Analysing Relations between dataset-III

2015

Author: Naseha Sameen

About: Introduction to Regression what it is, how many types of regression are there and how and when to use what regression

Introduction to Regression

LAB NOTEBOOK

NASEHA SAMEEN

Contents

Nasena's Lab Notebook

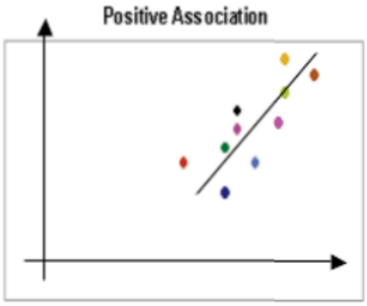
Regression

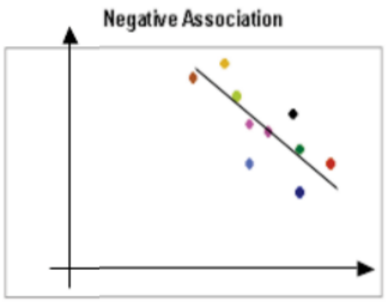
Definition :

- Regression tells us the exact kind of linear association that exists between those two variables. The term is used when one of the variables is a fixed variable, and the end goal is to use the measure of relation to predict values of the random variable based on values of the fixed variable
- Regression does not explain causation
- In particular, it indicates the extent to which you can predict some variables by knowing others, or the extent to which some are associated with others.

Direction

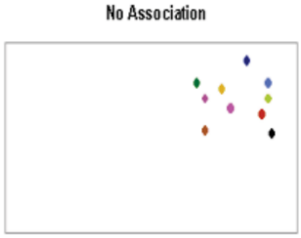
- Positive gradient: When the larger values of the horizontal (explanatory) variable are associated with larger values of the vertical (response) variable.
- Negative gradient: When the larger values of the explanatory variable are associated with smaller values of the response variable. As the explanatory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

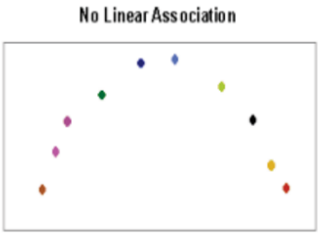


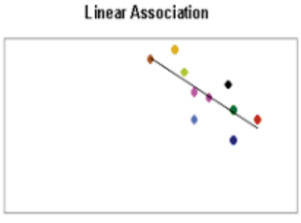


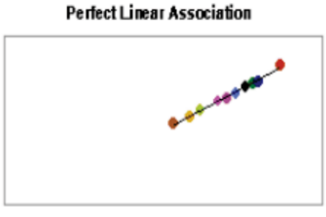
Form

- 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.



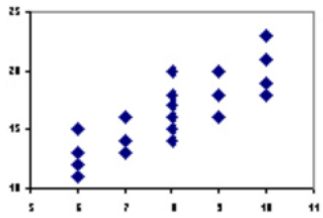




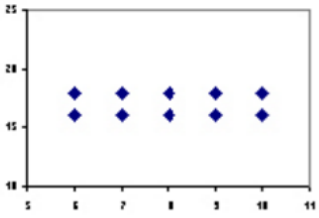


Strength

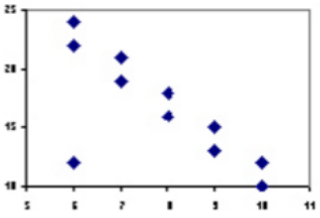
- They are the correlation term we use, like weak, strong, moderate.



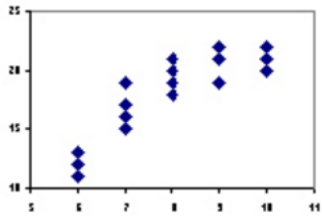
Linear, positive slope, weak



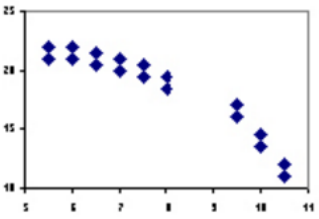
Linear, zero slope, strong



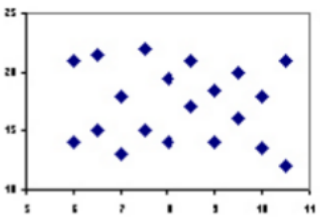
Linear, negative slope, strong, with outlier



Nonlinear, positive slope, weak



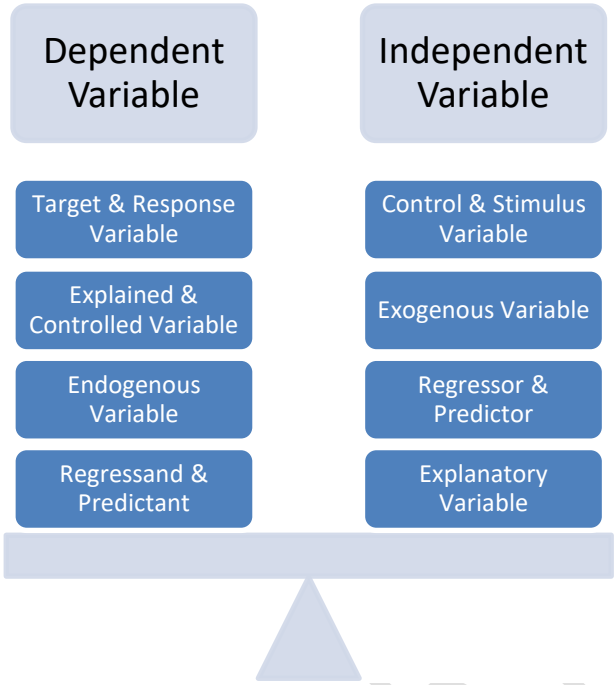
Nonlinear, negative slope, strong, with gap



