HAD 5744: Quantitative Methods I
Dr. Hoagland

Assignment #1

Late assignments will be penalized at the rate of 10% per day

Please submit a knitted R-Markdown file with text and code for this assignment. Don't forget to acknowledge your group members (although each of you must submit your own assignment) and to include your AI statement.

For the problem indicated with an asterisk (*), please submit a video of no more than 3 minutes explaining both how you got to your solution and what it means. Refer to the syllabus for complete instructions here.

Note: For this problem, we will use a rich, realistic administrative dataset (healthyR_data) with 187,721 rows covering hospital visits, patient demographics, charges, payments, and quality metrics. This data is available from the "healthyR" package and described here. Use this page for instructions to load the data and access data documentation.

Problem 1: Data Cleaning and Univariate Regression. We are interested in understanding the impact of length of stay on health costs. Our dependent variable will be the "total_charge_amount" variable indicating the total costs associated with a stay, and our primary indepdent variable will be "length_of_stay." The additional independent variables included in the data we will use in our summary table and subsequent assignment are: place of service ("ip_op_flag"), type of service ("service_line"), payer, admission day of week ("visit_start_date_time") and readmission expectation ("readmit_expectation").

a. Why would you care about this question? Write 1-2 paragraphs (total) that motivates understanding these determinants from (a) a physician's perspective, (b) a researcher's perspective, and (c) a policy-maker's perspective.

b. Draw a preliminary DAG (not necessarily in R! Can add in a scan of this later) documenting hypothesized relationships between this outcome and the independent variables included in the dataset that you think are relevant. Justify your DAG with text. Where do you think this DAG is imperfect or failing you?

c. Before running any regressions, you will need to clean and summarize your data. Convert all variables to either a numeric variable or dummy variables (either as binary indicators or appropriate factor variables) and choose how to deal with any missing data in your variables. Construct indicators for (i) inpatient, (ii) outpatient, (iii) payers, (iv) service types, (v) weekend admissions, and (vi) expected readmission within 30 days. For payers and service types, you do not have to include dummy variables for each individual category, you can (and should!) aggregate these as you see fit. Once this is done, produce a simple summary table reporting (a) the means and standard deviations for continuous variables and (b) the counts and percentages for categorial variables. This table should be well-formatted. Does anything stand out about your sample?

Note: I am indifferent as to whether you use R to construct this table or build it manually in Word.

d. Visualize the relationship between LOS and costs in a way that you think is compelling and/or interesting. Report the figure and tell me the story it is showing.

e. * Now we will run regressions. First regress total charges only on LOS. Report the coefficients, standard

errors, confidence intervals, p-values, R^2, and sample size in a well-formatted regression table. Interpret the table, noting the economic and statistical significance of the relationship. How much would an additional week of a hospitalization stay increase total health costs?