# Linear Regression Assumptions

Residual is the error of a model

## Linear relationship

there exists a linear relationship between x and y

a scatter plot can be used to determine linearity. if there is not then a non linear transformation like taking the log, sqrt, or reciprocal of one of the variables might help. another technique is adding another independent variable.

## Independence

the residuals are independent. there is no correlation between consecutive residuals in time series data.

we don’t want a pattern to emerge among consecutive residuals. ie if residuals grow larger as time goes on.

Durbin-Watson test is to formally test it, otherwise plot the time series of the data and the residuals

## Homoscedasticity

the residuals have constant variance at every level of x.

this means that the residuals should all fall in a similar range from the baseline no matter what the value of x is.

this means there is more variance in the regression coefficient estimates, leading to the model declaring a term is more significant when it is not. I think this is because the coefficient is more likely to change meaning the algorithm finds it more important to tune? this might not make sense.

we can detect this by plotting the residual as a function of the fitted value, ie the model error as a function of the x value.

one technique to eliminate this is to use the log of the dependent variable y in the regression. another is to redefine the dependent variable. one example of this is rate. if we were predicting the number of shops in a city based on population data we may use the per capita shop total as the y value. we could also use weighted regression, this assigns weight to each data point based on the variance of its fitted value. This means data with high variance has a lower weight on the regression.

look into the weighted regression more

## Normality

the residuals of the model are normally distributed.

use Q-Q plots to check that residual or error is distributed normally.

we basically plot quantile or z value of the residual versus the quantile or residual of the input value x.

if not good we can verify no outliers exist or apply nonlinear transformation to independent or dependent variable.

# Linear Regression

We fit a slope and y intercept to data

We square each term because it makes the math much easier and ensures that we get positive values

Sum of square residuals = sum((yp – y) ^ 2) where yp is the predicted y value aka the y value of the regression line for a given value of x

We want to minimize the sum of squared residuals, we can write yp s m\*x + b

Distance from the line to the datapoint is a residual, we can take the sum of squared residuals to calculate the error. The squared residuals are the loss function.

We call this method least squares to fit a line to the data.

R^2 tells us how much of the variation in the data can be explained by taking into account whatever variable we plot the new line against. Some natural variation exists, but we can say whatever percent variance is reduced taking the variable into account, or that our model removes whatever percent of the variation in making predictions

Chart

Description automatically generated

If we have a number of variables and some are not important then sum of squares will reduce the weight given to these variables to zero. While adding weird data may confuse the model, or have it take into account things that are merely coincidental, its more likely sum of squares will reduce these terms to a weight of zero, making it as if they simply did not exist.

This is important because if the p value is large it means there is a higher chance that this outcome could have been achieved simply with random chance. If its super low then its more likely we achieved this good variance reduction outcome because our model statistically models the relationship we are trying to analyze.

Diagram

Description automatically generated