MULTIPLE LINEAR REGRESSION (MLR)

In most business problems the response variable is not just a function of a single predictor variable, but multiple predictor variables **'may'** affect it. E.g., The Sales of a company may depend upon marketing spend, festive season, general economic condition, competitor launching a new product and many other factors. Even in the 'Soccer dataset' we considered in SLR, there were many other variables which may affect the 'Score' of a player apart from the most correlated variable 'Cost' that we considered. To account for many (say k) predictor variables (and say N observations) we must use **multiple linear regression(MLR)** which has equation that gives the predicted value of i[th] response variable Y_i as below:

$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \in_i$ **OR**

$$Y_i = \beta_0 + \sum_{j=1,i=1}^{k,N} \beta_j X_i + \in_i$$

Again $\in_i \sim N(0, \sigma^2)$ **i.e. error terms are normally distributed with mean 0 and variance** σ^2**.**

For 3-D case the equation reduces to:

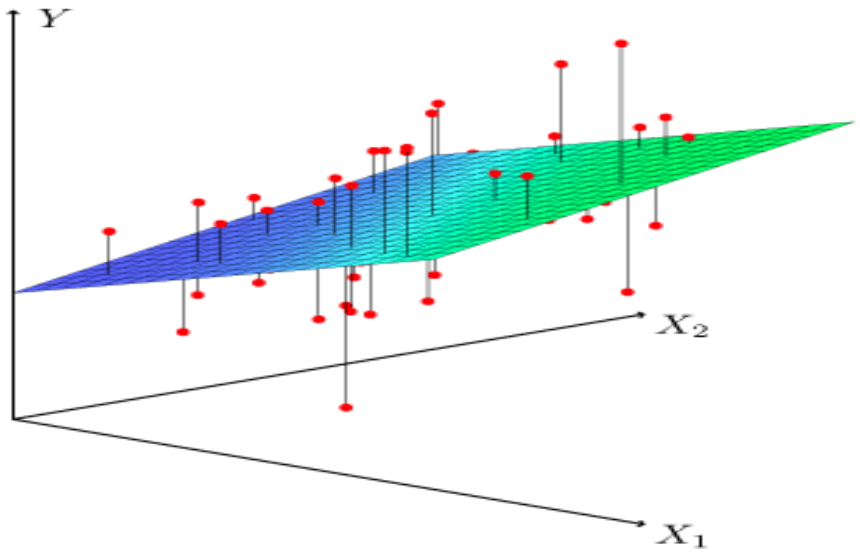
$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \in_i$ **and** expected value of Y:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + 0(mean\ of\ error - terms)$$

Which is the equation of a plane.

Also expected value of Y in this 3D

And the task of MLR is to find the **hyperplane** that is closest to all data-points in a **K-dimensional** space. *A hyperplane is a subset of space having one less dimension than K(no of predictor variables in MLR)*. It is not possible to draw such a hyperplane on paper but at least for a 3-D space the diagram below gives some idea of this hyperplane. Here **Y** is the response variable and $\mathbf{X_1}$ and $\mathbf{X_2}$ are the independent variables. Red dots are data points in 3D and the blue-green plane is the '**plane of closest fit**'. Diagram also shows residuals.



To find such a hyperplane, in K dimensional space we need to estimate values of $\beta_0, \beta_1, \ldots\ldots, \beta_k$. Which can be found from the use of matrix algebra and the method remains the same as SLR **i.e. minimizing the sum of squared error(SSE)**. Again we will just discuss the intuition behind the solution. The equation for MLR is: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_N + \in$

OR $Y_i = \beta_0 + \sum_{j=1,i=1}^{k,N} [\beta_j X_{i,j}] + \in_i$

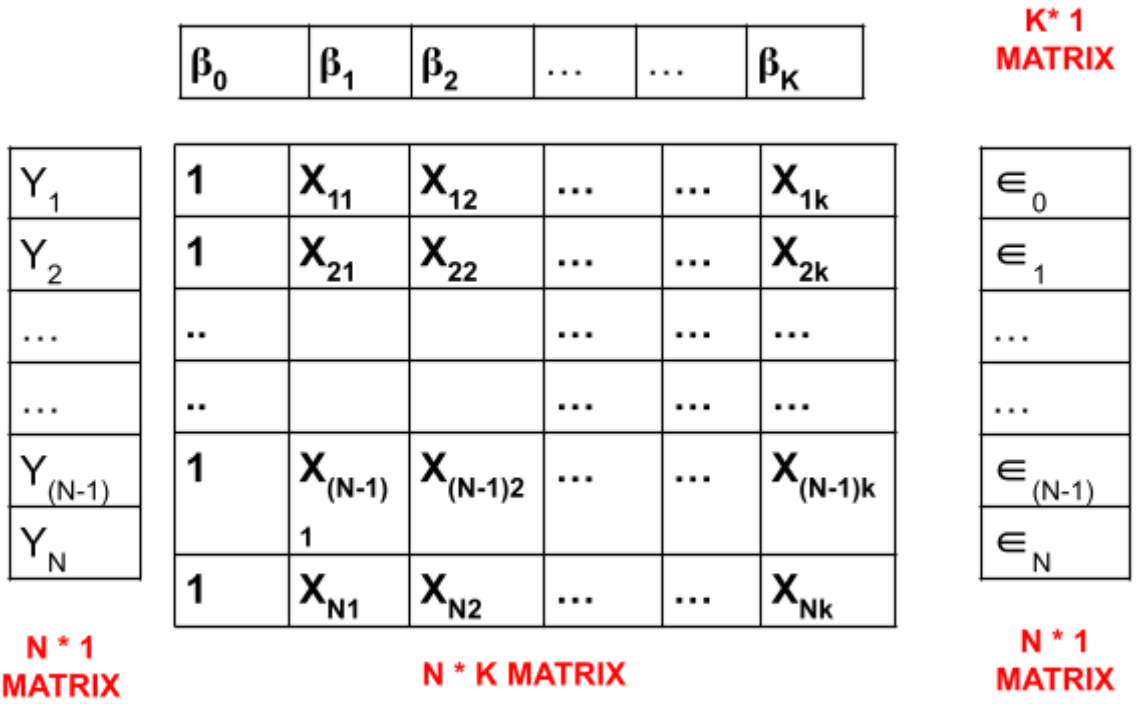
The above equation may look like the equation of SLR, but actually the meaning of each term in it is now different.

