A brief primer on linear regression – Part I

clevertap.com/blog/a-brief-primer-on-linear-regression-part-i

Pushpa Makhija

May 26, 2016

Prediction has always been a curious topic in life due to a key attribute – the extreme human desire to know what is coming next.

Let's ponder over our thoughts to answer a simple question – "Where is prediction most relevant in your life today?"

Predictions are central to every aspect of our life, whether we realize it or not. During school days, it was predicting what we would love to do in the future to choose a career path, checking the weather today to determine how should I dress, evaluating inventory numbers for the next day, to less important predictions made daily during our interactions with other people – like doing time management and getting into classes for a student, to dining, socializing, etc.

So, what's a prediction?

A prediction or forecast, is a statement about the future. It's a guess, sometimes based on knowledge or experience, but not always.

Now, let's consider a popular and common use case of the speed of an object to understand how predictions play an important role in our real world; in shaping our lives in ways and instances that we aren't aware of at first, and thereby help us to make informed decisions.

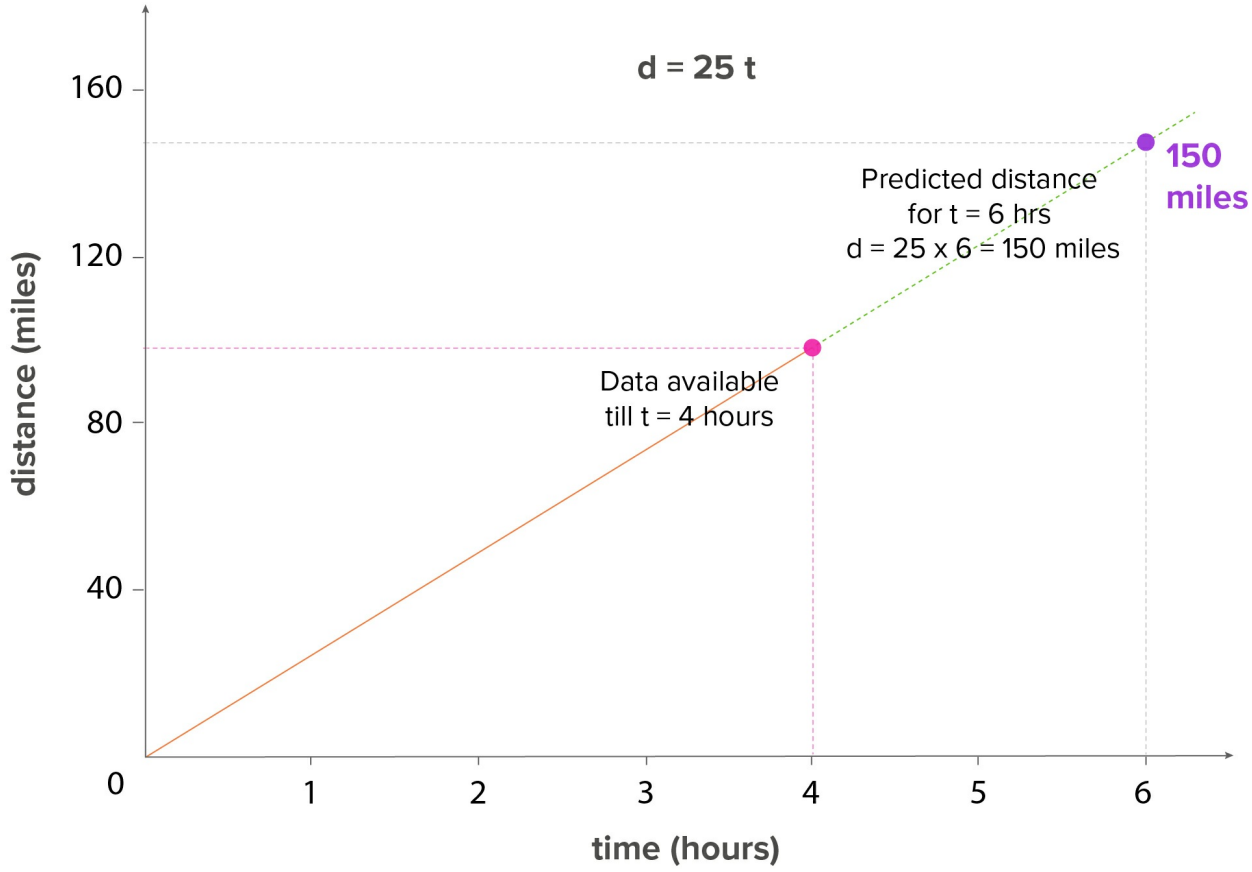
Have you ever thought of or answered day-to-day questions like –

- How fast vehicles such as cars and trains can go and how their speeds are calculated?
- When a police officer gives someone a speeding ticket, how does she know for sure if the person was speeding?

These questions force us to recollect our old learnings and refresh the concept of the physical system – the speed of an object is the magnitude of its velocity (the rate of change of its position) i.e. the speed of a certain object is calculated by dividing the distance travelled by the time taken to travel that distance.
At first glance, the above equation states that speed is a function of two quantities – distance and time. But, it really is a simple linear relationship called the rate formula because at least one of the 3 variables will always be a constant depending on the problem on hand.

$$Speed \sim f\left(distance, time\right) = \frac{distance}{time}$$

Graph showing relationship between distance and time when rate is a constant 25 miles per hour



The perfect linear relationship, as prevalent in the physical systems due to their inherent nature, are termed as Deterministic (or functional) relationships – comprising of an equation, that exactly describes the relationship between the two variables. Other examples could be

- Fahrenheit degree and Celsius degree (Fahr = 9/5 Cels + 32),
- Circumference (Circumference = pi x diameter) and
- Exchange rate conversion formula (new currency = (exchange rate) x (your currency))

Generally, we do come across scenarios depicting Statistical relationships – where the relationship between the two variables is not perfect, but there could be negative or positive relationship between the variables. Some examples could be

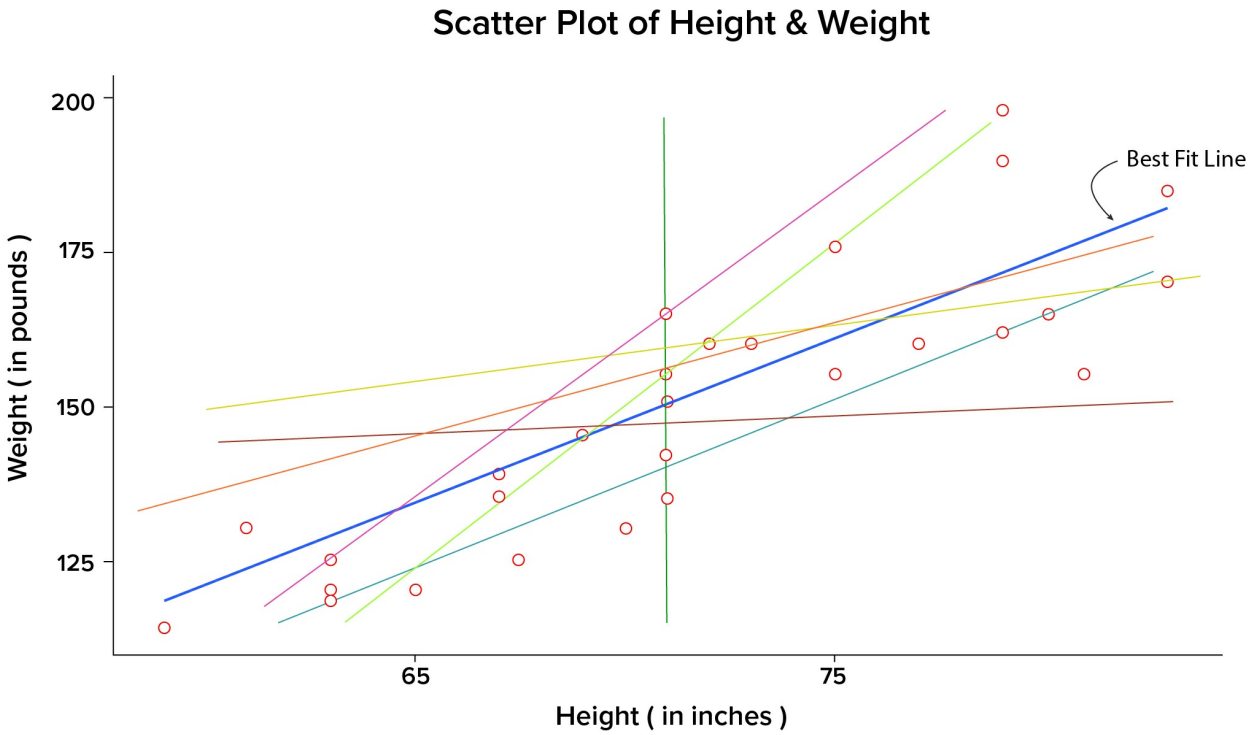
- Height and weight – as height increases, weight might increase but not perfectly
- Driving speed and gas mileage – as driving speed increases, gas mileage is expected to decrease, but not perfectly

These relationships, not of perfect kind nature, when graphed, gives a scatter plot of points, as seen from the plot of height-weight information of 30 adults as below:

Table-Plot

The above scatter plot reflects the relationship between height and weight as linear, depicting a positive increase in weight with height. We can thus fit a straight line to this data which would provide the best estimate of the observed trend.

The data points in the above scatter plot could be summarized in many ways as shown by various lines in the below plot:



Now the question arises – What is the best fit line that summarizes the relationship between height-weight, amongst all possible lines?

The best fit line is the one in blue color, and termed as *regression line,* which is actually the plot of the predicted score on y, for each possible value of x.

But, the next question comes – how to arrive at this best line?

The best line fitting the given data is obtained by "***minimizing the residual variation***" as below

where y_i is the actual observed value of response variable,

$$min \sum_i (y_i - \hat{y}_i)^2$$

$\hat{y}_i$ is the predicted value of response variable (as obtained from the model), and

$(y_i - \hat{y}_i)$ is the residual variation – the variation between the observed and predicted value of y

The closer the regression line comes to all the data points on the scatter plot, the better it is. This means that the minimum variation of points around the line results into low prediction error.

The best fit straight line to summarize the data, as described above, could be obtained by using a prediction method such as Simple Regression.

What is Simple Linear Regression?

Simple Linear Regression is a statistical technique that allows us to summarize and study relationships between two continuous i.e. numeric variables:

- The variable we are predicting is called the *criterion* or *response* or *dependent* variable, and
- The variable we are basing our predictions on is called the *predictor* or *explanatory* or *independent*

Simple linear regression gets its adjective ‘simple’, because it concerns the study of only one predictor variable.

For example, the height-weight information of 100 randomly selected people, aged between 20 and 60, can be quantified in terms of the equation or model, considering the response variable as weight and one predictor variable as height. Here, the inherent assumption, though quite unrealistic, is that “weight” can be measured by a single attribute – height. The model to fit this data could be written as

Weight (continuous)∼ Height (continuous)

In contrast, multiple linear regression, gets its adjective ‘multiple’, because it concerns the study of two or more predictor variables.

Extending our classic example of height-weight, we include other predictor variables, say, calorie intake, exercise level that would affect the person’s weight. The model to fit this data could be written as

Weight (continuous)∼ Height (continuous) + Calorie Intake (continuous) + Exercise Level (categorical)

A sample dataset of 10 rows pertaining to height-weight example along with other factors affecting the prediction is displayed below:

Both height and calorie intake individually are linearly related to weight as seen below in their scatter plots.

