistic Regression should be used for categorical variable.

Example

Purpose to determine the relationship

To determine relationship between years in schools, motivation and earning.

Independent Variable 1 (X1) Years in School	Independent Variable 2 Dependent Variable (X2) Motivation as measured by Higgins Motivation Scale	Dependent Variable (X2) (Y) Annual Sales in Dollars
12	32	$350,000
14	35	$399,765
15	45	$429,000
16	50	$435,000
18	65	$433,000

Correlation School and Motivation	0.968	(rx1,x2)
Correlation School and Sales	0.880	(rx1,y)

The Formula for R

$$R = \sqrt{\frac{\left[(r_{y,x1})^2 + (r_{y,x2})^2\right] - (2r_{y,x1}r_{y,x2}r_{x1,x2})}{1-(r_{x1,x2})^2}}$$

$$R = \sqrt{\frac{((.880)^2 + (.772)^2) - (2(.880)(.772)(.968))}{1-(.968)^2}}$$

Correlation Sales and Motivation	0.772	(rx2,y)

R = Sqrt(8762), R=0.9360

Same result can be obtained by using Data Analysis -> Regression. Multiple R value is used here.

Making Predictions:

$Y' = a + b_1X_1 + b_2X_2$

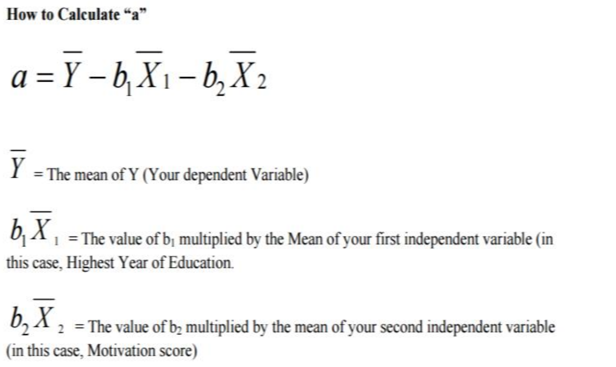
Y' = A predicted value of Y (which is your dependent variable)

a = The "Y Intercept"

b1 = The change in Y for each 1 increment change in X_1 (In our case, this is Highest Year of School)

b2 = The change in Y for each 1 increment change in X_2 (In our case, this is Motivation)

X = an X score (X is your Independent Variable) for which you are trying to predict a value of Y)



How to Calculate "a"

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$$

$\overline{Y}$ = The mean of Y (Your dependent Variable)

$b_1\overline{X}_1$ = The value of b_1 multiplied by the Mean of your first independent variable (in this case, Highest Year of Education.

$b_2\overline{X}_2$ = The value of b_2 multiplied by the mean of your second independent variable (in this case, Motivation score)

$$b_1 = \left(\frac{r_{y,x1} - r_{y,x2}r_{x1,x2}}{1-(r_{x1,x2})^2}\right)\left(\frac{SD_y}{SD_{x1}}\right) \qquad b_2 = \left(\frac{r_{y,x2} - r_{y,x1}r_{x1,x2}}{1-(r_{x1,x2})^2}\right)\left(\frac{SD_y}{SD_{x2}}\right)$$

$$b_1 = \left(\frac{(.880)-(.772)(.968)}{1-(.968)^2}\right)\left(\frac{36{,}116.693}{2.236}\right) \qquad b_2 = \left(\frac{(.772)-(.880)(.968)}{1-(.968)^2}\right)\left(\frac{36{,}116.693}{13.164}\right)$$

Calculating "a"

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$$

$$a = 409{,}353 - (34{,}356.085)(15) - (-3{,}657.213)(45.4)$$

Easier Still, the co-efficeints column in Regression output give us the Intercept and b_1, b_2, etc

	Coefficients
Intercept	56200.49309
Years in School (b1)	**34108.62442**
Motivation (b2)	-3490.679724

Make a Prediction for potential salesperson with 13 years of education 49 on Motivation. How much money in sales this person would bring in on an annual basis?

Years in School = 13, Motivation = 49

$Y' = a + b_1X_1 + b_2X_2$

= 56200.49309 + 34108.62442*13 + -3490.679724*49 = 328569.3041

The appropriateness of the multiple regression model as a whole can be tested by the F-test in the ANOVA table. A significant F indicates a linear relationship between Y and at least one of the X's.

Once a multiple regression equation has been constructed, one can check how good it is (in terms of predictive ability) by examining the coefficient of determination (R^2). R^2 always lies between 0 and 1.

R^2 - coefficient of determination

All software provides it whenever regression procedure is run. The closer R^2 is to 1, the better is the model and its prediction. We need to take care of the assumptions.

- We need to check if the independent variables individually influence the dependent variable significantly by **t-test**.
- **If the t-test of a regression coefficient is significant**, it indicates that the variable is in question influences Y significantly while controlling for other independent explanatory variables.

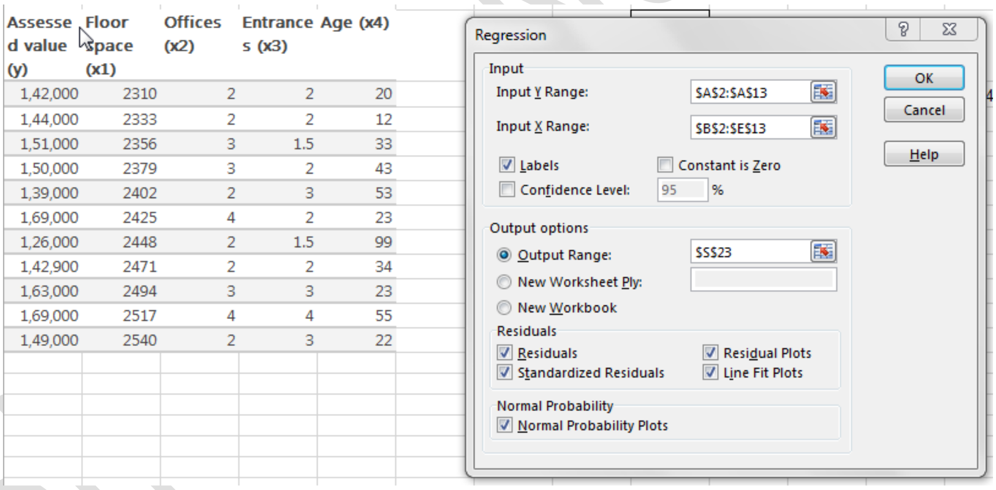
- **Nonexistence of multicollinearity**- the independent variables are not related among themselves. At a very basic level, this can be tested by computing the correlation coefficient between each pair of independent variables.
- The independent variables that statistically significant are indicated by
    - calculated t-statistics that exceed the critical values, and
    - the calculated p-values that are less than the significance level of 5%.