- The dependent variable Y is now a vector, Y, with a row for every observation: **(Y_0, Y_1, ..., Y_N)**
- The independent variables have been combined into a feature-matrix, X, with a column for each feature plus an additional column of '1' values for the intercept term.
- The regression coefficients (betas) make another vector **(β_0, β_1, ..., β_k), one member of the vector for each feature and one for intercept.**
- And error-term $\in$ also make a new vector.**($\in_0$, $\in_1$, ..., $\in_N$)**. **One Error term for each observation**.

- **<u>The * sign indicates matrix/vector multiplication.</u>**

This matrix multiplication is shown in the diagram below in an easy to remember format.

$$Y = \beta_0 * 1 + \beta_1 * X_1 + \beta_2 * X_2 + \ldots\ldots + \beta_K * X_K + \in_N$$



$$Y_i = \beta_0 + \sum_{j=1,i=1}^{k,N} [\beta_j X_{i,j}] + \in_i$$

Using equation, $Y_i = \beta_0 + \sum_{j=1,i=1}^{k,N} [\beta_j X_{i,j}] + \in_i$, We can find all coefficients, i.e. vector β_K **= (β_0, β_1, ..., β_k).** Again

the goal is to minimize the $SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$.

$$Y_i = \beta_0 + \sum_{j=1,i=1}^{k,N} [\beta_j X_{i,j}] + \epsilon_i$$

$$Y = X\beta + \epsilon$$

(N*1) = (N * K) * (K *1) + (N * 1)

Expected value of Y is: $E(Y) = X\beta$

Also, Covariance matrix is given by:

$$\sigma^2(\epsilon) = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & \dots & 0 \\ 0 & 0 & \sigma^2 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \\ 0 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

It can be proved by matrix operations that predicted values of beta-coefficients (i.e. the **K * 1** matrix above) $\widehat{\beta_k}$ which satisfy above equation is given by:

$$[X^T X]\,\widehat{\beta_k} = [X^T Y]$$

$$\widehat{\beta_k} = \left(X^T X\right)^{-1}\left(X^T Y\right)$$

The above solution uses a pair of matrix operations. X^T **stands for "transpose" of feature matrix X, and the -1 in** $\left(X^T X\right)^{-1}$ **indicates an "inverse" matrix.**

Again**,** most software packages like Python and R have readymade methods to calculate β_0, β_1, ..., β_k **hence one mostly need not bother even about the above short summary of the mathematical approach to find the solution.** Still, if someone is interested in derivation of equation $\widehat{\beta_k} = \left(X^T X\right)^{-1}\left(X^T Y\right)$ then one can follow the link below. **It also introduces the basics of matrices which is a must for any data-scientist.**

https://bookdown.org/ripberjt/qrmbook/introduction-to-multiple-regression.html

THREE TYPES OF PREDICTOR VARIABLES

1. Qualitative Predictor Variables (E.g. Male/Female, Open/Close, Yes/No)
2. Polynomial Terms E.g. No ice-cream cones sold may depend upon temp or $temp^2$, $temp^3$ etc.
3. Interaction effects E.g. no of ice cream cones sold depends on product of temperature and foot traffic: (t *n)

ANOVA TABLE FOR MLR

Source Of Variation	SS	df	MS
Sum Of Square Regression	**SSR** $= \boldsymbol{\beta}^T X^T \mathbf{Y} - \frac{1}{N}(Y^T J \mathbf{Y})$	**K** (NO OF PREDICTORS)	$MSR = \frac{SSR}{K}$
Sum Of Square Error	**SSE** $= Y^T \mathbf{Y} - \boldsymbol{\beta}^T X^T \mathbf{Y}$	**N-1-K**	$MSE = \frac{SSE}{N-1-K}$
Sum Of Square Total	**SSTO** $= Y^T \mathbf{Y} - \frac{1}{N}(Y^T J \mathbf{Y})$	**N-1**	**----------**
IN ALL ABOVE J IS AN (N * N) MATRIX WHOSE ALL ELEMENTS ARE 1 ONLY			

As we did in SLR, the importance of getting MSR and MSE is that from it we can do the F-test and obtain the coefficient of Multiple determination **R^2**.

F-TEST:

$$F^* = \frac{MSR}{MSE} = \frac{\frac{SSR}{K}}{\frac{SSE}{N-1-K}}$$

Null Hypothesis H_0: $\beta_1 = \beta_2 = \ldots\ldots\ldots\ldots = \beta_k = 0$

I.e., there is no, linear information, in any of the predictor variables $X_1, X_2, \ldots\ldots\ldots\ldots, X_k$.

Alternative Hypothesis H_1: At least one of $\beta_1, \beta_2, \ldots\ldots\ldots\ldots, \beta_k$ is non-zero. i.e., at least one of the predictor variables contains some linear information about the response variable.