Ht-Wt Table2

Scatterplots

However, both height and calorie intake together may affect the weight of an individual linearly in a multi-dimensional cloud of data points, but not in the same manner as they affect alone, in the above scatter plots.

The general mathematical model for representing the linear relationships (termed as *regression equations*) can be written as:

SLR-MR

Here, for simple regression – b, the slope of the linear equation indicates the strength of impact of the variable, and a, the intercept of the line. And for multiple regression – b_i (i =1, 2, …,n), are the slopes or regression coefficients, indicates the strength of impact of the predictors, and a, is the intercept of the line.

The regression coefficient estimates the change in the response variable Y per unit increase in one of the x_i (i =1, 2, …,n) when all other predictors are held constant i.e. for our height-weight example, if x_1 differed by one unit, and both x_2 and x_3 are held constant, Y will differ by b_1 units, on an average.

The intercept or Y-intercept of the line, is the value you would predict for Y if all predictors are 0 i.e. when all $x_i = 0$ (i =1, 2, …,n). In some cases, the Y-intercept really has no meaningful interpretation, but it just helps to anchor the regression line in the right place.

Conclusion

In this part, we introduced simple linear regression model with one predictor variable and then extended it to the multiple linear regression model with at least two predictors.

A sound understanding of regression analysis, and modeling provides a solid foundation for analysts to gain deeper understanding of virtually every statistical and machine learning technique. Although regression analysis is not the fanciest learning technique, it is a dominant and widely used statistical technique to establish a relationship model between two or more variables.

In the ensuing part, we will delve into the steps and methodology to develop multiple linear regression model.

A brief primer on linear regression – Part II

clevertap.com/blog/a-brief-primer-on-linear-regression-part-ii

Data Science

Pushpa Makhija

In the first part, we had discussed that the main task for building a multiple linear regression model is to fit a straight line through a scatter plot of data points in multidimensional space, that best estimates the observed trend.

While building models to analyze the data, the foremost challenge is, the correct application of the techniques– how well analysts can apply the techniques to formulate appropriate statistical models to solve real problems.

Furthermore, before proceeding to analyze the data using multiple regression, part of the process encompasses to ensure that data you want to analyze, can actually be analyzed using multiple regression. Therefore, it is only appropriate to use multiple regression if you understand the key assumptions underlying regression analysis and check whether your data "passes" the required assumptions to give a valid result.

Usually, it's plausible for one or more of the assumptions being violated, while analyzing real-world data. Even when the data fails certain assumptions, there is often a solution to overcome this. First, let's look at the assumptions, and then learn how to check / validate the assumptions and also discuss about the proposed solutions for correcting these violations, if any.

We would be using IVs for independent variables and DV for dependent variable interchangeably while going through listing and validating assumptions, exploring data, building the model and interpreting the model output.

Assumptions of Regression:

Number of Cases/Sample Size

When conducting regression analysis, the cases-to-Independent Variables ratio should ideally be 20 cases for every independent variable in the model. For instance, the simplest case with two IVs – would require that n>40. However, for qualitative i.e. categorical variables with many levels of values, we might require more than ideal 20 cases for this variable to have sufficient data points for each level of categorical variable.

In this age of Big Data, we don't need to worry about dealing with small samples. But, this assumption violation does result in generalizability issue of not being able to apply the model's valuable insights and recommendations to other similar samples or situations.

Type of the Variables

The dependent variable should be measured on a continuous scale (i.e. an interval or ratio

variable). Examples include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg or pounds), and so on.

The two or more independent variables can be either continuous (i.e. an interval or ratio variable) or categorical (i.e. an ordinal or nominal variable).

- Examples of ordinal variables include Likert items – a 7-point scale from "strongly agree" to "strongly disagree" or other way of ranking categories – a 3-point scale to explain the liking of the product, from "Yes", "No" and "May be".
- Examples of nominal variables include gender (2 groups: male and female), ethnicity (3 groups: Caucasian, African-American and Hispanic), physical activity level (5 groups: sedentary, slightly active, moderately active, active, and extremely active), profession (5 groups: surgeon, doctor, nurse, dentist, therapist) and so forth.

Revisiting our weight–height example, we notice two of the independent variables to be continuous and one as categorical – exercise level with 3 levels. Hence, for carrying out regression analysis, we need to create new variable(s) or recode the categorical variable – exercise level – into numerical values as the regression algorithm doesn't work with non-numeric variables. Exercise level for each person can be recoded as (1=Sedentary, 2=Moderately Active, 3=Very Active) based on their lifestyle and attitude towards exercise.

Linearity

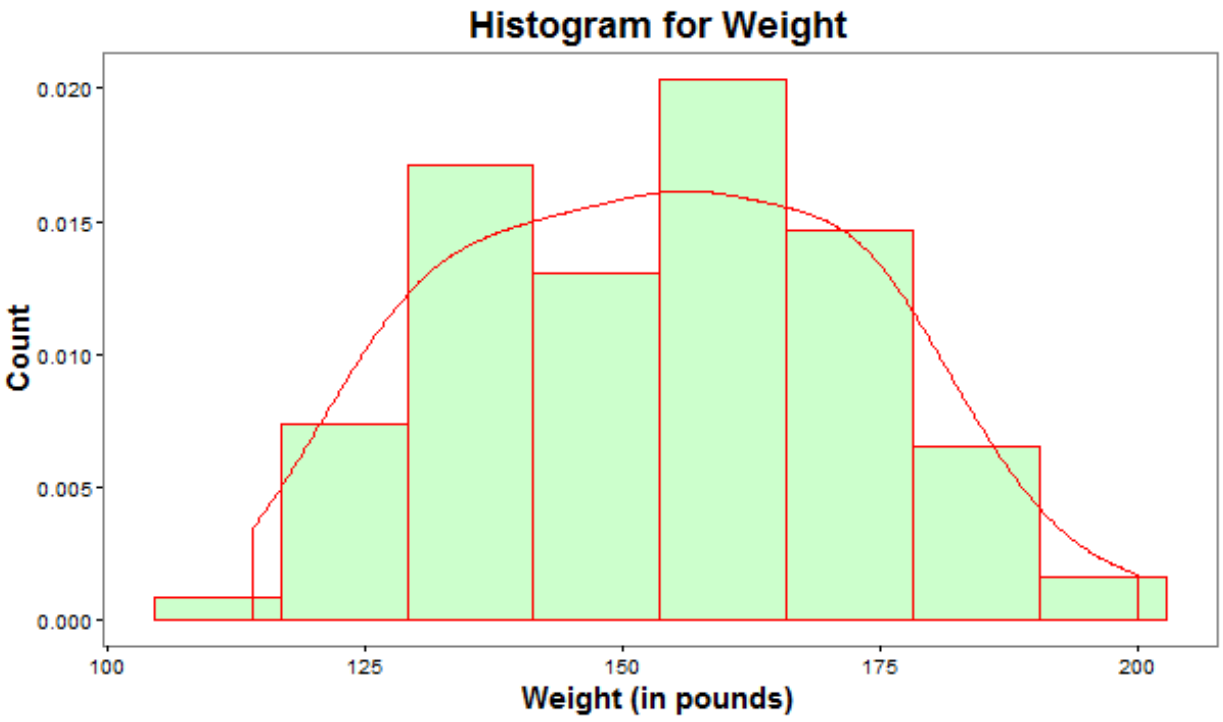
Multiple regression technique does not test whether the data is linear. Instead, it requires the existence of a linear relationship between – the dependent variable and each of the independent variables, and the dependent variable and the independent variables collectively (assessed from the model fit or from 3rd scatterplot as shown below).

Linearity

The above plots help us to visually answer: Are the two variables linearly related? We infer that each of the IVs (height, calorie intake) *in first 2 plots*, plotted one at a time, with the dependent variable (weight) and even *the last plot* (effect of IVs collectively via Predicted values of DV) signifies linear relationship between the variables.

Normality

Multiple Regression Analysis requires that variables are normally distributed. In practice, the distribution of the variables, close to normal distribution is acceptable. There are various ways to check the normality assumption. Histogram is a quick way to check normality.

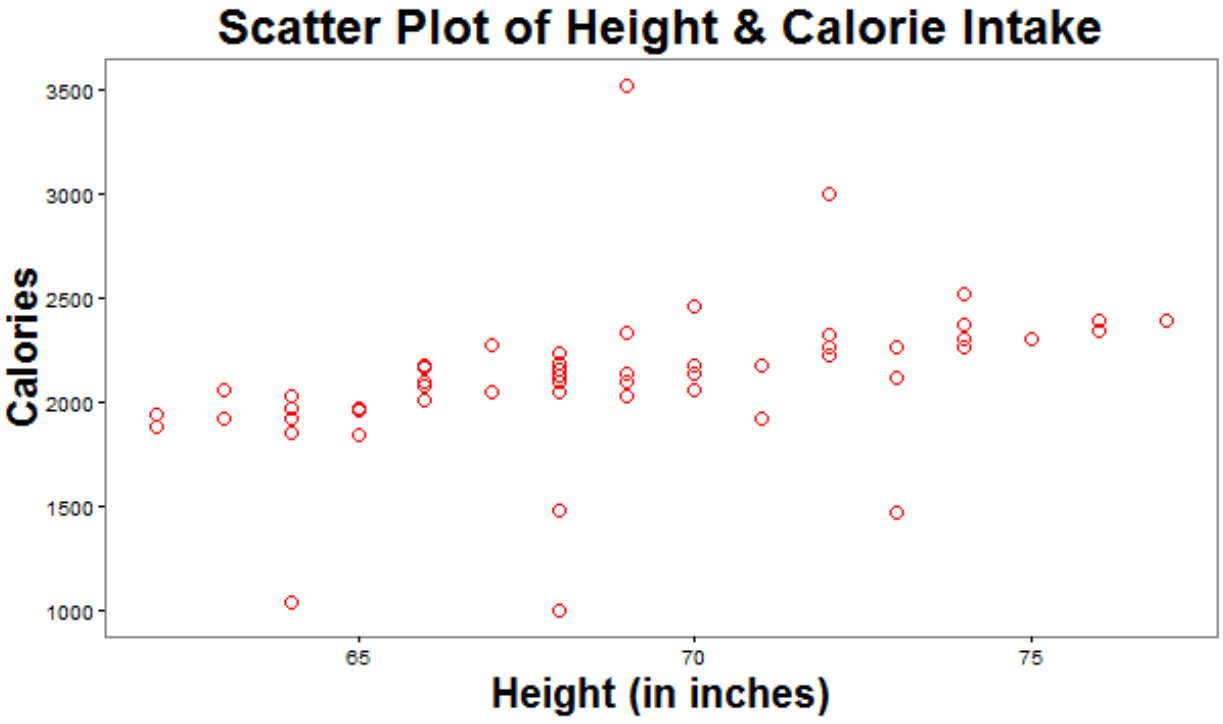


The above histogram plot includes a density curve that closely depicts the bell-shaped curve of normal distribution.

Absence of MultiCollinearity

Multicollinearity pertains to the relationship among IVs. exists when the IVs are highly correlated with each other or when one IV is a combination of one or more of the other IVs.