How to Calculate Multiple Regression:

There are three ways to calculate r in Excel

- Using the formulae
  1. LINEST - Syntax(List of variables for known y, List of variable for known x, optional value for stats and optional value for b - const)
     - Select g+1 cells where g is number of independent variables.
     - Type LINEST and Select the ranges and then press Ctrl+Shift+Enter
     - The last value is the intercept b, and the other values from 1 to g are the slope for variable x1 to xn
     - Replace them in the formulae Y = m1*x1 + m2*x2 + m3*x3 + m4*x4 + mn*xn + b + e
     - to find out value of Y
     - If n is the number of data points and const = TRUE or omitted, then v1 = n - df - 1 and v2 = df. (If const = FALSE, then v1 = n - df and v2 = df.)

- Using Statistical Tool in Excel
  1. Under the Tab "DATA", click on the Option, "Data Analysis"
  2. Click on Regression
  3. Select the range in Input for X and Y, Output and confident level
  4. Results are in two boxes



Assessed value (y)	Floor space (x1)	Offices (x2)	Entrances (x3)	Age (x4)
1,42,000	2310	2	2	20
1,44,000	2333	2	2	12
1,51,000	2356	3	1.5	33
1,50,000	2379	3	2	43
1,39,000	2402	2	3	53
1,69,000	2425	4	2	23
1,26,000	2448	2	1.5	99
1,42,900	2471	2	2	34
1,63,000	2494	3	3	23
1,69,000	2517	4	4	55
1,49,000	2540	2	3	22

Regression

Input

Input Y Range: $A$2:$A$13

Input X Range: $B$2:$E$13

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

● Output Range: $S$23

○ New Worksheet Ply:

○ New Workbook

Residuals

☑ Residuals ☑ Residual Plots

☑ Standardized Residuals ☑ Line Fit Plots

Normal Probability

☑ Normal Probability Plots

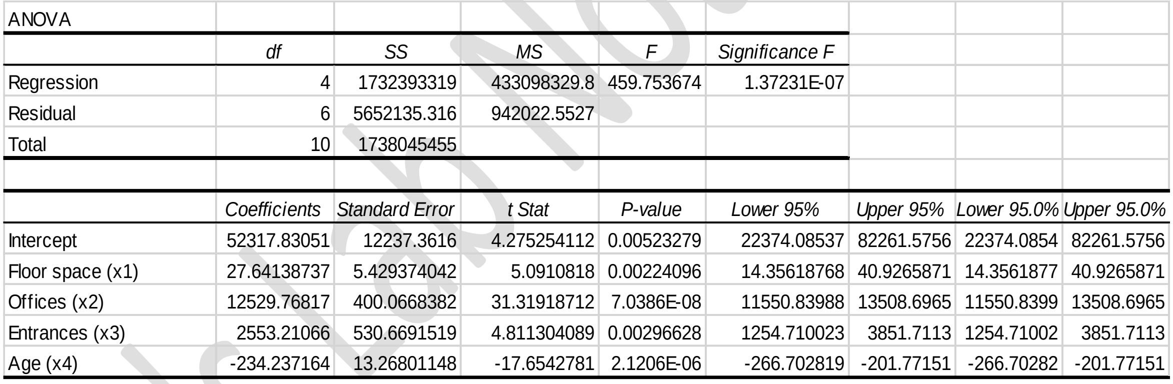
SUMMARY OUTPUT		
Regression Statistics		*Explanation*
Multiple R	0.998373	Correlation Coefficient
R Square	0.996748	Coefficient of Determination – Sq of R. Look at Result fits, rational for the model, & P values for interpretation
Adjusted R Square	0.99458	The adjusted R^2"penalizes" you for adding the extra predictor variables that don't improve the existing model.
Standard Error	970.5785	Standard Error calculated on Residue. a measure of the statistical accuracy of an estimate
Observations	11	No of Samples

v. Summary Output

- Multiple R: It is sqrt of R^2, which is noted by correlation between 0 and 1, and closer to 1 indicates stronger relation. It won't be negative.

- R Square: The term R-squared refers to the fraction of variance explained by a model. A low R square doesn't negate a significant predictor or change the meaning of its coefficient. **Look at Result fits, rational for the model, & P values for interpretation**
  Adjusted R square: The adjusted R^2"penalizes" you for adding the extra predictor variables that don't improve the existing model. It is always lower than the R-squared. **Select the model with variables giving higher Adjusted R^2**

- Standard Error: a measure of the statistical accuracy of an estimate, equal to the standard deviation of the theoretical distribution of a large population of such estimates. SQRT(RSS/(T – 2)) where T is the sample size. We reduce 2 from the sample size to account for the loss of two degrees of freedom, one for the regression estimate itself, and the second for the explanatory variable.