Let's define $F_{critical} = F_{(1-\alpha;\, k;\, N-1-k)}$, **which is usually obtained from** F-Distribution table as we saw in chapter on SLR, then if:

1. $F_{critical} \leq F_{(1-\alpha;\, k;\, N-1-k)}$, then we fail to reject (accept) null hypothesis. So, <u>there is no, linear information, in any of the predictor variables</u> $X_1, X_2, \ldots\ldots\ldots\ldots, X_k$. $\beta_1 = \beta_2 = \ldots\ldots\ldots\ldots = \beta_k = 0$.

2. $F_{critical} > F_{(1-\alpha;\, k;\, N-1-k)}$, then we reject the null hypothesis. i.e., So, at least one of $\beta_1 , \beta_2 , \ldots\ldots\ldots\ldots , \beta_k$ is non-zero. <u>i.e., at least one of the predictor variables contains some linear information about the response variable.</u>

However, one issue arises here that we can only know that at least one of the predictor variables is meaningful **but the question is which one**? Should we include all variables in our model? That will lead to another issue called multicollinearity. Actually, we always use the least no: of predictor variables in a model which can give a fair amount of predictive power.

R-squared and Adjusted R-squared

Here we need concept of coefficient of Multiple determination $\mathbf{R^2}$

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}\ ,\ R^2 \epsilon\ [0, 1]$$

R^2 is a measure of how much variance in unknown-data can be explained by predictor variables we included in our model**.**

We can plot pair-plots (**scatter plots of response variable Vs each predictor variable**) and if the scatter plot shows trend for a predictor **X_1** (fig-1) below, we can include that predictor. But if it shows a nearly horizontal line of fit passing through average values, for a predictor, we can discard the predictor (fig-2) **X_2**.

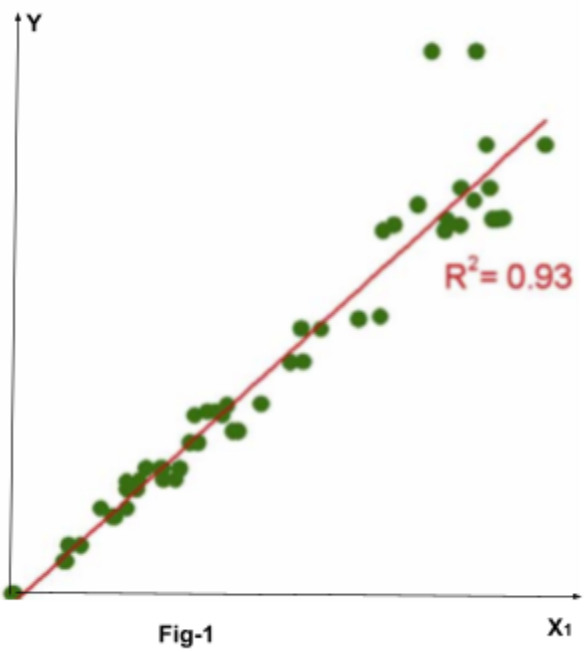


Fig-1

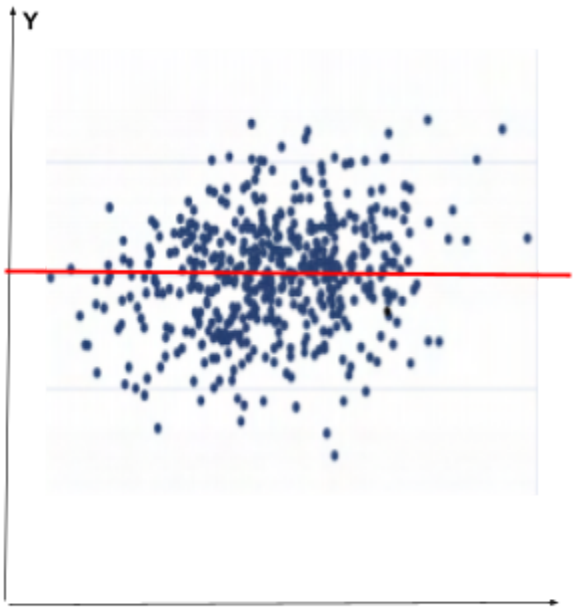


Fig-2

Now, in SLR, as there was only one predictor variable, adjusted-**R^2** does not have much role but in

MLR, suppose there are 5 predictor variables in a model and suppose the dataset has 10-15 observations only then we are using a lot of degrees of freedom and we need a term that penalizes us for that. In such a case we must use adjusted-$\mathbf{R^2}$ rather than $\mathbf{R^2}$. It is called adjusted $\mathbf{R^2}$ as it is adjusted for degree of freedom.

$$R_a^{\ 2} = 1 - \left(\frac{N-1}{N-1-K}\right)\frac{SSE}{SSTO}$$

For a long data-set where N>>>K usually the **adjusted- $\mathbf{R^2}$ i**s not needed**.**

We must also know two terms **AIC** and **BIC** seen on the output of **ANOVA** regression**.**

AIC stands for (*Akaike's Information Criteria*), a metric developed by the Japanese Statistician, Hirotugu Akaike, 1970. **The basic idea of AIC is to penalize the inclusion of additional variables to a model.** It adds a penalty that increases the error when including additional terms**. The lower the AIC, the better the model.**

