Homework 5

Aiden Kenny
STAT GR5205: Linear Regression Models
Columbia University

Novermber 25, 2020

Question 1

Collaborators: None

(a) Let Y be the number of nurses in the hospital and let X be the available faculty and services. The left and middle panels of Figure 1 show the histograms of Y and X, respectively. We see that Y is skewed right, while X appears to be normally-distributed. In addition, the scatterplot of Y vs. X, which is in the third panel of Figure 1, shows that there is a nonlinear relationship between Y and X. All of these indicate that Y is suitable for a data transformation. Specifically, we would like to perform a power transformation on Y in order to make it closer to a normal distribution.

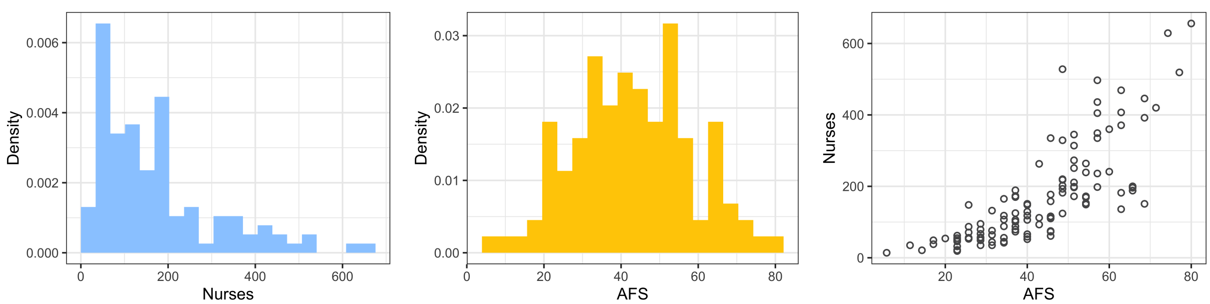

Figure 1: Histograms of Y and X and a scatterplot of Y vs. X.

(b) The power transformation function and its scaled counterpart are defined as

$$f_\lambda(Y) = \begin{cases} Y^\lambda & \text{if } \lambda \neq 0 \\ \log Y & \text{if } \lambda = 0. \end{cases} \quad \text{and} \quad g_\lambda(Y) = \begin{cases} (Y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log Y & \text{if } \lambda = 0. \end{cases}$$

When transforming Y, we first use the scaled power transform to fit the model $g_\lambda(Y) = \beta_0 + \beta_1 X + \epsilon$ in order to determine an optimal value of λ via maximum likelihood estimation. We then use f_λ to make Y closer to a normal distribution and perform any subsequent inferences. In vector form, our model is $\mathbf{g}_\lambda(\mathbf{y}) = \beta_0\mathbf{1} + \beta_1\mathbf{x} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and $\mathbf{g}_\lambda(\mathbf{y})$ is the transformed response (so $[\mathbf{g}_\lambda(\mathbf{y})]_i = g_\lambda(y_i)$) for some unknown λ. It can be shown (see Appendix A) that the log-likelihood function can be expressed as a function of only λ,

$$m(\lambda) := -\frac{n}{2}\log\left(\frac{2\pi \mathrm{e}}{n}\right) - \frac{n}{2}\log\left(\mathbf{g}_\lambda^T(\mathbf{y})(\mathbf{I} - \mathbf{H})\mathbf{g}_\lambda(\mathbf{y})\right) + (\lambda - 1)\sum_{i=1}^{n}\log(y_i),$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the hat matrix. $m(\lambda)$ has been plotted in the left panel of Figure 2, where we can see that the MLE is maximuzed at $\lambda = 0.085$.

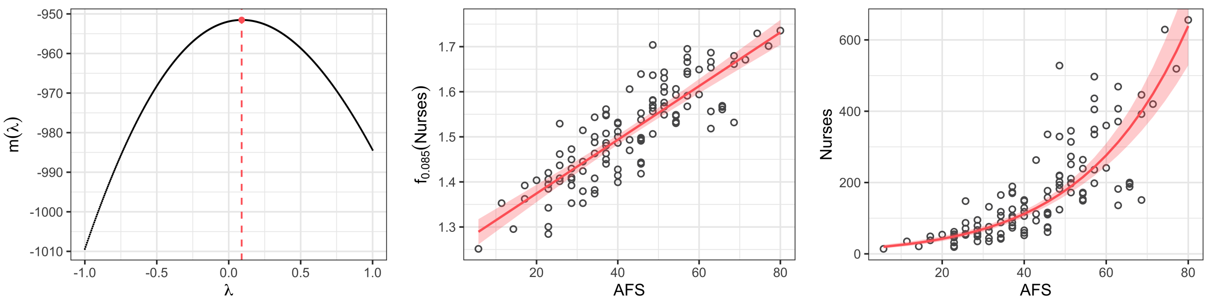

Figure 2: Relevant plots for the power transformation of Y.

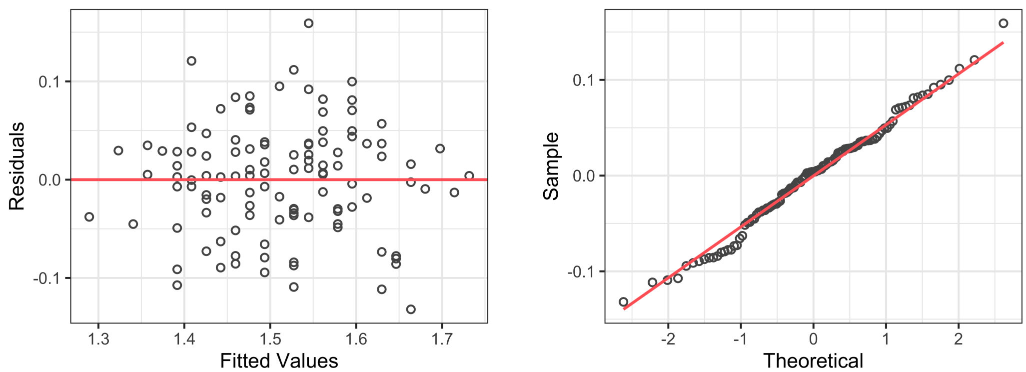

Figure 3: Diagnostic plots for the power transformation model.

(c) Now that we have our optimal λ, we will fit the model $f_{0.085}(Y) = \beta_0 + \beta_1 X + \epsilon$. We have $\hat{\beta}_0 = 1.255$ and $\hat{\beta}_1 = 0.005953$. A plot of this model, along with a 95% confidence interval, has been printed in the middle panel of Figure 2. This means that when X increases by 1 unit, $f_{0.085}(Y)$ is expected to increase by 0.005953 units. We can also make inferences about Y itself; we have $Y = f_{0.085}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 X) = \left(1.255 + 0.005953X\right)^{1/0.085}$. A plot of this model has been printed in the right panel of Figure 2. We must keep in mind that this model is non-linear, so it's expected rate of change will depend on the value of X.

(d) Both diagnostic plots can be found in Figure 3, and the data seems to confirm the model assumptions very well.

(e)