ANOVA									
	df	*SS*	*MS*	*F*	*Significance F*				
Regression	4	1732393319	433098329.8	459.753674	1.37231E-07				
Residual	6	5652135.316	942022.5527						
Total	10	1738045455							
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	
Intercept	52317.83051	12237.3616	4.275254112	0.00523279	22374.08537	82261.5756	22374.0854	82261.5756	
Floor space (x1)	27.64138737	5.429374042	5.0910818	0.00224096	14.35618768	40.9265871	14.3561877	40.9265871	
Offices (x2)	12529.76817	400.0668382	31.31918712	7.0386E-08	11550.83988	13508.6965	11550.8399	13508.6965	
Entrances (x3)	2553.21066	530.6691519	4.811304089	0.00296628	1254.710023	3851.7113	1254.71002	3851.7113	
Age (x4)	-234.237164	13.26801148	-17.6542781	2.1206E-06	-266.702819	-201.77151	-266.70282	-201.77151	

vi. M1 (slope is indicated in Coefficient X1 to Xn, and intercept b is 52317)

vii. So our regression equation is Y = 27.67(X1)+12529(x2)+2553(x3)+(-234(x4))+52317

viii. **Adjusted R square gives idea of the goodness of fit measure**, here it says 99.4% of Y is determined by X

ix. If **Significance F should be lesser than 0.05**. If is greater than 0.05, it's probably better to stop using this set of independent variables. Delete a variable with a high P-value (greater than 0.05) and rerun the regression until Significance F drops below 0.05.

x. Most or all P-values should be below 0.05.

xi. T-stat: use sample data to test hypotheses about an unknown population mean. the t-statistic is a ratio of the departure of an estimated parameter from its notional value and its standard error.

xii. Evaluating the Fitness of the Model Using Confidence Intervals – Lower and Upper 95%, it indicates the upper and lower bound within which 95% of the data exists.

- Word of caution, when you use a regression equation, do not use values for the independent variable that are outside the range of values used to create the equation. That is called **extrapolation**, and it can produce unreasonable estimates.

  Do not use Slope and intercept formulae or use Correl and RSQ method for each component and replace them in the equation.

Excel restricts the number of regressors up to 16 and all the regressor variables be in adjoining columns.

Creating a Model

Multiple Regression with Scenario and Hypothesis testing and Interpretation from table

Goals

- To minimize the sum of the squares of the errors between variable
- To understand the influence of each variable on one another and to whole system
- To model and predict
- Example: A researcher is attempting to create a model that accurately predicts the total annual power consumption of companies within a specific industry. The researcher has collected information from 21 companies that specialize in a single industry. The four pieces of information collected from each of the 21 companies are as follows: It is easy to mark out the non-relevant ones in the data collected
    1) The company’s total power consumption last year in kilowatts.
    2) The company’s total number of production machines.
    3) The company’s number of new employees added in the last five years.
    4) The company’s total increase in salary paid over the last five years.

Check for Potential Problems with linear regression which needs to be avoided

- Some of our assumptions may be violated
- Non-linearity of the response-predictor relationships
- Correlation of error terms
- Non-constant variance of error terms (Heteroskedasticity)
- Outliers
- High-leverage points
- Collinearity

Step 0 – Variables

Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?

The appropriateness of the multiple regression model as a whole can be tested by the F-test in the ANOVA table. A significant F indicates a linear relationship between Y and at least one of the X's.

F Test

An F test is used to determine if the relationship can be generalized to the population represented by the sample.

Another method of determining the best model for prediction is to test the significance of adding one or more variables to the model using the *partial F-test*. In general, the *partial F-test* is similar to the *F-test* used in analysis of variance. It assesses the statistical significance of the difference between values for R^2 derived from 2 or more prediction models using a subset of the variables from the original equation.

Significant F

There is also a **significance level for the model as a whole**. This measures the likelihood that the model as a whole describes a relationship that emerged at random, rather than a real relationship.

- **The lower the significance F value, the greater the chance that the relationships in the model are real.**

- This indicates the probability that the Regression output could have been obtained by chance. A small Significance of F confirms the validity of the Regression output. For example, if Significance of F = 0.030, there is only a 3% chance that the Regression output was merely a chance occurrence.

This tests the full model against a model with no variables and with the estimate of the dependent variable being the mean of the values of the dependent variable.

Significance F gives us the probability at which the F statistic becomes 'critical', ie below which the regression is no longer 'significant'.

This is calculated (as explained in the text above) as =FDIST(F-statistic, 1, T-2), where T is the sample size. In this case, =FDIST(9.126559714795,1,8) = 0.0165338014602297

If F has a value of 0.01000 then there is 1 chance in 100 that all of the regression parameters are zero. This low a value would imply that at least some of the regression parameters are nonzero and that the regression equation does have some validity in fitting the data (i.e. the independent variables are not purely random with respect to the dependent variable

F Value or Ratio

The F Value or **F ratio is the test used to decide whether the model as a whole has statistically significant predictive capability**. F value tests the full model against a model with no variables and with the estimate of the dependent variable being the mean of the values of the dependent variable. It is a part of Test statics and F Test

Under the null hypothesis that the model has no predictive capability, which means that all the results are purely due to chance and they cannot be predicted. **The null hypothesis is rejected if the F ratio is large**. High F ratio or Value means predictions can be done.

F ratio tests whether the regression SS is big enough, considering the number of variables needed to achieve it. Its value will range from zero to an arbitrarily large number.

- The F value = mean regression sum of squares/ mean error sum of squares.

Larger an F-value indicates a consistent pattern that is unlikely due to chance. It also indicates that the model is robust and has significant effect (the denominator which has error is lesser). The simple rule in most research situations is: the higher the F value, the better…

- F = Effect Variance (or "Treatment Variance") / Error Variance Or,
- F = Between-group Variance / Within-group Variance
- F = Explained Variance / Unexplained Variance

F value is used with p value for interpretation.

- Larger R^2 produce bigger values of F. That is, the stronger the relationship is between the DV and the IVs, the bigger F will be.

- Larger sample sizes also tend to produce bigger values of F. The larger the sample, the less uncertainty there is whether population parameters actually differ from 0. Conversely, in a small sample, even large effects may not be statistically significant.

  If additional variables do not produce large enough increases in R2, then putting them in the model can actually decrease F. It may be difficult to detect important effects. The F statistic does not tell you which effects are significant, only that at least one of them is. FR could be also used to compare two models describing the same experimental data: the higher FR the more adequate the corresponding model.

  A very large F-value means that the between-group variance (the effect variance) exceeds the within-group variance (the error variance) by a substantial amount. In such cases p-value then just gives a number to how likely a particular F-value is going to occur, with lower p-values indicating that the probability of obtaining that particular F-value is pretty low. This is actually why the degrees of freedom influence the F distribution.
  We will cover more of this when we create model and interpret results.

