L-1 SIMPLE LINEAR REGRESSION

FIRST THREE LECTURES

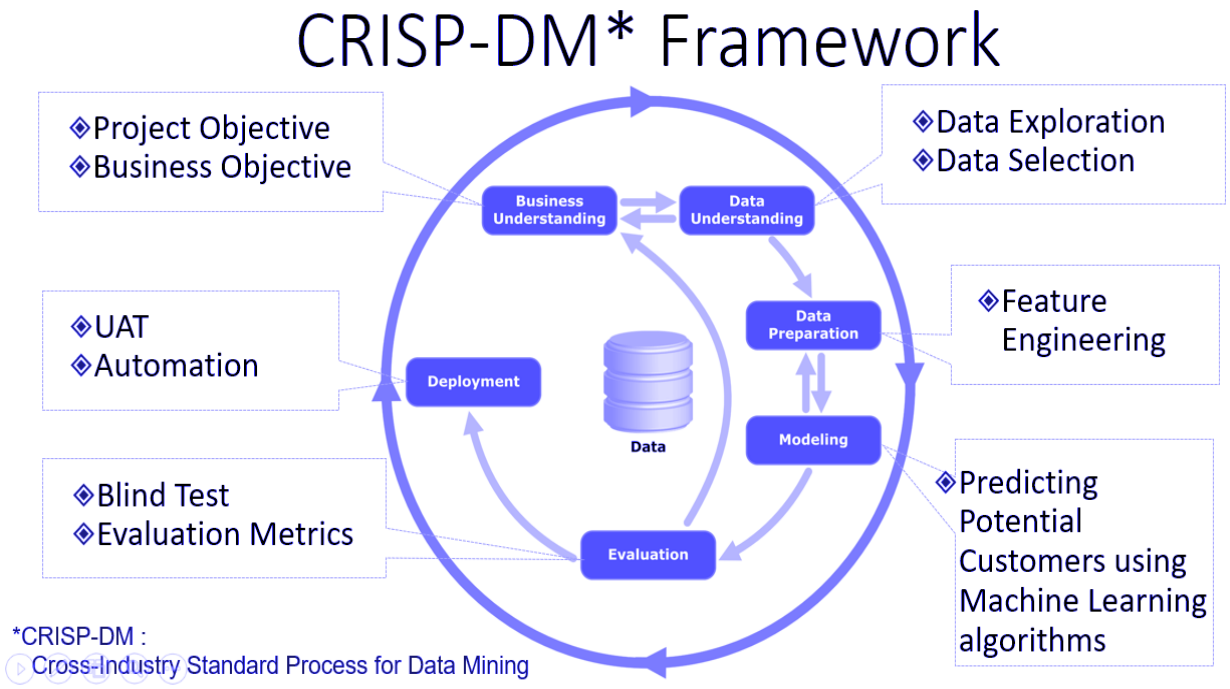
DATASET OF SOCCER PLAYERS. RESPONSE VARIABLE IS **'SCORE'**. THE DISCUSSION IS ABOUT OLS METHOD.

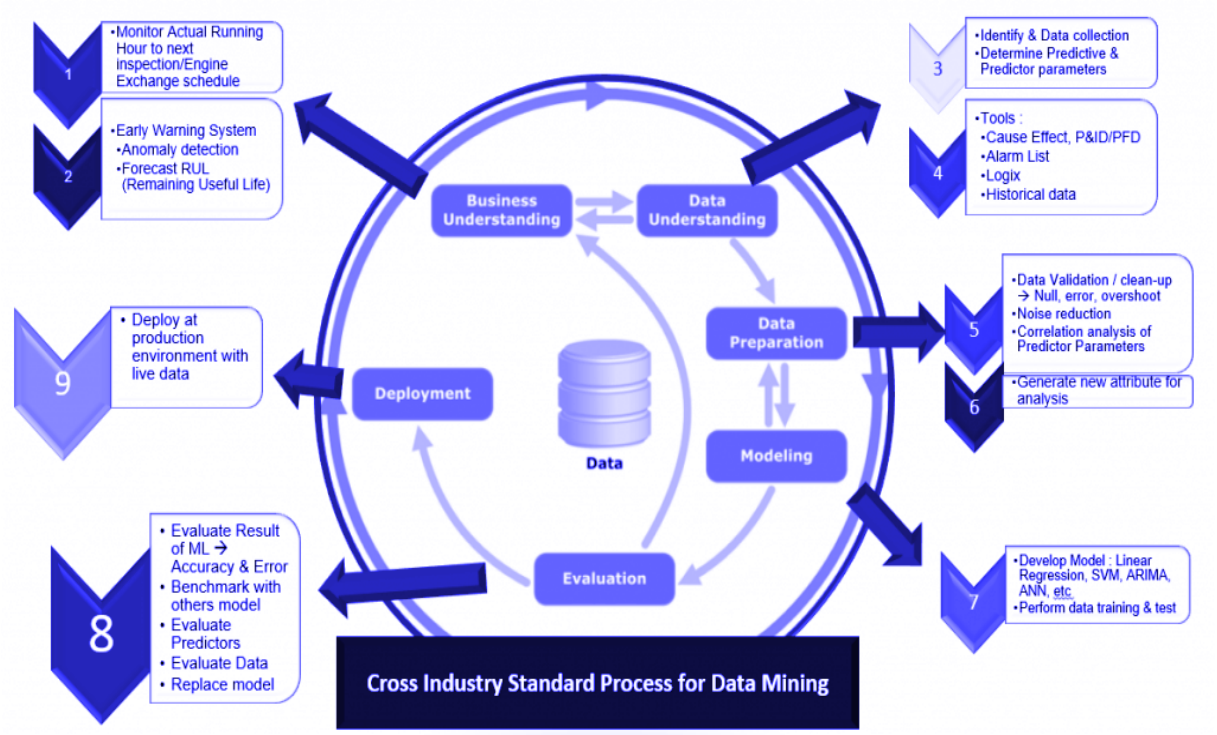
WHAT DATA SCIENTISTS DO

DATA-SCIENTISTS USE DATA TO **UNDERSTAND, PREDICT AND ACT.**

BASIC DATA ANALYTICS FRAMEWORK-CRISP DM

Cross Industry Standard Process For Data Mining





A COMMON EXAMPLE OF APPLICATION OF OLS

The example Of ice cream sales. **Let's assume that** the quantity of ice cream cones sold depends upon temperature. A graph like below was plotted. So we can make a hypothesis that there exists an almost linear relation between the number of ice creams sold (response/dependent variable) and temperature (

independent variable). The 'Line of fit' is shown by a dashed blue color line. This was a highly simplified example. The question is how to obtain the 'line of fit' if we have millions of observations and if the number of predictor(independent) variables are in hundreds. To

answer that, one needs to have a rigorous statistical model called **OLS** (Ordinary Least Square) Regression. If there is only one predictor variable the regression is called **SLR** (Simple Linear Regression). If there are 2 or more predictor variables the regression is called **MLR** (Multiple Linear Regression) E.g. Fuel efficiency (response variable) of a car depends on multiple predictor variables (car weight, engine size, engine type, fuel type, road conditions,..... etc.)

SR NO	Temperature	Ice cream cones Sold
1	50 ^{0}F	0
2	75 ^{0}F	100
3	100 ^{0}F	200

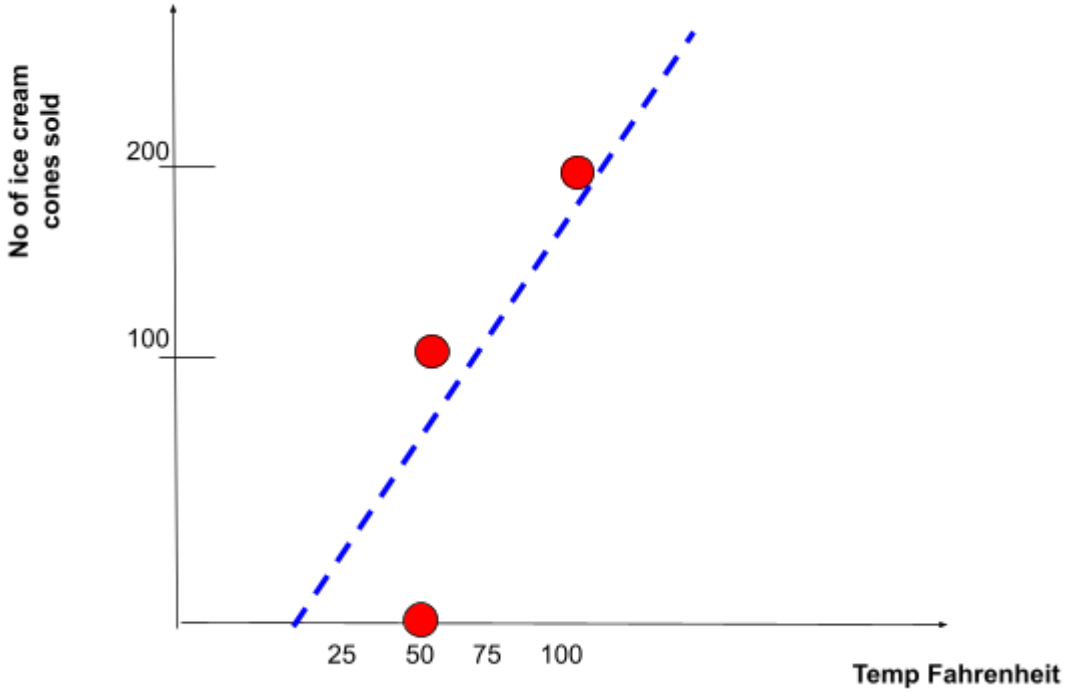


Figure 1

SOME MATHEMATICAL CONCEPTS

1. Mean/ Average:

If x= { 1,2,3 } then mean $\overline{X}$ = $\frac{1+2+3}{3}$= 2 . In general, for N observations:

$$\overline{X} = \frac{x_1 + x_2 + \ldots\ldots\ldots\ldots + x_N}{N} = \sum_{1}^{N} \frac{Xi}{N}$$ **…… Equation (1)**

2. Variance:

Variance for n observations:

$$\sigma^2 = \frac{\sum_{i=1}^{N} \left(Xi - \overline{x}\right)^2}{N-1}$$ **…… Equation (2)**

Variance, in a simple sense, gives the spread of observations/data-points w.r.t. Mean $\overline{X}$. Here n-1 appears as one degree of freedom is used up in calculating mean $\overline{X}$.

3. Standard Deviation:

The square root of variance is called standard deviation.

