

Analysing Relations between dataset-IV

Author: Naseha Sameen

2012

About: Linear Regression Model Building

Creating Regression Model

LAB NOTEBOOK
NASEHA SAMEEN

Contents

Linear Model Building:	2
Assumptions	2
Linearity	2
Homoscedasticity	4
Weak Exogeneity	7
Independence of errors (AKA No Auto correlation)	8
No Multi Collinearity	8
What Regression Does	10
Equation	10
Goal of Regression is	11
Method of Ordinary Least Squares (OLS)	11
Check for Potential Problems with linear regression which needs to be avoided	12
Validate Your Model	13
Linear Relationship	13
Fitting the Model	13
R^2	14
References:	16

Creating Regression Model

Linear Model Building:

Assumptions

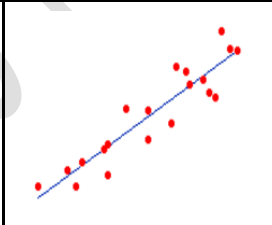
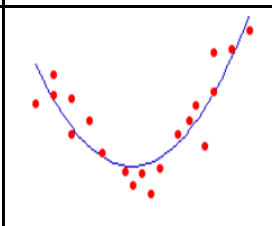
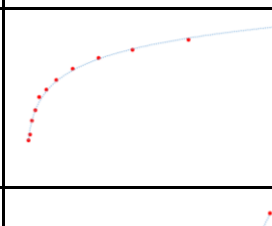
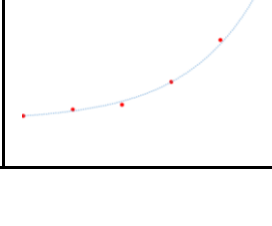
All of linear Regression should be followed. A quick review of the assumptions

Linearity

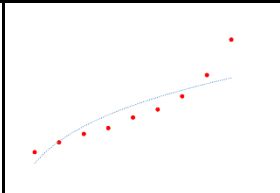
We have already gone through the requirement for an equation to be [linear](#):

You can skip these if you remember the Assumptions

- 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
- A quick guide to determine if the equation is Linear

Model	Dep.	Ind.	Graph	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		% Change in Y from an absolute change in X

Creating Regression Model

Power	Y	Log X		% Change in Y from an absolute change in X where power is >1
-------	---	-------	--	--

Non-linear relationships in Linear Model

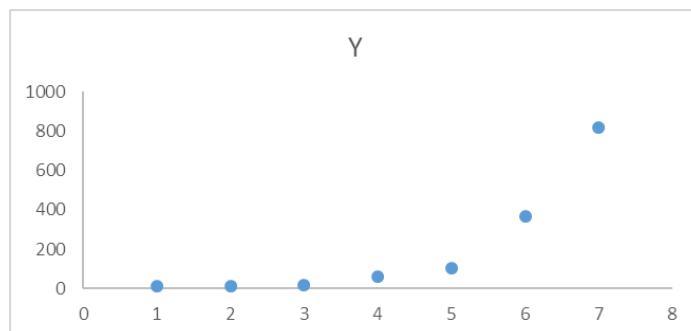
In some cases, the true relationship between the response and the predictors may be nonlinear. In such cases we transform the data

How to Transform Non-linear Variables to Make It Linear

- For the variables, complete regression analysis
- Check if the residual plot is not random, then transform the data. Random residual means, linearity, so no transformation is required
- From the residual plot see what kind of regression it is
- Select the formula for transformation
- Transform the variables
- Complete regression analysis with new variables
- Check the R^2 with the previous regression model
- If new R^2 is better than old, the transformation is good.
- Else check the model and continue with best fit

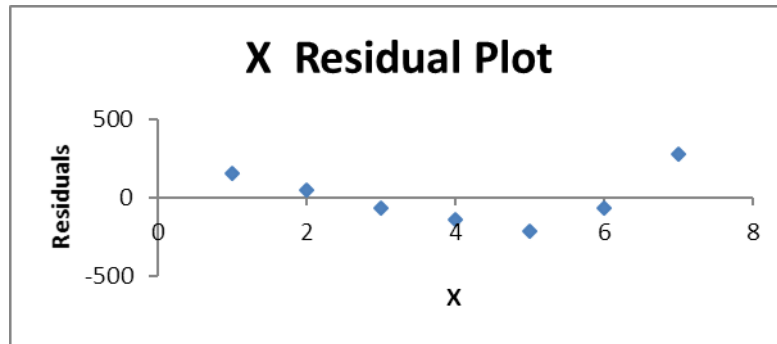
Model	Regression Eq	Transformation Formula	Predicted value (\hat{y})
Standard linear regression	$y' = y$	NA	$\hat{y} = b_0 + b_1x$
Exponential model	$y' = \log(y)$	Dependent variable = $\log(y)$	$\hat{y} = 10^{b_0 + b_1x}$
Quadratic model	$y' = \text{sqrt}(y)$	Dependent variable = $\text{sqrt}(y)$	$\hat{y} = (b_0 + b_1x)^2$
Reciprocal model	$y' = 1/y$	Dependent variable = $1/y$	$\hat{y} = 1 / (b_0 + b_1x)$
Logarithmic model	$y' = y = b_0 + b_1\log(x)$	Independent variable = $\log(x)$	$\hat{y} = b_0 + b_1\log(x)$
Power model	$y' = \log(y) = b_0 + b_1\log(x)$	Dependent variable = $\log(y)$ Independent variable = $\log(x)$	$\hat{y} = 10^{b_0 + b_1\log(x)}$

