Contents

[Simple Linear Regression 2](#_Toc518726055)

[Line of Best Fit 2](#_Toc518726056)

[Least Square Method 2](#_Toc518726057)

[Assumptions of linear regression 2](#_Toc518726058)

[What if these assumptions get violated 3](#_Toc518726059)

[Correlation of error terms 4](#_Toc518726060)

[2. Non Linearity of the data : 5](#_Toc518726061)

[Heteroskedasticity 6](#_Toc518726062)

[Normal Distribution of error terms 7](#_Toc518726063)

[Multicollinearity 7](#_Toc518726064)

[Outlier 8](#_Toc518726065)

[Residual Plot is used to detect the outlier 8](#_Toc518726066)

[High Leverage Points 8](#_Toc518726067)

[Different Plot to be looked 9](#_Toc518726068)

[Normal Q-Q Plot 9](#_Toc518726069)

[Residuals vs Leverage Plot 10](#_Toc518726070)

[Interpreting Regression Results 10](#_Toc518726071)

[Interpreting data Coefficient : 10](#_Toc518726072)

[Interpreting the Constant 10](#_Toc518726073)

[Std Error 11](#_Toc518726074)

[t-Ratio 11](#_Toc518726075)

[Prob>|t|: 11](#_Toc518726076)

[F Test for Overall Significance of the Model 11](#_Toc518726077)

[Individual Variable Significance – T test 12](#_Toc518726078)

[Interview Questions 12](#_Toc518726079)

[How would homo/heteroskedasticity affect regression analysis? 12](#_Toc518726080)

[What are control variable in the regression 13](#_Toc518726081)

# Simple Linear Regression

## Line of Best Fit

In general, when we use y^i=b0+b1xi to predict the actual response *yi*, we make a prediction error (or residual error) of size:

ei=yi−y^i

* **It can be of any shape depending on the number of independent variables (a point on the axis, a line in two dimensions, a plane in three dimensions, or a hyperplane in higher dimensions).**

***Least Squares Method:* The *best fit* is done by making sure that the sum of all the distances between the shape and the actual observations at each point is as small as possible. The fit of the shape is “best” in the sense that no other position would produce less error given the choice of shape.**

The equation of the best fitting line is: y^i=b0+b1xi

We just need to find the values *b*0 and *b*1 that make the sum of the squared prediction errors the smallest it can be.That is, we need to find the values *b*0 and *b*1 that minimize: Q=∑i=1n(yi−y^i)2

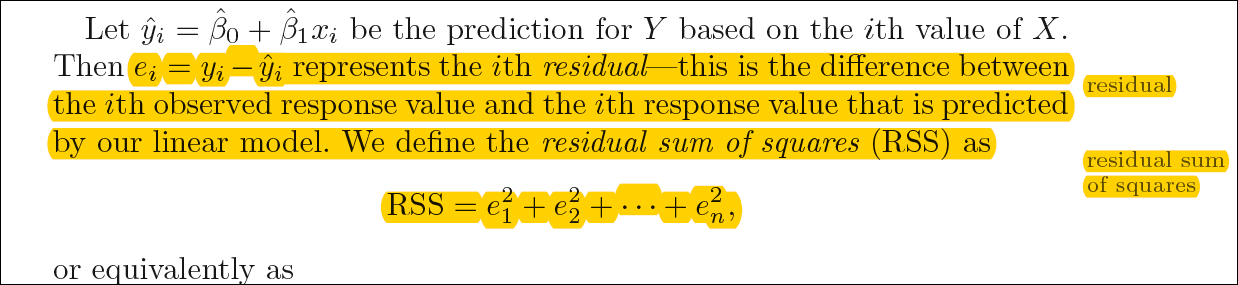
Here's how you might think about this quantity *Q*:

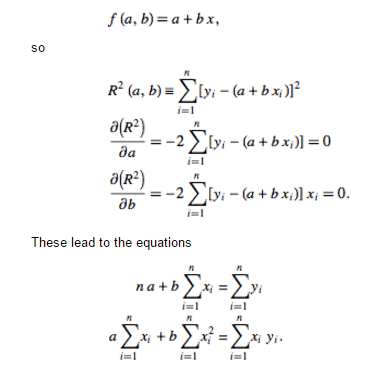
* The quantity  ei=yi−y^i is the prediction error for data point *i*.
* The quantity  e2i=(yi−y^i)2 is the squared prediction error for data point *i*.
* And, the symbol  ∑ni=1 tells us to add up the squared prediction errors for all *n* data points