Thus std deviation for a data-set of N observations

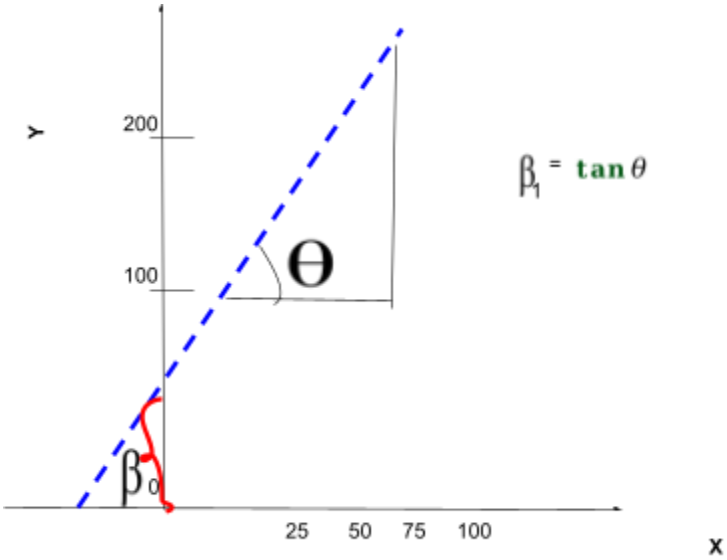
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} \left(Xi - \overline{x}\right)^2}{N-1}}$$ **……..Equation (3)**

4. Equation Of a straight line in 2-D

In X-Y plane the equation of a straight line is

$y = \beta_0 + \beta_1 X$. Here β_0 is intercept on Y axis and

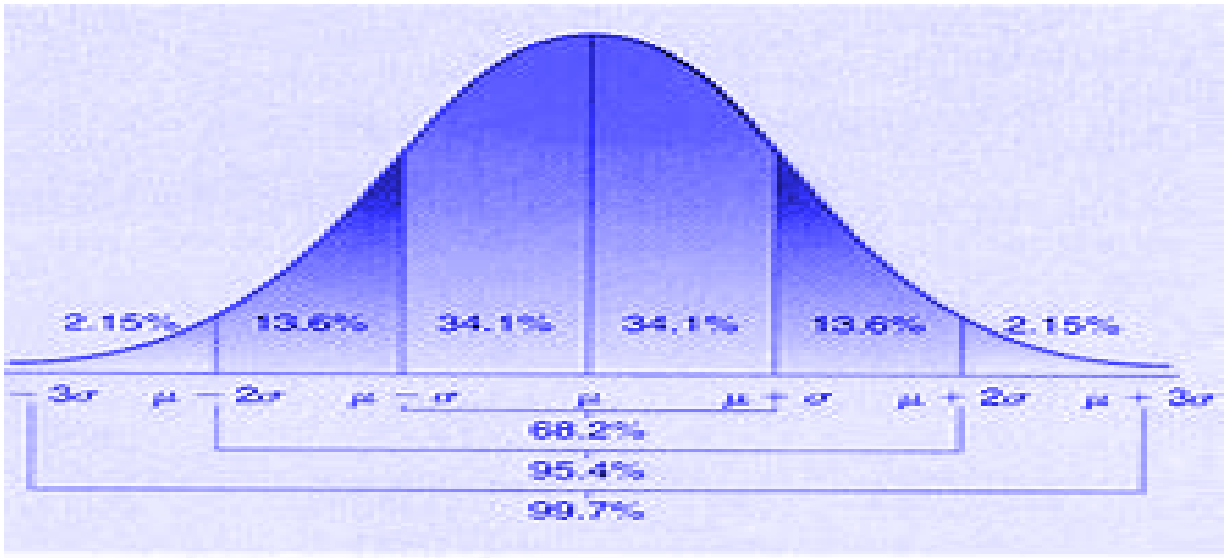
β_1 is called Slope

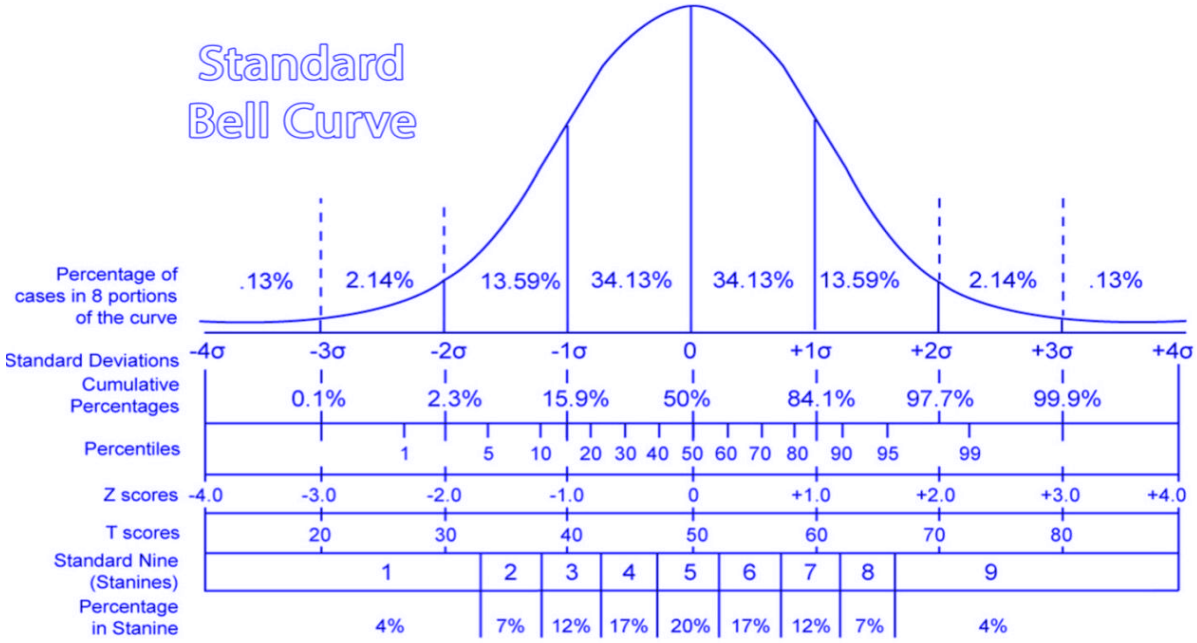


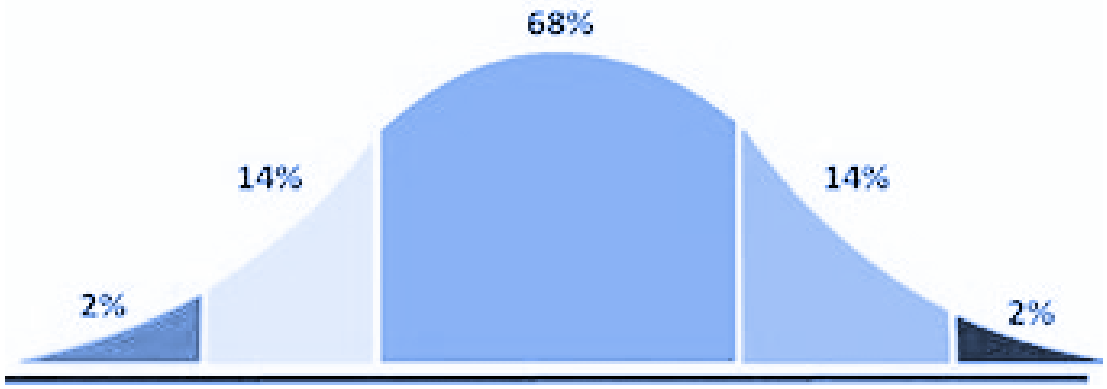
5. Normal/Gaussian Distribution OR Bell Curve

A bell curve is the informal name of a graph that depicts a normal probability distribution. The term obtained its name due to the bell-shaped curve of the **normal probability distribution graph**. However, the normal probability distribution is not the only probability distribution whose graph shows a bell-shaped curve. For example, the graphs of the **Cauchy** and **logistic** distributions also demonstrate a bell-shaped curve.

Characteristics of a Bell Curve:


2.15%
13.6%
34.1%
34.1%
13.6%
2.15%
68.2%
95.4%
99.7%

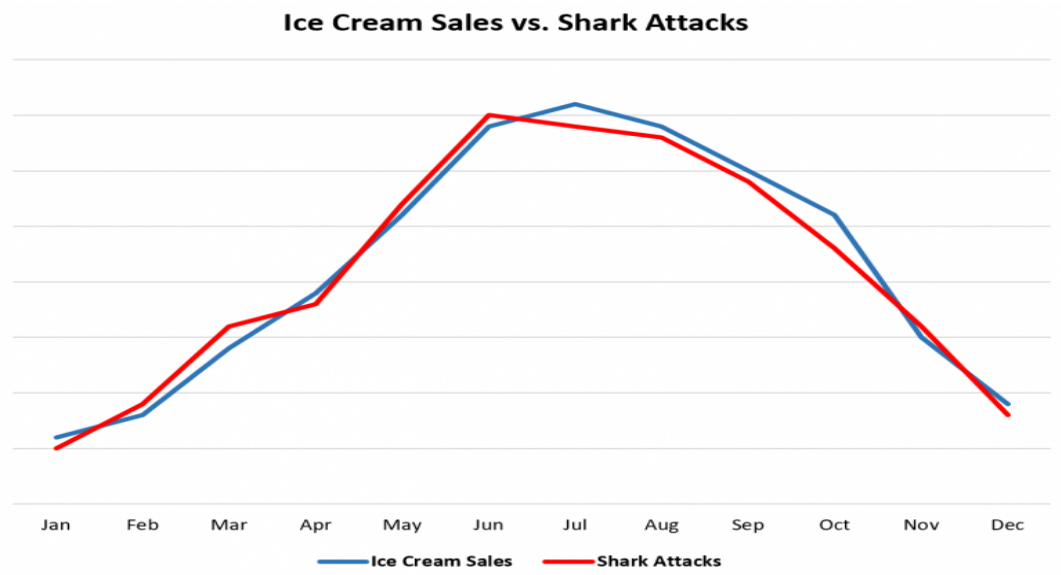

Standard
Bell Curve
Percentage of cases in 8 portions of the curve
.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
.13%
Standard Deviations
-4σ
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ
+4σ
Cumulative Percentages
0.1%
2.3%
15.9%
50%
84.1%
97.7%
99.9%
Percentiles
1
5
10
20
30
40
50
60
70
80
90
95
99
Z scores
-4.0
-3.0
-2.0
-1.0
0
+1.0
+2.0
+3.0
+4.0
T scores
20
30
40
50
60
70
80
Standard Nine (Stanines)
1
2
3
4
5
6
7
8
9
Percentage in Stanine
4%
7%
12%
17%
20%
17%
12%
7%
4%


68%
14%
14%
2%
2%

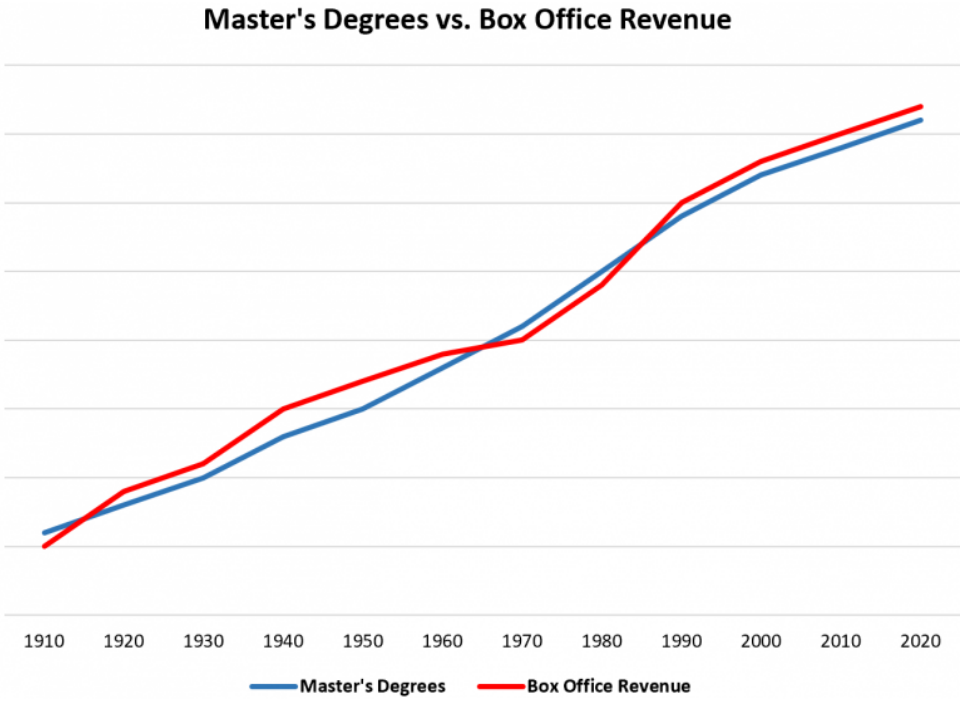
6. **<u>Correlation Does Not Imply Causation:</u>**

In SLR/MLR model building it is important to know which out of several possible independent variables should be included as predictor variables? Normally for this we find correlation of response variable and various independent predictor variables and **<u>just assume</u>** that **the independent variables with highest correlation with response variable should be included as 'predictor' variables in the model.** But there is a word of caution here: "**Correlation Is Not Causation**". Here are two examples for it:

(A) If we collect data for monthly ice cream sales and monthly shark attacks around the United States each year, we would find that the two variables are highly correlated. Does this mean that consuming ice cream causes shark attacks? Not quite. The more likely explanation is that more people consume ice cream and get in the ocean when it's warmer outside, which explains why these two variables are so highly correlated. **Although ice cream sales and shark attacks are highly correlated, one does not cause the other**.



(B) If we collect data for the total number of Master's degrees issued by universities each year and the total box office revenue generated by year, we would find that the two variables are highly correlated.