BIC (or *Bayesian information criteria***)** is a variant of AIC with a stronger penalty for including additional variables to the model**. The lower the BIC, the better the model.**

t-TEST IF INDIVIDUAL COEFFICIENT IS ZERO

Null Hypothesis H_0: $\beta_j = 0$

Alternative Hypothesis H_A: $\beta_j \neq 0$

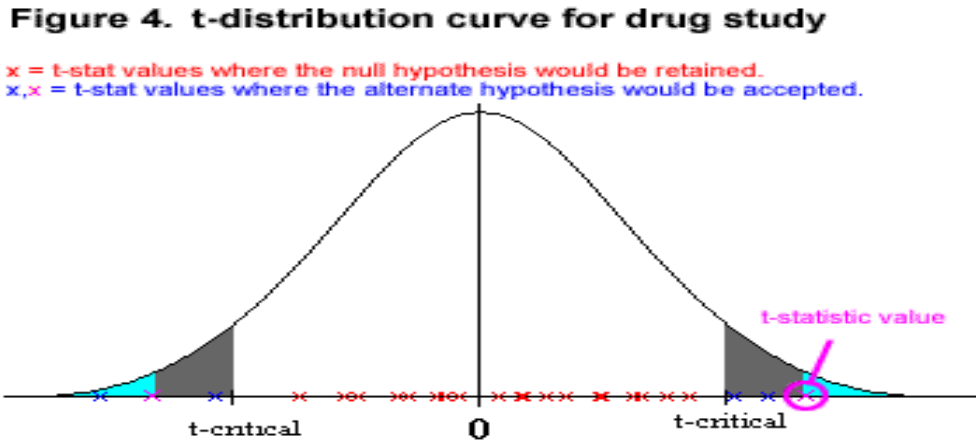
We first define a quantity $\mathbf{t_{critical} = t_{(1-\frac{\alpha}{2},\ N-1-K)}}$ **(found from t-distribution table)**

In t-test we first find,

$$t^* = \frac{\beta_j}{S(\beta_j)} = \frac{Coefficient-estimate}{Standard\ Error\ Of\ Coefficient\ Estimate}$$ and then

1. If $t^* > t_{critical}$ meaning that it is beyond it on the x-axis (**a blue x**), then the null hypothesis is rejected and the alternate hypothesis is accepted.
2. If the t-statistic had been less than the t-critical value (**a red x**), the null hypothesis would have been retained

Figure 4. t-distribution curve for drug study



NOTE: 1

Note that we can also use the T Score to P Value Calculator to calculate corresponding p-value and then use some threshold (say 0.05) and say that **if p<0.05 then reject the null hypothesis** $\beta_j = 0$, and **if p>0.05 then we accept the null hypothesis** $\beta_j = 0$.