For example, when we look into the pricing of house flat, both the variables – area in square feet and area in square cm or square inches doesn't contribute much to the price prediction as these 2 variables give the same information, though in a different way and are highly correlated as evident from the conversion formula.



As indicated in the above plot, height and calorie intake used for predicting weight reflects no discernible pattern i.e. the data demonstrates an absence of multicollinearity.

The other criteria that could be used to detect multicollinearity are Tolerance, Variance Inflation Factor (VIF), or Condition Index.

Absence of Significant Outliers among variables

There should be no significant outliers for both– among IVs and on DV. Outliers are points which lie outside the overall pattern of the data. The removal of these influential observations can cause the regression equation to change considerably and may improve correlation.

Potential outliers could be identified from the plots of each of the IVs and DV for weight – height example as below:

Outliers

The remedial measures for treating outliers could be:

- An outlier for a particular IV can be tackled either by deleting the entire observation, counting those extreme values as missing and then treat missing values, or retain the outlier by reducing the extremity by assigning a high score/value for that variable, but not too different from the remaining cluster of scores.
- Outliers on DV can be identified readily on residual plots since they are cases with very large positive or negative residuals (errors). In practice, standardized residual values greater than an absolute value of 3.3(i.e. above 3.3 or less than -3.3) are considered outliers.

Normality, Linearity, Homoscedasticity and Independence of Residuals

Residuals are the errors in prediction–the difference between observed and predicted DV scores.

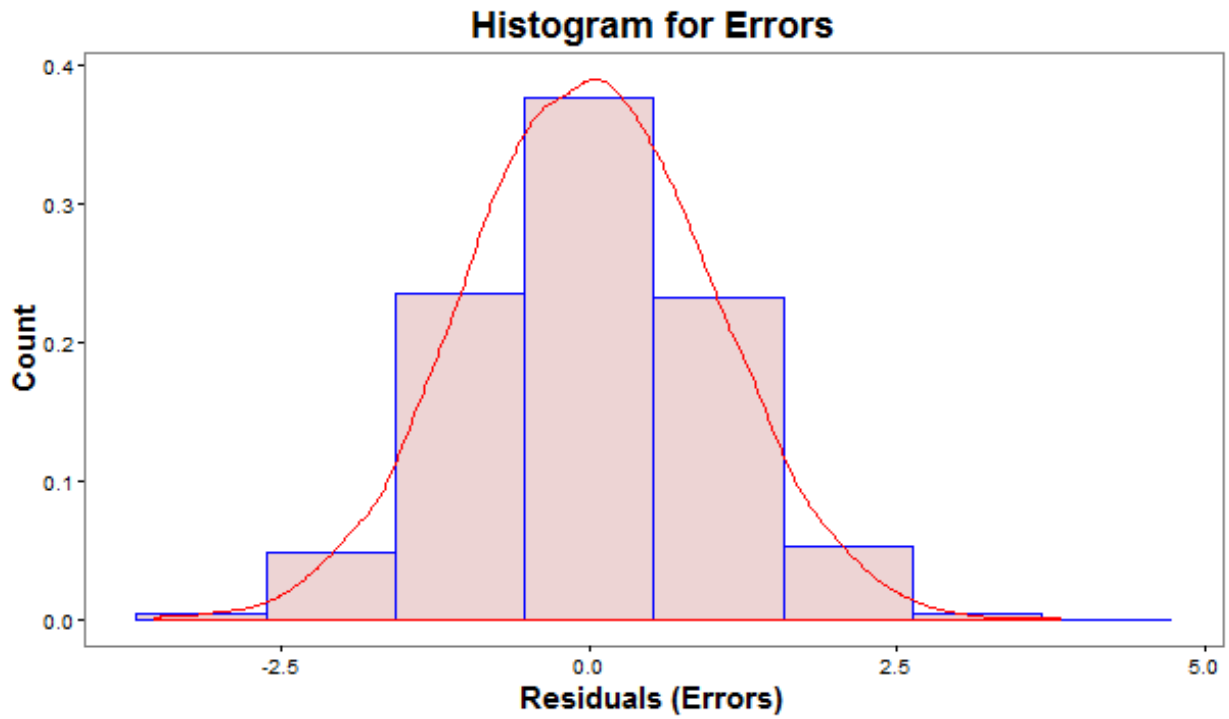
These characteristics of Residuals illustrates the nature of the underlying relationship between the variables, which can be checked from residuals scatter-plots.

The residual scatter-plots allow you to check

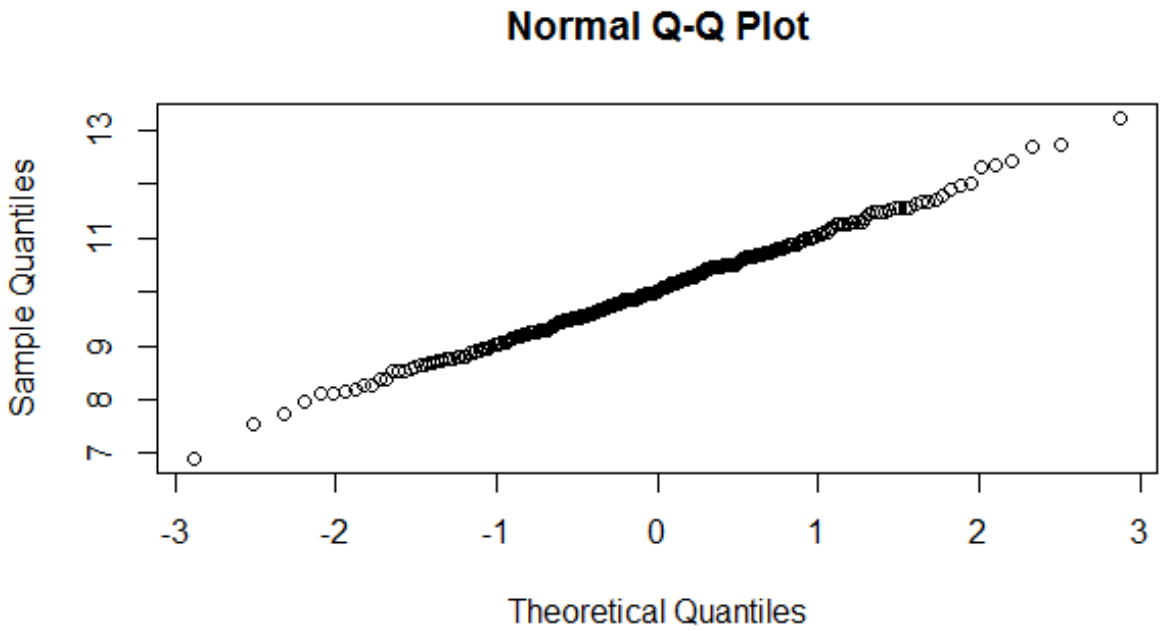
> Normality: The residuals should be normally distributed. Though, in practice, the distribution of errors, close to normal is acceptable.