## Least Square Method

Least square method is used to estimate the parameters

**Cost Function** – Sum of squared error (RSS) – Residual sum of square







https://cdn-images-1.medium.com/max/1080/1*_Rqrmx-PZy56idZZe1nKxg.png

**Least Squares Method for Simple Linear Regression.**

# Properties of estimator

Unbiased

# Assessing the accuracy of the Coefficient estimates

**P value**

Null hypothesis – no relation

Alternative hypothesis – there is a relation ( p value should be small)

Roughly speaking, we interpret the p-value as follows: a small p-value indicates that

it is unlikely to observe such a substantial association between the predictor

and the response due to chance, in the absence of any real association

between the predictor and the response. Hence, if we see a small p-value,

then we can infer that there is an association between the predictor and the

response. We *reject the null hypothesis*—that is, we declare a relationship

to exist between *X* and *Y* —if the p-value is small enough. Typical p-value

cutoffs for rejecting the null hypothesis are 5 or 1%. When *n* = 30, these

correspond to t-statistics (3.14) of around 2 and 2.75, respectively.

# Assumptions of linear regression

**Common to all regressions**

1. **Independence of errors**: the errors of the response variables are uncorrelated with each other

**Specific to linear regression (lcn)**

1. **Linearity**: response variable is a linear combination of regression coefficients and the predictor variables (a.k.a., mean model assumption )
2. **Constant variance** (a.k.a. homoscedasticity): different values of the response variable have the same variance in their errors, regardless of the values of the predictor variables (between rows)
3. **Normality**: error terms follow a normal distribution

## How to check these assumption validity

1. **Independence of error –**
2. **Lienarity**
3. **Constant Variance**

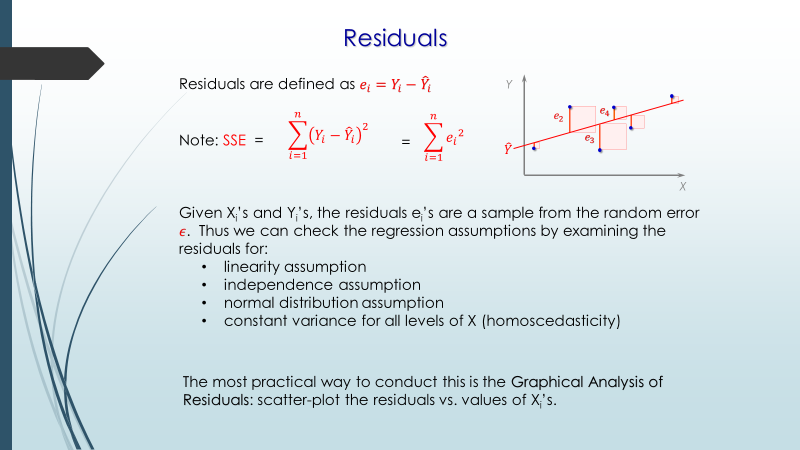
You can look at residual vs fitted values plot. If heteroskedasticity exists, the plot would exhibit a funnel shape pattern (shown in next section). Also, you can use Breusch-Pagan / Cook – Weisberg test or White general test to detect this phenomenon.

1. **Normality** –

Try to examine a histogram of the  to see if it appears to be bell- shaped . Although shape of a histogram may be difficult to judge unless the sample size is large.

By examining a normal probability plot of the residuals. Essentially, the ordered (standardized) residuals are plotted against theoretical expected values for a sample from a standard normal curve population. A straight-line pattern for a normal probability plot (NPP) indicates that the assumption of normality is reasonable.

Do a hypothesis test in which the null hypothesis is that the errors have a normal distribution. Failure to reject this null hypothesis is a good result. It means that it is reasonable to assume that the errors have a normal distribution.