Does this mean that issuing more Master's degrees is causing the box office revenue to increase each year?

Not quite. The more likely explanation is that the global population has been increasing each year, which means more Master's degrees are issued each year and the sheer number of people attending movies each year are both increasing in roughly equal amounts.

Although these two variables are correlated, one does not cause the other.

Thus our assumption that the most correlated independent variables are the predictor for response variable **is just a hypothesis** which needs to be tested by statistical tests and by getting good model performance.

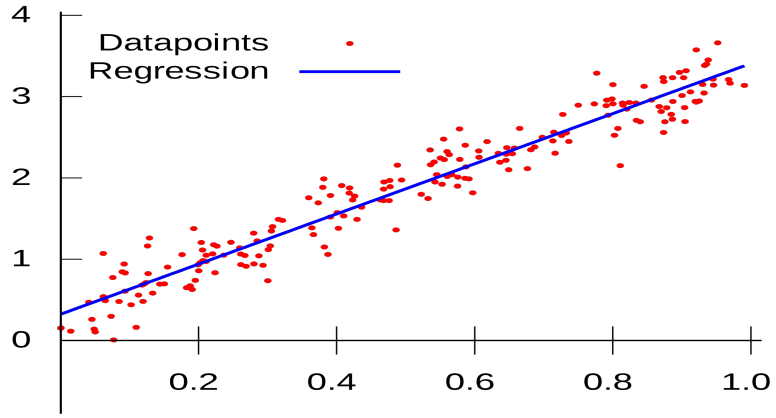
FurtherReference:
https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation#:~:text=The%20phrase%20%22correlation%20does%20not%20imply%20causation%22%20refers,of%20an%20observed%20association%20or%20correlation%20between%20them.

Another thing which one has to look for before assigning a variable as **'predictor'** of some **'response'** variable is the presence of a confounding

variable (lurking variable). E.g. It is seen that higher the foot traffic , higher the crime rate. Does that mean that 'foot-traffic' is responsible for 'crime' OR can we ban people from walking in order to reduce crime? Of course not! Looking deeper it was found that there is a third **'hidden/confounding/lurking'** variable **'temperature'** which is simultaneously causing an increase in foot-traffic **and** (due to more crowd) causing an increase in crime-rate.

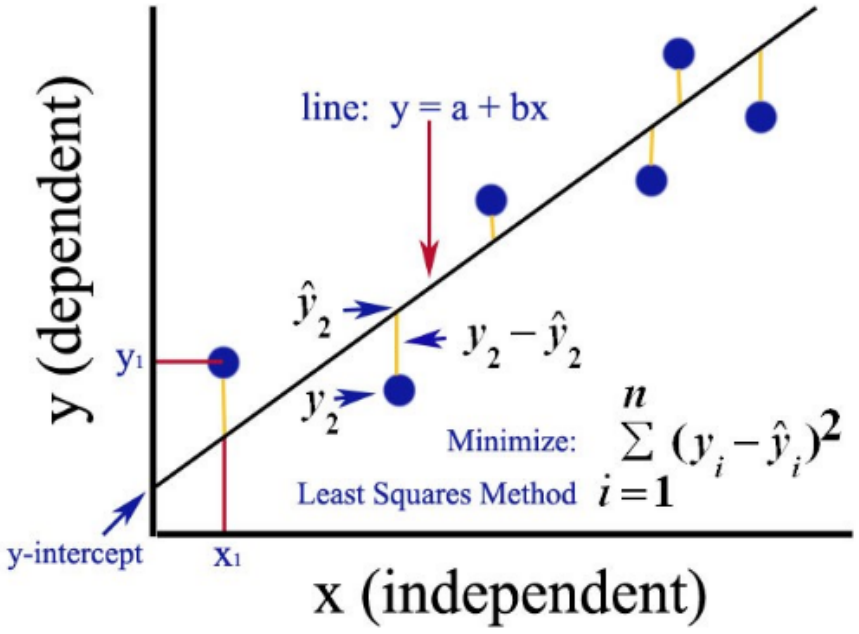
Formula For Linear Regression:

Suppose the data points are spread in the X-Y plane as shown in the figure by red dots. Our task here is to find the **regression line** or **Line Of Best Fit.** (blue color).



'Finding' the line means finding the values of coefficients β_0 **and** β_1 in the equation $y = \beta_0 + \beta_1 X$

One method to find values of coefficients β_0 **and** β_1 **(**or a and b in diagram below**)** is **<u>minimizing the residuals</u>** or **<u>minimizing the least square error</u>** (Hence the name **ordinary least square regression**). So first we need to understand what is meant by residual.



In the diagram above blue dots are **actual** data points. Suppose y coordinates of these points (i.e. actual values of response variable) are y_1, y_2,....., y_N.

Suppose the black line is the line of best-fit that we have obtained. All points on this line are predicted values of the response variable. Suppose they are $\hat{y}_1$, $\hat{y}_2$,....., $\hat{y}_N$ **. Then the residual error in each**

observation is $(y_1 - \hat{y}_1)$, $(y_2 - \hat{y}_2)$, ……., $(y_N - \hat{y}_N)$ **respectively.** From the diagram above it is clear that these error-terms in individual observations can be positive or negative. But we can define a term **SSE(S**um **O**f **S**quared **E**rror**) which is always positive.** It is defined as:

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

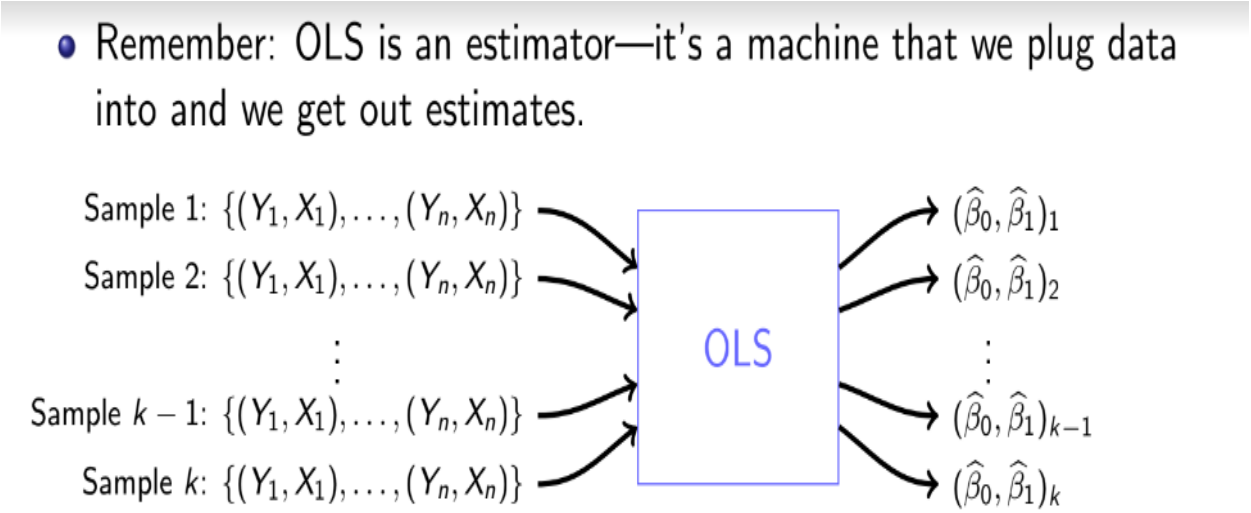
The **goal of ordinary least square regression is estimating the values of coefficients** β_0 **and** β_1 **by minimizing the value of SSE**.

NOTE:- whether we **can** really estimate the values of β_0 **and** β_1 is really a big debate in statistics and it divides statistics itself into two branches: **Bayesian statistics (we can not accurately estimate them)** Vs.**Frequentist statistics (we can estimate them with adequate - but not 100%- accuracy).**

For more interesting discussion on two approaches refer two links below:

https://cxl.com/blog/bayesian-frequentist-ab-testing/

https://analyticsindiamag.com/a-hands-on-guide-to-frequentist-vs-bayesian-approaches-in-statistics/



Before discussing how to estimate the values of coefficients β_0 and β_1 by minimizing the value of SSE let's define another term '**MSE**' (**M**ean **S**quared **E**rror) as below:

MSE = $\in = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N-2}$.

The equation of OLS regression in 2-D is often written including this error term as:

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 X + \in$$

Now it can be proved by partial derivative that

Estimated Value of β_0 **=** $\widehat{\beta}_0$ **=** $\dfrac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^2)-(\sum x)^2}$

Estimated Value of β_1 **=** $\widehat{\beta}_1$ **=** $\dfrac{(\sum y)(\sum x^2)-(\sum x)(\sum xy)}{n(\sum x^2)-(\sum x)^2}$

Ref:https://www.cuemath.com/data/regression-coefficients/

Alternatively the formulae for estimated values of coefficients β_0 **(i.e. intercept of regression line)** and β_1 **(i.e. slope of regression line) are given by:**

(This is more popular format)

$$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}$$

Derivation Ref:
https://are.berkeley.edu/courses/EEP118/current/derive_ols.pdf

Coefficients can be estimated by python.Please refer to the following jupyter notebook.

https://drive.google.com/file/d/1hNq6ljdo_9THGxfC21bLPuU8aRb6ycPz/view?usp=share_link

Here we have used a soccer data-set in which 'Score' is the response variable and there are 4-5 independent/predictor variables. The predictor that is having maximum correlation with score is 'Cost' (of player). So in OLS model we make equation:
Score = $\widehat{\beta}_0 + \widehat{\beta}_1 Cost$ **(Ignoring error term** $\in$**)**

Suppose the estimated value of β_0 **i.e.** $\widehat{\beta}_0$ = 0.6439 and the estimated value of β_1 is $\widehat{\beta}_1$= 0.2345. Thus we have **Score = 0.6439 + 0.2345 * Cost.**

Using the above formula **we can forecast 'Score' for any given player whose 'Cost' is known.** However at

this stage one must know the difference between interpolation and extrapolation.

Interpolation is the act of estimating a value within two known values that exist within a sequence of values. For example, in above soccer dataset **the minimum value of 'Cost' is 32** and **the maximum value of 'Cost' is 200** then the forecast works in the band of 'Cost' $\epsilon[32, 200]$ (Plus a small margin on either side). But **If we are forecasting outside this band of 'Cost'** $\epsilon[32, 200]$ **then the results of the OLS regression model become less and less reliable** as we are now extrapolating.