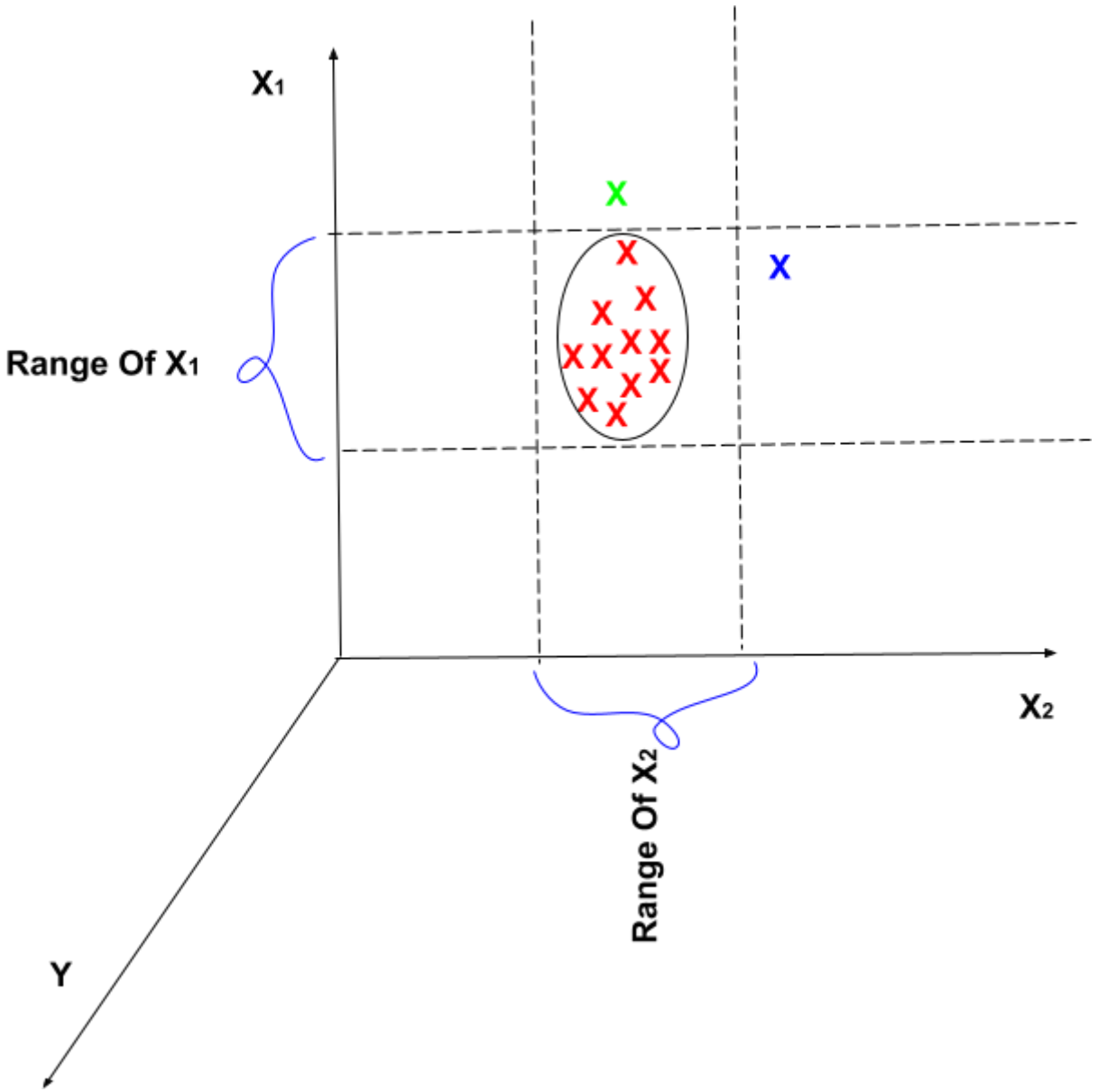
NOTE: 2

One more point is that actually in using t-test for coefficient of jth individual predictor i.e., β_j We are ignoring the effect of all other predictor variables which means we are assuming that all predictor variables are independent, **but this assumption is actually not true**.

NOTE:3

We also have to guard against hidden extrapolation. Let's consider just a simple case of two predictor variables X_1 and X_2. Red coloured data-points are actual data-points. Suppose we are making a prediction for the blue coloured data point, then, even though it is within the range of $\mathbf{X_1}$, it is beyond the range of $\mathbf{X_2}$. For the green point, it is the

vice-versa. Thus, we are doing 'hidden extrapolation' here. **As no: of predictors increase to X_1, X_2, ……..,X_K deciding whether we are doing interpolation or extrapolation becomes increasingly difficult.**



With this much background, we once again consider the soccer data-set that can be downloaded from: **https://drive.google.com/file/d/1cLFGRApehOCmbhnvkHRmQ9DkWZZ6f4QX/view?usp=share_link**

The jupyter notebook in which all three MLR models described below are obtained can be downloaded at.

https://drive.google.com/file/d/1T7rvze9-StxkTe5EJsNyoOCd_tw5e15t/view?usp=share_link

The quick EDA done using Pandas Profiler can be downloaded from link below:

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

We have also done detailed EDA. Here there are 12 independent variables: 'PlayerName', 'Club', 'DistanceCovered(InKms)', 'Goals', 'MinutestoGoalRatio', 'ShotsPerGame', 'AgentCharges', 'BMI', 'Cost', ‘PreviousClubCost', 'Height', 'Weight' for response variable 'Score'.

We discard some independent variables outright:

1. PlayerName has clearly nothing to do with Score. Remove PlayerName.

2. Club is a categorical variable. We **initially** removed it. We can consider including it later.
3. Score and Weight have very weak correlation -0.00016. So, we removed ‘Weight’ from predictor variables.
4. Next we checked for multicollinearity and found that MinutestoGoalRatio and ShotsPerGame have 0.95 correlation. So we kept only ShotsPerGame. Similarly, PreviousClubcost and height have 0.8 correlation. So we kept only PreviousClubcost i.e. removed the ‘height’ column.

   So our first model, considering **only numerical predictors,** had:

   X = df[['DistanceCovered(InKms)', 'Goals',

   'ShotsPerGame', 'AgentCharges', 'BMI',

   'Cost','PreviousClubCost']]

   Y = df ['Score']

   This model gave R^2=0.959, adjusted R^2=0.957,Prob (F-statistic): 9.69e^{-96}, AIC: 521.8 and BIC: 546.0.

   We got our model:

   Score = - 0.6790 * DistanceCovered(InKms)

   + 0.0279 * Goals

- 0.1151 * ShotsPerGame

- 0.0023 * AgentCharges

+ 0.1841 * BMI

+ 0.1612 * Cost

- 0.0953 * PreviousClubCost

+ 9.205

5. We know that the best model is the one that uses the least number of predictors, without sacrificing much predictive power. So we Now we tried making a second model by removing 'Goals', 'ShotsPerGame', 'AgentCharges' as their correlation with 'Score' was respectively 0.108,-0.532, -0.183 only. Thus now,

   x = df['DistanceCovered(InKms)', 'BMI', 'Cost', 'PreviousClubCost']]

   y = df ['Score'].

   This model gave R^2=0.96, adjusted R^2=0.959,Prob (F-statistic): 1.14e^{-100}, AIC: 506.7 and BIC: 521.8. Thus by reducing predictors, we have actually got a better model.

   Score = - 0.5930 * DistanceCovered(InKms)

   + 0.0780 * BMI

+ 0.1661 * Cost

- 0.0953 * PreviousClubCost

+ 8.1797

6. PlayerName and Club are categorical features. PlayerName is just a random variable but 'Club' **may** affect score. So obtain dummy variables from 'Club' and then include it in our model.

   This model gives **R^2= 0.966, adjusted R^2=0. 0.965,Prob (F-statistic): 1.79e^{-100}, AIC: 489.2 and BIC: 510.3.** This makes it the best model of the three models considered. Thus we can use our final model as:

   Score = - 0.3125 * DistanceCovered(InKms)

