Spring 2023 **Lealand Morin**

University of Central Florida
College of Business

QMB 6912
Capstone Project in Business Analytics

Problem Set #11: Take-Home Examination

Due Date: Sunday, 30 April 2023, at 11:59 PM.

The data engineer for the realtors has gathered more information on home sales, and has organized another dataset concerning an additional 2,473 sales of detached houses. As before, the data engineer scraped data from similar real estate Websites and combined it with the records from the realtors to identify the houses that were available for rent during the year after each sale. Around one third of the houses were identified as rental units, while the other two thirds were not observed for rent and were presumably occupied by the owner. These data are contained in the file `HomeSales2.dat`, which is available in the `Data` folder of the course repository and on the course Webpage under Module 11.

Each house in the dataset is a row, while the columns correspond to the variables whose names and definitions are the following:

Variable	Definition
`year_built`	the year in which the house was constructed
`num_beds`	the number of bedrooms in the house
`num_baths`	the number of bathrooms in the house
`floor_space`	the area of floor space in the house, in square feet
`lot_size`	the area of lot on which the house was built, in square feet
`has_garage`	an indicator for whether the house has a garage
`has_encl_patio`	an indicator for whether the house has an enclosed patio
`has_security_gate`	an indicator for whether the property is accessed through a security gate
`has_pool`	an indicator for whether the property includes a pool
`transit_score`	an integer to represent the convenience of transportation options
`school_score`	an integer to represent the quality of the schools in the county
`type_of_buyer`	a categorical variable to indicate the type of buyer, either “Owner-Occupied” or “Rental”
`price`	the price at which the home was sold

The first several variables are self-explanatory but some of the variables above warrant some description. The last two integers, `transit_score` and `school_score` describe the amenities available to the residents in each house. The `transit_score` is an integer from one to ten that indicates the convenience of transportation options at the location of the house. It factors in the availability of public transportation options, as well as the proximity to business districts and major highways, with a higher score indicating a higher level of convenience. Similarly the `school_score` is an integer from one to ten that indicates the quality of the schools in the county. It is compiled from indicators of college attendance, test scores, as well as other measures of school and student performance. As with the `transit_score`, a higher `school_score` indicates better school quality. The variables listed above also include the dependent variable. The `price` variable is the price at which the home was sold.

Download the file `HomeSales2.dat`, and replicate the analyses of the previous ten problem sets. Write this up as a report (using LaTeX) that you could send to your supervisor. The written report in your submission should have sections mirroring the content of Problem Sets #1 to #10, as follows.

- An **Introduction** which should describe the business problem, the relevant economic literature relating to asymmetric information in the real estate market, the statistical framework relating to sample selection, and the relevance of characteristic theory and hedonic pricing models.
- An **Analysis of the Dependent Variable** including histograms and kernel-smoothed estimates of the distribution.
- **Transforming the Dependent Variable:** An investigation of the potential to transform the dependent variable using the Box–Cox transformation.
- **Summary Statistics:** A summary of the data in tabular form.
- **Data Visualization** to produce graphical displays with multivariate analysis relevant to your business question.
- **Linear Regression Models:** An analysis using the standard linear regression model, which may include any customized variables or interaction terms.
- **Transforming the Explanatory Variables:** A consideration of the Box–Tidwell transformation to explore any nonlinearity in the relationships between your explanatory variables and the dependent variable. .
- **Nonparametric Regression Models:** An investigation of the potential for arbitrary nonlinear functional forms within your modeling framework and its application within a Generalized Additive Model, to extend the Box–Tidwell specifications that you found useful.
- **Sample Selection Model:** A final model incorporating your functional form decisions into a model that accounts for sample selection.
- **Summary and Conclusions:** A brief summary of your analysis, your main findings, and your recommendation for the managers of the real estate agency.

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, 30 April 2023, at 11:59 PM.