Regression Model Assumptions

We make a few assumptions when we use linear regression to model the relationship between a response and a predictor.

These assumptions are essentially conditions that should be met before we draw inferences regarding the model estimates or before we use a model to make prediction.

Because we are fitting a linear model, we assume that the relationship really is linear, and that the errors, or residuals, are simply random fluctuations around the true line. We assume that the variability in the response doesn't increase as the value of the predictor increases. This is the assumption of equal variance. We also assume that the observations are independent of one another. Correlation between sequential observations, or auto-correlation, can be an issue with time series data -- that is, with data with a natural time-ordering.

How do we check regression assumptions? We examine the variability left over after we fit the regression line. We simply graph the residuals and look for any unusual patterns. If a linear model makes sense, the residuals will have a constant variance, be approximately normally distributed (with a mean of zero), and be independent of one another. The most useful graph for analyzing residuals is a residual by predicted plot. This is a graph of each residual value plotted against the corresponding predicted value. If the assumptions are met, the residuals will be randomly scattered around the center line of zero, with no obvious pattern. The residuals will look like an unstructured cloud of points, centered at zero. If there is a non-random pattern, the nature of the pattern can pinpoint potential issues with the model. For example, if curvature is present in the residuals, then it is likely that there is curvature in the relationship between the response and the predictor that is not explained by our model. A linear model does not adequately describe the relationship between the predictor and the response. In this example, the linear model systematically over-predicts some values (the residuals are negative), and under-predicts others (the residuals are positive).

If the residuals fan out as the predicted values increase, then we have what is known as heteroscedasticity. This means that the variability in the response is changing as the predicted value increases. This is a problem, in part, because the observations with larger errors will have more pull or influence on the fitted model. An unusual pattern might also be caused by an outlier. Outliers can have a big influence on the fit of the regression line. In this example, we have one obvious outlier. Many of the residuals with lower predicted values are positive (these are centered above the line of zero), whereas many of the residuals for higher predicted values are negative. The one extreme outlier is essentially tilting the regression line. As a result, the model will not predict well for many of the observations.

In addition to the residual versus predicted plot, there are other residual plots we can use to check regression assumptions. A histogram of residuals and a normal probability plot of residuals can be used to evaluate whether our residuals are approximately normally distributed. However, unless the residuals are far from normal or have an obvious pattern, we generally don't need to be overly concerned about normality.

Note that we check the residuals for normality. We don't need to check for normality of the raw data. Our response and predictor variables do not need to be normally distributed in order to fit a linear regression model. If the data are time series data, collected sequentially over time, a plot of the residuals over time can be used to determine whether the independence assumption has been met. But this generally isn't needed unless your data are time-ordered.

So what do we do if we see problems in the residuals? How do we address these issues? We can use different strategies depending on the nature of the problem. For example, we might build a more complex model, such as a polynomial model, to address curvature. Or we might apply a transformation to our data to address issues with normality. Or we might analyze potential outliers, and then determine how to best handle these

outliers. For the most part, these topics are beyond the scope of this course, and we recommend consulting with a subject matter expert if you find yourself in this situation. However, we will discuss one approach for addressing curvature in an upcoming video.

Let's return to our cleaning example.

We fit a model for Removal as a function of OD. The bivariate plot gives us a good idea as to whether a linear model makes sense. The observations are randomly scattered around the line of fit, and there aren't any obvious patterns to indicate that a linear model isn't adequate.

Let's take a look at the residual plots. We'll see how to generate these plots in the next video demonstration. In the residual by predicted plot, we see that the residuals are randomly scattered around the center line of zero, with no obvious non-random pattern. And, although the histogram of residuals doesn't look overly normal, a normal quantile plot of the residual gives us no reason to believe that the normality assumption has been violated. The residual by row number plot also doesn't show any obvious patterns, giving us no reason to believe that the residuals are auto-correlated. Because our regression assumptions have been met, we can proceed to interpret the regression output and draw inferences regarding our model estimates. We see how to conduct a residual analysis, and how to interpret regression results, in the videos that follow.

Close