An example: For this data set, the equation is Exponential.

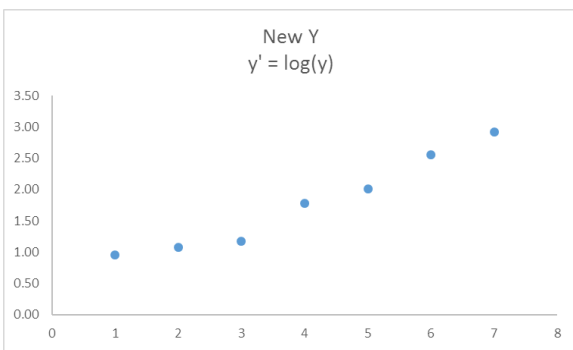


X	Y
1	9
2	12
3	15
4	60
5	102
6	364
7	820

Creating Regression Model

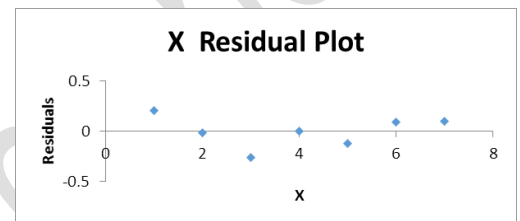


The regression analysis shows $R^2 = 0.67$
 The formulae to transform y is $y' = \log(y)$



X	New Y $y' = \log(y)$
1	0.95
2	1.08
3	1.18
4	1.78
5	2.01
6	2.56
7	2.91

The regression analysis shows $R^2 = 0.95$

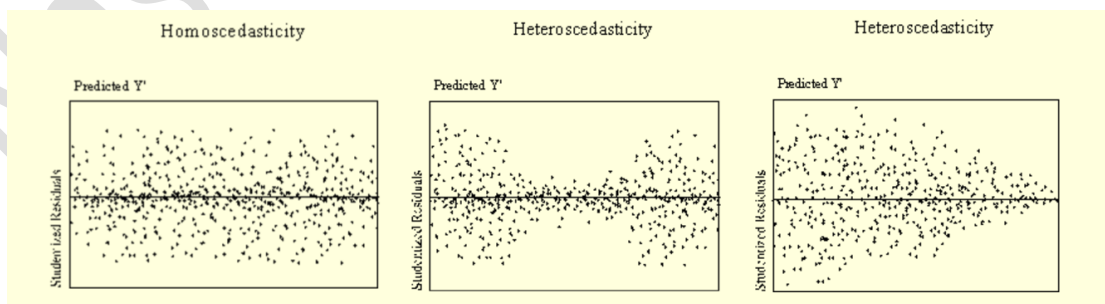


Homoscedasticity

- Variables should be Homoscedasticity which 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.
- 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. How

How to Identify Homoscedasticity

Residuals should be randomly scattered around 0 (the horizontal line) providing a relatively even distribution. Opposite is true for Heteroscedasticity where residuals are NOT evenly scattered around the line such as a bow-tie or fan shape.



Hypothesis used:

H_0 : the error term is homoscedastic

H_1 : the error term is heteroskedastic

Creating Regression Model

Breusch-Pagan-Godfrey

- This test measures increase of errors across y , the explanatory variable. The assumption that this test takes in is: linear function of 1 or more explanatory variable causes the error in variance. ([Example](#))
- Complete the Regression for $Y=mx+b+e$
- Plot all Residual e_i^2 .
- Find the variance of the regression residuals by squaring all residual, e_i
- Summing them up and dividing it by number of observations
- Now standardize the residual, by dividing each residue by the variance of the regression residue we obtained and then squaring it
- s_i^2 = Square of Standardized Residuals.
- Now run another regression with this s_i^2
- Check SS of this new Regression
- Divide the value by 2 for each tail of the hypothesis
- This number is our χ^2
- List degree of freedom, d
- Now check critical values of chi-square distribution with d degrees of freedom, for significance level
- Reject the Hypothesis H_0 : **The variances are equal (i.e., homoscedastic) if $X^2 > X_{Cr}$**
- Goldfeld-Quandt
 - Sort the independent variable in ascending order
 - Remove the mid-range, dividing the set into two set A & B (if the data range is small – do not remove any data)
 - Run regression on A & B
 - Note ESS – Or Residue SS (Error Sum of Square) & Degree of Freedom (df) of the residues
 - Residue SS of B/Residue SS of A
 - Go to [F Table](#) of required significant level (95% usually) with the required degree of freedom (df of residue)
 - Reject the Hypothesis H_0 : **The variances are equal (i.e., homoscedastic) if $F_{value} > F_{Cr}$**

White's test

- Complete the Regression for $Y=mx+b+e$
- Plot all Residual e_i^2 .
- Find the variance of the regression residuals by squaring all residual, e_i
- Summing them up and dividing it by number of observations
- Now standardize the residual, by dividing each residue by the variance of the regression residue we obtained and then squaring it
- s_i^2 = Square of Standardized Residuals.
- Run another regression with this s_i^2 – **lets name the R^2 obtained as $R_{\hat{u}2}^2$**
- n = sample size, df = degree of freedom
- $\chi^2 = nR_{\hat{u}2}^2$
- Now check critical values of chi-square distribution with d degrees of freedom, for significance level
- Reject the Hypothesis H_0 : **The variances are equal (i.e., homoscedastic) if $X^2 > X_{Cr}$**
- Or Sum s_i^2 = Square of Standardized Residuals.
- Create a Numerator in the table = Residue² * s_i^2
- Calculate Sum of this number
- White $Var(b) = \text{Numberator} / (\text{Sum } s_i^2)^2$
- Calculate square root of $Var(b)$
- Check standard Error of the variable from Regression table

Creating Regression Model

- Reject the Hypothesis H_0 : **The variances are equal (i.e., homoscedastic) if $\sqrt{\text{var}(b)}$ \leftrightarrow standard Error of the variable**

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 Regression with Natural Log of Residue as Y, and log of X as X
- Check t ratio,
- It is compared to a T distribution with $(n-k)$ degrees of freedom where here $n =$ observations and $k =$ variables.
- New Equation is $\ln(\text{new } y^2) = \text{New } m^*(\ln \text{ Original } X) + \text{New } b$
- Reject the Hypothesis H_0 : **The variances are equal (i.e., homoscedastic) if t ratio is significant, i.e. t ratio $>$ t crit**

Rectify the Heteroscedasticity

WLS Method – Weighted Least Square

Method 1 – When Variance is Known

- Divide each observation by its std. dev.
- Perform Regression again to get a new model.

Method 2 – When Variance is Not known

- Create a new Dependent variables – by dividing original dependent variable/Original independent variable
- Get the new Independent variable by inverting the independent variable.
- Inverse is nothing but inverse or reciprocal of a number which is obtained by $1/\text{number}$ or number^{-1} . Inverse of fraction x/y is y/x
- Run Regression on new X and Y
- Compare the previous model's Y's t ratio with New Y's t ratio (Follow the label, position of the variables would interchange in both the models)
- Stronger t ratio and lesser Standard Error indicates successful correction

Logarithmic Transformation

- Such transformation is used to correct Seasonal change in trends.
- Quick Fix to this is
- Log transformations is applied and then additive seasonal adjustment is performed to remove the seasonality effect
 - To get seasonal adjustment, find out the yearly average of the variable
 - Divide monthly data with this yearly variable to get the seasonal index
 - Add or subtract seasonal index to get seasonal adjusted data
- Use seasonally adjusted variable instead of the original

Box-Cox Transformation

- Box-Cox helps in transforming data to make a normal distribution, to remove the Left hand or right-hand skewness
- It is a process to identify right or optimum exponent λ to transform data.
- λ indicates the power to which all data should be raised.
- Box-Cox power transformation searches from $\lambda = -5$ to $+5$ until the best value is found.
- Formula: $\text{New } Y \text{ or } Y' = (Y^\lambda - 1) / \lambda$

Creating Regression Model

- Y is the response variable and λ is the transformation parameter.

