Midterm Two - Statistics 151A, Spring 2020

Due by midnight PDT on April 12, 2020

April 8, 2020

1 Overall Guidelines

You are allowed to use textbooks, R documentation, notes, scientific papers, and other written materials in print or online to help you complete this assignment. However, **you must not consult any other person** about any aspect of your project. If you have clarification questions about the assignment, you may contact Prof. Pimentel and the course staff by private bCourses message.

You will have four entire days to complete the assignment. However, the assignment itself should require only a minority of that time to complete. As a result, we do not plan to give extensions, and you should budget your time carefully in accordance with this policy, especially if you have other assignments or other unavoidable conflicts during the period of the assignment.

2 The Data

You are required to analyze the Bike Sharing dataset. The dataset is uploaded on bCourses under Files > Exams > Midterm_2 > BikeSharingDataset.csv.

The dataset contains information on bike rentals for two years (2011 and 2012) from Capital Bikeshare System, Washington D.C. Bike sharing systems are bike rentals where the whole process from membership, rental and return is automatic. Through these systems, the user is able to easily rent a bike from a particular position and return at another position.

This dataset is collected to address the problem of predicting the number of bike rentals in a given hour given the environmental and seasonal conditions for that hour. The dataset contains 17379 observations with each observation corresponding to one particular hour. The dataset contains the following 17 variables:

1. **instant** : Unique observation number.
2. **dteday**: Date.
3. **season**: Categorical variable (1: Spring, 2: Summer, 3: Fall, 4: Winter).
4. **yr**: Stands for year. Binary variable (0 stands for 2011 and 1 stands for 2012).
5. **mnth**: Stands for month. Takes the values $1, 2, \ldots, 12$.
6. **hr**: Indicates the hour of the day (takes values $0, \ldots, 23$).
7. **holiday**: Indicates whether the day is a holiday or not
8. **weekday**: Day of the week, numbered 0 (Sunday) through 6 (Saturday)

日期

9. **workingday**: Takes the value 1 if the day is neither weekend nor holiday and takes the value 0 otherwise.
10. **weathersit**: Takes four values:
    (a) 1 if the weather is Clear, Few clouds, Partly cloudy, Partly cloudy.
    (b) 2 if the weather is Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist.
    (c) 3 if the weather is Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.
    (d) 4 if the weather is Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog.

weather

11. **temp**: Normalized temperature in Celsius. The values are divided by 41 (the maximum temperature).
12. **atemp**: Normalized feeling temperature in Celsius. The values are divided by 50 (the maximum temperature).
13. **hum**: Normalized humidity. The values are divided by 100 (the maximum humidity).
14. **windspeed**: Normalized wind speed. The values are divided by 67 (maximum wind speed).
15. **casual**: The number of bikes rented by casual (unregistered) users for that hour.
16. **registered**: The number of bikes rented by registered users for that hour.
17. **cnt**: The number of bikes rented by both casual and registered users for that hour (this is the sum of **casual** and **registered**).

3 Research questions

Your assigned task is to use the data to answer the following two questions:

- What is the best regression model you can find for predicting total number of bikes rented in a particular hour (variable **cnt**) using environmental conditions and other available explanatory variables?
- What is the interpretation of your chosen model, and how much do you trust that interpretation? Is it reasonable to view any of your fitted parameters as causal effects? Why or why not? What guidance does the form of your model provide for future studies of bikeshare use?

Some of you may be tempted to use sophisticated machine learning methods or time series approaches (since the data has a time series structure) for your analysis. Please remember that this course is about linear regression and know that points will not be awarded for analysis involving complex model classes not covered through this point in the smester. You are welcome to suggest more complicated models as possibilities for future analysis if you see evidence in the data that they might perform well, but please refrain from fitting and analyzing such models as part of your assignment.

4 Submission

You are required to submit a short Rmarkdown report describing your analysis. This will be due on Gradescope. **It is your responsibility to make sure your document can knit properly, and extensions will not be given because of knitting errors**.

The report should contain an introduction, brief description of the data, a description of the analyses and results and general discussion/conclusions. Your R code should be included as an appendix at the end of the

report (code should not appear at any other place in the report). Figures should be carefully chosen, labeled, and referred to in the text. The maximum number of pages allowed in the report is **eight** (**excluding the appendix**). Also note the following points:

- The report should be written in paragraphs and sentences.
- The text should form a logical narrative, and not be just a series of plots or statistics. The text should be written in the best order to make a coherent report, i.e., it should not simply follow the order in which you did the analysis.
- Include plots to supplement the narrative, and again they should be a logical part of the narrative of your results and appropriate to the analysis.
- Run spell check and grammar check on your report.
- Comment the R code appropriately so that it is easy to understand.

5 Assessment

In contrast to previous assignments in this course, this midterm will be assessed holistically. There are no specific requirements to include particular plots or test particular methods; instead, you are free to construct your own convincing analysis and writeup in response to the research questions. Grades will be determined using the five-part rubric below, with each part receiving an equal weight in the grade.

1. **Choice of analysis**: Is an appropriate research question chosen and clearly defined? Is the model and analysis chosen relevant to the research question? Does it make effective use of the data in hand? Are modeling choices and assumptions properly justified with reference either to the data itself (e.g. through diagnostics or exploratory data analysis) or to the broader scientific context?
2. **Interpretation and evaluation**: Is model output interpreted correctly? Are conclusions justified by specific results from the analysis? Are weaknesses of the approach and alternate explanations or methodologies considered and discussed?
3. **Computations**: Are computations and calculations performed correctly? Are they documented clearly but concisely? Are computational portions of the report bloated with extraneous or unnecessary calculations or inefficient and repetitive code?
4. **Visual presentation**: Do plots and figures convey important information that contributes to the central arguments of the report? Are plots and figures labeled appropriately and referred to effectively in the text?
5. **Writing**: Is the report organized logically with a clear structure? Are individual steps in the analysis explained clearly? Do conclusions drawn from the data analysis respond meaningfully to the research question? Is writing concise and grammatically correct? Are length and formatting requirements observed?