## 

## What if these assumptions get violated

Let’s dive into specific assumptions and learn about their outcomes (if violated):

|  |
| --- |
| **Independence of errors**: the errors of the response variables are uncorrelated with each other |

Correlation of error terms **:** The presence of correlation in error terms drastically reduces model’s accuracy. This usually occurs in time series models where the next instant is dependent on previous instant. If the error terms are correlated**, the estimated standard errors tend to underestimate the true standard erro**r(estimated standard error are low)

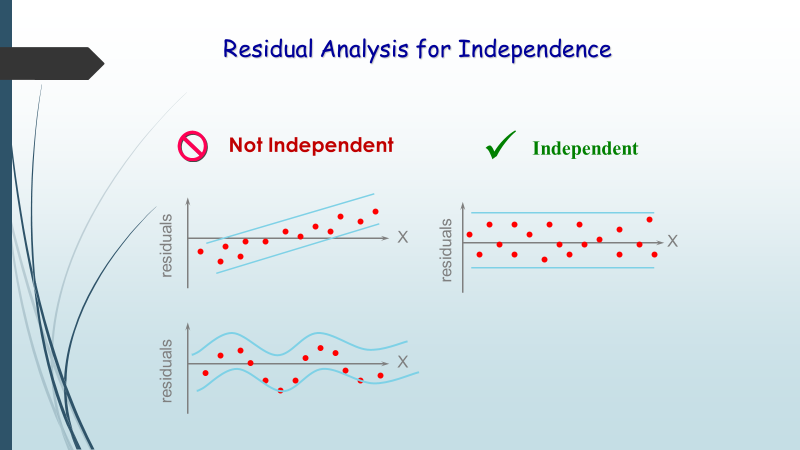
And also it causes confidence intervals and prediction intervals to be narrower. **Narrower confidence interval means that a 95% confidence interval would have lesser probability than 0.95** that it would contain the actual value of coefficients.

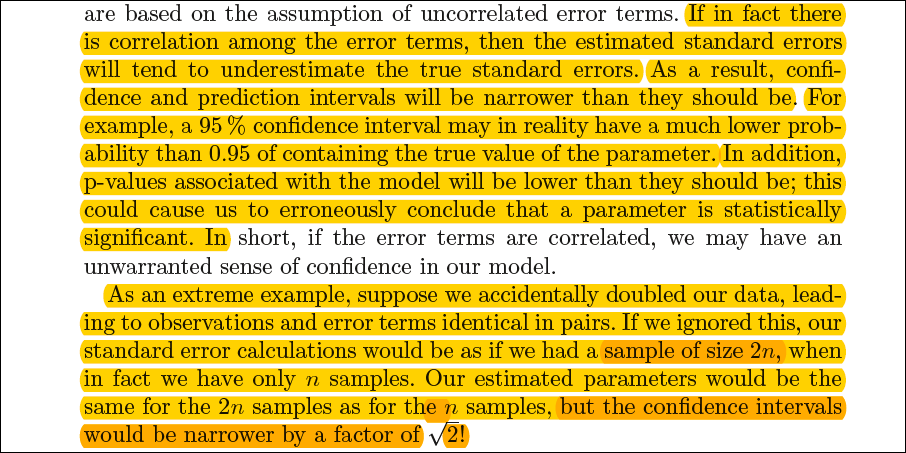
Let’s understand narrow prediction intervals with an example:

For example, the least square coefficient of X¹ is 15.02 and its standard error is 2.08 (without autocorrelation). **But in presence of autocorrelation, the standard error reduces to 1.20.** **As a result, the prediction interval narrows down to (13.82, 16.22) from (12.94, 17.10).**

Also, lower standard errors would cause the associated p-values to be lower than actual. This will make us incorrectly conclude a parameter to be statistically significant.