Nonlinear, zero slope, weak

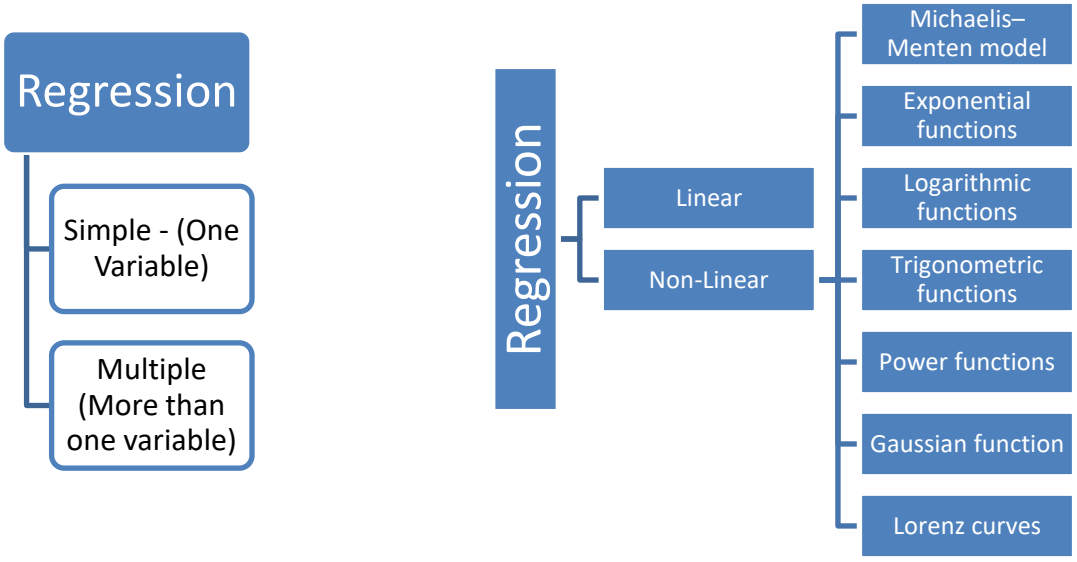
Nasehа's Lab Notes

Dependent & Independent Variable



- Dependent Variable – (Define)
  - Regressand/Predictand – the dependent variable in a regression that we are trying to find out
  - Endogenous Variable – Variable that is not attributable to any external or environmental factor.
  - Explained Variable – measures the proportion to which a mathematical model accounts for the variation (dispersion) of a given data set.
  - Controlled – A control variable is the variable that is not changed throughout an process, it's non changing attribute allows the relationship between the other variables to be explained.
  - Target/Response Variable – variable that is being predicted. It is a dependent variable, also called a output/response/outcome variable.
- Independent Variable
  - Explanatory Variable – It is a type of independent variable which might not be completely unaffected by other
  - Regressor/Predictor – In an equation y = mx + b, x is the independent variable or the regressor.
  - Exogenous Variable - Independent variable that affects a model without being affected by it
  - Control - A variable which is constant throughout process.
  - Stimulus Variable – is variable that evokes a response in the process

Classification of Regression



Linear & Non Linear Regression

The relationship can be linear or non-linear.

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
- 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

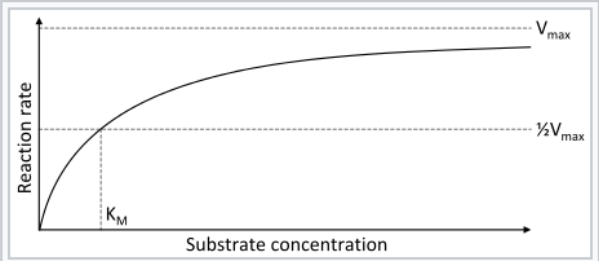
Nashi's Lab Notes

A look into Non-linear Regressions

Michaelis–Menten model

This model is helpful in explaining enzyme kinetics. The equation relates reaction rate to the substrate concentration (*S*)and describes the rate of enzymatic reactions

$$v = \frac{d[P]}{dt} = \frac{V_{\max}[S]}{K_{\mathrm{M}} + [S]}.$$



Michaelis–Menten saturation curve for an enzyme reaction showing the relation between the substrate concentration and reaction rate.

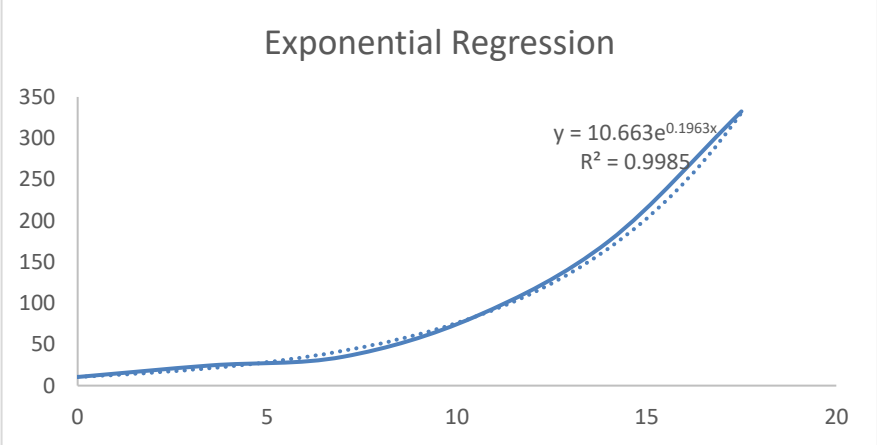
Exponential Regression:

This is the process to find the best suited exponential equation. Mostly seen where the growth is exponential like bacterial growth. The variable growth for power regression is multiplied by a fixed number < 1 in each equal unit of time.

Y = ab^x where a <> 0
Plot the point and opt for Exponential Equation.