F-Critical Value

First, use the Critical Significance Level (α: alpha) chosen in Step 2 and the between treatment (df_1= number of treatments – 1) and within treatment (df_2= total sample size – number of treatments) degrees of freedom calculated in Step 3 to find the Critical Value of F ($F_{critical}$) using a Critical Value Table such as the one below e.g., if α = 0.05, df1 = 2 and df2 = 12 then $F_{critical}$ = 5.096.
Table of Critical Values for a Critical Significance Level (α: alpha) = 0.05 for the F statistic where df_1 is the degrees of freedom between treatments (the numerator; the number of treatments - 1) and df_2 is the degrees of freedom within treatments (the denominator; the total sample size – number of treatments) for Anova.

- F-critical value is the F value above which 100% of the null sampling distribution occurs. This type of logic is used in many other types of statistical tests, which compare averages and other measures, instead of variances.

  Note that this value can be obtained from a computer before the experiment is run, as long as we know how many subjects will be studied and how many levels the explanatory variable has. Then when the experiment is run, we can calculate the observed F-statistic and compare it to F-critical. If the statistic is smaller than the critical value, we retain the null hypothesis because the p-value must be bigger than alpha, and if the statistic is equal to or bigger than the critical value, we reject the null hypothesis because the p-value must be equal to or smaller than alpha

  If Fcalculated > Fcritical, H0 is rejected.
  If Fcalculated < Fcritical, H0 cannot be rejected.

- **Compare your f-value with your f-critical value. If the f-critical value is smaller than the f-value, you should reject the null hypothesis.**

- Please refer to **F Critical Table**

Do all the predictors help to explain Y, or is only a subset of the predictors useful?

We need to examine the interaction between the predictor variables with respect to prediction. Only the variables that matters should be selected. Views differ as how this should be accomplished.

One school, **hierarchical regression** - argues that **theory should drive the statistical model** and that the decision of what and when terms enter the regression model should be determined by theoretical concerns.

Interaction terms – hierarchical principle

The additive assumption may not hold because we may have synergistic effects also known as interaction effects. To relax the additive assumption, we include interaction terms like X_1X_2.

The hierarchical principle states that if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.
In other words, if the interaction between X1 and X2 seems important, then we should include both X1 and X2 in the model even if their coefficient estimates have large p-values
Interaction applies also

- to qualitative variables
- to a combination of quantitative and qualitative variables.

Second school - **stepwise regression**, argues that the data can speak for themselves and allows the procedure to select predictor variables to enter the regression equation.

Stepwise Regression

These are -

Step Up or Forward Selection

- Start with the Null model.
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
- Then add to that model the variable that results in the lowest RSS for the new two-variable model.
- This approach is continued until some stopping rule is satisfied.

Step Down or Backward Regression

- We start with all variables in the model,
- Remove the variable with the largest p-value — that is, the variable that is the least statistically significant.
- The new (p – 1)-variable model is fit, and the variable with the largest p-value is removed. This procedure continues until a stopping rule is reached.
- For instance, we may stop when all remaining variables have a p-value below some threshold.

Mixed Selection

- This is a combination of forward and backward selection.
- We start with no variables in the model,
- As with forward selection, we add the variable that provides the best fit.
- We continue to add variables one-by-one.
- Of course, as we noted with the Advertising example, the p-values for variables can become larger as new predictors are added to the model.
- Hence, if at any point the p-value for one of the variables in the model rises above a certain threshold, then we remove that variable from the model.

- We continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

Null model

- A model that contains an intercept but no predictors

- How well does the model fit the data? (how strong is the relationship) → **RSE, R^2**
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Step 1 - Ascertain the Equation is Linear

We have already gone through the requirement for an equation to be linear: The relationship between the predictors and response are **additive** which means the effect of changes in a predictor X_j on the response Y is independent of the values of the other predictors.

- The linear assumption states that the change in the response Y due to a one-unit change in X_j is constant, regardless of the value of X_j.
- The outcome Y takes on continuous values
- Use Scatter plots to determine if the equation is linear or not and check direction, form and strength of the relationship
- The model will remain linear if it is linear in the parament vector β, even if one of the regressors can be a non-linear to another regressor

Model	Dep.	Ind.	Coefficient	Interpretations
Linear	Y	X		Constant slope
Quadratic	Y	X, X^2		Slope changes with X
Log-Linear	Ln Y	Ln X		Slope is elasticity
Exponential	Ln Y	X	Growth Model	% Change in Y from an absolute change in X
Semi-log	Y	Ln X	% Change in Money supply	Absolute change in mean of Y from % change in X

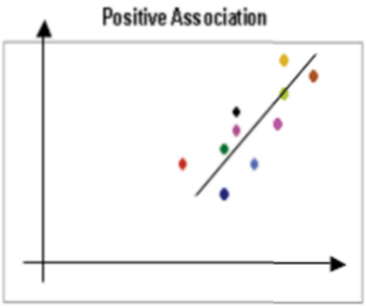
			effects on GNP by $	

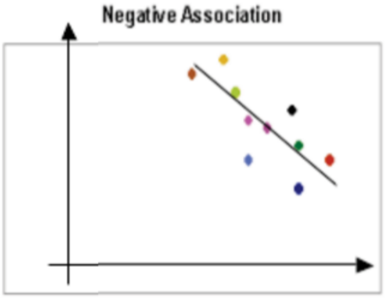
The relationship is said to be linear when:

- The relationship between the predictors and response are **additive** which means the effect of changes in a predictor X_j on the response Y is independent of the values of the other predictors.
- The linear assumption states that the change in the response Y due to a one-unit change in X_j is constant, regardless of the value of X_j.
- The outcome Y takes on continuous values
- Use Scatter plots to determine if the equation is linear or not and check direction, form and strength of the relationship

Direction

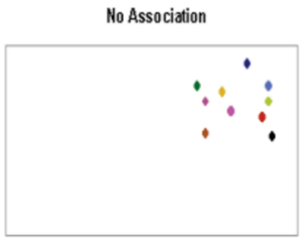
- Positive gradient: When the larger values of the horizontal (explanatory) variable are associated with larger values of the vertical (response) variable.
- Negative gradient: When the larger values of the explanatory variable are associated with smaller values of the response variable. As the explanatory variable increases, the response variable decreases.

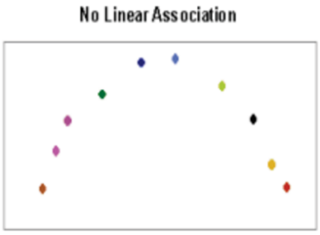


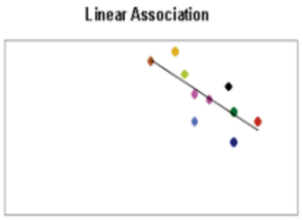


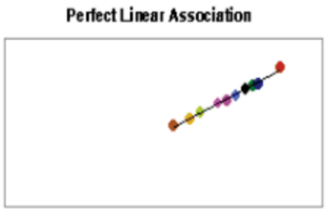
Form

- Linear or Non Linear. The relationship might be linear or curved or there might be no underlying form. In this course we will mainly concentrate on linear relationships, but we must be aware of the existence of non-linear ones.



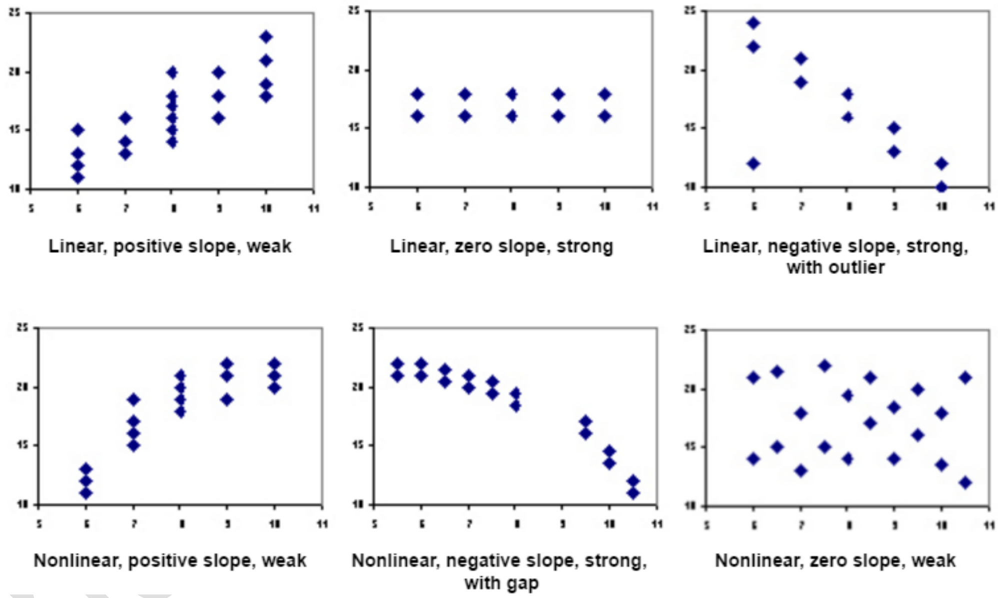






Strength

- They are the correlation term we use, like weak, strong, moderate.



Linear, positive slope, weak

Linear, zero slope, strong

Linear, negative slope, strong, with outlier

Nonlinear, positive slope, weak

Nonlinear, negative slope, strong, with gap

Nonlinear, zero slope, weak

Lurking Variable

- If non-linear trends are visible in the relationship between an explanatory and dependent variable, there may be other influential variables to consider. A lurking variable exists when the relationship between two variables is significantly affected by the presence of a third variable which has not been included in the modeling effort. Since such a variable might be a factor of time (for example, the effect of political or economic cycles), a time series plot of the data is often a useful tool in identifying the presence of lurking variables.

Data not to be Extrapolated

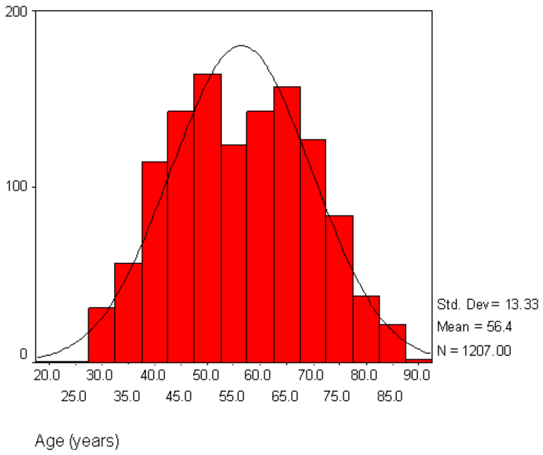
- **Do Not Extrapolate Regression Beyond Existing Data**
- The major purpose of linear regression is to create a Regression Equation that accurately predicts a Y value based on a new set of independent, explanatory X values. The new set

of X values should not contain any X values that are outside of the range of the X values used to create the original regression equation. The following simple example illustrates why a Regression Equation should not be extrapolated beyond the original X values.

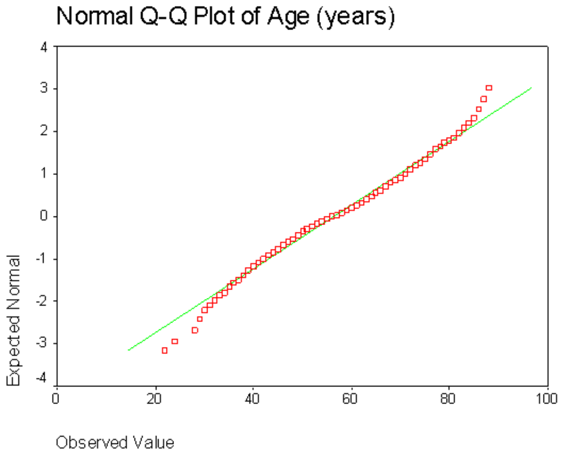
Error in Linearity will lead to model misspecification