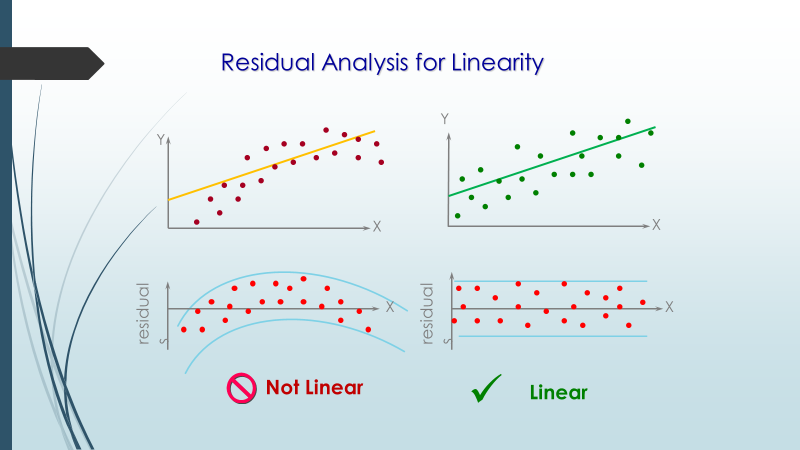
**How to check:** Look for Durbin – Watson (DW) statistic. It must lie between 0 and 4. If DW = 2, implies no autocorrelation, 0 < DW < 2 implies positive autocorrelation while 2 < DW < 4 indicates negative autocorrelation. Also, you can see residual vs time plot and look for the seasonal or correlated pattern in residual values





1. Non Linearity of the data :If you fit a linear model to a non-linear, non-additive data set, the regression algorithm would fail to capture the trend mathematically, thus **resulting in an inefficient model**. Also, this will result in erroneous predictions on an unseen data set.

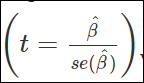
**How to check:** Look for residual vs fitted value plots (explained below). Also, you can include polynomial terms (X, X², X³) in your model to capture the non-linear effect.



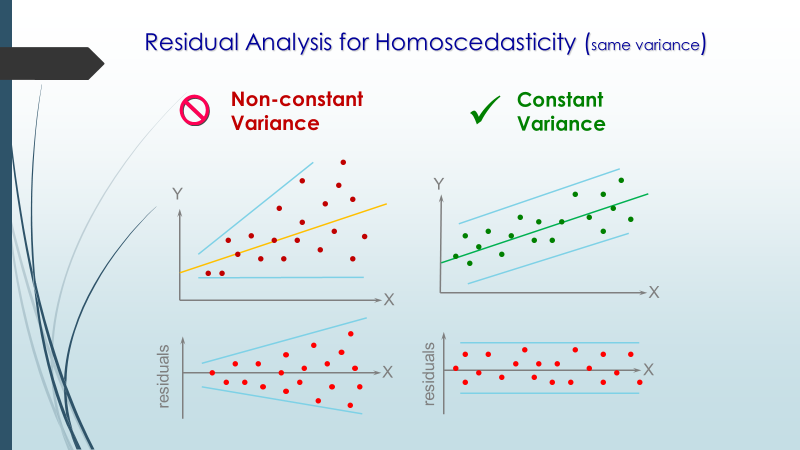
**3.** Heteroskedasticity**:**

|  |
| --- |
| different values of the response variable have the same variance in their errors, regardless of the values of the predictor variables |

The presence of non-constant variance in the error terms results in heteroskedasticity. Generally, non-constant variance arises in presence of outliers or extreme leverage values. Look like, these values get too much weight, thereby disproportionately influences the model’s performance. When this phenomenon occurs, the confidence interval for out of sample prediction tends to be unrealistically wide or narrow.

If the variance of an estimate is higher compared to best estimate’s variance (i.e. minimum variance estimate), t-statistic value will be smaller and make the coefficient insignificant. What appear to be insignificant coefficient, may be significant if we obtains best estimate from OLS.

**How to check** : You can look at residual vs fitted values plot. If heteroskedasticity exists, the plot would exhibit a funnel shape pattern (shown in next section). Also, you can use Breusch-Pagan / Cook – Weisberg test or White general test to detect this phenomenon.