The normality of errors could be gauged through:

(i) Histogram of Errors– should be mound shaped around 0.

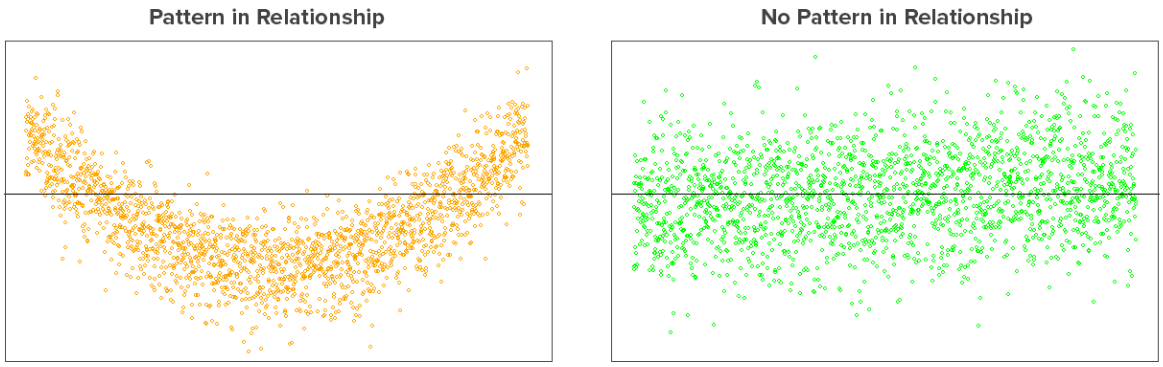


(ii) Normal Probability Plot (Q-Q plot)– is a scatter-plot created by plotting 2 sets of quantiles (often termed as “percentiles”) against one another. For example, the 0.3 (or 30%) quantile is the point at which 30% of the data fall below and 70% fall above that value. Q-Q plot help us to access if a dataset probably came from some theoretical distribution such as Normal, or other distribution.



(iii) Statistical tests like Correlation test, Wilks-Shapiro test etc

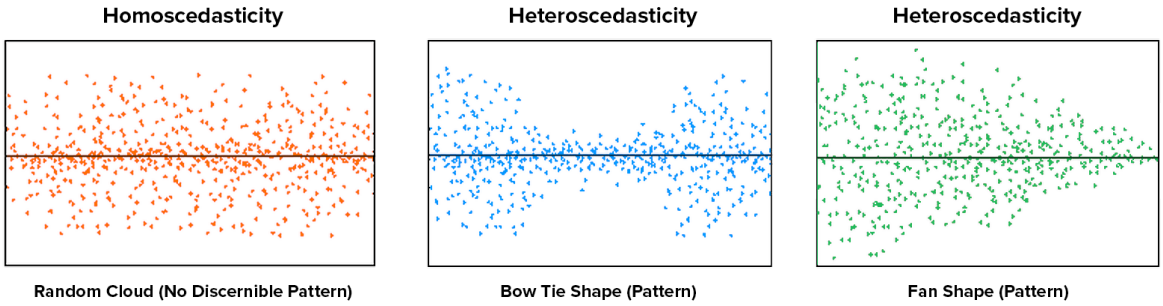
Linearity: The residuals plot should reflect a random scatter of points. A non-random pattern suggests that a linear model is inappropriate, and that data may require some transformation of the response or predictor variables or add a quadratic or higher term in the equation.



As seen in the above residuals plot – first one shows a pattern i.e. the relationship between IVs and DV is not linear. Therefore, the results of the regression analysis would *under-estimate* the true relationship.