$y = 10.663e^{0.1963x}$

Value of e is approximately 2.71828



X	Y
0	10.5
3.5	24.5
7	35
10.5	84
14	175
17.5	332.5

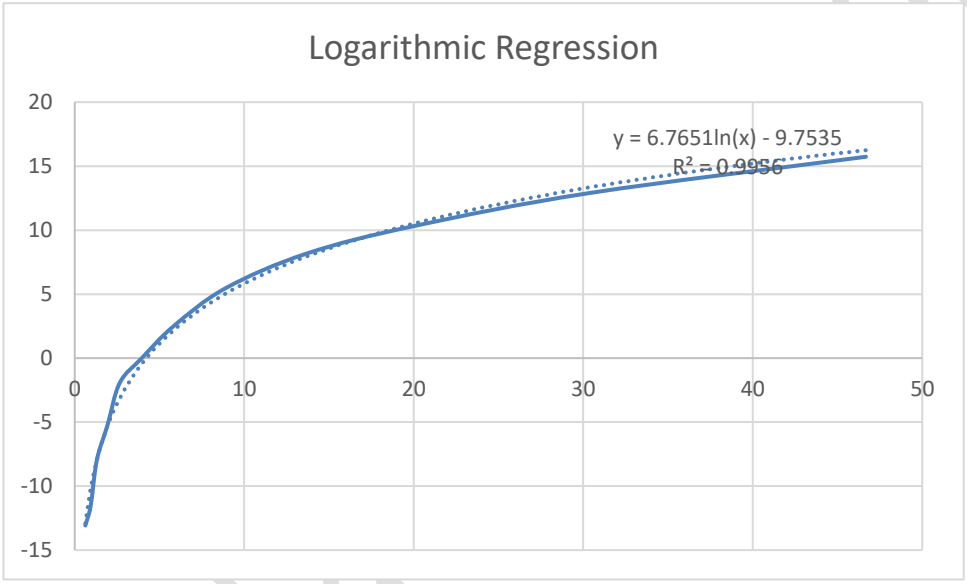
Logarithmic Regression:

This is the process to find the best suited logarithmic equation. Where changes in 1 unit or percentage of x leads to constant semilog value of y. Graph of y on log of x will show a straight line. Generally used in show growth over the years scenario, like growth of trees

Y = m ln(x) +b
Plot the point and opt for Exponential Equation.

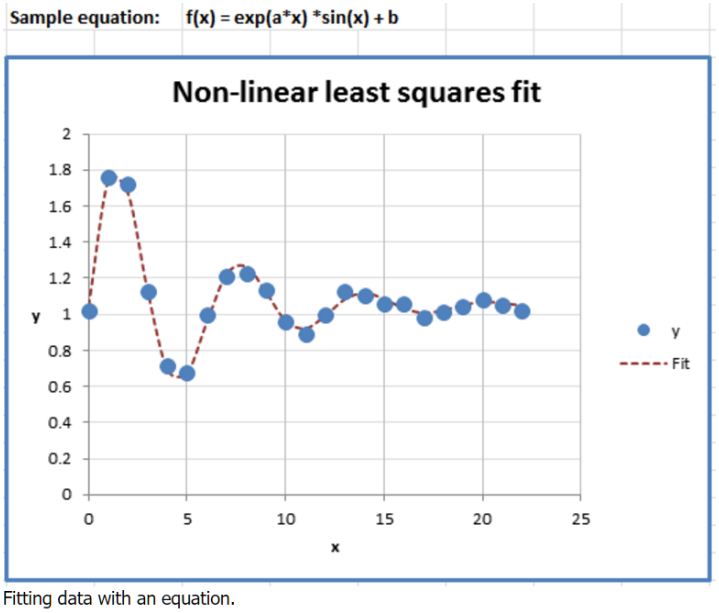
y = 6.7651 ln(x) - 9.7535

X	Y
0.6252	-13.0771
0.9378	-11.5662
1.3025	-7.9713
1.9277	-5.2621
2.6571	-1.9277
3.9596	0
5.8352	2.5008
8.9091	5.4705
13.7023	8.1797
20.0064	10.3158
30.4785	12.9208
46.6816	15.7342



Trigonometric functions

Fitting a periodic regression function to the data where usually time is involved. Such equations are used in annual cycle, the high tide cycle, change in temperature with variation of day length