****

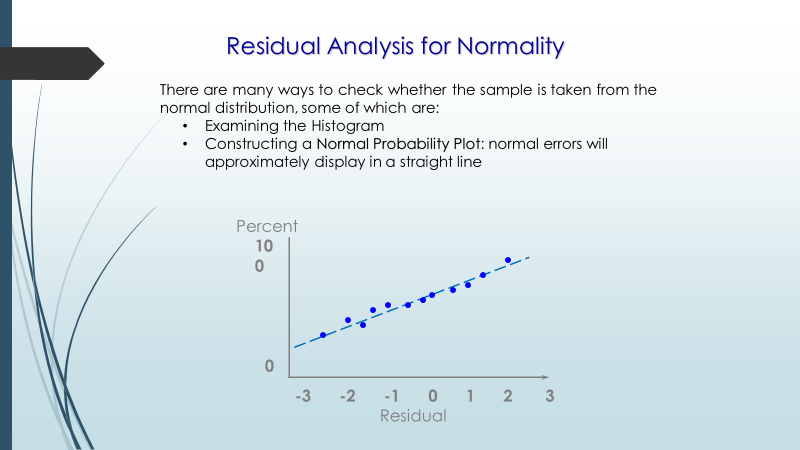
**What to do when this happen-** When this happens we can transform the variable such as log Y ot square root Y

* *How does heteroscedasticity affect the coefficient estimates and why? What are some fixes for heteroscedasticity?*

**4.**

Normal Distribution of error terms**:** If the error terms are non- normally distributed, **confidence intervals may become too wide or narrow**. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares. Presence of non – normal distribution suggests that there are a few unusual data points which must be studied closely to make a better model.

**How to check:** You can look at QQ plot (shown below). You can also perform statistical tests of normality such as Kolmogorov-Smirnov test, Shapiro-Wilk test.



**5**

Multicollinearity**:** This phenomenon exists when the independent variables are found to be moderately or highly correlated. In a model with correlated variables, it becomes a tough task to figure out the true relationship of a predictors with response variable. In other words, it becomes difficult to find out which variable is actually contributing to predict the response variable.

Another point, with presence of correlated predictors, the standard errors tend to increase. And, with large standard errors, the confidence interval becomes **wider** leading to less precise estimates of slope parameters.

Also, when predictors are correlated, the estimated regression coefficient of a correlated variable depends on which other predictors are available in the model. If this happens, you’ll end up with an incorrect conclusion that a variable strongly / weakly affects target variable. Since, even if you drop one correlated variable from the model, its estimated regression coefficients would change. That’s not good!

**How to check:** You can use scatter plot to visualize correlation effect among variables. Also, you can also use VIF factor. VIF value <= 4 suggests no multicollinearity whereas a value of >= 10 implies serious multicollinearity. Above all, a correlation table should also solve the purpose.

Or we can look the correlation matrix

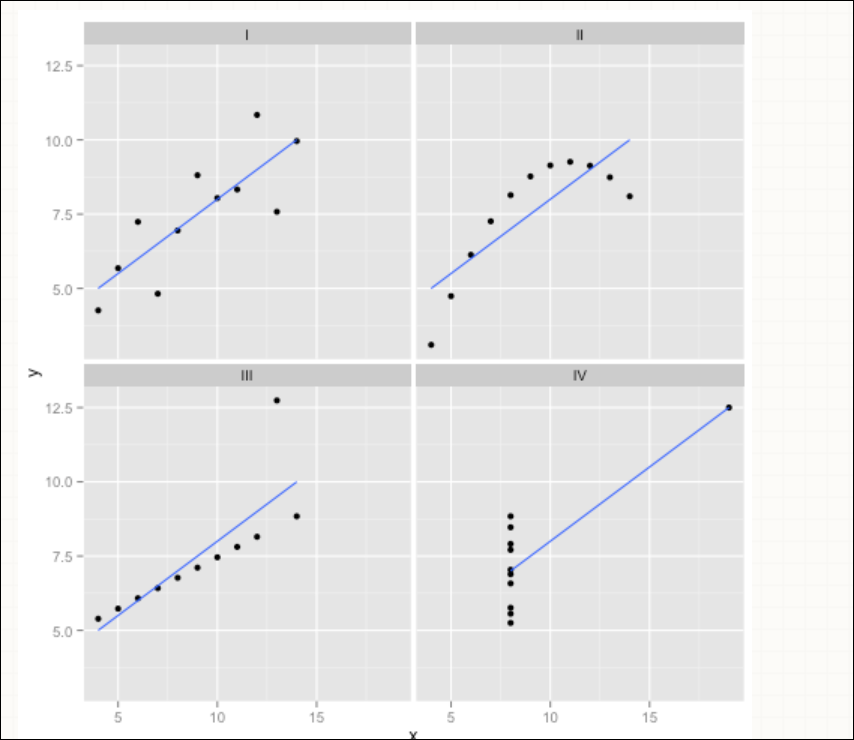
|  |  |  |
| --- | --- | --- |
|  |  |  |
| Multicollinearity |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Outlier