λ	Y'
-2	$Y' = 1/Y^2$
-1	$Y' = 1/Y^1$
-0.5	$Y' = 1/(\text{Sqrt}(Y))$
0	$Y' = \log(Y)$
0.5	$Y' = \text{Sqrt}(Y)$
1	$Y' = Y$
2	Y'

- Since Box-Cox can be used only in positive number, we add $1 - \min$ to all the sample values to convert min to 1 and change the other numbers accordingly
- Sort the data. That is your x . In excel you can also use the formula Small
- Transform x to get $y \rightarrow \text{IF}(\lambda=0, \text{LN}(\text{new } x), (\text{New } x^\lambda - 1)/\lambda)$
- Get the probability Curve, each sequence of the number like 1, 2, 3, divided by max number.
- Now calculate the inverse normal values, which is nothing but standardized Z value corresponding to 1-tailed Probability. The value here is between 0 to 1
- The distribution has a mean of zero and a standard deviation of one.
- Get Correlation R^2 from value y and new inverse normal number z ,
- Now we Goal Seek – Set the Value of the Cell which has Correlation value
- Our Goal is Max R^2 , which is 1. Set Target as 1
- The cell that we will change is λ
- The λ is the exponential that we would require to transform the data
- Our new y is y^λ that we got

Weak Exogeneity

- If a variable is **not** affected by other variables in the model, it is exogenous variable. This variable could be affected by factors *outside* of the model. The regression model should show a weak exogeneity. These variables can be fixed variables, or given, or out of scope of the model, not determined or explained by the model or they can influence endogenous variables.
- Endogenous variables will cause OLS (ordinary least squares) estimators to fail. As one of the prerequisite of OLS is - no correlation between an predictor variable and the error term.

How to Identify Endogeneity

- The Hausman Test detects endogenous regressors. It is also known as a test for model misspecification, a model that does not account for every variable that matters and may have biased coefficients, error, parameters etc.
- Hausman test helps to identify between fixed effects model or a random effects model.
- The null hypothesis - Preferred model is random effects, there is no correlation between the unique errors and regressors.
- The alternate hypothesis - Model is fixed effects.

Least Square Estimate

- Use Regression analysis

Creating Regression Model

- Check for Significance F. If the value is less than 0.05, for 5%, the model is OK. If Significance F 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.
- Most or all P-values should be below 0.05.

Hausman test

- Take the variable you think is endogenous (this is essentially X) as Y and all other explanatory variables as X and run regression
- Take the residuals from the regression test above as an additional explanatory variable in the regression.
- For 2nd regression, use usual Y as Y and the variables including the tested ones and the residue as X.
- Check for t-scores for particular alpha levels.
- Compare the alpha level (5% usually) to the p-value in the output. If the **p-value in the output is smaller** than the alpha level you chose, **reject the null hypothesis**. p-values to determine which terms to keep in the regression model.
- Compare the t-critical value in the output with the t-value. If the **t-value is smaller than the t-critical** value, **reject the null hypothesis**.
- If unsure between one-tailed test or a two-tailed test, go for two-tail t critical value.

Independence of errors (AKA No Auto correlation)

- Prerequisite of a regression model is that error terms are independent. A common violation of this assumption occurs when each error term is related to its immediate predecessor (ϵ_i is related to ϵ_{i-1}).
- One of the Test is Breusch-Pagan test which we covered earlier
- Another Test is Durbin-Watson statistic

Durbin-Watson statistic

- This test measures the correlation between the error term and the immediate predecessor.
- The statistic D ranges from 0 to 4. If ρ is 0, the D value is 2, close the D value is to 2, more independent the errors terms are.
- If $\rho = 1$, Durbin-Watson statistic = 0
- Generally, values of d within $1.5 < d < 2.5$ show that there is no auto-correlation in the multiple linear regression data.
- To reject a hypothesis, d is calculated and checked with Durbin-Watson Table
- $d = \text{sum of squared difference} / \text{sum square}$
 - if $d > \text{upper bound}$, fail to reject the null hypothesis of no serial correlation,
 - if $d < \text{lower bound}$, reject the null hypothesis and conclude that positive autocorrelation is present,
 - if $\text{lower bound} < d < \text{upper bound}$, the test is inconclusive.

No Multi Collinearity

Multicollinearity can impact on the quality and stability of the model by leading to

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

Creating Regression Model

- 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.

Collinearity can be categorized into 3 kinds

- Perfect Collinearity
 - It is the case when the one predictor can be predicted from either 1 or in some combination of 1 or more predictors. These are generally error of commission. Like addition of average score in Science and total score in the model, when one is a subset of another. In such cases, most of the times regression will return an error
- Almost Perfect Collinearity
 - The correlation value r is very close to 1 though not 1. In such cases the standard error would be very high
- Multicollinearity
 - Multicollinearity is a state when independent variables are not independent and there exists a relationship between them. It can cause high standard errors for the LSQ Estimation, leading to t-tests suggesting that the parameters are not significantly different from zero.