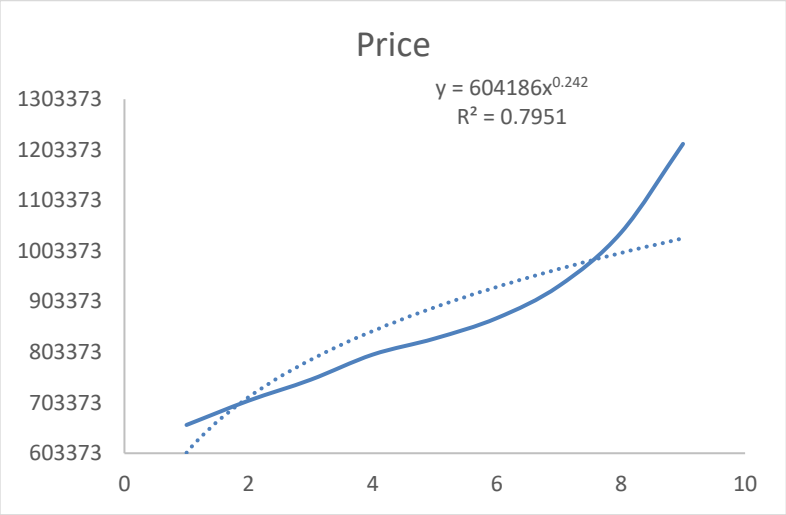
Fitting data with an equation.

$Y_i = g(t_i) + e_i$ where g(t) is sine or cosine wave with amplitude, angular frequency, and phase angle.

Power functions

Power regression is like the exponential function which is based on growth or decay which changes drastically with time. The variable growth for power regression is multiplied by a fixed number > 1 in each equal unit of time. The response variable is proportional to explanatory variable raised to a power.

$$Y = ax^{b}$$

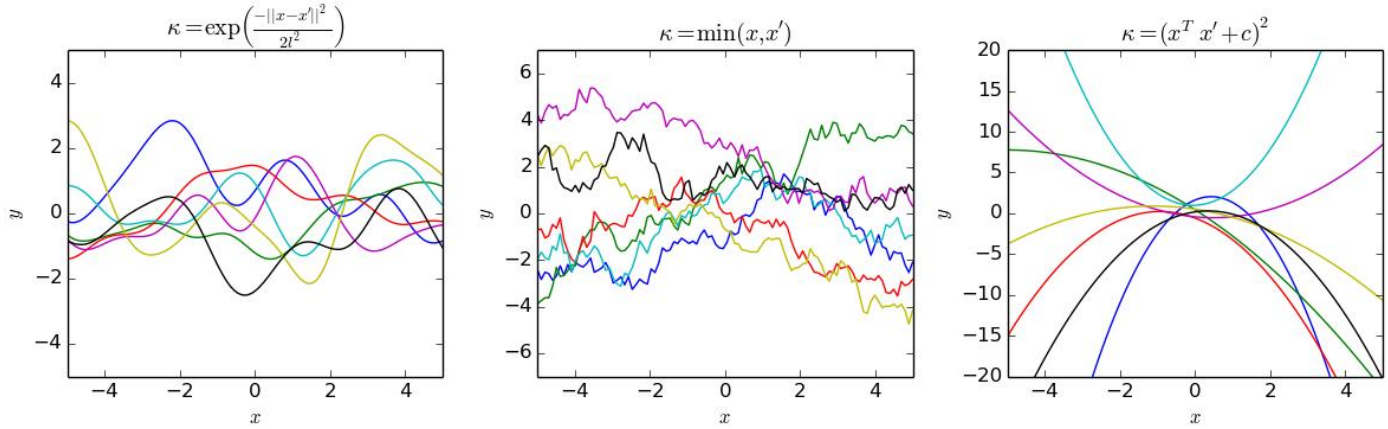


Year	Price
1	659303
2	707271
3	748759
4	798414
5	830192
6	870540
7	934046
8	1039456
9	1215012

Gaussian function

In this model in the continuous domain like time or space observations are recorded. Every observed point is related with a normally multivariate distributed random variable. It is a combined distribution of random variable which are infinite in number.

