Homework 8 key

1. The paradox is that extreme data values tend influence the model in a way that causes the model to adapt and minimize error for those extreme values when estimating the parameters. We are asking whether an extreme value is unusual with respect to the model but that model was made by estimating that same extreme value. The authors recommend using studentized deleted residuals (AKA just studentized residuals) to solve this. The three reasons are: studentized residuals solves the scaling problem and eliminates the paradox, the studentized residual has a natural interpretation in the model comparison framework, and it is very unlikely the studentized residuals would miss an unusual value that other transformed residuals would catch.
2. Most problematic model assumptions?

When you narrow it down to the assumptions you can test when assessing the residuals, then the book says normally distributed residuals and constant variance because of increased risk of both Type I and Type II error when these assumptions are violated. The book recommends transformations as a solution. If you answered this question based on what we explicitly said in lab and lecture you may have said Independence of errors because violation of this assumption gives inaccurate SEs, and GLM is not robust at handling non-independence. In that case you could use a repeated-measures design or multi-level modeling as a solution. You may have also answered exact X as violations of this increase standard error and lead to inaccurate regression coefficients.

1. Shape of these distributions, logic of a quantile-quantile plot



**A**:

Thin-tailed distribution. Extreme scores show a flatter slope; this indicates that scores on the ends of the distribution are not as extreme as they should be.

**B**:

Thick-tailed distribution. Extreme scores show a steeper slope; this indicates that scores on the ends of the distribution are more extreme than they should be.

**C**:

Normal. The line is straight; scores are where they should be for a normal distribution.

A Q-Q plot shows the percentile distribution of the data’s residuals against the percentile distribution that would be expected if the distribution were normally distributed. The vertical axis shows the rank-ordered percentile scores for each observation; the horizontal axis shows the z-scores for each percentile from the normal distribution. If the data are normally distributed, points in the Q-Q plot lie along a straight line.

**Data analysis**

****

****

In order to test the relationship between national GDP and infant mortality rates, it was necessary to transform the data to better adhere to the assumptions of the general linear model. Namely, both variables were positively skewed, but to different extents. The residuals of a model fit using the untransformed variables had very non-normal and inconsistently distributed residuals, and clearly indicated a nonlinear relationship. However, the relationship was simple and monotonic and could be addressed by log transforming both variables. Before interpreting this model, however, we noticed that a handful of nations had excessive influence on the model's parameter estimates: namely, Sao Tome, Sudan, Bosnia, Iraq, and Tonga. Sao Tome, Sudan, Bosnia, and Tonga each had markedly lower rates of infant mortality than would be expected given their GDP, which biases the relationship between GDP and infant mortality to be less than it would otherwise. We reported the model with these five nations removed, however whether these nations were included or excluded did not change the overall relationship. A nation's GDP was found to be a strong predictor of that nation's infant mortality rate overall: approximately, every time a nations' GDP quadruples, the model predicted that infant mortality would be halved (95% CI[-.57,-.48] , t(201) = -22.05, p < .001). In our sample of nations, log GDP accounted for 72% of the variance in log infant mortality ($η\_{p}^{2}$ = .72).