Step 2 - Data is Normally distributed

- **Normality:** in the population, the data on the dependent variable are normally distributed for each of the possible combinations of the level of the X variables; each of the variables is normally distributed
  - This can be tested by: Eyeballing data in histogram

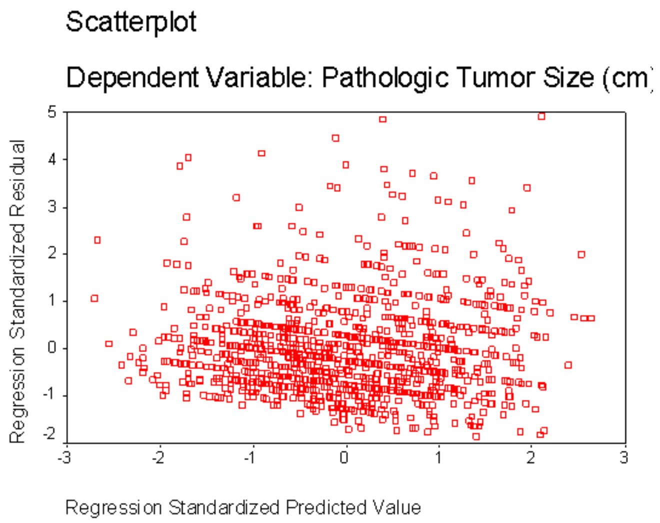


  - Construct a normal probability plot, Q plot. In this plot, the actual scores are ranked and sorted, and an expected normal value is computed and compared with an actual normal value for each case. The actual values lining up along the diagonal that goes from lower left to upper right. This plot also shows that age is normally distributed:



  - By looking at a plot of the "residuals." Residuals are the difference between obtained and predicted DV scores. If the data are normally distributed, then residuals should be normally distributed around each predicted DV score that is the majority of residuals is at the center of the plot for each value of the predicted score, with some residuals trailing off symmetrically from the center. Residual plot is recommended before graphing each variable separately because if this residuals plot looks good, then separate plots is not needed. Below is a residual plot of a regression where age of patient and time (in months since diagnosis) are used to predict breast tumor size. These data are not perfectly normally distributed in that the residuals about the zero line appear slightly more spread out than those below the zero line. Nevertheless, they do appear to be fairly normally distributed.

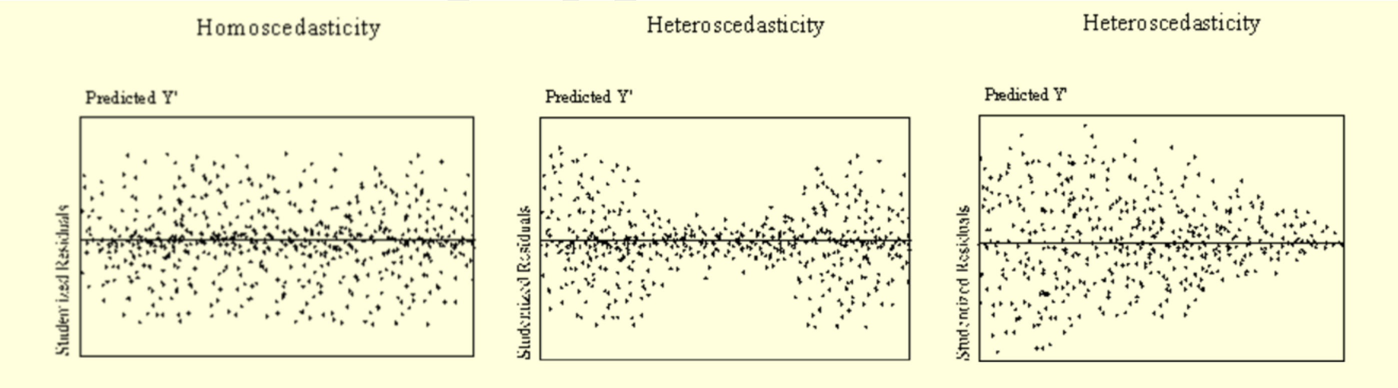
- Normality test can be done using Shapiro-Wilk Original Test for dataset 5000
- But since we may have more data than 5000, we would use Extended Shapiro-Wilks test
  - Rearrange the data in ascending order so that x1 ≤ … ≤ xn.
  - Define the values m_1, …, m_n by m_i = NORMSINV((i − .375)/(n + .25))
  - Let M = [mi] be the n × 1 column vector whose elements are these mi and let
    $$m = M \cdot M = M^T M = \sum_{i=1}^{n} m_i^2$$
  - If M is represented by the n × 1 range R1 in Excel, then =SUMSQ(R1) calculates the value m.
  - Set u = 1/$\sqrt{n}$ and define the coefficients *a1*, …, *an* where
  - $a_n = -2.706056u^5 + 4.434685u^4 - 2.071190u^3 - 0.147981u^2 + 0.221157u + m_n m^{-0.}$
  - $a_{n-1} = -3.582633u^5 + 5.682633u^4 - 1.752461u^3 - 0.293762u^2 + 0.042981u + m_{n-1}$
  - *ai* = *mi* /$\sqrt{\epsilon}$ for 2 < *i* < *n* − 1
  - *a*2 = −*an*-1 *a*1 = −*an*
  - where
    $$\epsilon = \frac{m - 2m_n^2 - 2m_{n-1}^2}{1 - 2a_n^2 - 2a_{n-1}^2}$$
  - It turns out that *ai* = −*an*-*i*+1 for all *i* and that
    $$1 = A \cdot A = A^T A = \sum_{i=1}^{n} a_i^2$$
  - where *A* = [*ai*] is the *n* × 1 column vector whose elements are the *ai*.
  - 5. The *W* statistic is now defined by
    $$W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
  - Because of the above properties of the coefficients *a1*, …, *an* it turns out that *W* = the square of the correlation coefficient between *a1*, …, *an* and *x1*, …, *xn*. Thus the values of *W* are always between 0 and 1.
  - It also turns out that for values of *n* between 12 and 5,000 the statistic ln (1−*W*) is approximately normally distributed with the following mean and standard deviation:
  - $\mu = 0.0038915(\ln n)^3 - 0.083751(\ln n)^2 - 0.31082 \ln n - 1.5861$
  - $\sigma = e^{0.0030302\ (\ln n)^2 - 0.082676 \ln n - 0.4803}$
  - 6. Thus we can test the statistic
    $$z = \frac{\ln(1 - W) - \mu}{\sigma}$$
  - using the standard normal distribution. If the p-value ≤ *α* then we reject the null hypothesis that the original data is normally distributed.
-



Error in Normality or Normality bias leads to incorrect probability coverage

Step 3 - Homoscedasticity is exhibited.