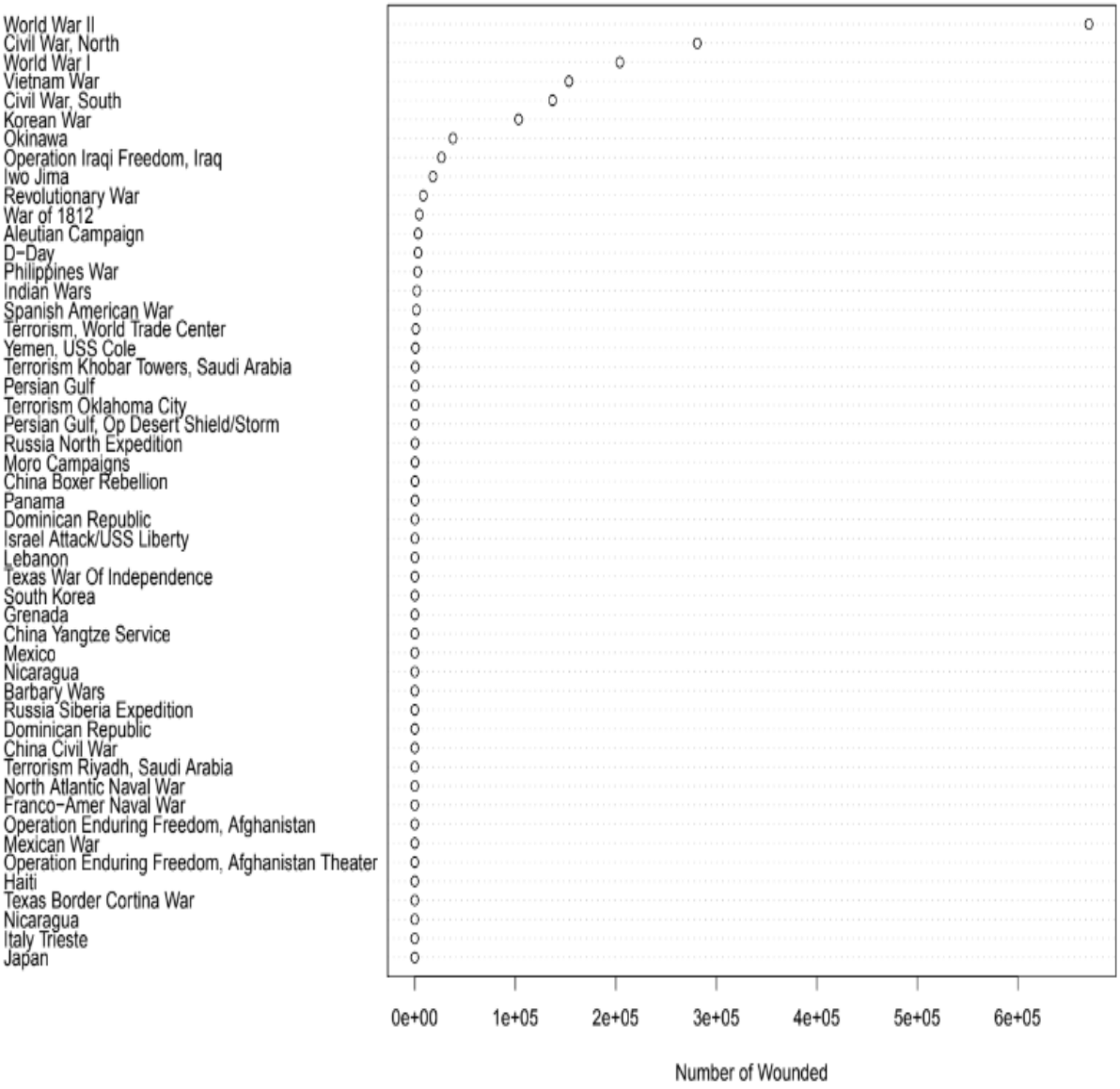
Extrapolation refers to estimating an unknown value based on extending a known sequence of values or facts. To extrapolate is to infer something not explicitly stated from existing information.

Even while interpolating, one must know, that OLS model requires assumptions, as below, to be correct:

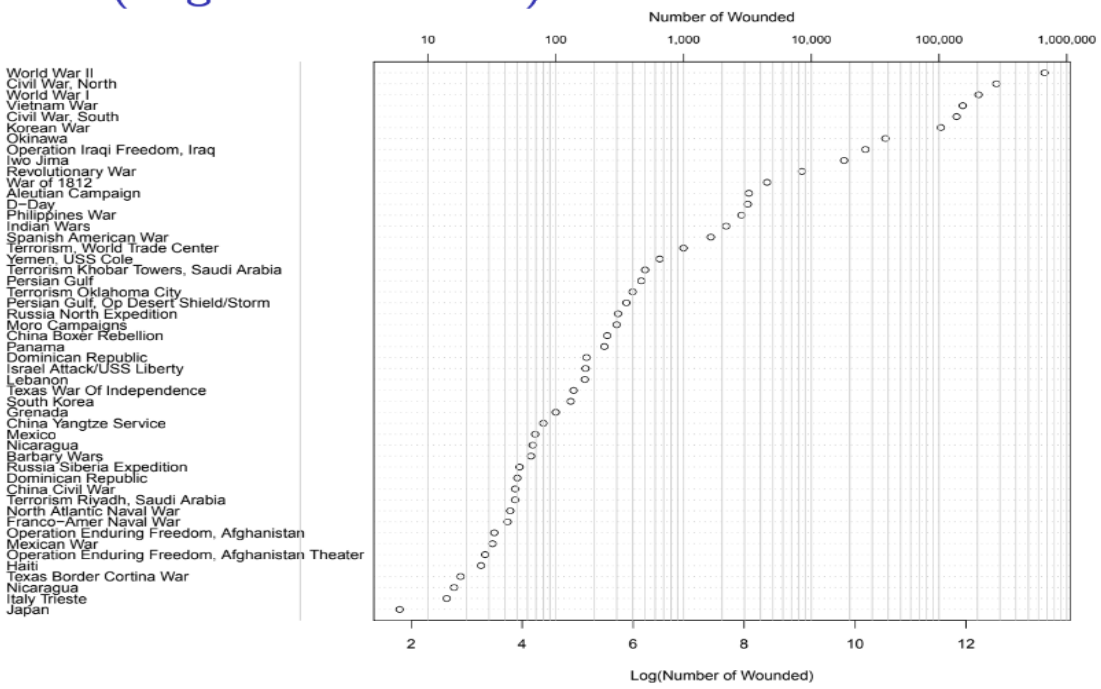
OLS Assumptions

1 Linearity in Parameters: The population model is linear in its parameters and correctly specified. In simple language "y increases or decreases (approximately) linearly with x"

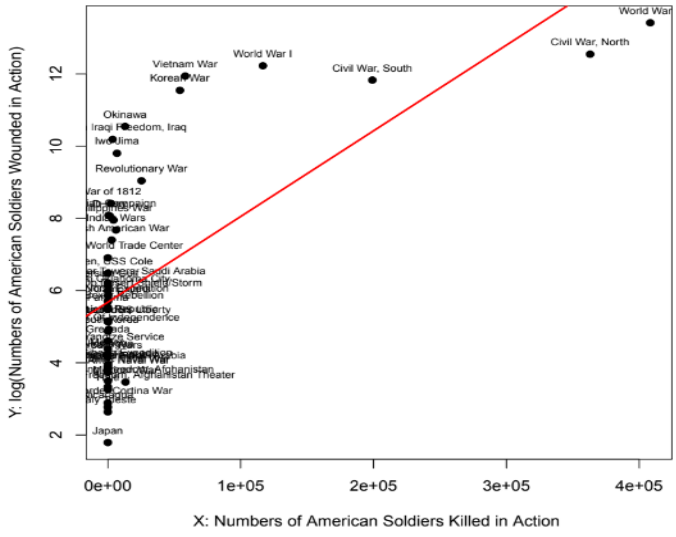
It is important to note that not all is lost if the scatter plot shows non-linear distribution. One can still use variable-transformation to obtain a nearly linear distribution. As an example consider a graph showing no of wounded on X axis in various big wars (on Y axis).



Wounded (Logarithmic Scale)

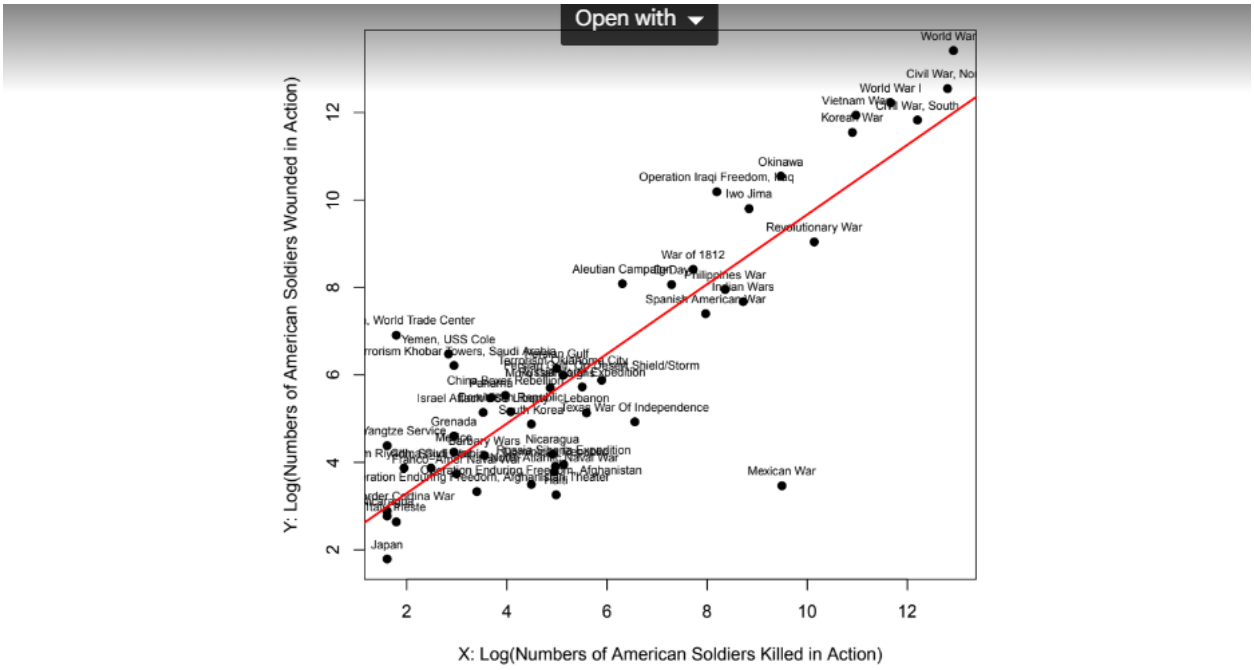


The regression line fitted for even logarithmic scale (logY Vs. X) is not satisfactory.



$\hat{\beta}_1 = 0.0000237 \longrightarrow$ One additional soldier killed predicts 0.0023 percent increase in the number of soldiers wounded on average

But the regression line fitted for log-log scale(LogY Vs Log X) is highly satisfactory. Similarly one can use **square-root transformation**,**Box-Cox transformation** etc.



$\hat{\beta}_1 = 0.797 \longrightarrow$ A percent increase in deaths predicts 0.797 percent increase in the wounded on average

<u>2 Random Sampling:</u> The observed data represent a random sample from the population described by the model.

- ❖ A population is the entire group that you want to draw conclusions about. It can be a group of anything you want to study, such as

persons, objects, events, organizations, countries, species, organisms, etc.

- A sample is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.

3 Variation in X: There is variation in the explanatory/predictor variable.

4 Zero conditional mean: Expected value of error term is zero conditional on all values of explanatory variable

5 Homoscedasticity: The error term has the same variance conditional on all values of the explanatory variable.

6 Normality: The error term is independent of the explanatory variables and normally distributed. **Below is detailed discussion of OLS assumptions and some possible cases where the assumptions can be violated.**

A MORE DETAILED EXPLANATION OF ABOVE ASSUMPTIONS IS BELOW

OLS Assumption 1. Linearity in Parameters

Population regression model is **linear in its parameters** and correctly specified as:

$Y = \beta_0 + \beta_1 X_1 + \in$

Note that it **can be nonlinear in variables**

Examples below **DO NOT** violate linearity assumption

$Y = \beta_0 + \beta_1 X + \in$ or

$Y = \beta_0 + \beta_1 X^2 + \in$ or

$Y = \beta_0 + \beta_1 \log(X) + \in$

Examples below **do not have linearity in parameters**

$Y = \beta_0 + \beta_1^2 X + \in$ or

$Y = \beta_0 + \exp(\beta_1) X + \in$

β_0, β_1: Population parameters — fixed and unknown

$\in$: Unobserved random variable with E[u] = 0. It captures all other factors influencing Y other than X

OLS Assumption 2: Random Sampling

The observed data: (y_i, x_i) for i = 1, ..., n

represent an **i.i.d. (independent and identically distributed)** random sample of size n following the population model.