   + 0.2677 * BMI

   + 0.1459 * Cost

   - 0.0984* PreviousClubCost

   + 0.9915 * CHE

   + 0.3328 * LIV

   + 2.5777 * MUN

   + 3.9020

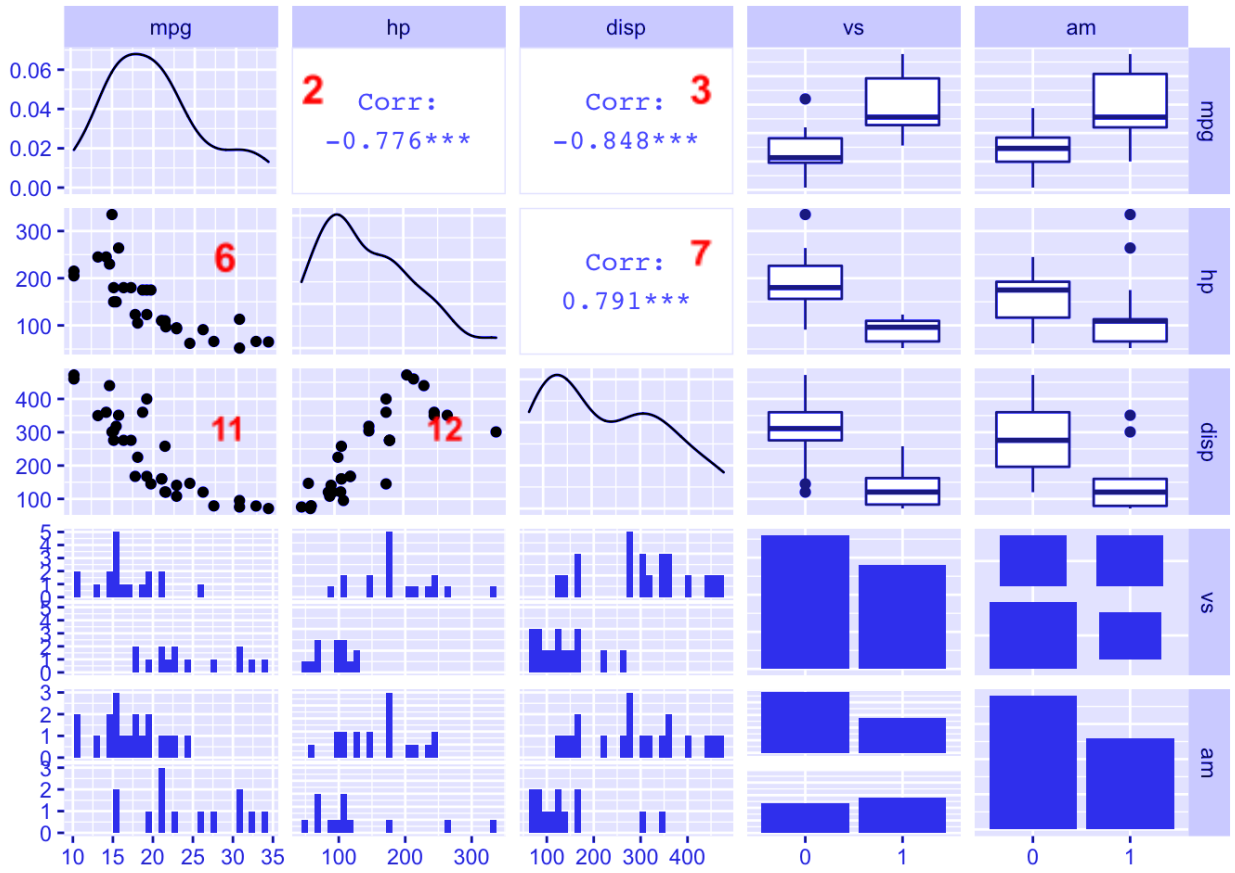
We also note down the ranges in which each predictor variable is present in the dataset**.** This can help us avoid extrapolation and give the idea that within which range the model gives accurate results.

Predictor	Min Value	Max Value
DistanceCovered (InKms)	**3.80**	**6.72**
BMI	**16.75**	**34.42**
Cost	**28.00**	**200.80**
PreviousClubCost	**34.360**	**106.00**

MLR: Diagnostics, Remedial measures & Multicollinearity

I. Scatter-plot matrix (pair-plot)

Let's consider the well-known example of predictor variables for response ‘mileage’ of a vehicle. Then the pair-plot as below shows scatter plots of various pairs of variables:



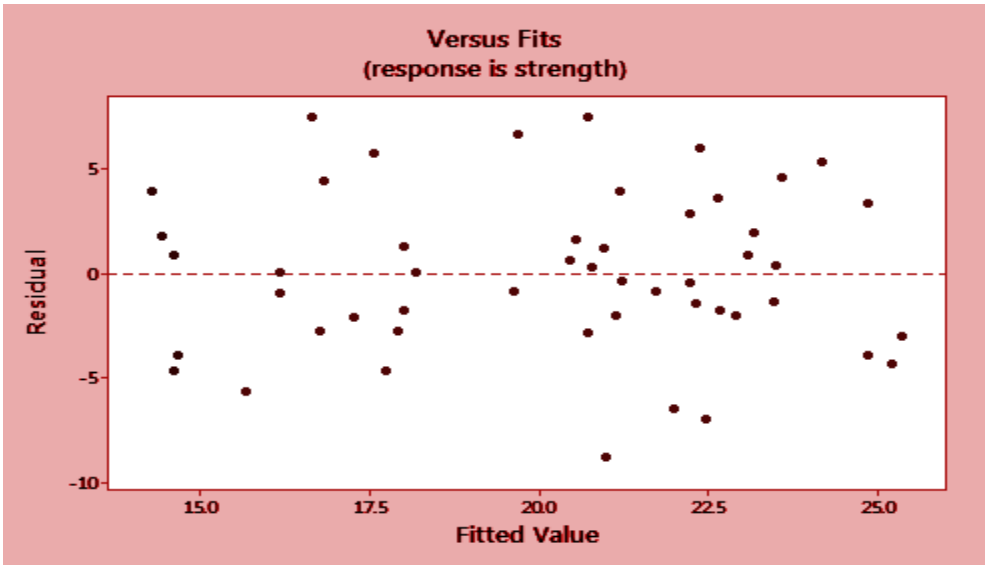
Pair Plot No	Y Vs X_i	Nature Of Pair Plot
3,11	mpg Vs engine displacement	Clear Negative nearly linear correlation -0.848
2,6	mpg Vs horsepower (hp)	Clear Negative nearly linear correlation -0.776
Study mutual correlation of predictors for multicollinearity		
7,12	Displacement vs horsepower	Clear positive correlation 0.791

