Spring 2023 **Lealand Morin**

University of Central Florida
College of Business

QMB 6912
Capstone Project in Business Analytics

Problem Set #8

Due Date: Sunday, 26 March 2023, at 11:59 PM.

Revisit the empirical specifications you considered in Problem Set #7. Now consider some variables as candidates for arbitrary nonlinear specifications. As before, continuous variables are candidates for this sort of specification but ordinal variables such as `school_score` and `transit_score` are also worth considering, since these numeric variables also lie within a continuous scale.

To introduce your analysis, first compare the method of using a Generalized Additive Model with a model using the Box–Tidwell transformation.

Start with your regression specification from Problem Set #6 and estimate your model with all other variables included but leave out a candidate variable—a variable you will investigate for an arbitrary nonlinear relationship with house prices. Next, calculate the predictions from this adjusted model and obtain the residual house prices. Then, repeat the prediction exercise using the candidate variable as the dependent variable and all other explanatory variables to predict the candidate variable. Now perform a nonparametric regression of the residual house prices on the residuals of the candidate variable. Produce a plot of the nonparametric regression function. Repeat this exercise for each of your candidate variables, then observe the estimated relationships and determine whether any important nonlinearities arise.

Now analyze a generalization of the single-index function in the GLM, which results in the Generalized Additive Model. Estimate the Generalized Additive Model with nonlinear functions for any variables that you determine are appropriate from your nonparametric analysis. The `mgcv` package is useful for this analysis. Do any of these models materially outperform either your recommended linear model from Problem Set #6 your recommended nonlinear model from Problem Set #7? Do the results suppport the conclusions of the linear model? Does the GAM expose any important constraints of the linear model that should be relaxed?

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/
- Tables/
- Text/
- 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. Similarly, keep any LaTeX scripts for tables in the Tables/ folder. You may put any written text in the Text/ folder, if not already included in a tex file in your Paper/ folder. 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, 26 March 2023, at 11:59 PM.