36-402 DA Exam 1

Chae Yeon Ryu (chaeyeor)

4/3/2020

INSTRUCTIONS – REMOVE BEFORE SUBMITTING

This is a template for your data analysis report. It should work for anyone who is using R Markdown to generate PDFs through LaTeX. If you make the PDFs in some other way, as long as you can get 12 point fonts and reasonable margins, it will be fine.

Marking your answers

Each section below contains numbered questions. You are **required** to mark the sentence in your report that most closely answers each question.

For example, for question 2 in the EDA section, mark the sentence with **(2)**. For question 3 in Modeling & Diagnostics, mark the sentence with **(3)**.

You are writing a report, not simply writing answers to the questions, so you should not leave the questions in your report.

Figure captioning/sizing tips

Here is an example of creating a plot, setting its size (in inches), and giving it a caption. Figures with captions "float" on the page, usually appearing at the top or bottom; don't try to fight LaTeX and force them to appear in a certain place, because you will lose.

Notice that the figure appears in the PDF, but the code does not.

You can only make one figure per code chunk.

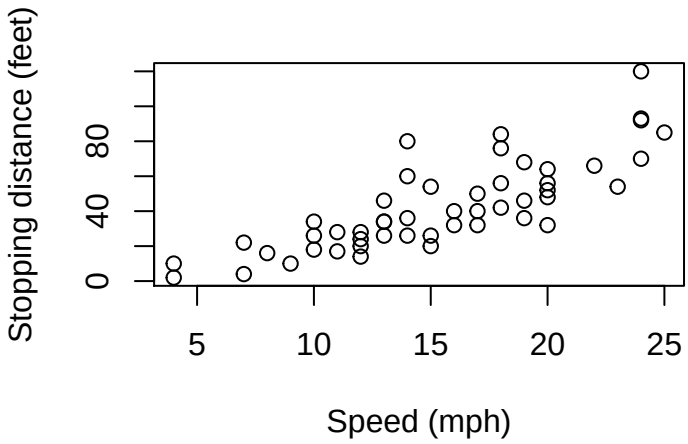

Figure 1: Distance it took cars at each speed to stop.

Additive models package

Use the **mgcv** package, **not** the gam package, for this exam. mgcv has a `gam` function that works in much the same way as the one you've already used, with minor differences. **Warning:** do not load *both* packages; since they define functions with the same names, this quickly becomes very confusing.

In mgcv, to make a fit that smooths with 4 effective degrees of freedom, write

since otherwise mgcv will automatically use GCV to pick the best edf. In this exam, use 4 effective degrees of freedom for all continuous variables.

To fit with an interaction between a continuous variable and a factor variable, use

This fits a separate smoothing spline to `Petal.Length` for each level of the `Species` factor.

To do an F test between two models, use

Introduction

The Department of Education is interested in knowing whether there's a relationship between students' earnings after graduation and their prior education, economic status, and the tuition. This study utilizes the College Scorecard data from 1,300 American colleges and universities. The goal of this analysis is to quantify the statistical relationship with these variables and see if students who attend more expensive schools earn more money after graduation on average.

It was found that students who attend more expensive schools tend to earn more money after graduation to some extent, and the relationship between blahhhhh.

1. Clearly state the research questions and objectives of your study.
2. Briefly mention your final findings.

Exploratory Data Analysis

The College Scorecard data consist of 1,300 American colleges and universities with measured scores and values such as thier average net price, mean SAT score, fraction of federal Pell grant, fraction of federal student loan, and median earnings. The dependent variable is Median earnings (dollars) of students working and not enrolled, 10 years after entry, and independent variables are Average net price, