Exam 2 — Part 2 — 4/7/2023

Instructions

This part is worth 40 points total. The exam (both parts) is worth 100 points total.

You have until the end of the class period to complete this part of the exam.

You may use your plebe-issue TI-36X Pro calculator.

You may refer to notes that *you have handwritten*, not to exceed *one side* of an 8.5" × 11" piece of paper.

You may *not* use any other materials.

No applications except for JupyterLab may be open on your laptop during the exam.

No collaboration allowed. All work must be your own.

Do not discuss the contents of this exam with any midshipmen until it is returned to you.

Type your answers **directly in this Jupyter notebook**, and submit this notebook (just the `ipynb` file) using the submission form on the course website.

Problem 0

For this exam, we will revisit the `RailsTrails` dataset from the `Stat2Data` package.

Run the cell below to load and preview the data.

In [1]:
```
# library(Stat2Data)
# data(RailsTrails)
# head(RailsTrails)
```

Problem 1

For this problem, we will focus on the following variables in `RailsTrails`:

Variable	Description
Price2014	Price estimate from 2014 (in thousands of dollars)
SquareFeet	Square footage of interior finished space (in thousands of square feet)
BikeScore	Bike friendliness (0-100 score, higher scores are better)

Consider the following model, which we will call Model A:

$$\text{(Model A)} \qquad Price2014 = \beta_0 + \beta_1 SquareFeet + \beta_2 BikeScore + \beta_3 BikeScore^2 + \varepsilon$$
$$\varepsilon \sim N(0, \sigma_\varepsilon^2)$$

a.

Fit Model A. Provide **only** the summary output for this part.

Feedback. See Example 1 in Lesson 18 Part 2 for a similar problem.

Remember that quadratic terms must be included in `I()` to ensure that R treats operators as math. See the beginning of Lesson 18 Part 2 for details.

b.

Is the overall model effective?

- Answer yes or no.
- Report (1) the name of the hypothesis test, (2) the test statistic and (3) the p-value you used to make your decision. Use a significance level of 0.05.

Feedback. For similar problems, see Example 3 in Lesson 14 Part 1 and Problem 3c in the Review Problems for Exam 2.

Note that the test statistic and p-value for the relevant test are given in the summary output.

In addition, note that the problem asks for ***the name*** of the hypothesis test you used. This should be something like "t-test for multiple linear regression coefficients" or "ANOVA F-test for multiple linear regression."

c.

Consider Model B, which only uses $SquareFeet$ to predict $Price2014$:

$$(\text{Model B}) \qquad Price2014 = \beta_0 + \beta_1 SquareFeet + \varepsilon \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

(i)

Conduct a nested F-test to compare Model A and Model B.

In particular, in the cell below, compute the test statistic and p-value for this test. You are encouraged to use any appropriate "shortcut" functions in R.

You will be asked to interpret the results of the test in part (ii).

Do not provide summary output for the fit of Model B.

Feedback. See Example 1 in Lesson 20 Part 2 for a similar problem. Also see STAT2 Exercise 3.47, assigned for homework.

In particular, see part f, which shows you how to use the `anova()` function to perform the nested F-test in just one line of R code (after fitting the reduced model).

Note that this "shortcut" was also used in Problem 5 in the Review Problems for Exam 2.

(ii)

Based on the nested F-test, which model is "better"? Briefly explain. Use a significance level of 0.05.

Feedback. For similar problems – in particular, problems about interpreting the nested F-test – see Example 1 in Lesson 20 Part 2 and Problem 5 in the Review Problems for Exam 2. Also see STAT2 Exercise 3.47, assigned for homework.

Problem 2

Suppose now you want to develop a model to predict $Price2014$ from the following variables:

Variable	Description
Bedrooms	Number of bedrooms
BikeScore	Bike friendliness (0-100 score, higher scores are better)
Distance	Distance (in feet) to the nearest bike trail entrance
NumFullBaths	Number of full bathrooms
SquareFeet	Square footage of interior finished space (in thousands of square feet)
WalkScore	Walking friendliness (0-100 score, higher scores are better)

a.

In the code cell below, run the best subsets regression procedure. In part b, you will be asked to interpret your output.

Feedback. See Example 2 in Lesson 21 and Problem 1a in the Review Problems for Exam 2 for similar problems about running the best subsets regression procedure. Also see STAT2 Exercises 4.5c and 4.7, assigned for homework.

Those examples (and this problem) requires the use of the `leaps` library. You can find instructions on installing the `leaps` library in Lesson 21, right before Example 2.

b.

What is the best model, using Mallows's C_p? Give your answer by (1) stating the Mallows's C_p of the model you have chosen, and (2) listing all the predictors in the model you have chosen.

Feedback. See Example 3 in Lesson 21 for a similar problem.

Problem 3

For this problem, we will consider the following model that predicts $Price2014$ from $NumFullBaths$, $SquareFeet$, and $WalkScore$:

$$Price2014 = \beta_0 + \beta_1 NumFullBaths + \beta_2 SquareFeet + \beta_3 WalkScore + \varepsilon \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

a.

(i)

Compute the VIFs of all the predictors in this model.

Feedback. See Example 3 in Lesson 19 for a similar problem about computing the *variance inflation factors (VIFs)* of each predictor in a model. Also see STAT2 Exercise 3.53, assigned for homework.

That example (and this problem) requires the use of the `car` library. For instructions on installing the `car` library, please see the email from me to the class on Friday 3/24.

(ii)

Should we be worried about multicollinearity? Briefly explain. State any rules of thumb you use.

Feedback. See Example 3 in Lesson 19 for a similar problem. In addition, see Lesson 19 for a rule of thumb for using the VIFs to detect multicollinearity. Also see STAT2 Exercise 3.53, assigned for homework.

Note that ***each predictor in a model has its own VIF***. It doesn't make sense to say, "the VIF", unless there is only one predictor in the model.

b.

Split your data, using the following code:

```
set.seed(2000)
train <- sample(104, 78)
```

Compute the cross-validation correlation, using `RailsTrails[train,]` as the training sample, and `RailsTrails[-train,]` as the holdout sample.

Feedback. For similar problems, see Problem 1b of the Review Problems for Exam 2, and Example 1 in Lesson 22. Also see STAT2 Exercise 4.9, assigned for homework.

c.

Suppose the shrinkage on cross-validation is 0.06. Should we be worried about the predictive performance of our model? Briefly explain. State any rules of thumb you use.

Feedback. See Example 1g in Lesson 22 for details on shrinkage, and a rule of thumb for using shrinkage to evaluate the predictive performance of a model.

Grading rubric

Problem	Weight
1a	0.4
1b	0.4
1c(i)	0.4
1c(ii)	0.4
2a	0.4
2b	0.4
3a(i)	0.4
3a(ii)	0.4
3b	0.4
3c	0.4
Max Score	**40**