Live Session Week 4 Agenda

1. Wage data: more methods to check assumptions
2. Analysis of customer churn data
3. Simulation exercise

You will likely need the R files for weeks three and four from the asynchronous materials (the estimation and inference scripts).

1 Wage Data Diagnostics

This activity uses the workspace *wage1.Rdata* from the asynchronous materials. The workspace contains the data frame *data* (the actual data) and the data frame *desc* (description of the data). This activity uses the linear model object *model1* from the week 4 R script.

1. Normality: extract the residuals from the linear model object with *model1$resid* or *resid(model1)*. Then, make a histogram (or density estimate) or Normal probability plot using the function *qqnorm* to make the plot and then *qqline* to add the reference line.

2. Constant variance: extract the fitted values from the linear model object with *model1$fitted* or *fitted(model1)*. Make a *plot* of the fitted values (x) and residuals (y). You can also use the function *ncvTest* in the library *car* to provide a p-value for a hypothesis test. You can also use the following types of residuals.

   - Standardized residuals: internally standardized residuals; extract with the function *rstandard*
   - Studentized residuals: externally standardized residuals; extract with the function *rstudent*

   When using these, add to the plot reference lines at 2 and -2 (with the function *abline*). Observations outside of these bounds are called outliers. You can use the function *which* to identify the case/row/observation number. You can also use the function *outlierTest* in the package *car* to do this. Note that when the errors exhibit heteroskedasticity, use the functions *coeftest*, *waldtest* and *linearHypothesis* for inference.

3. Independence: if your data are time-ordered, use the function *acf* to obtain an auto-correlation plot of the residuals. The second half of the course addresses this is detail. Are these data time ordered?

4. Plot the residuals vs. the variables in the model. If there are any clear trends, patterns or groupings, this is an indication the form of the model is not correct for that particular variable.

5. Plot the residuals vs. the variables not in the model (in particular grouping variables). Look for groups of residuals that do not have a mean of zero. Use the function *tapply* to apply the function *mean* to the residuals by another (categorical) variable. For example, try this with the female and married variables.

6. Leverage: leverage measures how far away a given observation is from the center of the predictor variables a given, values far from the center often have a lot of influence in the estimation process (e.g. think about a teeter-totter). Use the function *hatvalues* to extract the leverage values. Observations with leverage or hat values larger than $\frac{2*(k+1)}{n}$ (where k is the number of predictors and n is the sample size) are said to have large leverage. You can also use the function *influencePlot* in the library *car*.

7. Influence: an observation is influential if, when the observation is removed, the estimated model changes significantly. Cook's Distance combines leverage and residual information. Use the function *cooks.distance* to extract the values and then make a plot of them. Identify values that stand out relative to the majority of values. You can also use the function *influencePlot* in the library *car*.

2 Customer Churn

The R workspace *customerChurn.Rdata* contains the data frame *customerChurn*. These data are a sanitized use case from IBM Watson Analytics. Here you will explore the relationship between the amount of time individuals are customers (tenure, in months) and their total charges over the lifetime of their accounts (TotalCharges, in dollars).

1. Examine the structure of the data and conduct an EDA (recall our last live session).
2. Fit a regression model that uses tenure to predict the total charges. Identify R^2 (reported as a proportion). Make sure that you can interpret the model coefficients.
3. Check the model assumptions and look for unusual observations (outliers, large leverage, influence).
4. Are you able to assess independence by considering an autocorrelation plot of the residuals? Explain.
5. Which of the omitted categorical variables may explain any violation of the zero-mean condition?
6. I just bought a house and I don't plan on going anywhere. Assuming this model is correct, what can I estimate/predict my total charges to be in 20 years? Hint: use the function *predict.lm* and maybe get a confidence and prediction interval. Is this an example of extrapolation or interpolation.
7. What is a proxy for the average charges per month? Use a hypothesis test to determine if the average monthly charge is more than \$75.
8. Fit a regression model that uses tenure, gender, and if they have a partner to predict the total charges. Check the model assumptions and look for unusual observations.
9. Use a hypothesis test to determine if the additional variables can be removed from the model.

3 Simulating Problematic Data

The asynchronous materials contain examples of how to simulate data from a regression model. Here, you will simulate data where the conditions are not satisfied. Simulate data in each of the below conditions. Then fit a regression model and verify the condition is violated.

1. Non-Normal residuals
2. Heterskedastidicy
3. Zero mean: non-linear and missing a grouping variable
4. Endogientiey
5. Serial correlation (hing: consider an autoregressive model)