Identifying Multicollinearity

Multicollinearity identified in 4 ways:

Quick and Easy Indications

- High R^2 but few significant t ratios.
- F-test rejects the null hypothesis, but none of the individual t-tests are rejected.
- Correlations between pairs of X variables are more than with Y variables.

Variance Inflation Factor

- Variance Inflation Factor (VIF) – VIF is inverse of Tolerance = $1/T$. Tolerance is $1 - R^2$
- $VIF = 1/1-R^2$
- $VIF > 10$ indicates presence of multicollinearity
- $VIF > 100$ indicates a certainty of multicollinearity

Tolerance

- Tolerance – is the measure of the influence of one independent variable on all other independent variables;
- Tolerance $T = 1 - R^2$ of original regression.
- $T < 0.1$ indicates multicollinearity
- $T < 0.01$ indicates a certainty of multicollinearity

Correlation matrix

- Correlation coefficient that varies from 0 to 1, both in negative and positive scale need to < 1 and closer to 0
- Closer the number is to 1, stronger is the relationship

Condition Index

- The condition index is calculated using a factor analysis on the independent variables. Values of 10-30 indicate a mediocre multicollinearity in the linear regression variables, values > 30 indicate strong multicollinearity.

Creating Regression Model

Removing Multicollinearity

- Easiest way to reduce multicollinearity is to 'centering the data' which is subtracting the mean from the independent variables.
- Increasing the sample size
- transform the highly correlated variables into a ratio
 - Usual method is using factor analysis or manual methods involve adding the related predictors together or averaging them or subtracting the predictors with some weights
- Removing the variable that does not look essential. Check the correlations between variables remove the variable that are highly correlated but after taking into account the importance of the variables.
 - Removal of variable might lead to The Omitted variables bias
 - If an explanatory variables that should be present in the regression is not taken into the model, then the estimated coefficients will be not accurate.
 - The omitted variable might try to enter in the only way it can – through its positive correlation with the explanatory variable
 - The consequence of omitted variables bias is that all those explanatory variables that could affect the dependent variable should be included.

What Regression Does

Regression takes a significant sample size from the population and estimates the relationship between the parameters. It

- Identify and explanatory variables
- Establish a relationship between dependent & explanatory variables

Equation

- $Y = m_1X_1 + m_2X_2 + \dots + M_nX_n + b + \text{error}$

Y = Dependent Response Variable

m = Slope of the population

x = Explanatory or Independent variable

b = Y Intercept of the Population

error = random error

We know that the random errors are nothing but the difference between the observed and predicted values.

b = mean value of the dependent variable Y, when the independent variable x = 0

m = the change in the value of the dependent variable, Y, for unit change in the independent variable x.