Residual Plot is used to detect the outlier

Studentized residuals – To check outliers

High Leverage Points



# Different Plot to be looked

**1**. Residual vs Fitted Values

There are two major things which you should learn:

1. If there exist any pattern (may be, a parabolic shape) in this plot, consider it as signs of non-linearity in the data. It means that the model doesn’t capture non-linear effects
2. If a funnel shape is evident in the plot, consider it as the signs of non constant variance i.e. heteroskedasticity.

**Solution**:

To overcome the issue of non-linearity, you can do a non linear transformation of predictors such as log (X), √X or X² transform the dependent variable. To overcome heteroskedasticity, a possible way is to transform the response variable such as log(Y) or √Y.

Also, you can use weighted least square method to tackle heteroskedasticity.

## 

## Normal Q-Q Plot

This q-q or quantile-quantile is a scatter plot which helps us validate the assumption of normal distribution in a data set. Using this plot we can infer if the data comes from a normal distribution. If yes, the plot would show fairly straight line. Absence of normality in the errors can be seen with deviation in the straight line. If you are wondering what is a ‘quantile’, here’s a simple definition: Think of quantiles as points in your data below which a certain proportion of data falls. Quantile is often referred to as percentiles. For example: when we say the value of 50th percentile is 120, it means half of the data lies below 120.

**Solution:** If the errors are not normally distributed, non – linear transformation of the variables (response or predictors) can bring improvement in the model.

## Residuals vs Leverage Plot

It is also known as Cook’s Distance plot. Cook’s distance attempts to identify the points which have more influence than other points. Such influential points tends to have a sizable impact of the regression line. In other words, adding or removing such points from the model can completely change the model statistics.

But, can these influential observations be treated as outliers? This question can only be answered after looking at the data. Therefore, in this plot, the large values marked by cook’s distance might require further investigation.

**Solution:** For influential observations which are nothing but outliers, if not many, you can remove those rows. Alternatively, you can scale down the outlier observation with maximum value in data or else treat those values as missing values.

# Interpreting Regression Results

Interpreting data Coefficient : (Parameter Estimate) for every one unit increase in the price of a house, - .793 fewer houses are sold.

## Interpreting the Constant

The constant is often defined as the [mean](http://statisticsbyjim.com/glossary/mean/) of the [dependent variable](http://statisticsbyjim.com/glossary/response-variables/) when you set all of the [independent variables](http://statisticsbyjim.com/glossary/predictor-variables/) in your model to zero

Std Error**:** if the regression were performed repeatedly on different datasets (that contained the same variables), this would represent the standard deviation of the estimated coefficients

t-Ratio**:** the coefficient divided by the standard error, which tells us how large the coefficient is relative to how much it varies in repeated sampling. If the coefficient varies a lot in repeated sampling, then its t-statistic will be smaller, and if it varies little in repeated sampling, then its t-statistic will be larger.

Prob>|t|:the p-value is the result of the test of the following null hypothesis: in repeated sampling, the mean of the estimated coefficient is zero. E.g., if p = 0.001, the probability of observing an estimate of that is at least as extreme as the observed estimate is 0.001, if the true value of is zero. In general, a p-value less than some threshold , like 0.05 or 0.01, will mean that the coefficient is “statistically significant.”

there is a difference between statistical significance and actual significance. Statistical significance tells us how sure we are about the coefficient of the model, based on the data.

Actual significance tells us the importance of the variable – a coefficient for X could be very close to 0 with very high statistical significance, but that could mean it contributes very little to Y. Conversely, a variable could be very important to the model, but its statistical significance could be quite low because you don’t have enough data

# Interview Questions

## How would homo/heteroskedasticity affect regression analysis?

If this assumption fails (Not equal variance across the levels of independent variable - Heterosedasticity), the estimate produces by OLS (Ordinary Least Square) will be no longer minimum variance estimate.