- Homoscedasticity means that the variance of errors is the same across all levels of the Independent Variables. When the variance of errors differs at different values of the IV, heteroscedasticity is indicated. For multiple regression, homoscedasticity should be there.
- Classic Definition of Homoscedasticity: In the population, the variances of the dependent variable for each of the possible combinations of the levels of the X variables are equal.



- Residuals should be randomly scattered around 0 (the horizontal line) providing a relatively even distribution. Heteroscedasticity is indicated when the residuals are not evenly scattered around the line such as a bow-tie or fan shape.
- Possible tests for this are the Goldfeld-Quandt test when the error term either decreases or increases consistently as the value of the dependent variables increases as shown in the fan shaped plot or the Glejser tests for heteroscedasticity when the error term has small variances at central observations and larger variance at the extremes of the observations as in the bowtie shaped plot (Berry & Feldman, 1985). In cases where skew is present in the Independent Variables, transformation of variables can reduce the heteroscedasticity. Or Breusch-Pagan test
- The formal methods that we consider are all based on statistical tests of the following general null and alternative hypotheses
  - H_0: *the error term is homoskedastic*
  - H_1: *the error term is heteroskedastic*

- Breusch-Pagan test:
  - Estimate the population regression model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$ and obtain the residuals, e_i.
  - Square the residuals or e_i^2.
  - Estimate the population regression model $e_i^2 = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \cdots + \gamma_k x_{ki} + \varphi$
  - Perform an *F*-test for overall significance to see if the squared residuals are statistically related to any of the independent variables.
- Park Test:
  - Run regression $Y_i = a + bX_i + e_i$ despite the heteroskedasticity problem (it can also be multivariate)
  - Obtain residuals (e_i), square them (e_i^2), and take their logs (ln e_i^2)
  - Run a spurious regression: $\ln e_i^2 = g_0 + g_1 \ln X_i + v_i$
  - Do a hypothesis test on $\hat{g}_1$ with H_0: $g_1 = 0$
  - Look at the results of the hypothesis test:
  - reject the null: you have heteroscedasticity, fail to reject the null: homoskedasticity, or $\ln e_i^2 = g_0$ which is a constant

Non-linear relationships in Linear Model

In some cases, the true relationship between the response and the predictors may be nonlinear.

A simple extension is to use polynomial regression, here we include terms like X^2
note: $Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times \underbrace{X_1^2}_{X_2}$.
This model is still linear in the base X1 and X2!
Idea: Try out a lot of different models, each containing a different subset of the predictors, and then use appropriate statistics to select the best model. These include the AIC, AICc, BIC, Adjusted-R-squared,
The subsets of p variables are $\mathbf{2^p}$ so we can't try them all. Instead we have some systematic ways that we choose some models for consideration using Forward/Backward/Mixed/Null Selection

- Forward Selection
- Backward Selection
- Mixed Selection

Error in Homoscedasticity leads to model misspecification and inflated standard errors

Step 4 – Data Check

Number of Cases

- When doing regression, the cases-to-Independent Variables (IVs) ratio should ideally be 20:1; that is 20 cases for every IV in the model. The lowest your ratio should be is 5:1 (i.e., 5 cases for every IV in the model).

Accuracy of Data

- Check maximum and minimum to check if data is in range

Missing Data

- If the missing data is not much, opt not to include those variables in analyses.
- If only a few cases have any missing values, then you might want to delete those cases.
- If removing leads to data being lost then replace the missing ones with average or trimmed average

Remove Extreme Outliers

- This is done to understand relationship border. The minimum border of the relationships will be the bivariate correlations of all possible predictor variables with the dependent measures. The maximum border will be a linear regression model with all possible predictor variables in the regression model.
- This can be done by getting Sorting the Data or can be found by dividing the variable y/mean
- As a "rule of thumb", an extreme value is considered to be an outlier if it is at least 1.5 interquartile ranges below the first quartile (Q1), or at least 1.5 interquartile ranges above the third quartile (Q3).
- Calculation of Interquartile Range
  - Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively. Or It is defined as the difference between the largest and smallest values in the middle 50% of a set of data.
  - Q1 is the "middle" value in the first half of the rank-ordered data set.
  - Q2 is the median value in the set.
  - Q3 is the "middle" value in the second half of the rank-ordered data set.
  - The interquartile range is equal to Q3 minus Q1.
  - The implications of removing/retaining the outlier must be clearly stated (it is unethical to simply erase a data point because it is not in the mainstream pattern!). Reasons and justification for any action must be clearly enunciated.
  - If outlier were to be removed, we would have a data set with a high level of association. As it is, the outlier has a significant effect on the level of association.

Identification and Impact of Influencer

- If a point lies far from the other data in the horizontal direction, it is known as an influential observation. The reason for this distinction is that these points have may have a significant impact on the slope of the regression line.
- The impact is to be studied and Influencer are not to be removed

Error due to outliers or outlier biased results in inflated standard errors

Step 5 – Create a Correlation Matrix

- The purpose of this step is to identify independent variables that are highly correlated which would cause an error called multicollinearity.
- Multicollinearity does not reduce the overall predictive power of the model but it can cause the coefficients of the independent variables in the regression equation to change erratically when small changes are introduced to the regression inputs.
- Multicollinearity can drastically reduce the validity of the individual predictors without affecting the overall reliability of the regression equation.
- When highly correlated pairs of independent variables are found, one of the variables of the pair should be removed from the regression. The variable that should be removed is the one with the lowest correlation with the dependent variable, Y.