$Y = 690 \times \text{Value of Gold} + 1120 + \text{error}$

Here error = 0

It means if x = 0, then the value of Y = 1120

690 is the change in value of Y for every unit change of x

Same loop continues for the number of variables n

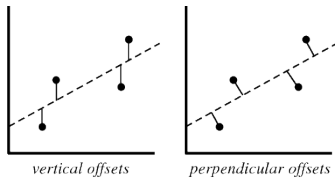
Goal of Regression is

- 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

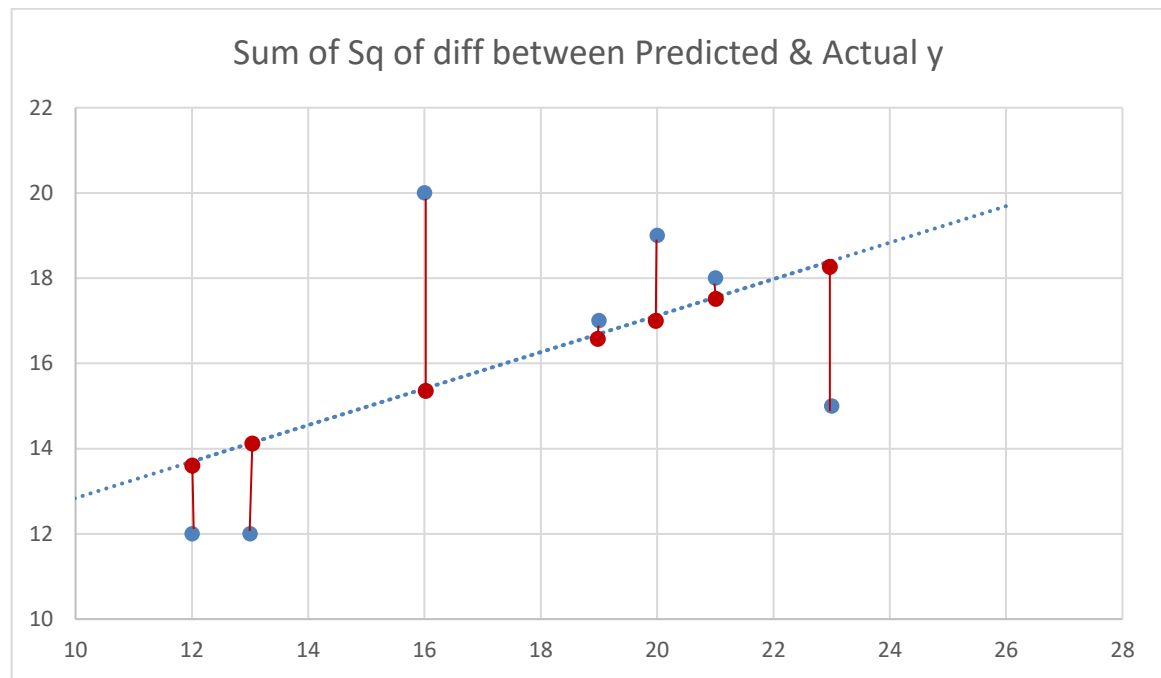
Method of Ordinary Least Squares (OLS)

Unless the R^2 is 100%, there will be some amount of variation in y which remains unexplained by x . The unexplained variation is the error component of the regression equation. That error is the sum of the differences between each observed value and its value as predicted by the regression equation.

Regression Analysis, attempts to fit a line through a set of plotted points that minimizes the residuals to give us a line that is called line of best fit. This line is fit in such a way that the sum of the squared residuals is minimized.



- There are two type of offsets to calculate the least distance.
- In practice, the vertical offsets from a line (polynomial, surface, hyperplane, etc.) are almost always minimized instead of the perpendicular offsets. This provides a fitting function for the independent variable x that estimates y for a given x (most often what an experimenter wants), allows uncertainties of the data points along the x - and y -axes to be incorporated simply, and also provides a much simpler analytic form for the fitting parameters than would be obtained using a fit based on perpendicular offsets.
- Ordinary least squares (OLS) is used to estimate the unknown parameters, with the goal of minimizing the sum of the squares of the differences actual and predicted values. Residual SS = sum of squares of the **differences between the values of y predicted by equation and the actual values of y** . In other words, it is derived from the cumulative addition of the square of each residual, where a residual is the distance of a data point above or below the fitted line



Advantage of Ordinary Least Squares (OLS)

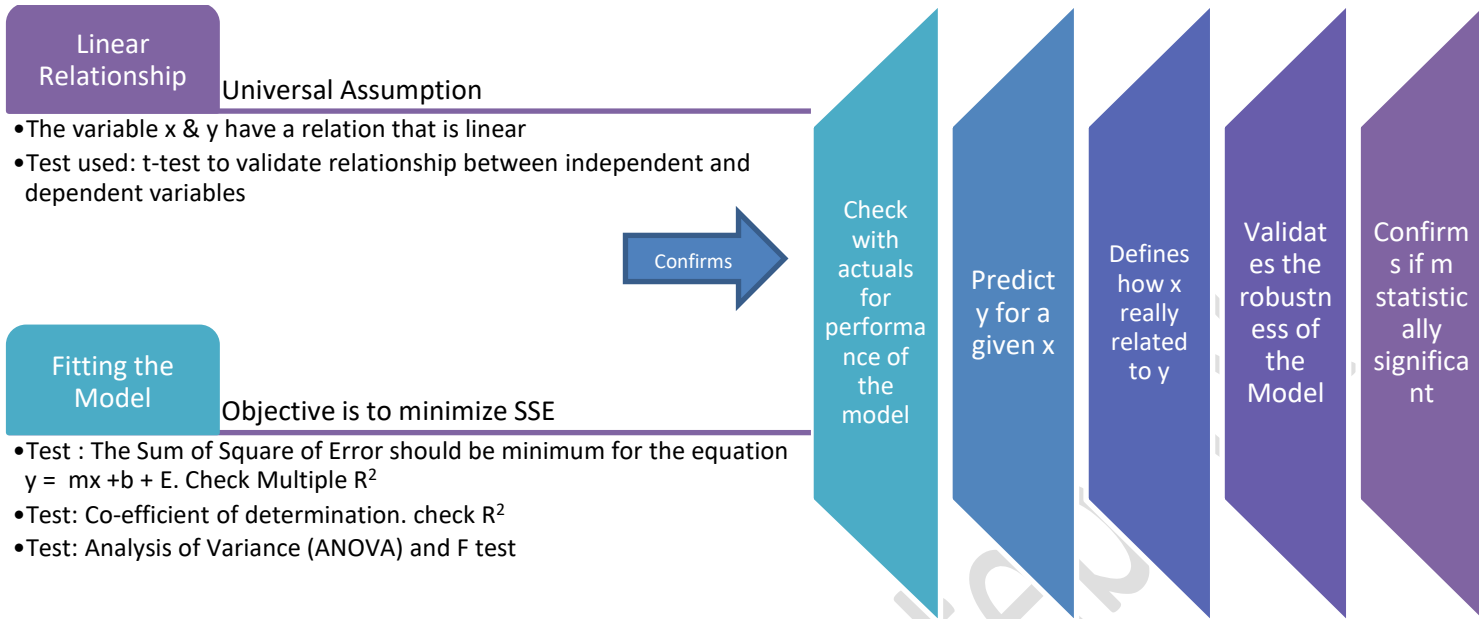
When error terms are not correlated as prerequisite of a linear OLS are BLUE, “Best Linear Unbiased Estimates. They have

