WEEK ONE UPDATE

LINEAR REGRESSION

Consider we need to determine the frequency of breakdown versus the number of machines used per day in a manufacturing firm.

For different times where it broke down, we take readings for

* No of breakdowns
* No of machines available for production

We take readings for different days and then plot them on a graph.

Different line rotations are made on the graph with the plotted data points with the residuals measured and the sum of their squares taken. This gives the sum of squares for different lines across the data. These sums of squares are then plotted on the graph and the least sum of square data point on the graph is used as the line of best fit.

If the slope is not zero that implies there is a dependence of breakdowns on the no of machines available.

R2 is determined by using the variance at the mean and the best fit line which is divided by the variance at the mean. This R2 determines how much variation in the number of the breakdowns is explained by taking the no of machines available for production into account. Say we get 70%, it implies that we saw 70% reduction in variation once the no of machines available for production was put into account.

We need a way to determine if the R2 is statistically significant. This is done by using the p-value derived from F.

LOGISTIC DISTRIBUTION

This as linear regression are used in machine learning. Assuming same instance as above