# Initial Posts

|  |
| --- |
| **Plotting Residuals**  In linear regression analysis, plotting residuals can help determine if a linear regression model is appropriate. A residual is the distance between the predicted value and the observed value. If the residuals of a linear regression model are randomly scattered and do not seem to follow any sort of pattern then this means that a linear model would potentially be a good fit. If the residuals however seem to follow a pattern this means that a linear model is probably not a good fit. |
| **Goodness of Fit**  For linear regression, the best line that fits the data is determined by finding a straight line which produces the least amount of summed residuals. When just looking at 2 variables (simple regression), less summed residuals from the predicted line and observed data will produce a higher coefficient of determination (aka R^2). The coefficient of determination is often confused with Pearson's correlation and people think that you can just get the coefficient of determination by squaring Pearson's correlation which is not the case. They both measure to similar but different things. Pearson's measures how correlated 2 variables are whereas the coefficient of determination measures how much the model explains the data. |
| **Linear Least Squares**  In linear regression, the least squares is the straight line that minimizes the sum of the residuals from the model and the observed data. Although you can find the line that produces the least squares, this doesn't necessarily mean that the model is a good predictor for the outcome variable and predictor variable(s). Other testing is then done to determine if the best fitting line is actually a good predictor or not. |
| **p-value**  A p-value is a measure of the (p)robability of observing the data that you did. A critical value  identifies the point at which we would say that the data observed would not happen due to chance. Usually this is .05, or .01. Depending on whether the test is one-sided or two-sided indicates if we are testing whether there is any effect (two-sided) vs. an effect only in a particular direction (one-sided). |
| **Chi-Squared**  This is a proportion test which measures the difference between expected frequencies for values vs. the observed frequencies. In the example of the dice test, the expected values for each side of the dice should be equal if the dice is fair. When an experiment is ran, each side of the dice rolled is compared with the expected values to determine if the observed values happened by chance or if there is an apparent effect. |

# Replies

|  |
| --- |
| I agree that linear regression is a good place to start when first starting off in the field of Data Science. From other books and literature it seems that others feel the same. It is a simple model that helped me specifically, wrap my head around developing a model vs. testing the model for accuracy. |
| Choosing a test statistic depends on the observed data. For example, the distribution of your data may make median a better test statistic than mean. It really just depends on what question you're trying to answer and how good the test statistic would be at estimating a population. Various estimation methods and testing can help determine what test statistic would be good at estimating a population. |
| When I first started out in this program I must admit I was a bit confused about the hypothesis testing process. I assumed that the predicted (aka NULL) hypothesis would be the effect that we think we're seeing and that the alternative would be that the assumption is not true and that there is no effect. It makes sense now though especially in the realm of science where you never want to actually assume an effect to be true unless there is evidence. |
| Thanks Amelia, I always find good explanations at statisticshowto.com. The site itself is not the prettiest but they do a pretty good job at explaining statistics in a straightforward and simple way. |
| Good explanation Jeff. A residuals plot is a good way to determine if the model is not a good fit. For example, if you see that there is a pattern in the residuals vs. the predicted values, it probably means that the model is not a good fit for the data. |