**Multiple Linear Regression**

* Formula - y = b0 + b1\*x1 + b2\*x2 + …… + bn\*xn
* It is the same as simple linear regression, but with many variables.
* b0  one constant, and then many pairs of b and x.
* Here as well we have a dependent variable y, along with a few causations for it.
* If we take the example of years of work experience and salary, for multiple linear regression it could be like – how much work experience you have, years of education, how much profit you bring to the company and so on.
* And we also have the coefficient b, same as simple linear regression.

Caveat (Warnings) of Linear Regressions:

* Linearity – The relationship between X and the mean of Y is linear. The relationship between dependent and independent variable should be linear. It is also important to check for outliers, as linear regression models are sensitive to outliers. This assumption can be tested by scatter plots.
* Homoscedasticity – The variance of residuals is the same for any values of X.
* Multivariate Normality – For any fixed value of X, Y is normally distributed.
* Independence of Error – Observational errors are independent of each other. This is also called as Autocorrelation. And linear regression requires there is no autocorrelation in the data. Autocorrelation occurs when the residuals are not independent from each other.
* Lack of Multicollinearity – There should be little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are highly correlated to each other.

Before building a linear regression model we need to check if these assumptions are true, and only then we can proceed to build a model and be sure we are building a good linear regression model.

Dummy Variables –

* Let’s say we have a couple of numerical variables, and maybe one column for categorical variables. To create a regression model, we will need to change the categorical variable into a numerical variable, and that’s where the concept of dummy variable comes in.
* For instance, we have a “State” column in our data, which has categorical variables like “California” and “New York”. And these variables are important for the regression model. So, we simple create a dummy variable of the state column.
* Dummy variable is nothing but creating columns for every category in the categorical variable column.
* In the above-mentioned example, we shall create separate columns for California and New York. And if the row in the State column says “California”, we will simply set the California column as 1 and New York Column as 0.
* And while implementing the regression model we will only consider the California column, and not include the State column, nor the New York column.
* And the formula for this would be - y = b0 + b1 \* x1 + b2 \* x2 + b3 \* D1, where D1 is nothing but the dummy variable.
* We might think that if we do not include California for our regression model, then the equation for CA would result into 0, as there would be no coefficient for California. Because the only way we include CA in our regression model is by setting NY to 0.
* But that’s not the case, because the way regression model works is that they’ll make the dummy variable not included as default equation for the regression model.
* So, what it means is that, by default the coefficient for CA is included in constant (b0), and if D1 is set to 0, the equation for California would be y = b0 + b1 \* x1 + b2 \* x2.

Dummy Variable Trap –

In the example above, we omitted the California dummy variable. Now why is that? What will happen if we include second dummy variable as well?

* The intuition here is, that we are basically duplicating a variable here

y = b0 + b1 \* x1 + b2 \* x2 + b3 \* D1 + b4 \* D2.

* This is because D2 = 1 – D1.
* The phenomena where one or several independent variables in a linear regression predict another is called multicollinearity.
* As a result of this effect the model cannot distinguish between the effects of D1 from the effect of D2 and thus it won’t function properly.
* And this called the dummy variable trap.
* If we do the math, we can see that we cannot have b0, b3 and b4 in the equation at the same time, the constant and all the dummy variables (including the one we are supposed to exclude).
* So, in conclusion when building a model with multiple dummy variables, always exclude one dummy variable from the equation.

**Building a Model –**

* While building a model we may have a lot of independent variables that we could be potential predictors for the dependent variable. And we need to decide which independent variables we want to keep and which ones we should throw out.
* There are two main reasons why we need to throw out the independent variable columns:

1. Garbage in - garbage out – If we throw in a lot of stuff into our model, our model won’t be a good model, it won’t be reliable, it won’t be doing what it is supposed to do.
2. At the end of the day, we will have to explain these variables and understand not just the math behind it, but what it means that certain variables predict the behavior of our dependent variable.

* So, we only want to keep the variables that are useful and predict something.

**5 Methods of Building Models:**

1. **All-in** –

Basically, what it means is that just throw in all the variables. We can do that if we have prior knowledge of the data/variables. If you know that these exact variables are the ones that are true predictors. We might know that from domain knowledge, we might have done this model before, or somebody just gave us the variables and asked to just build the model.

The other reason is to use this method is – you must.

And the last reason why we would use this method is to prepare for backward elimination.

1. **Backward Elimination** –

Step 1 – Select a significance level to stay in the model (e.g., SL = 0.05).

Step 2 – Fit the full model with all possible predictors. It will be the all-in approach.

Step 3 – Consider the predictor with highest p-value. If p-value > SL, go to step 4 for FINISH.

Step 4 – Remove the predictor with the highest p-value.

Step 5 – Fit the model without this variable\* (If we just remove the variable, we can’t say that we have got a new model. We must rebuild the model without those variables.)

And after step 5, we go back to step 3 and repeat the process until we get to the point where the highest p-value is still less than the defined SL.

1. **Forward Selection –**

Step 1 – Select a significance level to enter the model (e.g., SL = 0.05).

Step 2 – Fit all possible simple regression models y ~ xn. Select the one with the lowest P-Value. We take the dependent variable and create a regression model with every single independent variable that we have.

Step 3 – Keep this variable and fit all possible models with one extra predictor added to the one(s) you already have. It means that, we have selected simple linear regression with one variable, now we need to construct all possible linear regression with two variables, where one variable is the one that we have already selected.

Step 4 – Out of all this possible two (or multiple) variable linear regression, we consider the one with the lowest p-value. If P < SL, go to step 3, otherwise go to FINISH.

We keep repeating step 3 and 4 until the variables we add have a P > SL. The trick here is you keep not the current model, but the previous one, and that makes sense because we just added a variable with insignificant value.

1. **Bidirectional Elimination –**

Step 1 – Select a significance level to enter and to stay in the model. E.g., SLENTER = 0.05, SLSTAY = 0.05.

Step 2 – Perform the next step of Forward Selection (new variables must have: P < SLENTER to enter).

Step 3 – Perform ALL steps of Backward Elimination (old variables must have: P < SLSTAY to stay).

Step 4 – No new variables can enter, and no old variables can leave.

After step 4, our model is finished.

1. **All Possible Models (Score Comparison) –**

Step 1 – Select a criterion of goodness of fit (e.g., Akaike Criterion)

Step 2 – Construct all possible regression models: 2n – 1 possible combination (where n is the number of variables we have).

Step 3 – Select the one with the best criterion.

Note – Sometimes, methods 2, 3 and 4 are also referred to as **Stepwise Regression**. And sometimes just method 4 will be referred to as **Stepwise Regression**.