- However, we cannot conclude that any of these correlations are important until we test for significance. calculating the test statistic for each of the pair-wise correlations above, we see that there are many statistically significant correlations (indicated in green), suggesting that multicollinearity may be a problem.
- In the case of multicollinearity, we could either:
  - Increase the sample size (which will often reduce the correlation among the independent variables,
  - Re-specify the regression model, removing or restating the independent variables such that there is less correlation among them.
- If no VIF > 5, then perform best subsets regression with all variables; List all models with C_p close to or less than (k + 1); Choose the best model; Consider parsimony (Do extra variables make sense and make a significant contribution?); Perform complete analysis with chosen model, including residual analysis; check for linearity and violations of other assumptions
- If one or more VIF > 5, remove them from the model; Re-estimate the new model with the remaining variables, and repeat this step.
- Try all combinations and select the best using
- the highest adjusted r^2 and lowest standard error, OR
- The C_p Statistic $$C_p = \frac{(1-R_k^2)(n-T)}{1-R_T^2} - (n-2(k+1))$$
- Where k = number of independent variables included in a particular regression model
- T = total number of parameters to be estimated in the full regression model
- R_k^2 = coefficient of multiple determination for model with k independent variables
- R_T^2 = coefficient of multiple determination for full model with all “T” estimated parameters
- The best model are those with C_p values that are small and close to K+1.

Step 6 – Create the Regression Equation from the data

Step 7 – Calculate and Examine Appropriate Measure of Association & Tests of Statistical significant for each co-efficient and for the Equation as a whole

Step 8 – Detection of Error

Graphical

The first approach must be graphical, glaring problems can easily be detected this way – This involves basic scatterplots, Density/QQ plots of residuals, residuals vs. fitted, influence plots, component-residual plots etc.

Detection of Error Statistical

- Any normality test on the residuals (e.g. Shapiro-Wilks) –
- Breusch-Pagan test for heteroscedasticity –
- Durbin-Watson for autocorrelation1 –
- RESET test for linearity –
- Many measures of outliers (Cook’s distance, dfBetas, Mahalanobis’ distance etc.) –
- Variance Inflation Factor for collinearity

Step 9 – Accept or Reject Null Hypothesis

Hypotheses can be one-tailed or two-tailed, e.g. H0: β1 = 0 or H0: β1 = 0 HA: β1 ≠ 0 HA: β1 > 0 The first is an example of a two-tailed alternative.

- Sufficiently large positive or negative values of β1 will lead to rejection of the null hypothesis. The second is an example of a 1-tailed alternative. In this case, we will only reject the null hypothesis if β1 is sufficiently large and positive.
- If β1 is negative, we automatically know that the null hypothesis should not be rejected, and there is no need to even bother computing the values for the test statistics.

- You only reject the null hypothesis if the alternative is better.

Step 10 – Accept or Reject the Research Hypothesis

Step 11 – Explain the practical implication of the findings

Foot Note:

- **Linearity**: In the population, the relation between the dependent variable and the independent variable is linear when all the other independent variables are held constant.
- **Independence**: The data of any particular subject are independent of the data of all other subjects
- Random Variable: A random variable, usually written X, is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables, discrete and continuous.
- ANCOVA: Analysis of covariance. The method of ANCOVA allows the analyst to make comparisons between groups that are not comparable with respect to some important variable, often referred to as a covariate. This is done by making an adjustment based on fitting a particular kind of regression line. In addition to allowing for imbalances, the method removes variation due to the covariate and therefore provides a more precise analysis. A geometrical interpretation is that the 'unexplained variation' with respect to which the significances of group differences are ultimately assessed
- The null hypothesis here is that there is not a general relationship between the response (dependent) variable and one or more of the predictor (independent) variables, and the alternative hypothesis is that there is one. A big F, with a small p-value, means that the null hypothesis is discredited, and we would assert that there is a general relationship between the response and predictors (while a small F, with a big p-value indicates that there is no relationship).
- The null hypothesis here is that there is not a general relationship between the response (dependent) variable and one or more of the predictor (independent) variables, and the alternative hypothesis is that there is one. A big F, with a small p-value, means that the null hypothesis is discredited, and we would assert that there is a **general relationship between the response and predictors (while a small F, with a big p-value indicates that there is no relationship**).
-

https://www3.nd.edu/~rwilliam/stats1/x93.pdf

Multiple Classification Analysis

- **Multiple classification analysis:** Multiple Classification Analysis (MCA) is a technique for examining the interrelationship between several predictor variables and one dependent variable in the context of an additive model Independent variables may be measured on nominal or ordinal scales and the dependent variable may be interval scale or a dichotomy.

- **Additive model:** Such a model assumes that the dependent variable can be predicted from an additive combination of the independent (or predictor) variables. In other words, they assume that the average score on the dependent variable for a given set of individuals (objects or cases) is predictable by adding the effects of several predictors.

- **Eta:** Eta indicates the ability of a predictor, using the given categories, to explain variation in the dependent variable.

- **Eta square:** Eta^2 is the correlation ratio and indicates the proportion of the total sum of squares, explained by the predictor.

- **MCA Beta:** This is directly analogous to Eta statistic, but is based on the adjusted means rather than the raw means. Beta is a measure of the ability of a predictor to explain variation in the dependent variable, after adjusting for the effects of all other predictors. Note that this is not in terms of percentage of variance explained.

- **Multiple correlation coefficient squared**: This coefficient indicates the proportion of variance explained in this run of the program.

- **Adjustment for degrees of freedom:** This is the factor used to correct for capitalizing on chance in fitting the model in the particular sample being analyzed.

- **Multiple correlation coefficient squared (Adjusted):** This coefficient estimates the proportion of variance in the dependent variable, explained by the predictor variables.

https://www.riskprep.com/all-tutorials/36-exam-22/131-regression-analysis

https://www.cgc.maricopa.edu/Academics/LearningCenter/Math/Documents/AnalyzingLinearRegression.pdf

http://chemistry.oregonstate.edu/courses/ch361-464/ch464/RegrssnFnl.pdf

https://www.riskprep.com/all-tutorials/36-exam-22/131-regression-analysis

http://go.owu.edu/~deswartz/210/text_notes/ch09.htm