So, such pair-plots can help identify both:

1. Which predictor variables have high correlation with the response variable. They are candidates for including in the MLR model
2. Which predictor variables have high correlation among themselves. One from each such pair should be potentially removed to reduce multicollinearity and then R^2, AIC and BIC should be checked.

II. Residual Plot Analysis

One should also check the residual Vs fitted value plot and it should be without any pattern (see fig. below)



This plot is a classical example of a well-behaved residuals vs. fits plot. Here are the characteristics of a well-behaved residual vs. fits plot and what they suggest

about the appropriateness of the simple linear regression model:

1. The residuals "bounce randomly" around the 0 line. This suggests that the assumption that the relationship is linear is reasonable.
2. The residuals roughly form a "horizontal band" around the 0 line. This suggests that the variances of the error terms are equal.
3. No one residual "stands out" from the basic random pattern of residuals. This suggests that there are no outliers.

Caution: While above analysis is useful one should not overinterpret the residual Vs predicted value graph. It should be seen with other indicators. One would especially want to be careful about putting too much weight on residual vs. fits plots based on small data sets. Sometimes the data sets are just too small to make interpretation of a residuals vs. fits plot worthwhile. For detailed discussion on residual analysis one can see: https://online.stat.psu.edu/stat462/node/116/

One can also plot residual Vs Single Predictor (Say $\mathbf{X_1}$) and then residual Vs another predictor (Say $\mathbf{X_2}$). If these individual plots show some increasing/decreasing but roughly linear pattern then it indicates that with $\mathbf{X_1}$ **and** $\mathbf{X_2}$

probably some more work is to be done(may be introducing higher order residuals etc.).

III. Interaction Effects in residual plots

With two predictor variables $\mathbf{X_1}$ **and** $\mathbf{X_2}$**,** we also need to plot residual Vs ($\mathbf{X_1}$* $\mathbf{X_2}$**)** plot and with three predictor variables $\mathbf{X_1}$ **,** $\mathbf{X_2}$ **and** $\mathbf{X_3}$**,** we need to check more interactions,

$\mathbf{X_1}$* $\mathbf{X_2}$ **,** $\mathbf{X_2}$* $\mathbf{X_3}$**,** $\mathbf{X_1}$* $\mathbf{X_3}$ etc. and add one column in our dataset for each product and then plot residual Vs each product (now as a single variable) plot to do residual analysis as above.

IV Multicollinearity Problem

$$BMI = \frac{Weight\ in\ Kilogram}{(Height\ in\ meter)^2}$$

Now suppose a healthcare-prediction model is being made and the predictor variables include BMI, height(m) and weight(kg) all three. Then these 3 are actually correlated variables. Including such variables is not desirable due to reasons like:

1. **Numerical Stability:** If a model contains many correlated predictors then giving a small perturbation (say adding just one/two observations to the original 100 observation data set) produces a big change in estimates of coefficients.

2. **Partial coefficients can not be interpreted correctly:**

We interpret the coefficients of MLR partially (say keeping all others constant, what if we change variable whose coefficient is β_j). Clearly such discussion becomes meaningless if the variable whose coefficient is β_j is correlated with the variable whose coefficient is β_k as 'keeping all others constant' is impossible.

Having said this, multicollinearity is actually very common in MLR models so one must use the pair-plots, correlation matrix, residual analysis etc. to try to make a best performing model with the least number of predictors. Later we will use two more techniques Lasso and Ridge regression. The Ridge regression is particularly great to solve multicollinearity problems.

In the discussion on multicollinearity, we used the phrase that "The best model is the one that uses the least number of predictors, without sacrificing much predictive power". However, in the case of models with higher order terms(polynomial regression), this rule should not be blindly applied. Suppose we have a Y Vs X_i graph as a parabolic/cubic or even higher power graph. E.g.

If we have assumed that equation like below fits our data:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X^2_2 + \ldots\ldots + \beta_7 X^7_7$$

Then the best way to tackle this problem is Assuming that $X^2_2 = X_8$, $X^3_3 = X_9$, $X^4_4 = X_{10}$, $X^5_5 = X_{11}$, $X^6_6 = X_{12}$, $X^7_7 = X_{13}$ and then rewrite above equation as:

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_8 + \ldots\ldots + \beta_7 X_{13}$ and treat it as a simple MLR**.**