- Estimation is Unbiased
- Estimation has minimum variance
- As sample size increases the sample intercept tends towards population intercept

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

Validate Your Model



Linear Relationship

We have already covered Assumptions and how to test them.

We have also ruled out autocorrelation in Durbin-Watson Test.

In statistics, the t-statistic is a ratio of the departure of an estimated parameter from its notional value and its standard error. The only difference is that we have to estimate the population standard deviation, σ . Remember, if you know σ , then use the z-test; if you don't know σ , then estimate σ (find s)

$$t = \frac{\bar{X} - \mu}{s_X}$$

- The t value is compared with Critical T for the degree of freedom to accept or reject a Hypothesis
- If $t_{calc} > t_{critical}$, we reject the null hypothesis and accept the alternate hypothesis. Otherwise, we accept the null hypothesis.
- H_0 = There is no relationship between dependent and independent variable.

Fitting the Model

R Values - They indicate the accuracy of the equation

<i>Regression Statistics</i>	
Multiple R	0.772426647
R Square	0.596642925
Adjusted R Square	0.589566485
Standard Error	1.2109982
Observations	59

Creating Regression Model

R^2

- It is also called the coefficient of multiple determination. It is the square of the co-efficient of correlation between y and x. R-squared value indicates the proportion of variation in the y-variable that is due to variation in the x-variables.
- It ranges in value from 0 to 1. Closer to 1, indicates the percentage of variation in y that is due to x.
- In the above example 59% of the variation in y is explained by x

Multiple R

- The correlation coefficient between the observed and predicted values.
- It ranges in value from 0 to 1. A small value indicates that there is little or no linear relationship between the dependent variable and the independent variables.
- "Multiple R" in the "Regression Statistics" is close to 1, then the least-square regression line indeed fits the data points pretty well, and there is a linear (positive or negative) relation between the two variables.
- In above example, the model is fitted to 77%

Adjusted R Square:

- Adjusted R square: The Adjusted R^2 is attempting to account for statistical shrinkage. The 'Adjusted R Square' is computed because the 'R Square' tends to somewhat over-estimate the success of the model when applied to the real world.
- The adjusted R^2 "penalizes" you for adding the extra predictor variables that don't improve the existing model. It can be helpful in model selection. Adjusted R^2 will equal R^2 for one predictor variable. As you add variables, it will be smaller than R^2 .
- R^2 increases even when you add variables which are not related to the dependent variable, but adjusted R^2 take care of that as it decreases whenever you add variables that are not related to the dependent variable.
- Adjusted R square gives idea of the goodness of fit measure, here it says 58% of Y is determined by X

Anova

Anova details & Significance F - They indicate the probability that the output obtained was deliberate and was not a chanced occurrence

ANOVA					
	Df	SS	MS	F	Significance F
Regression	1	123.647878	123.647878	84.31399593	7.79023E-13
Residual	57	83.59144844		1.466516639	
Total	58		207.2393264		

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower
Intercept	-2.22106586	0.542526978	4.093927034	0.000135244	-3.307457394	1.134674327	3.30
Income	0.000296533	3.22941E-05	9.182265294	7.79023E-13	0.000231865	0.000361201	0.00

- Significance of F indicates the percentage the output was a chance occurrence. For example = 0.010, there is only a 1% chance that the Regression output was a chance occurrence. In this example 0.0000000008% of the output is a chance occurrence.
- 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.

