More Regression

When we have fitted an LS line to our data plot, there are a lot of things we can (and should!) do to test whether or not our model gives a good description of the linear relationship between our variables. We have talked about this in lecture, now let’s do it in JMP-IN.

1. **Test for Parameter Estimates**

   Last time we plotted our response variable versus our predictor variable, then fitted a LS line to the plotted data. This gives us a regression line, describing the linear relationship between predictor and response, and the equation for the line.

   In the output, we also find a table with the title “Parameter Estimates.” This table gives us the estimates for the parameters in the model, \( b_0 \) (the y-intercept) and \( b_1 \) (the slope of the line). In general, we are more interested in \( b_1 \) than in \( b_2 \), since the slope describes how the predictor affects the response.

   The parameter estimates both come with a t-statistic and a corresponding p-value for testing \( H_0 : \beta_1 = 0 \) (or \( H_0 : \beta_0 = 0 \)). We reject the null hypothesis if the p-value is less than \( \alpha \). If we fail to reject \( H_0 : \beta_1 = 0 \), we conclude that there is no relationship between predictor and response.

2. **Residual Plots**

   We talked about residual plots briefly last week. Whenever you do regression you should always check the residual plot.

   Residual plots are created to ensure that the model is not systematically over- or under estimating for certain values of the predictor variable or for the estimated values.

   To plot the residuals versus the predictor, click that red triangle next to “Linear Fit” and choose “Plot Residuals”. A scatter-plot will be displayed at the bottom of the window. Make sure that there is no systematic dependence of the sign or the magnitude of the residuals on the predictor.

   For simple linear regression, a plot of the residuals versus the predicted values will give the same information as the previous plot. To create this plot, choose “Fit Model” under “Analyze” in the main menu. Choose the response variable and click the button “Y.” Then choose every predictor variable in the model (in simple linear regression, there will be only one) and click the “Add” button. Finally, click “Run Model.” At the bottom of the new window, you will find the plot.
3. **Check Normality**

The normal assumption can be evaluated visually with a normal quantile plot of the residuals. To test more formally, you should use the Shapiro-Wilk test. Save the residuals under “Linear Fit” in the “Fit Y by X” window. Then use the “Distribution” platform to do the necessary tests.

4. **Check for Outlying and Influential Observations**

First of all, I want you to note the difference between an outlying and an influential observation. The response for an outlying observation will fall far from the fitted line. Hence, outliers will have large positive or negative residuals. Outliers can off course influence the regression line, but they are always poorly fitted by the model.

Influential points play a very important role in determining the position of the LS line. These data points can – but do not have to be outliers. An influential point will move the LS line so that the point is well fitted by the model.

A standard measure of the influence that individual cases have on the LS line is called “Cook’s Distance.” This method compares the full model to the same model with the suspect observation excluded. If the two models are similar, the point is not influential. If the models are significantly different, the observation may be influential.

To compute Cook’s D values, run the regression model as normal. Be sure to use the “Fit Model” platform, even if you only have one predictor. Click the red triangle at the top of the window and choose “Save Columns,” then “Cook’s D Influence” near the bottom of the list. Nothing will happen in the output window, but in the data table a new column appears containing the Cook’s D values.

Scan the list and find the highest value(s). Is it large compared to the others? If so, then the corresponding case has high influence. If there is not much difference between the highest value(s) and the majority of the data, then by the criteria set by the Cook’s D measurement, there are no influential points.

If you determine that a point has a great influence over the model, remove it from the analysis and run the model again. Now you decide; is the new model better than the first one? This is your choice!

If you choose to remove the influential point from the final model, make sure that you give some justification in the context of the problem. Using real-world reasons, why is it acceptable to toss the observation? It is not good practice to simply ignore outliers to make the model better.