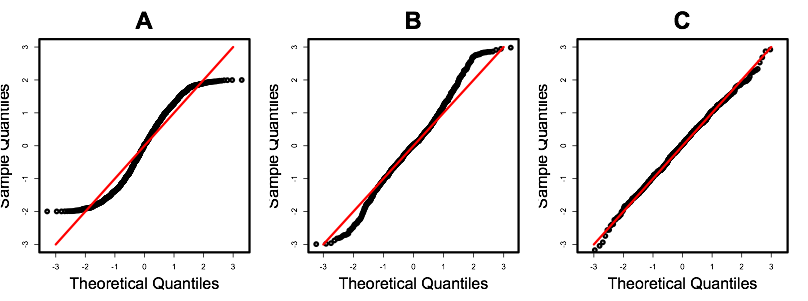
**Psych 610**

**Homework 11: Due November 29th 2023, 1:30pm**

**Reading questions**

1. Judd, McClelland, & Ryan talk of a “paradox” in using standardized residuals to detect outliers. Describe this paradox, present what the authors recommend using to solve it, and list the three reasons they recommend this solution.
2. Of the assumptions of linear regression that refer to residuals, which two are the most problematic to violate in large samples? Why are these particularly concerning? What are solutions to violating these assumptions?
3. 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 were adapted from a dataframe in the *car* package. For this HW, read in and use the data from the “HW11Data” file provided on Canvas.

Each row in the data represents a country in the United Nations (Country). Infant mortality (infantMortality) is the rate of death in infancy among live births. Per capita GDP (ppgdp) is a measure of a country’s wealth 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. Both GDP and per capita GDP should be positive numbers.

|  |  |  |
| --- | --- | --- |
| Variable name | Description | Values |
| infantMortality | Infant deaths/1000 live births | 1.92 – 124.53 |
| ppgdp | GDP per capita, in US dollars | 114.8 – 105095.4 |
| Country | Name of the country | … |

1. Explore the data!
   1. Read in the raw data and change the variable names to Snake Case.
   2. Remove unrelated variables, keep only *infant\_mortality*, *ppgdp*,and *country*.
   3. Obtain descriptive statistics for *infant\_mortality* and *ppgdp*.
   4. Create quick-and-dirty histograms of *infant\_mortality* and *ppgdp*.
   5. There are two obvious outliers with invalid values on variable *infant\_mortality* and *ppgdp*. Find them and assign NA to replace the invalid values.
   6. Create quick-and-dirty histograms of *infant\_mortality* and *ppgdp* once more. Are these variables normally distributed? If the *variables* are not normally distributed, would it always be a problem by itself? Why or why not?
2. Conduct the analysis WITHOUT data transformation or case analysis.
   1. Fit a model to predict infant mortality from per capita GDP.
   2. Report the results in one sentence, including relevant statistics.
3. Check for violations of model assumptions
   1. Make a quick-and-dirty scatterplot of the relationship between infant mortality and per capita GDP.
   2. Check for violations of model assumptions on the model you fit above. Write down whether you think each of the model assumptions was violated. Explain how you used the graphs from modelAssumptions to arrive at this conclusion.  
      *Note*: if the plots don’t show up, run the code in the console at the bottom of R.
4. It turns out that the model assumptions were rather severely violated. So, let’s transform the variable(s). What is the suggested Box-Cox transformation? According to what you learned in lecture and lab, what might be a better transformation?
5. Log-transformation of the outcome variable
   1. Apply log-transformation to *infant mortality* and refit the model. Summarize the results of the model in a sentence, *including a practical interpretation of the coefficient*.
   2. Check for violations of model assumptions again on the refitted model above. Did log-transforming infant mortality solve the problems?
6. Log-transformation of the predictor
   1. Apply log-transformation to *per capita* *GDP* and refit the model to predict *untransformed* infant mortality. Summarize the results of this model, *including a practical interpretation of the coefficient*.
   2. Check for violations of model assumptions again on the refitted model above. Did log-transforming per capita GDP solve the problems?
7. Log-transformation of the outcome variable and the predictor
   1. Fit a model where you use *log-transformed* per capita GDP to predict *log-transformed* infant mortality. Summarize the results of this model, *including a practical interpretation of the coefficient*.
   2. Check for violations of model assumptions again on the model you fit above. Did log-transforming both infant mortality and per capita GDP solve the problems?

Use the model you fit in question 7a to answer the questions below.

1. Conduct a full case analysis with the model you just picked. Specifically, write down which countries you would consider to have high leverages, to be model/regression outliers, to have excessive influence on the model as a whole, and to have excessive influence on the parameter estimate for log-transformed per capita GDP specifically.
2. Remove outlier(s)
   1. Remove any observations of countries which you think are outliers.
   2. Justify why you removed the observation(s). How would it/they impact your conclusion regarding the relationship between infant mortality rate and per capita GDP, should it/they be kept in your model?
3. Refit the model in 7a but with the outlier(s) removed. What changed in your results?

Use the model you fit in question 10 to answer the questions below.

1. Make two publication-quality graphs that show the effect of GDP on infant mortality. In one graph, plot both the transformed variables [a linear relationship between log2(infant mortality) and log2(ppgdp)]. In the other graph, plot both the untransformed variables [a curvilinear relationship between infant mortality and ppgdp].
2. Write a short paragraph reporting these results. Explain the transformation you made, and why you made it. Explain your case analysis, and why you chose to exclude the data that you did. Finally, interpret your model using all stats in APA format (i.e., report the effect of GDP on infant mortality, the F value, degrees of freedom, a variance-based effect size, and the p value).
3. How long did this week’s homework take you?