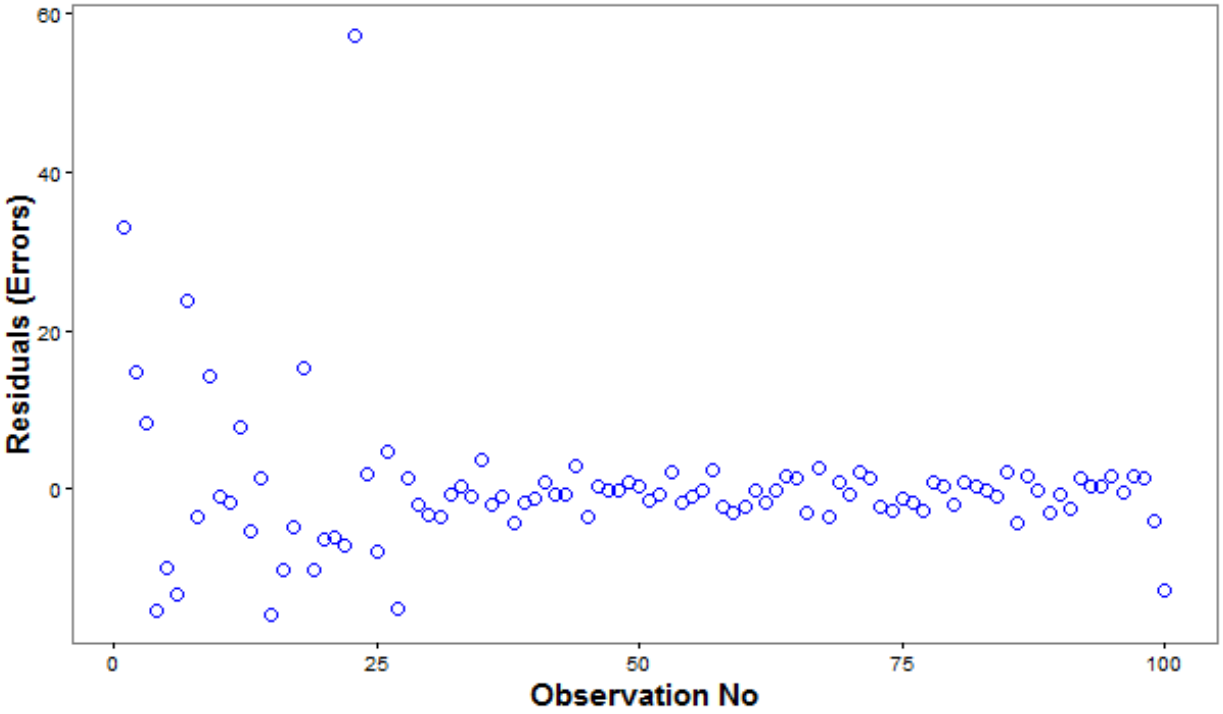
> Homoscedasticity: The scatter-plot is a good way to check whether homoscedasticity (i.e. the error terms along the regression are equal è constant variance across IV values) is given.

The homoscedasticity and heteroscedasticity plots of data reveals either no pattern or some pattern as shown below:



Heteroscedasticity i.e. non-constant variance of errors can lead to serious distortion in findings and weaken the analysis and increase the prediction errors. A non-linear transformation might fix this problem.

> Independence: The residuals should be independently distributed i.e. no correlation between consecutive errors. In other words, one of the error is independent of the value of another error(s).



A random pattern of Errors, as above indicates independence of errors.

Closing Thoughts:

It may happen that you get fascinated by the insights arising from your linear regression model, but you should force yourself to probe into the validity and conformance of the key assumptions underlying your regression model, so as to be able to apply it and get similar results from unseen or new data.

In the concluding part , we will learn how to build the regression model and interpret the model output to evaluate the quality of the model.

A brief primer on linear regression – Part III

clevertap.com/blog/a-brief-primer-on-linear-regression-part-iii

Data Science

Pushpa Makhija

In Part I, we learnt the basics of Linear Regression and in Part II, we have seen that testing the assumptions in simple and multiple regression before building a regression model is analogous to knowing the rules upfront before playing a fair game.

Building a regression model involves collecting predictor and response values for common samples, exploring data, cleaning and preparing data and then fitting a suitable mathematical relationship to the collected data to answer: *Which factors matter the most? Which factors we can ignore?*

In the linear regression model, the dependent variable is expressed as a linear function of two or more independent variables plus an error introduced to account for all other factors, as mentioned below:

$$Y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \ldots + b_nx_n + \epsilon$$

Linear Regression Model Building

Prior to building any predictive model, the data exploration and preparation stage ensures that every variable is in the form as desired by model. The next step is to build a suitable model and then interpret the model output to assess whether the built model is a good fit for the given data.

As a quick recap of our height – weight dataset, a sample of 10 rows of this dataset has been displayed below:

Let's work on fitting a linear model to the above dataset by using height, calorie intake, and exercise level as predictors for predicting the response variable – weight of an individual and then derive valuable insights by interpreting the model output.

Ht-Wt Table2

Interpreting the output

The model output for our height-weight example is displayed below:

Output Table

This output includes a conventional table with parameter estimates and their standard errors, t-value, p-value, F-statistic, as well as the residual standard error and multiple R-squared.

Now we define and explain briefly each component – marked as **#1 – #10** in the above model output.

#1 Residuals – are essentially the errors in prediction – for our example, the difference between the actual observed "Weight" values and the model's predicted "Weight" values.

The Residuals section of the model output breaks down into 5 summary point measures, to help us assess whether the distribution of residuals is normal i.e. bell shaped curve across the mean value zero (0).

In our case, we see that the distribution of the residuals do not appear to be strongly normal as median is to the left of 0. That means the model predicts certain points that fall far away from the actual points.

Residuals can be thought of as similar to a dart board. A good model is the one which will hit the bull's-eye some of the time. When it doesn't hit the bull's-eye, the miss should be close enough to the bull's-eye than on the outer edges of the dart board.

#2 Estimated Coefficients – are the unknown constants that represent the intercept and slope terms in the linear model. The estimated coefficient is the value of slope calculated by the regression.

As in the above table, the coefficient Estimate column contains two or more rows: the first one is the intercept and the rest rows are for each of the independent variable(s).

> The intercept is the base value for DV (weight) when all IVs (height, calorie, exercise level in our case) are zero. In this context, the intercept value is relatively meaningless since weight of 0 lbs is unlikely to occur for even an infant. Hence, we cannot draw any further interpretation from this coefficient.
>
> From the second row onwards there are slopes for each of the independent variables considered for building predictive model. In short, the size of the coefficient for each IV gives the size of the effect that variable has on the DV and the sign on the coefficient (positive or negative) gives the direction of the effect. For example, the effect height has in predicting weight of a person. The slope term of height indicates that for every 1 inch increase in the height of a person, the weight goes up by 3.891 pounds (lbs), *holding all other IVs constant*.