Advantage of this method is that the derived quantities are obtained explicitly. These are mostly average value of the process over a range of times which reduces the average estimation of error



Different kernels on the prior function distribution of the Gaussian process.
From left to right – 1. Squared exponential kernel, Brownian, Quadratic.

Sahas Lab Notebook

There are a number of common covariance functions:[9]

- Constant : $K_{\mathrm{C}}(x, x') = C$
- Linear: $K_{\mathrm{L}}(x, x') = x^T x'$
- Gaussian noise: $K_{\mathrm{GN}}(x, x') = \sigma^2 \delta_{x,x'}$
- Squared exponential: $K_{\mathrm{SE}}(x, x') = \exp\left(-\frac{\|d\|^2}{2\ell^2}\right)$
- Ornstein–Uhlenbeck: $K_{\mathrm{OU}}(x, x') = \exp\left(-\frac{|d|}{\ell}\right)$
- Matérn: $K_{\mathrm{Matern}}(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}|d|}{\ell}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}|d|}{\ell}\right)$
- Periodic: $K_{\mathrm{P}}(x, x') = \exp\left(-\frac{2\sin^2\left(\frac{d}{2}\right)}{\ell^2}\right)$
- Rational quadratic: $K_{\mathrm{RQ}}(x, x') = (1 + |d|^2)^{-\alpha}, \quad \alpha \geq 0$

Logistic Regression

In this regression type the dependent variable is categorical. If the dependent variable take binary state or 0 or 1, it is a simple logistic regression else it forms a multinomial logistic regression, ordinal logistic regression, if they are ordered. This regression is not just an analysis but is also a good predictor.

Data is collected, a predictive equation is created, the logit is calculated, prospect of each occurrence is calculated and variables are adjusted for maximize sum of P(X)Y * [1 - P(X)] (1-Y)

Nasena's Lab Notebook

Linear Regression

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

Equation

- Y = mx + b + error

Y = Dependent Response Variable
m = Slope of the population
x = Explanatory or Independent variable
b = Y Intercept of the Population
error = random error

Conditions

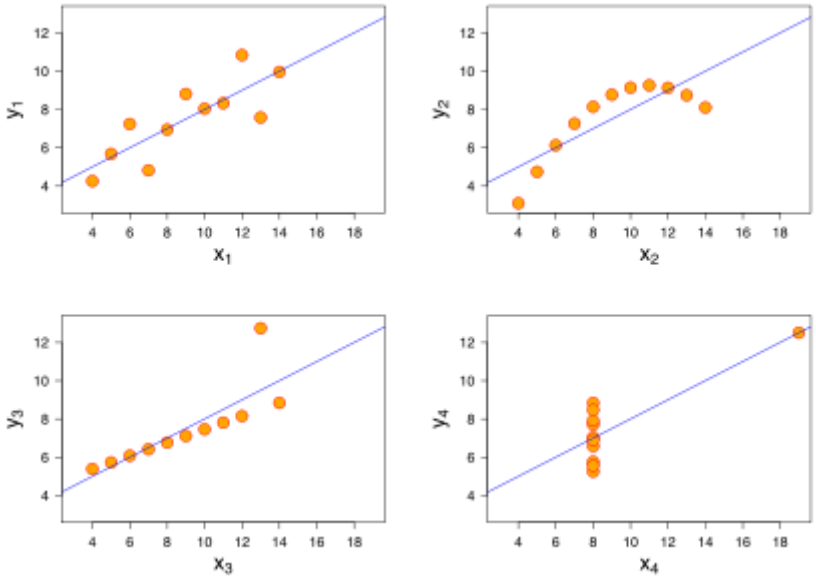
- **Linearity** – the function should be linear, even though some variable may or may not be non-linear in relationship with other variable.
- **Exogeneity –** Exogeneity indicates that the variable is set externally and changes are usually brought by external factors. For Linear regression the variables must display very weak exogeneity or a strong endogeneity.
- **Homoscedasticity -** The variance should be same and finite for all the random variables
- **Independence of errors** – Any errors creeping in the response variables are uncorrelated
- **No multicollinearity –**When 2 or more variable are highly correlated and inclusion of both would influence the model. We remove 1 or more highly correlated variables which influences each other till multicollinearity is removed
- **No autocorrelation –** It is also called Serial Correlation. It happens when the correlation between the values of the same variables is based on related objects.