In probability theory and statistics, a collection of random variables is ***independent and identically distributed*** *if each random variable has the same probability distribution as the others and all are mutually independent*

Data example consistent with this assumption is a cross-sectional survey where the units are sampled randomly

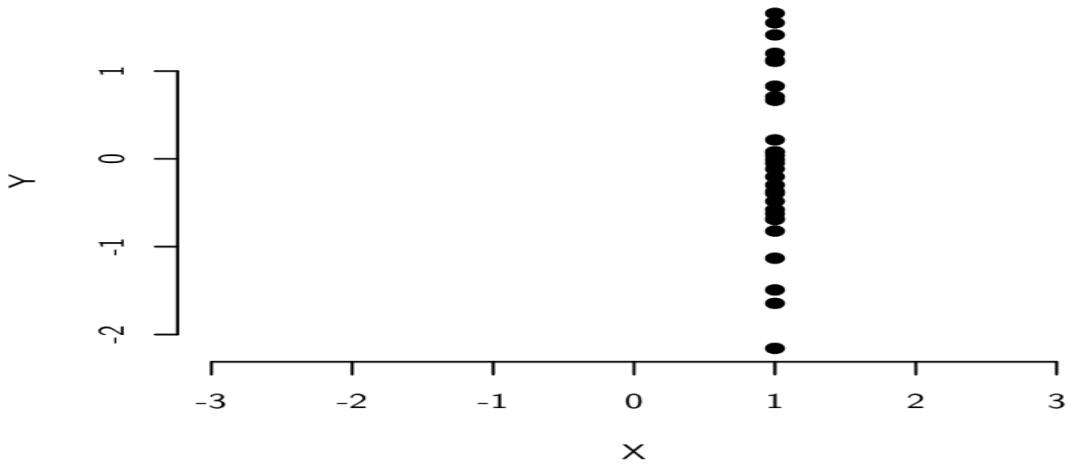
Potential violations of this assumption can be seen in:

1. Time series data (regressor values may exhibit persistence)
2. Sample selection problems (sample not representative of the population)

OLS Assumption 3: Variation in X; a.k.a. No Perfect Collinearity)

The observed data: x_i for i = 1, ..., n are not all the same value.

Why does this matter? How would you draw the line of best fit through the scatterplot below, which violates this assumption?



Here x_i and $\bar{x}$ are the same as x_i = constant. Thus we can not calculate if this assumption is not valid.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

<u>In fact, this is the only assumption needed for using OLS as a pure data summary.</u>

OLS Assumption 4: Zero conditional mean

The expected value of the error term is zero conditional on any value of the explanatory variable: **$E[\in|X] = 0$**

NOTE:

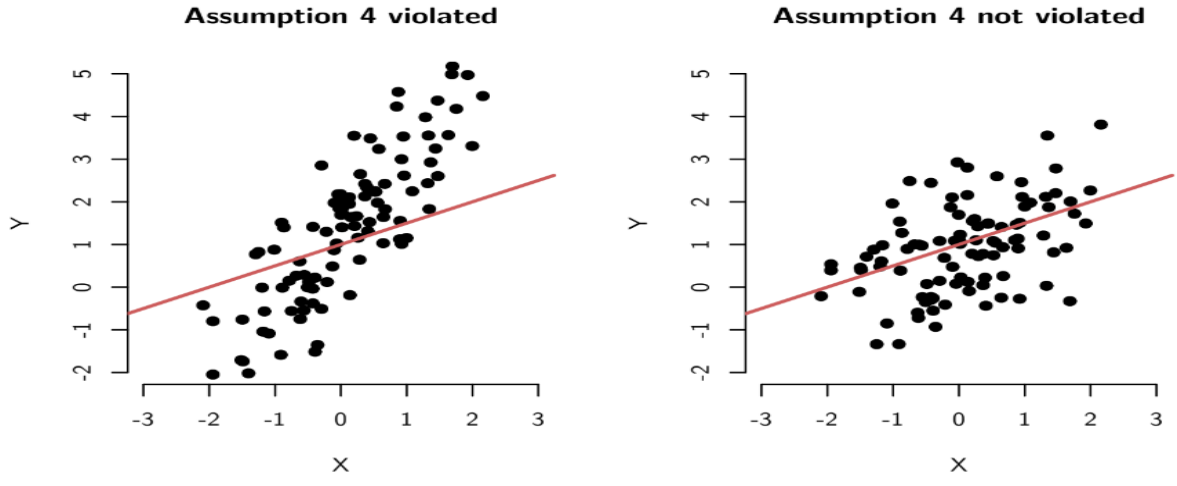
1. **E[∈|X] = 0** implies a slightly weaker condition **Cov(X, ∈) = 0**.
2. Given random sampling, **E[∈|X] = 0** also implies

   $E[\in_i|x_i] = 0$ for all i

Violations:

Recall that ∈ represents all unobserved factors that influence Y. If such unobserved factors are also correlated with X. **Hence Cov(X, ∈) ≠0**

In left diagram below, when x is negative the error ∈are negative and when x is positive, the error ε are positive. When x=0, ∈ =0 (nearly). Thus **Cov(X, ∈) ≠0. Hence assumption 4 is violated.** But in the right side figure, error terms are randomly scattered without having any relation with X. So, assumption-4 is valid



Now we know that, under Assumptions 1-4, we know that:

$$\widehat{\beta_1} \sim ?(\beta_1, ?)$$

In words this means that we know that **the sampling distribution is centered on the <u>true population slope</u>**, but we don't know the population variance. In order to derive the sampling variance of the OLS estimator, apart from assumptions (1) To (4) above, we need another assumption: **Homoscedasticity**.

OLS Assumption 5 Homoscedasticity:

How can we derive Var[$\widehat{\beta}_0$] and Var[$\widehat{\beta}_1$]? Let's make the following additional assumption: **The conditional variance of the error term is constant and does not vary as a function of the explanatory variable**:

- This implies **Var[$\in$|x] = $\sigma^2\in$**

→ **all errors have an identical error-variance** ($\sigma^2_{ui}=\sigma^2_u$ for all i)

- Taken together, Assumptions I–V imply:

E[Y |X] = $\beta_0 + \beta_1 X$

Var[Y |X] = $\sigma^2 u$

<u>Violation:</u>

Var[$\in$|X = x_1] $\neq$Var[$\in$|X = x_2] is called heteroskedasticity.

In short heteroskedasticity means non-constant variance and homoscedasticity means constant variance of error terms for all values of X.

The assumptions I to V are collectively known as **Gauss-Markov assumptions.**

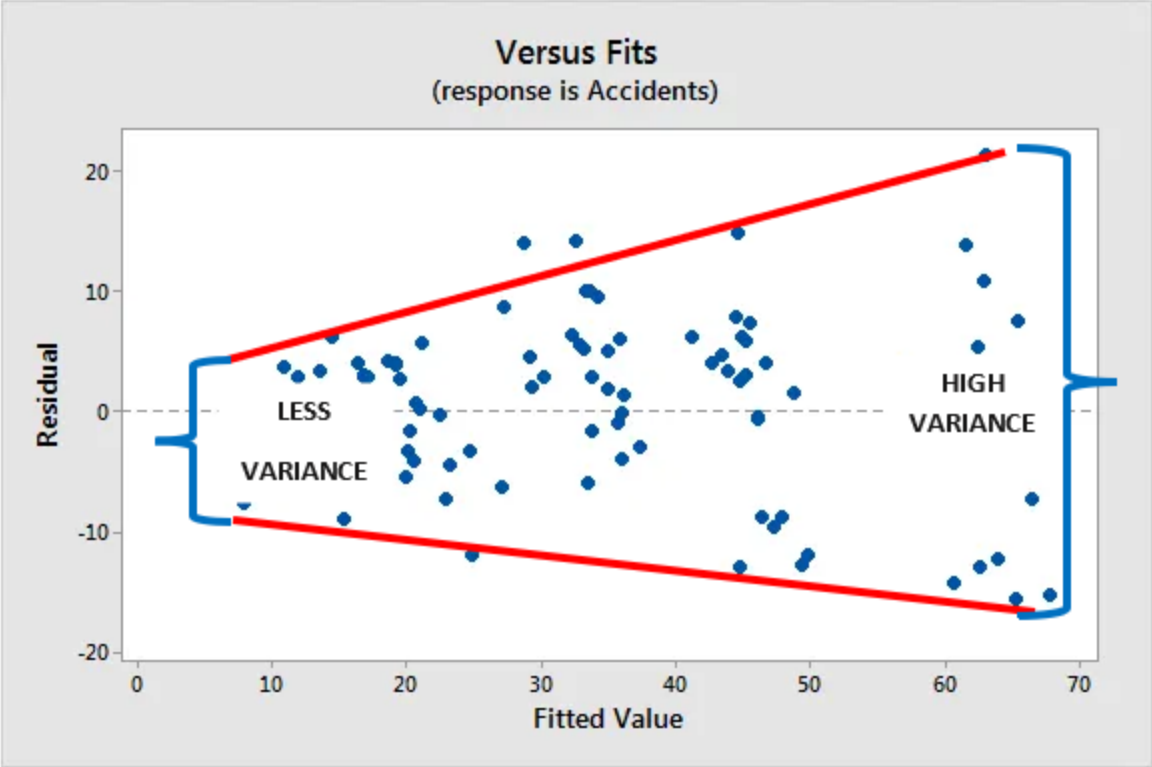
If these five assumptions are obeyed then:

$$Var[\hat{\beta}_1 \mid X] = \frac{\sigma_u^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma_u^2}{SST_x}$$

$$Var[\hat{\beta}_0 \mid X] = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right\}$$

where $Var[u \mid X] = \sigma_u^2$ *(the error variance).*

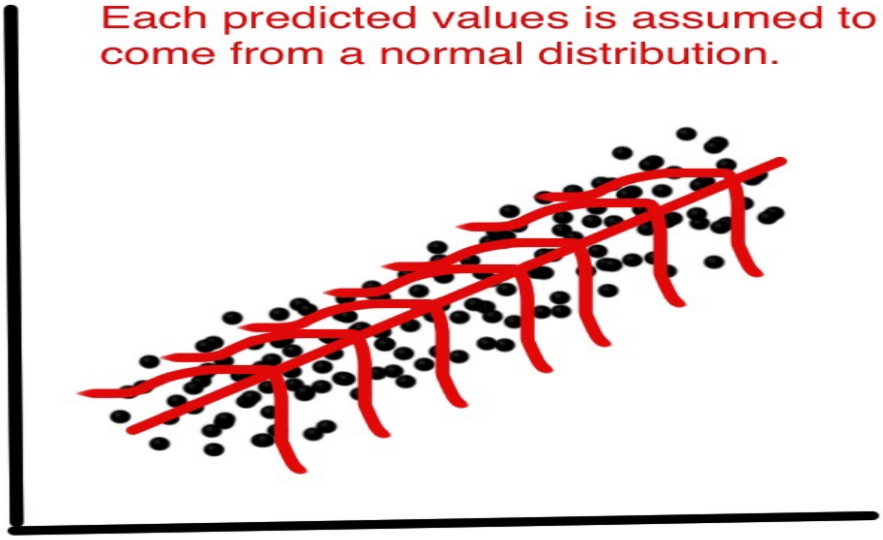
A quick method to see if there is **heteroskedasticity** is to examine the scatter plot of residual Vs fitted values. If this scatterplot has a megaphone like shape it indicates non-constant variance and assumption 5 is violated



Even in this case one can use linear regression but after 'fixing' heteroskedasticity.

OLS Assumption 6: Normality

In simple language the error terms(black) $\in$ are distributed normally(like bell curve) with zero mean and σ^2 variance i.e. $\in \sim \mathbf{N(0,\sigma^2)}$ w.r.t. the line of fit(red) as (diagram)



Regarding above assumptions we must note following things:

When Assumptions I–IV are all satisfied, we can estimate the structural parameters β without bias and **thus make causal inference.**

However, we can make **predictive inference even if some assumptions are violated**.

