Spring 2023 **Lealand Morin**

University of Central Florida
College of Business

QMB 6912
Capstone Project in Business Analytics

Problem Set #3

Due Date: Sunday, 5 February 2023, at 11:59 PM.

In this problem set, you are to analyze the dependent variable, before including it in statistical models that contain explanatory variables. The first transformation to consider is the logarithmic transformation. Revisit your analysis of the densities of these two alternatives for the dependent variable. To determine how far these densities differ from that of the normal distribution, create a pair of QQ-plots, one for the original variable and one for the logarithm of the variable. Which one is closer to the normal distribution? Do you notice any notable deviations from normality?

Now consider a family of transformations that contains the two alternatives above. The Box–Cox specification uses a transformation of the dependent variable to change the distribution of the variable. The Box–Cox transformation can produce a distribution that is similar to a Gaussian (normal) distribution, which for a variety of reasons is ideal for statistical inference. Under the Box–Cox transformation of r_n, the rate of return on vehicle n,

$$\Lambda(r_n) \equiv \begin{cases} \frac{r_n^\lambda - 1}{\lambda} & \text{if } \lambda > 0 \\ \log r_n & \text{if } \lambda = 0. \end{cases}$$

This transformation is implemented in the function `BoxCox_Trans` in the sample code found in the `demo_03` folder in the course repository.

The dependent variable can be decomposed into a location parameter μ^0 and an error U, so

$$\Lambda(P_n) = \mu^0(\lambda) + U_n,$$

where the U_ns are independent, mean-zero, constant-variance $\sigma^2(\lambda)$, Gaussian (normal) errors. In the above equation, for clarity, the dependence of μ^0 and $\sigma^2(\lambda)$ on λ is made explicit. The function `log_like_uni` calculates the logarithm of the likelihood function as a function of the parameter λ, when the dependent variable is defined by the Box–Cox transformation, and the errors are Gaussian (normal). This function centers the data around the mean $\mu^0(\lambda)$, as in the equation above, by estimating the appropriate μ^0 for each value of λ; do the same for $\sigma^2(\lambda)$. It also includes the Jacobian term, which is required when a random variable is transformed. Using the data in `HomeSales.dat`, graph the logarithm of the likelihood function with respect to λ and find the optimal value of λ. To simplify the analysis, use the data from both rental and owner-occupied sales in one sample.

Use the likelihood ratio statistic to decide between transforming the dependent variable `price` using either the linear or the logarithmic specification; that is, between $\lambda = 1$ or $\lambda = 0$. Also consider the possibility of choosing the maximum likelihood estimate of λ but consider the practical implication of complicating the regression modeling to follow with a nonstandard transformation.

Create another QQ-plot using the Box–Cox transformation with the optimal value of *lambda*. Include this plot in another pair of QQ-plots, one with the optimal transformation and the other with the next-best choice among the $\lambda = 0$ or $\lambda = 1$ alternatives considered above. Now which one is closer to the normal distribution? Do you still notice any notable deviations from normality? Which version of the dependent variable do you recommend?

Verify your solution using the `boxcox` functions in either the `MASS` package or the `car` package with the default formula `price ~ 1`. Extend your analysis by passing a formula to these functions that includes explanatory variables to predict `price`. For now, estimate only a rudimentary regression model with some variables included; you will improve this model with insight gained through the next several problem sets. In this specification, use the Box–Cox transformation to find a value of λ such that the *residuals* of the model are approximately normally distributed. Would you change your recommendations after this analysis?

Prepare and compile your work in LaTeX and include scripts for any of the calculations in R. In particular, create the following directory structure, separate from your existing work:

- `Code/`
- `Data/`
- `Figures/`
- `Paper/`
- `Misc/`

In a file called `README.md`, which should also live in the directory containing the above folders, provide the instructions concerning how to run the executable shell script `DoWork.sh` (in the same directory) that will execute the code that produced all of the answers collected and documented in your report, which will live in the subdirectory `Paper/`. In the subdirectory `Code/`, keep the R code; in `Data/` keep the raw data file you downloaded, so that `DoWork.sh` can load it into R, and in `Figures/` keep any figures you created for your answers. Put anything else in the subdirectory `Misc/`. I should then be able to replicate all of your work simply by typing

- `$ ./DoWork.sh`

on the command line of a terminal window.

To provide you a template, which makes preparation easier for you and grading easier for me, I have placed sample LaTeX and R code in the GitHub repository for the course: `QMB6912S23`, under my GitHub username `LeeMorinUCF`; pull this repository and use these files a framework within which to create the answers for this problem set. Push the files to a folder on your GitHub repository and I will pull your submissions to my computer for grading.

Be sure to support your calculations with descriptions of what you were trying to do (for example, in comments in your R code as well as in the LaTeX explanations) because partial credit will be given.

Due Date: Sunday, 5 February 2023, at 11:59 PM.