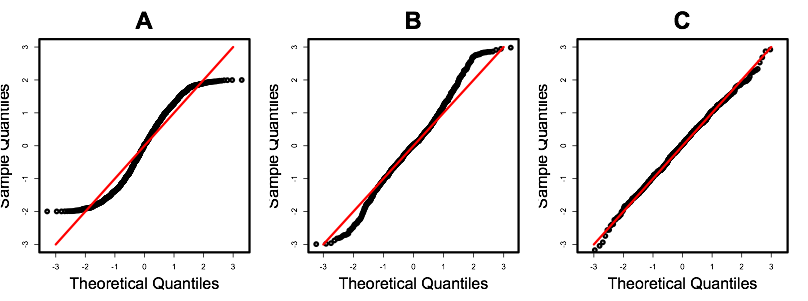
**Psych 632**

**Homework 7: Due 3 November 2020, 11:59pm**

**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.

**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 | InfantMortality | Infant deaths per 1000 live births | 2 – 169 |
| 2 | ppgdp | GDP per capita, in 2009-11 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. Hint: If you just used varPlot for this, then you would get an error. This is because there are missing values in the data frame. To get around this, you can use na.omit. Here is the code to plot each variable and have varPlot ignore the NAs:

varPlot(na.omit(d$infantMortality))

varPlot(na.omit(d$ppgdp))

1. Estimate a model to predict infant mortality from national GDP. Name this model ‘m1’.
2. 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. Save the full dataset into a new variable called “dFull” and remove outliers with high influence.
3. Refit the model with these outliers removed. Name this new model ‘m2’.
4. Check for violations of model assumptions.
5. It turns out that assumptions were violated. So, let’s transform the full dataset (i.e., dFull from 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?
6. Apply an appropriate transformation to infant mortality and refit the model. Name this model ‘mt1’. Then, check for violations of model assumptions. Summarize your findings in a sentence in your R script.
7. Apply an appropriate transformation to GDP and refit the model. Name this model ‘mt2’,Then, check for violations of model assumptions. Summarize your findings in a sentence in your R script.
8. Run a new model using the transformed versions of both variables. Name this model ‘mt3’. Check for violations of model assumptions and summarizing your findings in a sentence.
9. Normally you would want to check this new model (with the transformed variables) for outliers but you can skip this step by simply writing “OK” as a comment in your R script.
10. 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.
11. At the end of this document, there is example code you could use to plot the effect of GDP on infant mortality using the model where both variables are transformed. This code is for log2 transformation, but you could modify the highlighted code to work with whatever transformation you chose. We are not asking you to do this for this homework but write ‘OK’ in your R script to confirm that you are now aware of this code in case you need to use it in the future.
12. At the end of this document, there is example code you could use to create another graph, in which you include a curvilinear fit line based on the model in which only GDP has been transformed. This code is for log transformation but you could modify the highlighted code to work with whatever transformation you chose. We are not asking you to do this for this homework but write ‘OK’ in your R script to confirm that you are now aware of this code in case you need to use it in the future.
13. 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).

Code for question 12:

Yhat = seq(6, 16, length = 100)

dMortality <- data.frame(ppgdpLog2=Yhat)

dMortality <- modelPredictions(mt3, dMortality)

scatterplot <- ggplot() +

geom\_point(data=dFull, aes(x = ppgdpLog2, y = infantMortalityLog2), color = "black") +

geom\_smooth(aes(ymin = CILo, ymax = CIHi, x = ppgdpLog2, y = Predicted),

data = dMortality, stat = "identity", color="red") +

theme\_bw(base\_size = 14) +

scale\_x\_continuous("GDP (log2)", breaks = seq(6, 16, by=1)) +

scale\_y\_continuous("Infant Mortality (log2)", breaks = seq(0, 8, by=1)) +

labs(title = 'Effect of GDP on Infant Mortality')

scatterplot

Code for question 13:

scatterplot <- ggplot(data=d, aes(x = ppgdp, y = infantMortality)) +

geom\_point(color = "black") +

stat\_smooth(method = 'lm', formula = y~log(x)) +

theme\_bw(base\_size = 14) +

scale\_x\_continuous("GDP per capita", breaks = seq(0, 45000, by=10000)) +

scale\_y\_continuous("Infant mortality rate", breaks = seq(0, 200, by=25)) +

labs(title = 'Effect of GDP on Infant Mortality')

scatterplot