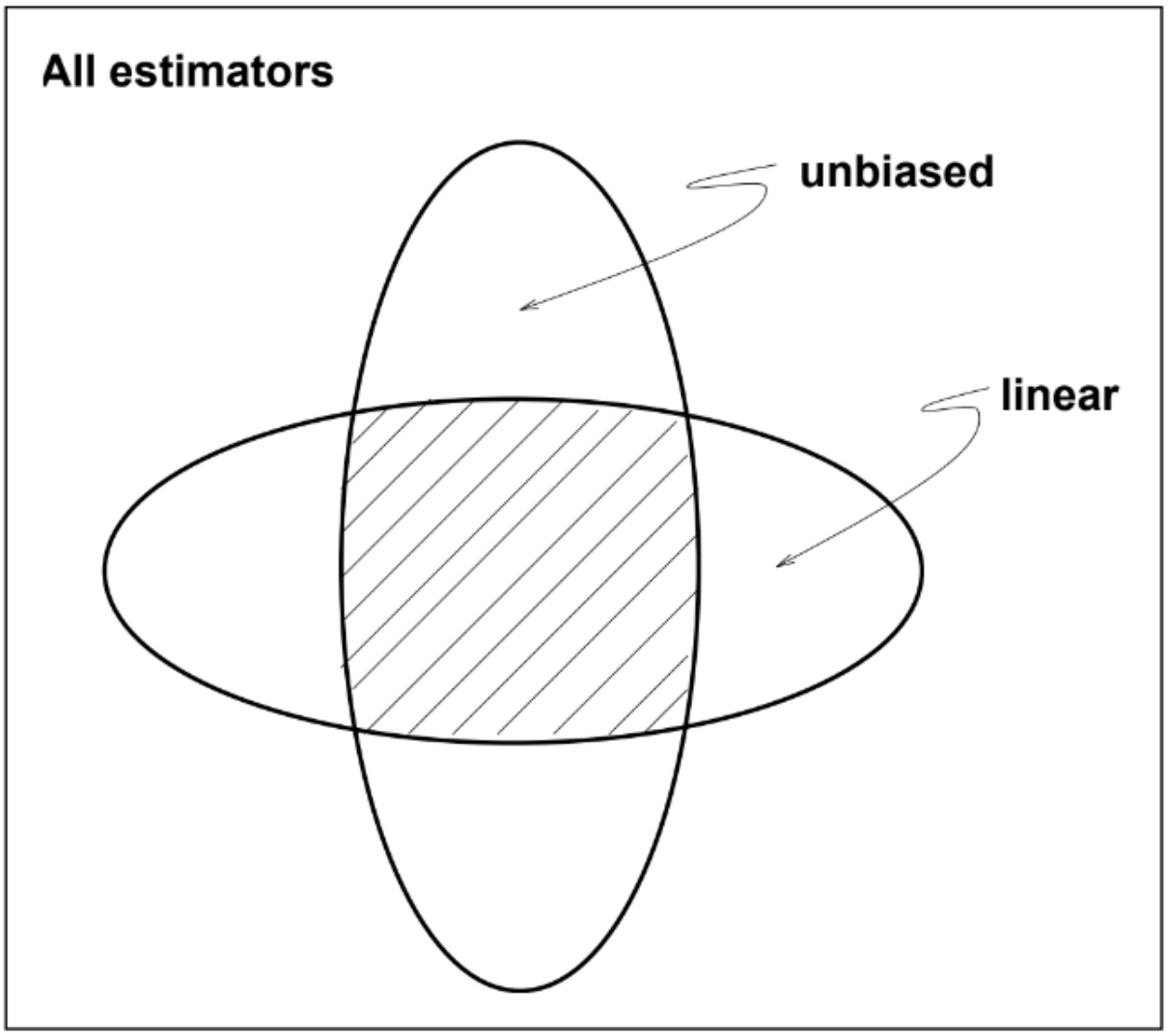
Further, there are ways to ‘fix’ things (not always) when one or more assumptions are violated. For details See links below:

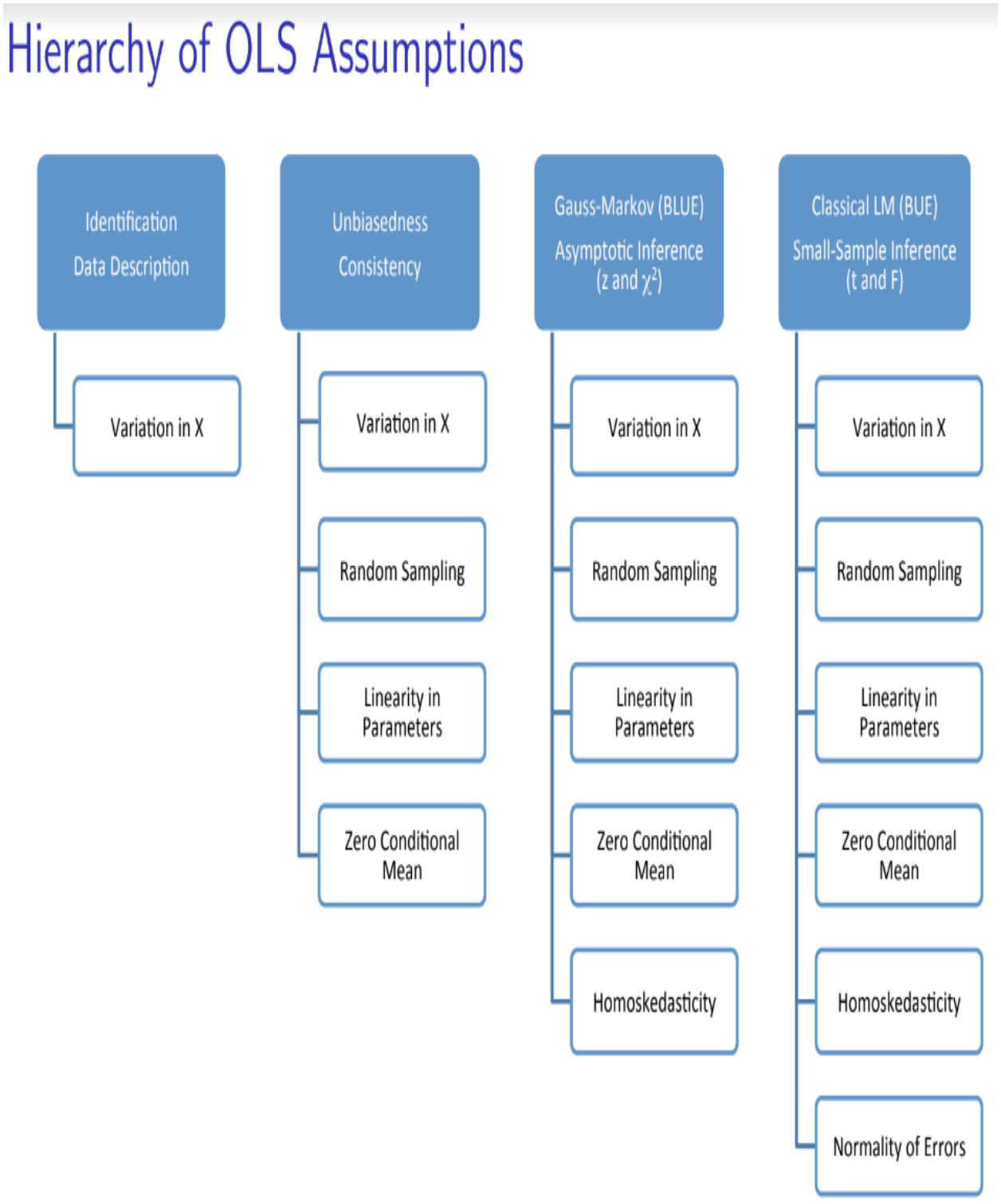
- Slide 10 of this pdf
- Link-2
- Link-3

Gauss-Markov theorem:

Given OLS Assumptions 1-5, the OLS estimator is **BLUE**, i.e. **B**est **L**inear **U**nbiased **E**stimator:

1. **B**est here means Lowest variance
2. **L**inear: Among Linear estimators
3. **U**nbiased: Among Linear Unbiased estimators





Properties of OLS/RegressionLine

1. **The residuals will be 0 on average**: $\frac{1}{n}\sum_{i=1}^{n} \widehat{\in}_i = 0$

   Where $\widehat{\in}_i = y_i - \widehat{y}_i$.

2. **The residuals will be uncorrelated with the predictor :** $\widehat{cov}$ **(X$_i$,** $\widehat{\in}_i$ $) = 0$

3. **The residuals will be uncorrelated with the fitted values:** $\widehat{cov}$ **(Y$_i$,** $\widehat{\in}_i$ $) = 0$

   In 2 and 3 above $\widehat{cov}$ is the sample covariance

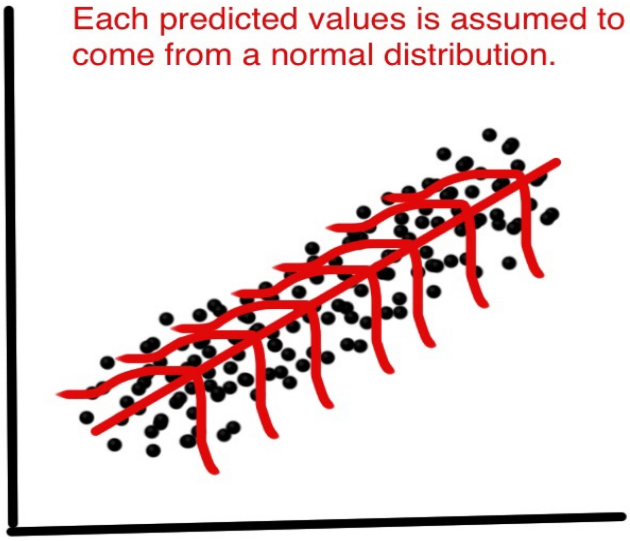
4. Sum of square of residuals $\sum_{i=1}^{N}\left(y_i - \widehat{y}_i\right)^2$ is minimum.

5. $\sum\widehat{y}_i$ **=** $\sum y_i$ **i.e. sum of fitted values = sum of actual values** for response variable.

6. Regression line always passes through point($\overline{X}$, $\overline{Y}$)

POINT ESTIMATE OF VARIANCE

We had earlier seen following diagram for regression line:



Let's see how we can make **point estimation of variance** . We know that variance is given by:

$$\sigma^2 = \frac{\sum_{i=1}^{N} \left(Xi - \overline{x}\right)^2}{N-1}$$

(N-1 because , we have used up one degree of freedom in calculation of $\overline{x}$)

We also know that Sum Of Squared Errors/residual sum of squares is given by:

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} r_i^2$$

Going one step further,

$$S^2 = \frac{SSE}{N-2} = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N-2} = MSE$$=Mean Squared Error

It can be shown that the expected value for MSE i.e.

$$E(MSE) = \sigma^2 = \frac{\sum_{i=1}^{N} \left(Xi - \overline{x}\right)^2}{N-1}$$ **is unbiased.**

So the standard deviation *S* is best estimated by the square-root of MSE.

$$S = \sqrt{MSE}$$

SAMPLING DISTRIBUTION OF $\hat{\beta}_0$ AND $\hat{\beta}_1$

Thus so far we have done following things:

1. We have found estimated values of coefficients β_0 **(i.e. intercept of regression line)** and β_1 **(i.e. slope of regression line) are**

$$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X}$$

And

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

2. We have shown that they are minimum variance unbiased linear estimators.

3. We have shown that we can estimate σ^2 using the mean squared error **(MSE) where**

$$MSE = \frac{SSE}{N-2} = \frac{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}{N-2}$$

ANOVA PARTITIONING

ANOVA= Analysis Of Variance In Regression Analysis

Some concepts required for this sections are:

SSTO: Sum Of Square Total. $SSTO = \sum_{i=1}^{N}(y_i - \overline{y})^2$

(N-1 DEGREES OF FREEDOM)

SSE: Sum Squared Error/ Residual Sum Of Squares

$SSE = \sum_{i=1}^{N}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{N} r_i^2$ **(N-2 DEGREES OF FREEDOM)**

SSR/RSS: SUM SQUARED REGRESSION

$$SSR = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$ **(1 DEGREES OF FREEDOM)**

It is clear that $SSR = SSTO - SSE$

Also $MSR = \frac{SSR}{1}$ **(BECAUSE 1 D.O.F.)**

AND $MSE = \frac{SSE}{N-2}$ **(BECAUSE N-2 D.O.F.)**

Consider table below which is often seen as output of many ANOVA packages:

Source	Sum Squared Value SS	DOF	Mean Squared Value MS	Expected (MS)
SSR	$\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$	**1**	**MSR**	$\sigma^2 + \beta_1^2 \sum_{i=1}^{N} (x_i - \bar{X})^2$
SSE	$\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$	**N - 2**	**MSE** $\frac{SSE}{N-2}$	σ^2

SSTO	$\sum_{i=1}^{N} (y_i - \bar{y})^2$	**N - 1**	**-**	**-**

Above SSR, SSE and SSTO are important **as they enable us to calculate important parameters** like **F-score**, which is useful in hypothesis analysis **and** R^2 which is a measure of how much linear information about response variable Y is captured by predictor variable X.

In most of the SLR i.e. $\mathbf{Y = \beta_0 + \beta_1 X_1}$ cases, we take null hypothesis that Y is **NOT** dependent on X_1 i.e. $\boldsymbol{\beta_1 = 0}$ and alternative hypothesis that Y **DOES** depend on X_1 i.e., $\boldsymbol{\beta_1 \neq 0}$.

