MSDS/MSIT 5043
Principles of Statistical Analysis and Decision Modeling

Mr. Eric Stephens
College of Computing and Technology
Lipscomb University
e-mail: eric.stephens@lipscomb.edu
Phone: (615) 584-2979

Regression Project

PLEASE READ THESE INSTRUCTIONS CAREFULLY!!

You are the owner of a data science consulting firm, and you also own the building that your company currently occupies. However, since your business has grown significantly in the last two years, you now have to move into a larger building, as you no longer have any additional capacity (that is, your building has no vacant space). At first, you considered selling your current property, but because of the increasing demand for office space in your city you are now seriously thinking about renting it out instead. However, your decision depends heavily on how much rent you can charge, as you need to be able to offset the costs of ownership (and hopefully make a small profit). If you determine that you cannot cover your total expenses, you will sell the building.

In order to estimate what your building would likely rent for, you have gathered data on 100 commercial properties that have been rented in your city within the last six months (found in the file **`comm_prop.csv`**). The dataset contains the following seven variables:

- **RentRate**: Monthly rental rate of the property (in thousands)
- **Age**: Age of the property in years
- **OperExp**: Total monthly operating expenses (in thousands)
- **VacRate**: Vacancy rate of the building
- **SqFt**: Total square footage of the building
- **Taxes**: Total monthly tax expense (in thousands)
- **W2MiDT**: Whether or not the building is located within 2 miles of downtown (1=Yes, 0=No)

To understand the relationship each of these variables has with the rental rate—as well as to be able to determine the rental rate for your specific property—you need to build a regression model using this data.

For this assignment, you will need to create an R markdown file that both shows the code you executed, as well as describes your analysis process and the findings from it (both from a statistical and a practical point of view). There is no specific length requirement for this assignment, but please try to keep it as brief as possible while still covering all of the required points.

Your analysis should consist of the following (see rubric on the following page for point details):

1) **An examination of the distributions of the variables**. You don't have to provide every element of the distribution for each variable, but do make note of any that look particularly skewed or highly unusual in some way. (5 points)

2) **An examination of the correlations for all of the continuous variables**. Review both the correlation matrix (with the actual figures) and the scatterplot matrix to help you identify the strength and direction of the relationships. Again, you don't have to provide every single correlation in your analysis, but do be sure to point out any that are moderately to very strongly correlated. (5 points)

3) **The identification and evaluation of a suitable regression model for predicting rental rates based on the other variables in the dataset**. Follow the guidelines in the document "Basic Regression Modeling Process" to build the model (Hint: finding an appropriate model for this project will require a couple of iterations of steps 6 through 8). For this assignment, only consider the individual variables as possible predictors (i.e., don't worry about interactions). As you iterate through the process, be sure to indicate any variables you drop from the model along the way, and why. ***[INSERT EXTRA CREDIT STEPS HERE (IF DESIRED); SEE PAGE 5]***

   Once you've settled on a model, report its adjusted R^2 value, its F statistic, the p-values of the t tests for each parameter estimate, your analysis of the residual plot, and the values of the parameter estimates (i.e., model coefficients) for each predictor variable. Concerning the adjusted R^2 value, provide your interpretation of it, as well as what important variables you think might be missing from your dataset. (50 points)

4) **The application of the model to your particular situation, and the resulting decision.** Once you have built the model, you need to apply it in order to determine an estimated rental rate for your property. Your building is 9 years old, has a total square footage of 40,000, incurs $13,000 in operating expenses and $540 in taxes per month, and is **not** located within two miles of downtown. Determine your estimated monthly rental rate based on your regression model (*including a prediction interval on your estimate*), and whether you decide to a) keep the building to rent, or b) sell it. [**Note**: You may or may not need to use either or both your operating expense amount or tax amount as inputs in your regression model, but you must take both costs into account when comparing your total monthly outlay against your estimated monthly rental income.] (25 points)

5) **A well-constructed R markdown file** that clearly outlines your modeling process. (15 points)

I will be looking for the following two elements in my evaluation and grading of your submission:

- Your ability to properly conduct and interpret the various statistical methods
- Your ability to clearly relate and apply the results of your analysis to the issue at hand

Once your report is complete, please publish the markdown file (as you did with the R exercise), enter the URL in the text box on Blackboard, and also attach your .Rmd file.

This project is due no later than Tuesday, May 2 at 11:59pm.

Project Scoring Rubric

Area of Focus	Element	Points Possible	Score
Variable distributions	Inclusion of meaningful plots	2	
	Comments	3	
Correlations	Inclusion of meaningful plots	2	
	Comments	3	
Building the model	Evaluation of potential predictor variables	10	
	Identification of evidence of potential multicollinearity	5	
	Review and interpretation of diagnostic plots and formal tests	10	
	Interpretation of R^2 and ANOVA for final model	10	
	Commentary regarding potentially important variables that are missing from the dataset	5	
	Evidence of iteration through the model building process	10	
Applying the model	Proper use of the final model (as a mathematical function)	10	
	Interpretation of the result within the context of the given business problem (i.e., providing your decision and your rationale behind it)	10	
	Additional comments (and potential shortcomings) regarding the suitability of the model for the problem	5	
Markdown file	Output is well designed (use of proper headers, plots are easy to interpret, etc.)	5	
	Thought process is easy to follow throughout the entire document	10	
TOTAL		**100**	

Extra Credit

You can gain up to **10 extra points** on your overall course grade by extending the project to consider other variable selection methods as described below. This should be considered an extension of the first paragraph of Part 3 of the assignment instructions, and thus any results should be incorporated into the remainder of Part 3, as well as Parts 4 and 5 as appropriate.

- Use backward stepwise regression with ***all single order terms (excluding any that may cause multicollinearity) and all interaction terms included***, and report the results, as well as how the resulting model compares to the previous one. For the resulting model, report the adjusted R^2 value, the F statistic, the p-values of the t tests for each parameter estimate, your analysis of the residual plot, and the values of the parameter estimates (i.e., model coefficients) for each predictor variable. Also perform formal tests to validate the assumptions of normally-distributed residuals and constant variance.
- Use best subsets regression with ***all single order terms (excluding any that may cause multicollinearity) and all interaction terms included***, and explain which variables the output indicates should be included. Before simply fitting a model that includes the suggested variables, *think critically* about what you already know concerning the relationship between some of these variables, as well as the principle of hierarchy. Once you've decided which variables to include, use them to fit a third model (removing any variables that are non-significant or are not needed to preserve hierarchy) and report the results, as well as how the resulting model compares to the previous two models you fit.

At this point, evaluate the three models you've fit (these two plus the initial model that was required for the project), and determine which of three you will decide to use. Resume with the second paragraph of the Part 3 instructions to complete the project.

The following rubric will be used to determine the total number of extra points. Fractions of points may be awarded as necessary.

Area of Focus	Element	Points Possible	Score
Building the model	Application and interpretation of stepwise regression result	2	
	Application and interpretation of all subsets result	2	
	Review and interpretation of diagnostic plots and formal tests for the additional models	2	
	Comparison of all three models and selection of final model	4	
TOTAL		**10**	