#3 & #8 Standard Error of the Coefficient Estimate & Residual Standard Error – are just the standard deviations of Coefficient Estimate and Residuals. The standard error of the estimate measures the dispersion (variability) of the coefficient, the amount it varies across cases and Residual standard error reflects the quality of a linear regression fit. Lower the standard error, the better it is, for accuracy of predictions. Therefore, you would expect to observe most of the actual values to cluster fairly closely to the regression line.

For example, to decide among the 2 datasets of 10 heights having same mean of 69 inches but different standard deviations (SD) – one with $\sigma = 2.7$ and the other one with $\sigma = 6.3$, we should select the dataset with $\sigma = 2.7$ to use height as one of predictor for creating predictive model.

In our example, the std error of the height variable is 0.262 which is far less than the height Coefficient – Estimate (3.891). Also, the actual weight can deviate from the true regression line by approximately 8.968 lbs, on an average.

#9 Multiple R-squared & Adjusted R-squared – The R-squared statistic (R^2), also known as Coefficient of determination, is a metric used to evaluate how well the model fits the actual data.

R^2 corresponds with the proportion of the variance in the criterion variable which is accounted for, by the model.

$$\mathbf{R^2} = \frac{Explained\ variation}{Total\ variation} = Proportion\ of\ variation\ in\ Y\ values\ explained\ by\ the\ linear\ relationship\ with\ X$$

R^2 always lies between 0 and 1. Hence, a number near 0 represents that a regression does not explain the variability in the response variable and a number close to 1 does explain the observed variance in the response variable.

R^2 tends to somewhat over-estimate the success of the model since it automatically and spuriously increases when extra explanatory variables are added to the model. Adj. R^2 corrects this value to provide a better estimate of the true population value by taking into account the number of variables and the number of observations that goes into building the model.

where p is the total number of variables in the model (excluding the constant – Intercept term), and n is the sample size.

$$\mathbf{Adjusted\ R^2} = 1 - \left(1 - R^2\right)\frac{n-1}{n-p-1}$$

Unlike R^2 – always increasing as more variables are included in the model, adjusted R^2 increases only if the new term improves the model more than what would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. It is always lower than R^2. Furthermore, adj. R^2 is the preferred measure to evaluate the model fit as it adjusts for the number of variables considered.

While choosing between two models, it's better to choose the one with higher adj. R^2. However, this higher value doesn't necessarily indicate the accuracy of the predictions and the adequacy of the regression model.

In our case, ~80% of the variance in the response variable (weight) is explained by the predictors (height and calorie intake). Intuitively also, by knowing these values – height, calorie intake – we would be able to predict the weight of an individual quite well, as also reflected in the obtained relatively strong R^2 value.

#4 – #7; #10 t – value of the Coefficient Estimate; Variable p – value, Significance Stars and Codes; F-statistic with Degrees of Freedom and p-value – are the terms used to assess the model fit and the significance of the model or its components through the statistical tests.

t – value of the Coefficient Estimate –

is a score to measure whether or not the regression coefficient for the variable is meaningful for the model i.e. the coefficient is significant and different from zero.

The t-statistic value is computed as:

$$t - statistic = \frac{Coefficient\ Estimate}{Std\ Error\ of\ the\ Estimate}$$

In our example, the t-values of height, calorie intake are relatively far away from zero and are large relative to the standard error, which could indicate an existence of the relationship.

Variable p-value & Significance Stars and Codes –

p-value indicates a probability that the variable is NOT relevant i.e. Pr(>(|t|) acronym in the model output. A small p-value indicates that it is unlikely that a relationship between a predictor (say, height) and response (weight) variables exists due to chance. Generally, a p-value of 5% or less is considered as cut-off point.

In our example, the p-values for height and calorie intake are very close to zero (indicated by '***' in the table), suggesting that it is likely that significant relationship exists between height, calorie intake and weight of the people – with the obtained coefficient estimates – different from zero.

F-statistic, Degree of Freedom and Resulting p-value –

are the metrics to evaluate the overall model fit of the data. F-statistic is a good indicator to assess whether there is a relationship between the dependent and independent variables. The further F-statistic is from 1, the higher the likelihood of the existence of relationship between dependent and independent variables.

To explain Degrees of Freedom, let's consider a scenario where we know 9 of the data points and the mean of 10 data points. We don't have freedom to choose the actual value of 10^{th} observation, as we can easily calculate the same by (mean * 10 – Sum of all 9 observations). This results in one data point going into estimating this actual value of 10^{th} data point, giving us choice of 9 degrees of freedom (d.f.) for these 9 known points.

In our example, the F-statistic is 98.53 which is much larger than 1 given in the 100 observations. The degrees of freedom are 4 (the number of variables used in the model (5 – including Intercept) – 1) and 95 (the number of observations included in the dataset (100) – the number of variables used in the model (5)). Also, the p-value is low and close to 0. Hence, a large value of F and small p-value indicates the overall significance of the model.

Closing Thoughts

It is often trickier to spot a bad model rather than identifying and selecting a good model.

Multiple regression analysis is not only the most widely used tool but also the most abused one. Furthermore, the sensible use of linear regression requires one to check for any errors in variables, treat outliers and any missing values, validate the underlying assumptions for any violation(s); determine the goodness of fit and accuracy of the model through statistical tests; deal with potential problems that may occur in the model and the difficulties involved in rigorously evaluating the quality and robustness of the model fit. Linear regression is important because it is the basic model used by many analysts to compare with other complex models to generate data insights.