H_0: $\boldsymbol{\beta_1 = 0}$ (There is **no linear** information in X_1 about Y)

H_1: $\boldsymbol{\beta_1 \neq 0}$ (There **is** linear information in X_1 about Y)

Now to check which Hypothesis is correct we need to find F-Score: $F^* = \frac{MSR}{MSE}$

Now According to Cochran's theorem, under the condition when H_0 is true, F^* has a typical distribution:

F^* **~ (N-2)** $\frac{\chi^2(1)}{\chi^2(N-2)}$

In words: F^* **follows F-Distribution when** H_0 **is true**.

In practice we use F^* and F-distribution curve as below:

$If\ F^* <= F_{1-\alpha,n-2} \Rightarrow H_0\ is\ true \quad If\ F^* > F_{1-\alpha,n-2} \Rightarrow H_1\ is\ true$

We already know how to find $F^* = \frac{MSR}{MSE}$ but the question is how to find $F_{1-\alpha,n-2}$ **(often also called F-critical).** Well, for that we have to use F-Distribution table:

Fortunately we never have to conduct the F-test manually as described above. The output of linear regression process in most known tools like R or Python readily prints all this (and many other) informations. So we just have to use it.

COEFFICIENT OF DETERMINATION: R^2

$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$

R^2 gives an estimation of how much **linear** information is contained in our predictor variable X about our response variable Y . R^2 is the most common quantity to determine validity of a regression model. Further, **0**$\leq R^2 \leq 1$ which gives us freedom to express it in percentage. So, in a model with R^2= 95%, there is more linear information in X_1 about Y than in a model with R^2= 75%**.** R^2=1 or 100% means the 'line of fit' passes through **all** the actual data points and R^2=0 means the 'line of fit' passes through mean Y, i.e. **what at the best the regression model can do is that it predicts mean value**.

Also it must be noted that **high** R^2 **does not mean** the model is good. It just means that there is more linear information in X_1 about Y but there may be another model (may be **non-linear**) which gives an even better fit. So, coefficient of determination R^2 is just about determining linearity (**How linear is Y about X**).

With this much theoretical background, now one can refer again to the jupyter notebook

https://drive.google.com/file/d/1hNq6ljdo_9THGxfC21bLPuU8aRb6ycPz/view?usp=share_link

It has all the steps properly explained and well-commented. The data-set can be downloaded from link below:

https://drive.google.com/file/d/1cLFGRApehOCmbhnvkHRmQ9DkWZZ6f4QX/view?usp=share_link

Using Pandas profiler we can do quick automated EDA. It requires only a few steps. Code in Italic below:

!pip install pandas-profiling

from pandas_profiling import ProfileReport

report = ProfileReport(df, title="Pandas Profiling Report For Soccer Data")

report # will display report with jupyter notebook

report.to_file("Pandas Profiling Report For Soccer Data.html") # will save report to disk

https://drive.google.com/file/d/1svRwwf8d73qbMlEzxYlVXCIcJoqRq2zo/view?usp=share_link

#download and open using a browser. Detailed explanation of what information is obtained by Pandas' Profiler is given in the Jupyter notebook.

The correlation matrix for the given data-set is as below:

Thus there is 0.96 correlation between Score And Cost.

So we start SLR/OLS modeling with:

Score = $\widehat{\beta_0} + \widehat{\beta_1} Cost$

After using statsmodel.api to make a linear model with constant/intercept term, We get $\widehat{\beta_0}$ = 0.9472,$\widehat{\beta_1}$ =0.1821

So, **Score = 0.9472 + 0.1821 * Cost**

Further R^2 = 0.93 indicates that our model will be able to explain 93% of the variance in new data, which means it is a very good model.

Further Prob(F-statistics) = $6.91e^{-88}$ is very small and it indicates that there is nearly zero chance that null hypothesis i.e., H_0: **β_1=0** (There is **no linear** information in X_1 about Y) ; is true. In other words X_1 contains a lot of linear information about Y. This is one more confirmation that linearity assumption is obeyed.

	Coef	Std Err	t	P >|t|
constant	**0.9472**	**0.325**	**2.916**	**0.004**
Cost	**0.1821**	**0.004**	**44.456**	**0.000**

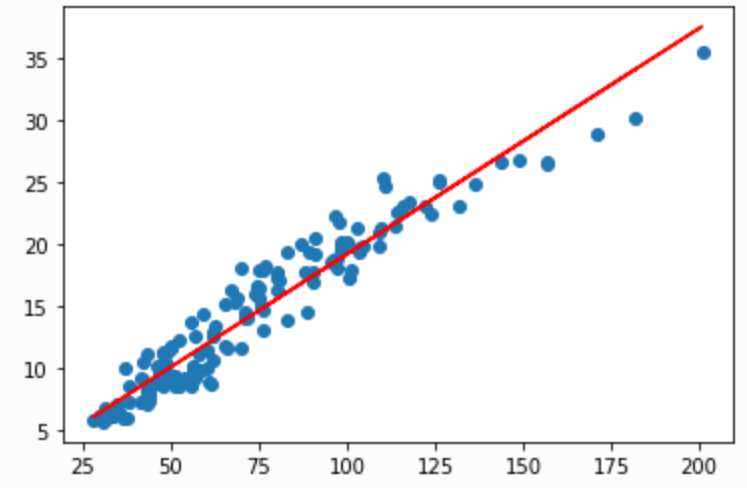
Thus chances of P >|t| for both constant and Cost variables are also miniscule. This means that our model is very good.

To see how good a fit this model provides on train data, we can plot a scatter plot of train data:

plt.scatter(X_train,Y_train)

And on this scatter plot we have to superimpose our line of fit

*plt.plot(X_train,b0+b1*X_train,'r')*



So, the line gives a good fit on train data. Though once Cost goes above 150 (on X axis) model tends to overestimate, how much would that player Score.

Let's do the same with testing data

get test data X_test with intercept

X_test_with_intercept = sm.add_constant(X_test)

predict on fitted data

Y_test_fitted = lr2.predict(X_test_with_intercept)

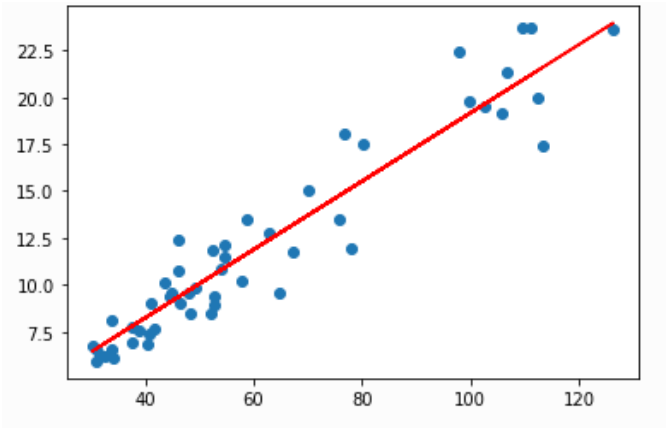
Plot scatter plot of X_test and Y_test

plt.scatter(X_test,Y_test)

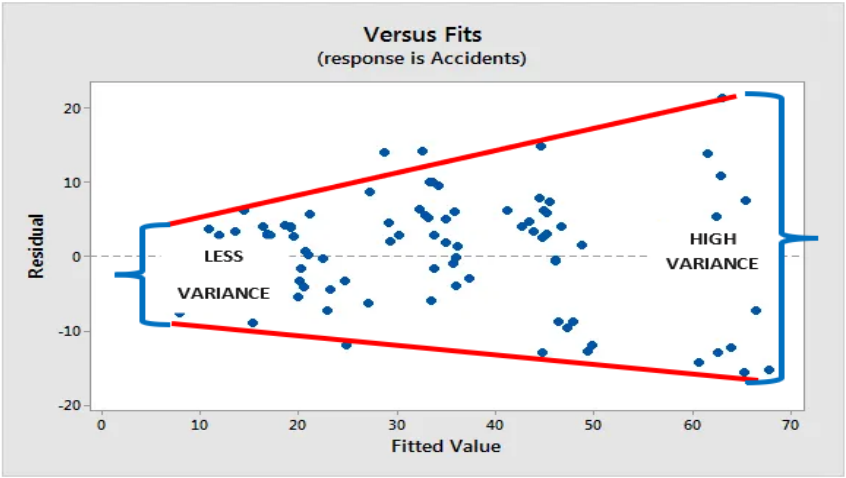
and fit the line as per model lr2

plt.plot(X_test, Y_test_fitted, 'r')

plt.show()

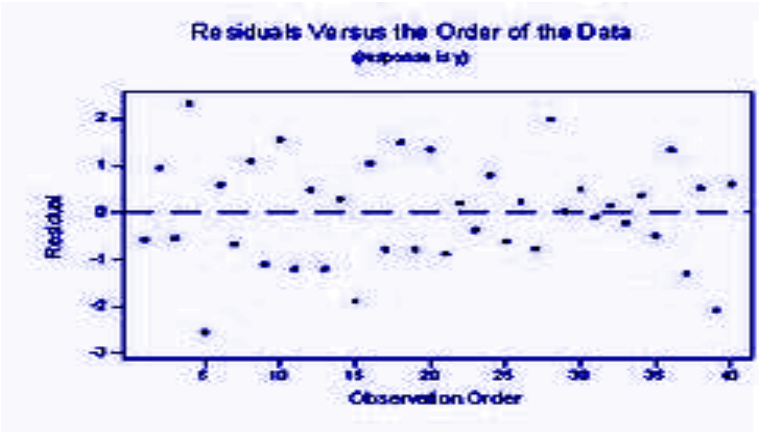


1. Non-linearity. In this case re-evaluate whether the linear model itself is possible or one should try fitting , say, a parabola. In some cases the nonlinearity can be 'fixed' by variable transformations as we saw earlier.
2. Heteroskedasticity (non-constant variance). E.g. The graph of residuals Vs fitted values shows a megaphone like shape. Here one can consider using different standard deviations σ.

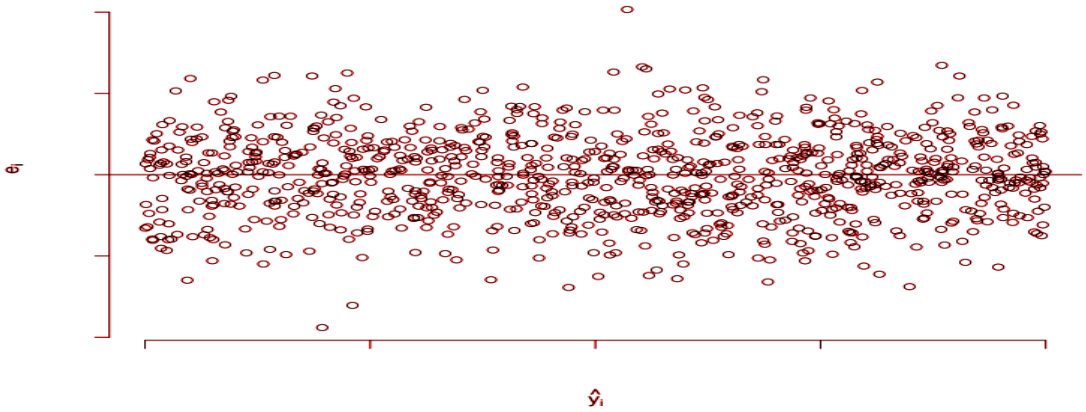


In this case, **consider using transformations** or **using 'weighted least squares'**

3. Sometimes it is helpful to plot **sequence plots of residuals** i.e. residuals Vs time plot to identify any dependency between the residual and time.. This is useful when the residuals for data points plucked at different times show cyclical patterns. **For these cases, time series analysis may be more suitable.**