Factors to Consider:

- Consistent and Unbiased:
  - For an estimation procedure to have a desirable sampling properties, the nature of statistical relationship between the regressors between and the error terms should be consistent and unbiased
- Precision of estimate
  - Care should be taken while collection of data, the arrangement, or probability distribution of the predictor variables to get a better precision in estimation of β.

Nasscom Lab Notebook

Interpretation

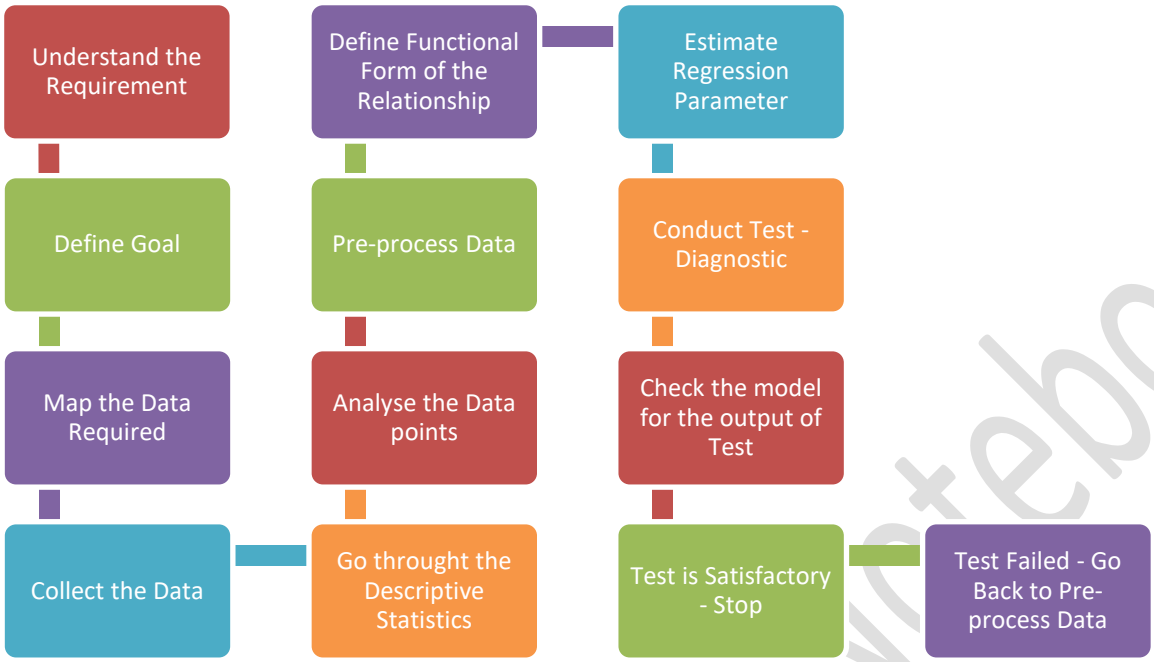


The data set is Anscombe's quartet. It has nearly identical simple descriptive statistics, however they create very different graphs. So interpretation is the key.

Be careful for some variables might not allow to be held fixed, or the partial effect might be high though total unique effect may be 0.

Nasehá's Lab Note

Regression Model Development



Nasena's Lab Notebook

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/RegrssnFnl.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/leastsquares.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

Naseha's iNotebook

Naseha Sameen | +91 95999 641 69| naseha@sevenslutions.in

sevenSolutions