If the variance of an estimate is higher compared to best estimate’s variance (i.e. minimum variance estimate), t-statistic (t=β/se(β^))value will be smaller and make the coefficient insignificant. What appear to be insignificant coefficient, may be significant if we obtains best estimate from OLS.

Heteroscedasticity would not affect your parameter estimates and your beta coefficients are still unbiased. The problem however lies in the standard errors. The standard errors no longer follow a consistent t/f distribution resulting in invalid inference or computing the confidence intervals. One can simply ignore the heteroscedaisty and assume none exists. Then the resulting confidence intervals will be narrower (false result).

## What are control variable in the regression

Control variables are usually variables that you are not particularly interested in, but that are related to the dependent variable. You want to remove their effects from the equation.

Control variables are factors that influence the outcome, but often arise from the experimental design, not the variables of interest. In other cases, they are confounding factors that cannot be catered for by adjusting the experimental design. Their impact can be minimal or great, but where they are known attempts must be made to neutralize them.

This is usually done by including them as independent variables, that are then ignored or just acknowledged. They can sometimes also be accommodated by changing the sample weights, or including them as offsets. If either of the latter are used, it will be the equivalent of assigning the variable a coefficient of one (at least in logistic regression, the way we did it).

Suppose, for example, you were interested in the difference in height of people from different ethnic groups.  You could gather a sample of people from different ethinc groups and measure them and compare the heights.  But you'd probably want to control for some other variables that are known to relate to height (e.g. sex).  It's known that men are taller then women. Even if your samples are random, the sample from each group will not have the same proportion of men and women - so you want to add that as a variable

## What is R2? What are some other metrics that could be better than R2 and why?

* goodness of fit measure. variance explained by the regression / total variance
* the more predictors you add the higher R^2 becomes.
  + hence use adjusted R^2 which adjusts for the degrees of freedom
  + or train error metrics

## You run your regression on different subsets of your data, and that in each subset, the beta value for a certain variable varies wildly. What could be the issue here?

The dataset might be heterogeneous. In which case, it is recommended to cluster datasets into different subsets wisely, and then draw different models for different subsets. Or, use models like non parametric models (trees) which can deal with heterogeneity quite nicely.

## Your linear regression didn’t run and communicates that there are an infinite number of best estimates for the regression coefficients. What could be wrong?

* p > n.
* If some of the explanatory variables are perfectly correlated (positively or negatively) then the coefficients would not be unique.

## How to interpret a Q-Q plot in a Linear regression model?

A Q-Q plot is used to check the normality of errors. In the above chart mentioned, Majority of the data follows a normal distribution with tails curled. This shows that the errors are mostly normally distributed but some observations may be due to significantly higher/lower values are affecting the normality of errors.

## Q24. You’ve got a data set to work having p (no. of variable) > n (no. of observation). Why is OLS as bad option to work with? Which techniques would be best to use? Why?

**Answer:** In such high dimensional data sets, we can’t use classical regression techniques, since their assumptions tend to fail. When p > n, we can no longer calculate a unique least square coefficient estimate, the variances become infinite, so OLS cannot be used at all.

To combat this situation, we can use penalized regression methods like lasso, LARS, ridge which can shrink the coefficients to reduce variance. Precisely, ridge regression works best in situations where the least square estimates have higher variance.

Among other methods include subset regression, forward stepwise regression.

## Q32. You have been asked to evaluate a regression model based on R², adjusted R² and tolerance. What will be your criteria?

**Answer:** Tolerance (1 / VIF) is used as an indicator of multicollinearity. It is an indicator of percent of variance in a predictor which cannot be accounted by other predictors. Large values of tolerance is desirable.

We will consider adjusted R² as opposed to R² to evaluate model fit because R² increases irrespective of improvement in prediction accuracy as we add more variables. But, adjusted R² would only increase if an additional variable improves the accuracy of model, otherwise stays same. It is difficult to commit a general threshold value for adjusted R² because it varies between data sets. For example: a gene mutation data set might result in lower adjusted R² and still provide fairly good predictions, as compared to a stock market data where lower adjusted R² implies that model is not good.