ST563 - Homework 2

Wenbin Lu

1 Instructions

Please follow the instructions below when you prepare and submit your assignment.

- Include a cover page with your homework. It should contain
  1. Full name
  2. Course #
  3. HW #
  4. Submission date
- Assignments should be submitted using **Gradescope**.
- Neatly typed work should be submitted. All R code/output should be well commented on, with relevant outputs highlighted and discussed.
- When you solve a particular problem, do not only give the final answer. Instead, show your work and the steps you used (with proper explanation) to arrive at your answer to get full credit.
- Submission in the PDF format is preferred; Please convert other formats such as doc or docx into PDF.

Problem 1 (20 points)

Do Chapter 2, Problem 7

Problem 2 (20 points)

Do Chapter 3, Problem 4

Prolem 3 (60 points)

Consider the Ames housing data available in the package `AmesHousing`.

```
# Create data
ames <- AmesHousing::make_ames()
dim(ames)
```

```
## [1] 2930   81
```

We are interested in predicting `Sale_Price` based on `Gr_Liv_Area` (Above grade (ground) living area square feet) and `Year_Built` (Original construction date).

Consider the following models

(a) A multiple linear regression with both `Gr_Liv_Area` and `Year_Built` as predictors but with no interaction.

(b) A multiple linear regression with both `Gr_Liv_Area` and `Year_Built` as predictors but with interaction.

(c) Same as model (a) but with log-transformed `Sale_Price` as response.

(d) Same as model (b) but with log-transformed `Sale_Price` as response.

Investigate the models above, and answer the following questions.

(A) Split the full `ames` data into training and testing sets (no need for repeated splits). Create a table showing performance of the four models in the training set. [Hint: Note that models (c) and (d) are on the log-scale. So we need to scale the predictions to the original unit before computing certain performance measures.]

(B) How do the models perform in the test set? Which model would you choose to be the best predictive model?

(C) For your choice of the final model in (B), fit the model to the full data (for the best accuracy) and provide estimates and corresponding standard errors for the regression coefficients in a table.

(D) For your final model, perform regression diagnostics to see whether there are concerns about model assumptions being met.

(E) Do you suspect that there are some influential points in the data? Explain.

(F) In your best model, consider also including `Garage_Area` and `Garage_Cars` as predictors. Do you see any issues with multicollinearity? Explain.