Creating Regression Model

- Most or all P-values should be below 0.05.
- P-Values of each is the possibility that the output did not occur by chance. The lower the P-Value, higher the possibility that Y-Intercept is valid.
- For example, a P-Value of 0.012 for a regression coefficient indicates that there is only a 1.2% chance that the result occurred only as a result of chance.
- In the above example - There is 0.000000000779% chance that the coefficient 0.000296533 (Income)

Residue and Errors

- Standard error is the estimate of the standard deviation of errors in the equation.
- Se, measures the scatter of the observed values around the regression line.
- Smaller the error, better the model fits. $Se = 0$, is a perfect fit
- Standardized Residual - It is the residual divided by the standard deviation of the residual; that is, it is a residual standardized to have standard deviation 1. If standardized residuals is larger than about ± 2.5 , should be investigated as a potential outlier. For very large samples, many observations could have standardized residuals outside ± 2.5 while not being outliers.
- Se for regression coefficient is the amount of error in a regression coefficient.

Creating Regression Model

References:

- <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/RegrsnFnI.pdf>
- <https://www.riskprep.com/all-tutorials/36-exam-22/131-regression-analysis>
- http://go.owu.edu/~deswartz/210/text_notes/ch09.htm
- <https://wantlearnmath.jimdo.com/statistics/logarithmic-regression/>
- https://en.wikipedia.org/wiki/Linear_regression
- www.statisticssolutions.com/autocorrelation/
- www.yourdictionary.com › Dictionary Definitions › regressand
- <https://www3.nd.edu/~busiforc/handouts/DataMining/dataminingdefinitions.html>
- <http://onlinelibrary.wiley.com/doi/10.1002/0470011815.b2a09052/abstract>
- <http://mathbits.com/MathBits/TISection/Statistics2/sinusoidal.html>
- <http://www.jkp-ads.com/articles/leastquares.asp>
- https://en.wikipedia.org/wiki/Gaussian_process
- <http://www.excelmasterseries.com/ClickBank/Thank You New Manual Order/ePUB Files/Advanced Regression/Text/Logistic Regression.html>
- <http://stattrek.com/regression/linear-transformation.aspx?Tutorial=AP>
- <http://www.statisticssolutions.com/homoscedasticity/>
- <https://datascienceplus.com/how-to-detect-heteroscedasticity-and-rectify-it/>
- <http://people.duke.edu/~rnau/testing.htm#homoscedasticity>
- <http://www.statisticshowto.com/breusch-pagan-godfrey-test/>
- <http://itfeature.com/heteroscedasticity/white-test-for-heteroskedasticity>
- <http://www.statisticshowto.com/park-test/>
- <https://analysights.wordpress.com/tag/breusch-pagan-test/>
- <https://analysights.wordpress.com/tag/heteroscedasticity/>
- <https://people.duke.edu/~rnau/411seas.htm#additive>
- <https://analysights.wordpress.com/2010/09/16/forecast-friday-topic-forecasting-with-seasonally-adjusted-data/>
- <http://www.itl.nist.gov/div898/handbook/eda/section3/eda336.htm>
- <https://www.spcforexcel.com/knowledge/basic-statistics/box-cox-transformation#box-cox-calculations>
- <https://www.isixsigma.com/tools-templates/normality/making-data-normal-using-box-cox-power-transformation/>
- <http://www.real-statistics.com/correlation/box-cox-transformation/box-cox-normal-transformation/>
- <http://www.real-statistics.com/correlation/box-cox-transformation/>
- <http://www.statisticshowto.com/hausman-test/>
- <https://masterofeconomics.org/2010/08/02/hausman-test-for-endogeneity-parents-education-as-iv-for-offspring-education-transmission-of-inate-ability/>
- file:///C:/Users/naseh/Desktop/using_excel_for_principles_of_econometrics3e.pdf
- file:///C:/Users/naseh/OneDrive/Studies/Business%20Analysis/Durbin_Watson_tables.pdf
- <http://www.real-statistics.com/multiple-regression/collinearity/>
- <http://www.statisticssolutions.com/assumptions-of-linear-regression/>
- <http://www.real-statistics.com/matrices-and-iterative-procedures/goal-seeking-and-solver/>
- file:///C:/Users/naseh/OneDrive/Studies/Business%20Analysis/3460_regression%20analysis%20II.pdf

Creating Regression Model

<http://psychologicalstatistics.blogspot.in/2013/11/multicollinearity-and-collinearity-in.html>

http://reliawiki.org/index.php/Simple_Linear_Regression_Analysis

http://excelmasterseries.com/ClickBank/Thank_You_New_Manual_Order/ePUB_Files/Advanced_Regression/Text/Regression_Output.html

Naseha Sameen