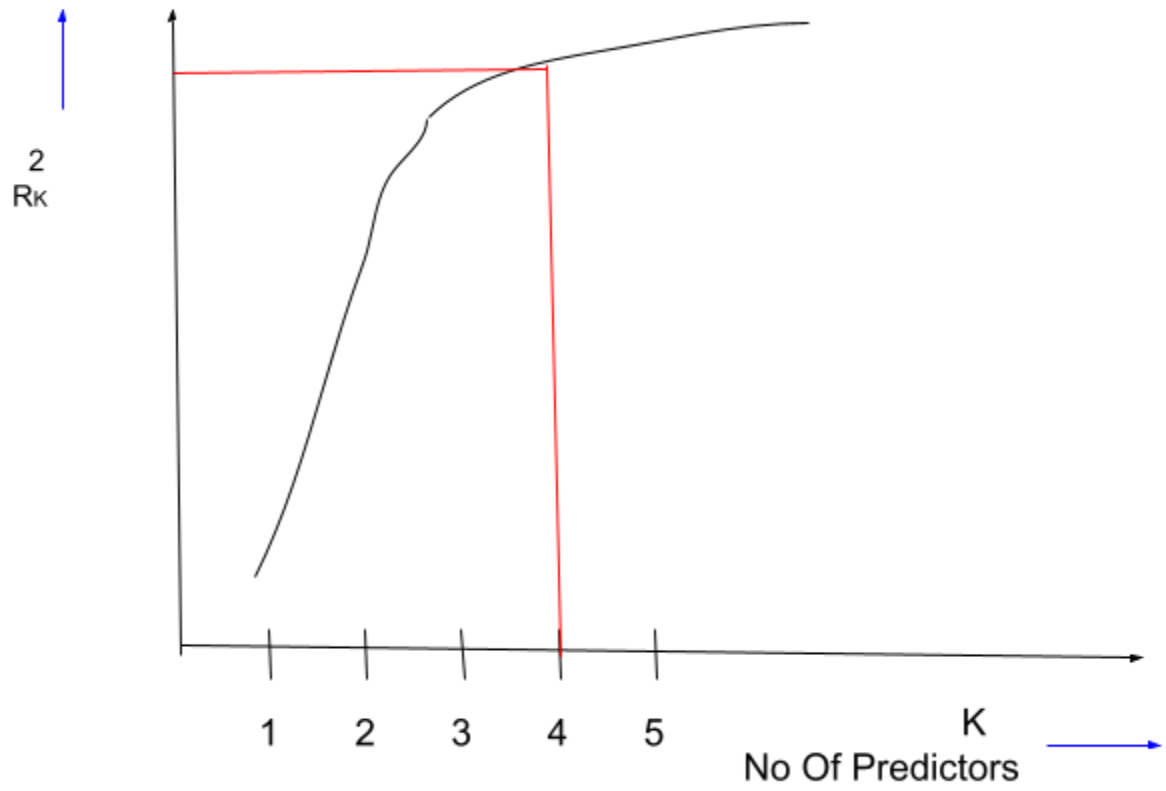
Also if we are including $X^7_7 = X_{13}$ In our model then the best practice is to include **<u>all</u>** lower order terms also irrespective of considerations for multicollinearity.

<u>MODEL SELECTION CRITERION</u>

If there are 'K' predictor variables then (even ignoring interaction effects and higher order polynomials etc.) we can still have **2^K potential regression models.**So what is the criteria for selecting the best model? There are following criteria:

1. R^2_K / SSE_K .

   Plot R^2_K Vs K **graph** and **look for elbow of the curve** at which point R^2_K is maximized.



E.g., in figure above there is not much advantage in increasing the number of predictors beyond 4 as improvement in R-squared would be marginal.

2. Mallows's C_k criteria

It essentially helps us to reduce the mean squared error and also helps us to reduce the bias as much as possible. According to this:

A. We want to have small C_K statistic which gives less MSE.
B. We also want to have a C_K near K which reduces bias.

One can refer to the details of Mallows's C_k criteria here. Mallows's C_k criteria is equivalent to Akaike information criterion(AIC) in the special case of Gaussian linear regression.

Here we try multiple values of C_K and try to find the best value which gives the best possible trade off between reducing RMSE and reducing bias.

3. **AIC$_K$ and BIC/SBC$_K$ criteria**.

   Models having small SSE also do better on these two criteria. At the same time, these two criteria penalize for including many predictor variables. These two parameters are used to Identify the best subset of predictor variables to be included in the model.

4. **Press$_K$ criteria**:

   The **prediction sum of squares** (or **PRESS**) is a model validation method used to assess a model's predictive ability that can also be used to compare regression models. For a data set of size *n*, PRESS is calculated by omitting each observation individually and then the remaining *n* - 1 observations are used to

calculate a regression equation which is used to predict the value of the **omitted response value** denoted by $\hat{y}_{i(i)}$. We then calculate the i[th] PRESS residual as the difference $y_i - \hat{y}_{i(i)}$. Then, the formula for PRESS is given by:

$$PRESS = \sum_{I=1}^{n} y_i - \hat{y}_{i(i)}$$

It is a measure of how well the fitted values predict the actual responses. **Thus models with smallest value of Press $_K$ have the smallest prediction error**.

PRESS can also be used to calculate the **predicted R^2** (denoted by R^2pred) which is generally more intuitive to interpret than PRESS itself. It is defined as:

$$R^2_{predict} = 1 - \frac{PRESS}{SSTO}$$

It is a helpful way to validate the predictive ability of model without selecting another sample or splitting the data into training and validation sets.

Together, PRESS and $R^2_{predict}$ **can help prevent overfitting** because both are calculated using observations not included in the model estimation. Overfitting refers to models that appear to provide a good fit for data set at hand, but fail to provide valid predictions for new observations.

For model selection, sometimes the approach used is: out of $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$, …..,$\mathbf{X_7}$, removing one predictor and making model of other predictors, and repeating same process iteratively. But it is a very tedious method. Instead, using one of four methods listed above is recommended. One can also use algorithms that help us to identify best subsets of variables **in an automated manner.** Some such algorithms are stepwise elimination, Forward Select, backward eliminate etc.. These are handy as just for with 10 predictors we will have $\mathbf{2^{10}=1054}$ possible models. We can not try each and every model.