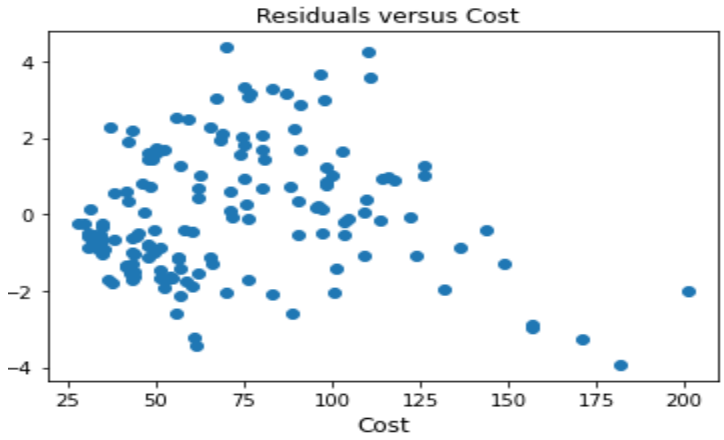


4. Check if errors are not i.i.d (Independent and identically distributed). Perfect i.i.d residual plot (with zero mean) is shown below



5. Missing predictor variable (i.e. using multiple linear regression or MLR, instead of SLR)

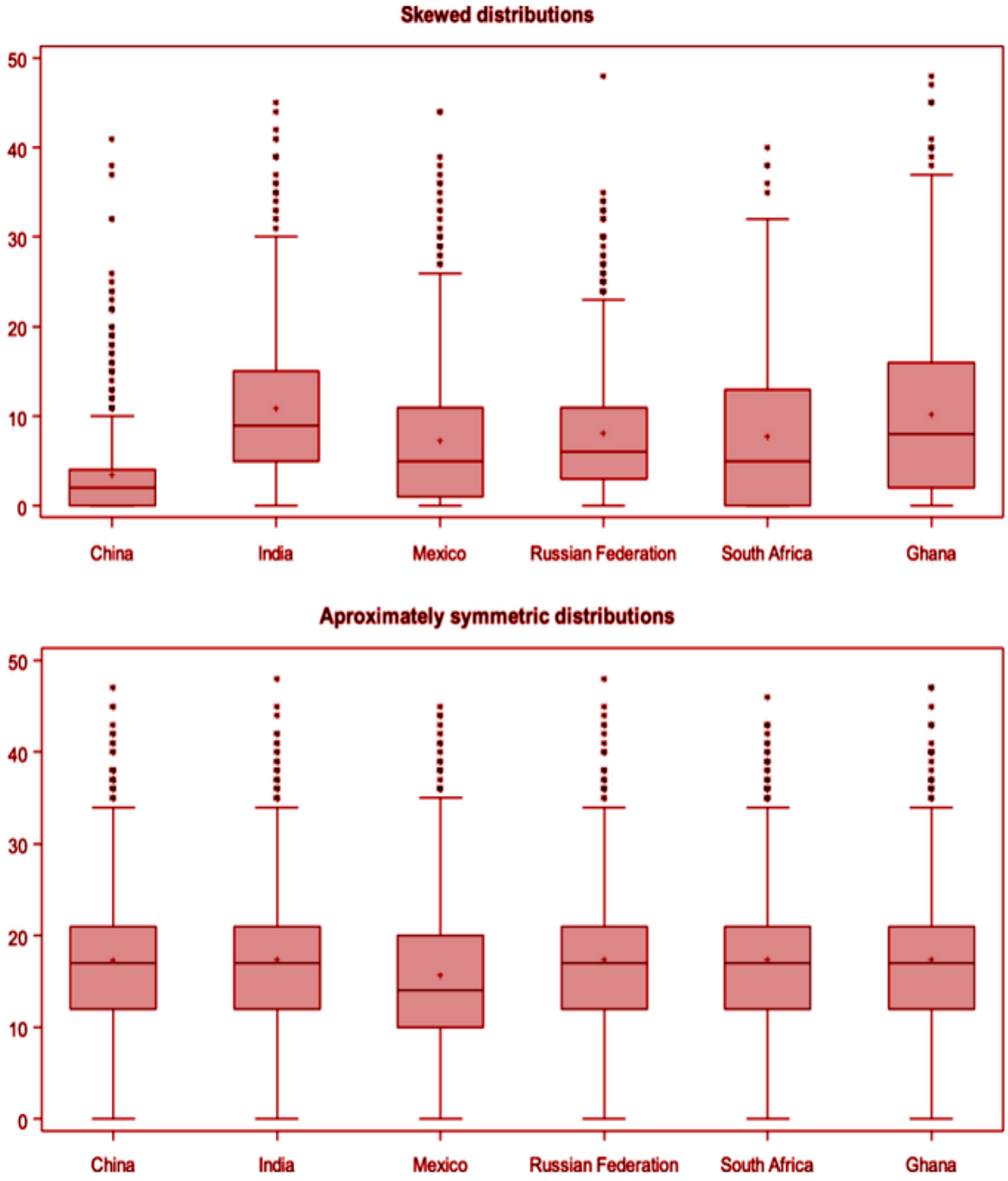
For checking 1 and 2 above we plotted a graph of **residual vs predictor for training data**. It also appears with nearly zero mean.



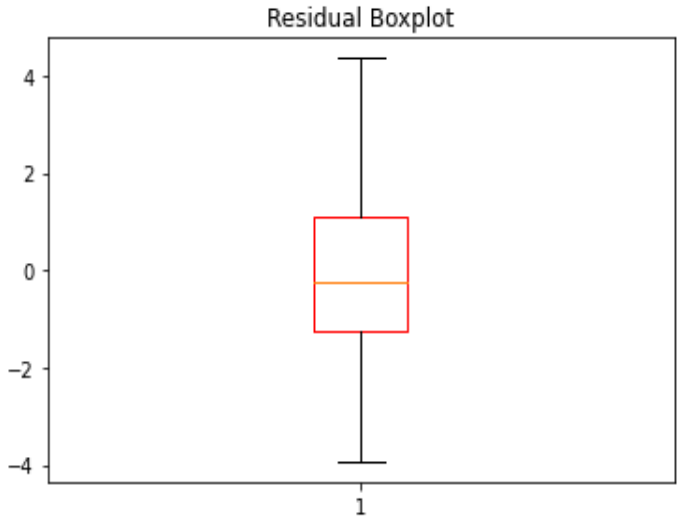
We also plotted a graph of **residual vs predictor for test data**. It also appears with nearly zero mean.

We can plot a box-plot of residuals. Following diagrams show how to recognize skewed box plots (non normal) of residuals

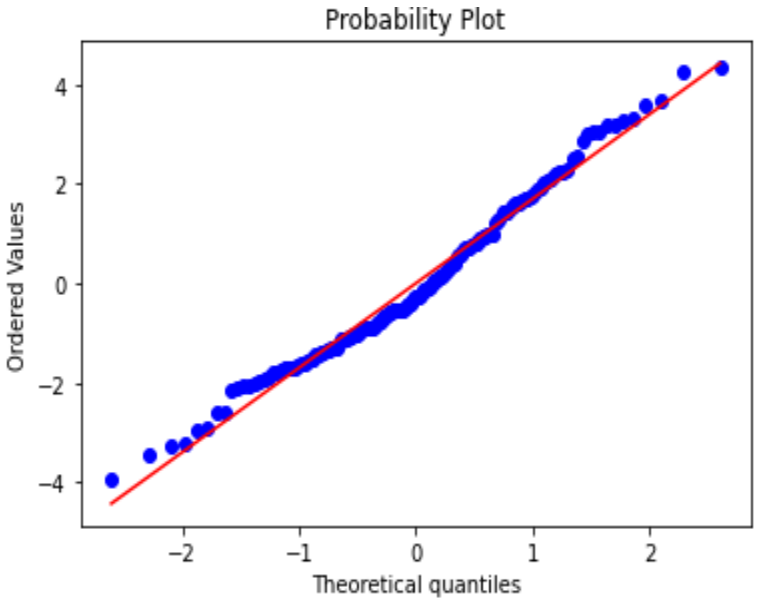
and approximately symmetric box plots (almost normal) of residuals.



In our case the residual boxplot is as below, indicating perfect normality. **Boxplot also showed that there are no outliers**

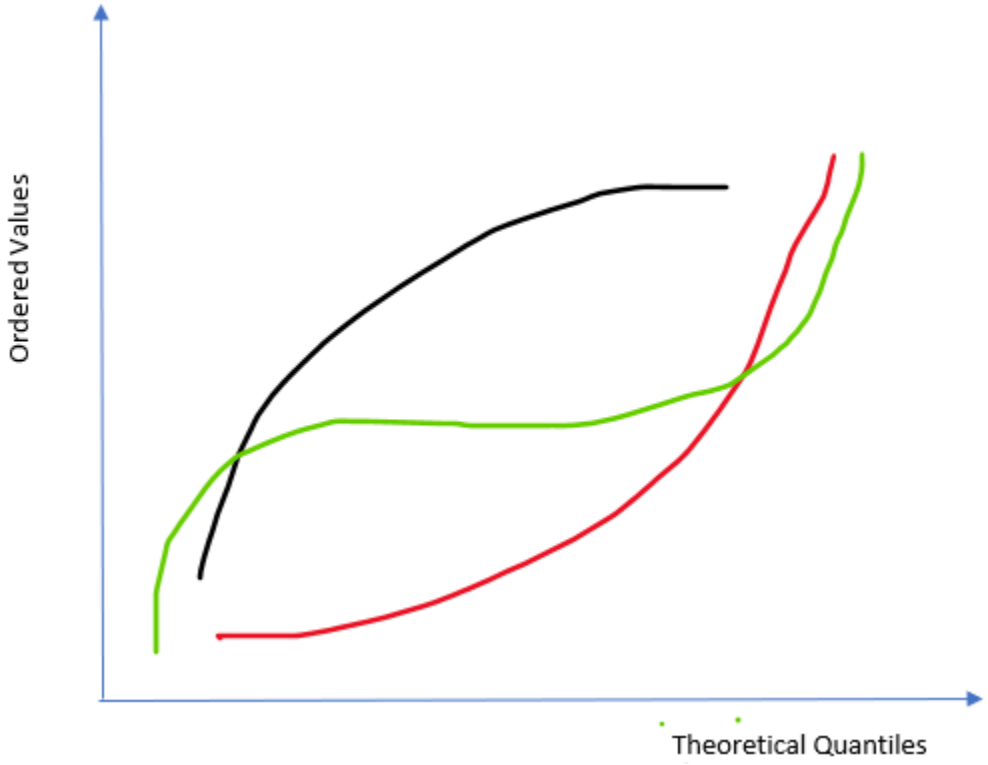


One can check for a **normal probability plot** (expected value Vs. residuals) **often also called Q-Q plot**. Here we got plot as below which again confirmed normality:



For skewed distribution the plots may be as below:

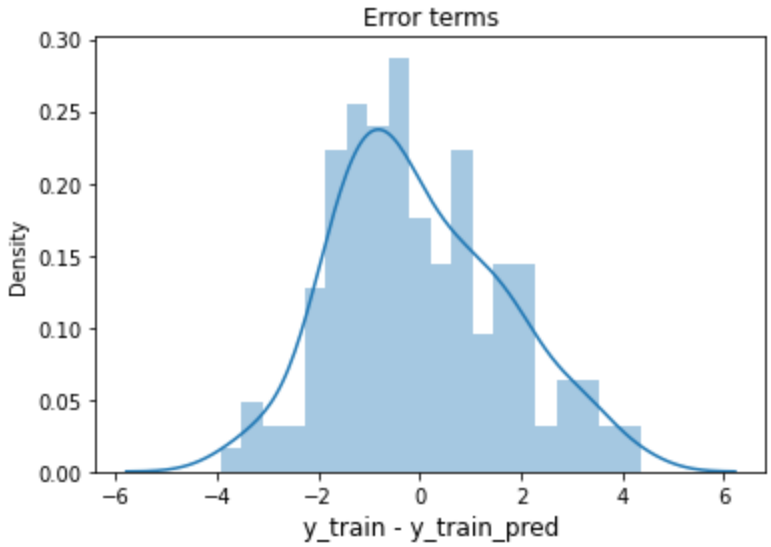
Red-skewed right, **black**-skewed left, **green**-heavy tail



In a right-skewed distribution, the mean is greater than the median because the unusually high scores distort it.

In a left-skewed distribution, **the mean is less than the median** because the unusually low scores distort it.

We also plotted a distplot (histogram showing the distribution of residuals). It showed nearly normal distribution as below:



One more thing that we need to see is are we missing any predictor apart from ‘Cost’ for our response variable ‘Score’. That discussion leads to **Multilinear regression**. This is continued in chapter MLR