**Psych 610/710**

**Homework 8: Due 1 November 2017, 5:00pm**

**Reading questions**

Answer these in your Word doc.

1. Judd, McClelland, and Ryan talk of a “paradox” in using standardized residuals to detect outliers. Describe this paradox, what the authors recommend using to solve it, and list the three reasons they recommend this solution.
2. Of the five major assumptions of linear regression, which two are the most problematic to violate? Why are these particularly concerning? What are solutions to violating these assumptions?



1. For each of these quantile-quantile plots (A-C), describe what the distribution of scores look like on the raw data. Then, in 3 sentences or less, explain the logic of a Q-Q plot.

**Data Analysis**

The data for this homework are located in the cars package, so make sure the cars package is loaded. The dataset is called UN. For details, ?UN.

Per capita GDP is a measure of the total output of a country that takes [gross domestic product](http://www.investopedia.com/terms/g/gdp.asp) (GDP) and divides it by the number of people in the country. It is a relative measure of a country’s wealth. Infant mortality is, unfortunately, exactly what it sounds like.

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Variable name | Description | Values |
| 1 | infant.mortality | Infant deaths per 1000 live births | 2 – 169 |
| 2 | gdp | GDP per capita, in 1998 US dollars | 36 – 42416 |

1. Get acquainted with the data through univariate descriptive stats and by plotting the distributions of each variable. Note anomalies, if they exist.
2. Estimate a model to predict infant mortality from national GDP.
3. Conduct a full case analysis of this model. Specifically, note in your script any points you consider to have high leverage, to be a model outlier, to have influence on the model as a whole, and to have influence on the parameter estimate of GDP specifically. Remove outliers with high influence.
4. Refit the model with these outliers removed.
5. Check for violations of model assumptions.
6. It turns out that assumptions were violated. So, let’s transform the full dataset (i.e., before you removed any outliers). What is the suggested Box-Cox transformation? According to what you learned in lecture and lab, what might be a better transformation?
7. Apply an appropriate transformation to infant mortality and refit the model. Then, check for violations of model assumptions. Summarize your findings in a sentence in your R script.
8. Apply an appropriate transformation to GDP and refit the model. Then, check for violations of model assumptions. Summarize your findings in a sentence in your R script.
9. Run a new model using the transformed versions of both variables. Check for violations of model assumptions and summarizing your findings in a sentence.
10. Let’s assume that transforming both variables is the best solution to model assumption violations. Repeat a case analysis of the model where both variables are transformed, and note any points with leverage, any model outliers, and any points that have undue influence.
11. Remove these outliers and refit the model. Interpret the coefficient associated with GDP (remember, *both* variables are transformed in this model).
12. Compare the standard error in this final model with transformed variables to the standard error in a model with raw variables. Why do these differ? Answer this question in your R script.
13. Plot the effect of GDP on infant mortality using the model where both variables are transformed. Insert this graph into your word document.
14. Create another graph, in which you include a curvilinear fit line based on the model in which only GDP has been transformed. Insert this graph into your Word document.
15. Write a short paragraph reporting these results. Explain the transformations you made, and why you made them. Explain your case analysis, and why you chose to exclude the data that you did. Finally, interpret your model using all stats (i.e., report the effect of GDP on infant mortality, the confidence interval, the t value, degrees of freedom, a variance based effect size, and the p value).