Northeastern University
College *of* Professional Studies

Regression Diagnostics with R

Purpose of Assignment (WHY)

It is important for you to be able to interpret and evaluate the models that you build. In this assignment, you will fit two regression models, interpret the results and implement diagnostic techniques to identify and correct issues with the model.

Program Competencies

Program Learning Outcomes (PLOs)

1. Statistics & Math Demonstrate the foundational knowledge and skills critical to pursue data analytics as a profession in relation to statistics and math.
2. Analytics Systems Technology (Tools)/Advanced Analytics Demonstrate the knowledge of advanced tools in data analytics.
3. Business Analytics Agility Apply the principles, tools and methods of analytics to a comprehensive real-world problem or project related to data analyses for tactical and/or strategic decision making.
4. Business Process Management Integrate the major theories, tools, and approaches in data analytics to identify data-driven insights for informed business process management.
5. Communicating with Data Design and deliver presentations, reports, and recommendations that effectively translate technical results/data solutions and are coherent and persuasive to different audiences.

Course Learning Outcomes

This assignment is directly linked to the following key learning outcomes from the course syllabus:

CLO1: Fit, interpret, and evaluate regression models using standard functions and diagnostic techniques.
CLO2: Correct issues with overfitting, linearity, multicollinearity and outliers.
CLO3: Select best model from multiple predictors using automated techniques.

Assignment Description (WHAT)

For this exercise, you will need to download the attached AmesHousing dataset. In this assignment, you will implement the skills you have learned to fit, interpret and evaluate a regression model. Once you have completed steps 1 through 14, prepare a report to document your findings.

Criteria for Success

Refer to the attached rubric for more details on the report. The report should contain a well written cover/title page, introduction, body, conclusion, and references. It must follow APA format and have at least 1000 words (excluding title page and references page. All R code

used for your report should be included in an appendix at the end of the report.

Graphs, figures, charts, and tables are very useful visual effects to communicate your results and impress your readers. However, such items should not be included in the report unless they are well described and interpreted. Please use subtitles to make your assignment more reader friendly as well.

Format & Guidelines

The report should follow the following format:

(i) Title page

(ii) Introduction

(iii) Analysis

(iv) Conclusion/Interpretations

(v) References

Deliverables (HOW)

1. Load the Ames housing dataset.
2. Perform Exploratory Data Analysis and use descriptive statistics to describe the data.
3. Prepare the dataset for modeling by imputing missing values with the variable's mean value or any other value that you prefer.
4. Use the "cor()" function to produce a correlation matrix of the numeric values.
5. Produce a plot of the correlation matrix, and explain how to interpret it. (hint - check the corrplot or ggcorrplot plot libraries)
6. Make a scatter plot for the X continuous variable with the highest correlation with SalePrice. Do the same for the X variable that has the lowest correlation with SalePrice. Finally, make a scatter plot between X and SalePrice with the correlation closest to 0.5. Interpret the scatter plots and describe how the patterns differ.
7. Using at least 3 continuous variables, fit a regression model in R.
8. Report the model in equation form and interpret each coefficient of the model in the context of this problem.
9. Use the "plot()" function to plot your regression model. Interpret the four graphs that are produced.
10. Check your model for multicollinearity and report your findings. What steps would you take to correct multicollinearity if it exists?
11. Check your model for outliers and report your findings. Should these observations be removed from the model?
12. Attempt to correct any issues that you have discovered in your model. Did your changes improve the model, why or why not?

13. Use the all subsets regression method to identify the "best" model. State the preferred model in equation form.

14. Compare the preferred model from step 13 with your model from step 12. How do they differ? Which model do you prefer and why?

Northeastern University
College *of* Professional Studies

Assignment Rubric

Category	Above Standards	Meets Standards	Approaching Standards	Below Standards
Introduction 15%	Clearly and briefly introduces the goals of the project, the question that needs to be answered and the methods used in the analysis. The goals, questions and methods outlined are consistent with one another.	Introduction provides a brief and intelligible overview of the goals and methods of the assignment.	Introduction provides an overview of the goals and methods of the assignment, but is ambiguous or not concise.	Does not introduce project goals, project questions or methods.
Analysis 25%	Incorporates R code and the outputs. Provides detailed analysis of the output focusing on significance results. Uses visualizations to make major points.	Provides all R code and the outputs. Includes interpretation of the output, graphs, figures, charts, and tables and the significance of the results in the analysis.	Provides R codes and outputs, but the R code does not match the outputs or is missing some code or outputs. Includes limited interpretations, charts, and tables and the significance of the results in the analysis.	Does not provide R code or its outputs or minimal R code is provided. Includes few interpretations, charts, or tables. Does not identify the significance of the results in the analysis.
Data Visualizations 25%	Data visualizations are appropriate for the level and type of analysis. . Uses graphs, figures, charts, and tables to increase visual effects of the main points being made based on the results.	Data visualizations are appropriate for the level and type of analysis. Graphs, figures and tables communicate insights and significance to the reader.	Data visualization are useful for the level and type of analysis, but graphs, figures and tables do not clearly communicate the significance of the results to the reader.	Data visualization are used minimally or not at all. If graphs, figures and tables are used, it is unclear what they are intended to communicate or why.
Interpretation & Conclusions 25%	Wraps up the findings in a conclusion that provides an answer to the question(s) posed in the introduction. Makes specific recommendations based on the data presented.	The conclusion summarizes and makes sense of the results, making good points that reflect clear understanding of the assignment material.	The conclusion summarizes and makes sense of the results, making good points that reflect a basic understanding of the assignment material.	The conclusion does not summarize or attempt to make sense of the results. Conclusions do not reflect an understanding or reflect a misunderstanding of the material.

Northeastern University
College *of* Professional Studies

Report: Writing Mechanics, Title Page, & References **10%**	There are no noticeable errors in grammar, spelling, and punctuation; and completely correct usage of title page, citations, and references. The report contains approximately 1,000 words.	There are one to three noticeable errors in grammar, spelling, and punctuation; and completely correct usage of title page, citations, and references. The report contains approximately 1,000 words.	There are three to five errors in grammar, spelling, and punctuation; and completely correct usage of title page, citations, and references. The report contains approximately 1,000 words.	There are more than five errors in grammar, spelling, and punctuation; or the usage of title page, citations, and references are incomplete; or the report contains far less than 1,000 words.