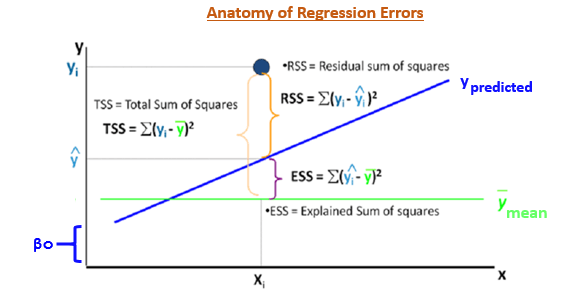
**Residual Sum of Squares (RSS)** – ?[Actual(y) – Predicted(y)]²

**Explained Sum of Squares (ESS)** – ?[Predicted(y) – Mean(ymean)]²

**Total Sum of Squares (TSS)** – ?[Actual(y) – Mean(ymean)]²

****

The most important use of these error terms is used in the calculation of the Coefficient of Determination (R²).

R² = 1 - (ESS/TSS)

R² metric tells us the amount of variance explained by the independent variables in the model. In the upcoming section, we’ll learn and see the importance of this coefficient and more metrics to compute the model’s accuracy.

**What if these assumptions get violated ?**

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

**1. Linear and Additive:**  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.

**2. Auto correlation:**

The error terms must be uncorrelated i.e. error at t must not indicate the at error at t+1. Presence of correlation in error terms is known as **Autocorrelation**. It drastically affects the regression coefficients and standard error values since they are based on the assumption of uncorrelated error terms.

**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.

**3. 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.

**4. Heteroskedasticity:** The error terms must possess constant variance. Absence of constant variance leads to **heteroskedestacity**.

**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.

**5. 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.

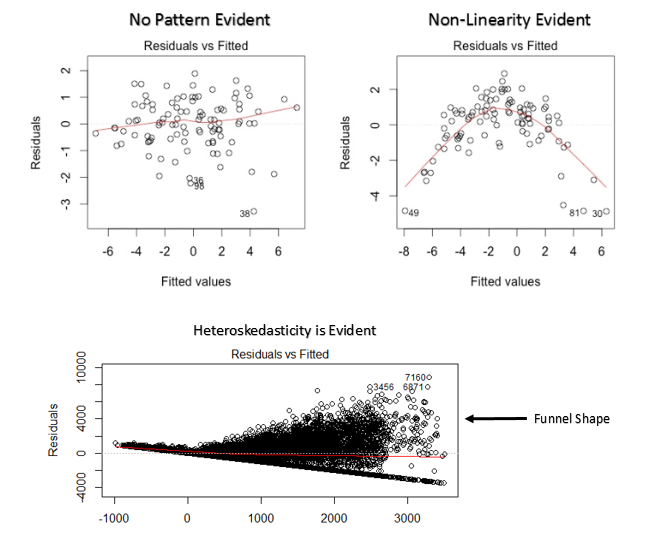
**How to check:** You can look at QQ plot (shown below). You can also perform statistical tests of normality such as Kolmogorov-Smirnov test, Shapiro-Wilk test.

## Interpretation of Regression Plots

Until here, we’ve learnt about the important regression assumptions and the methods to undertake, if those assumptions get violated.

But that’s not the end. Now, you should know the solutions also to tackle the violation of these assumptions. In this section, I’ve explained the 4 regression plots along with the methods to overcome limitations on assumptions.

1. **Residual vs Fitted Values**



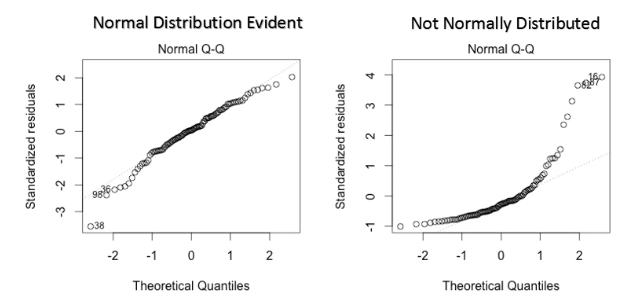
This scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values). It is one of the most important plot which everyone must learn. It reveals various useful insights including outliers. The outliers in this plot are labeled by their observation number which make them easy to detect.

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.

1. **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.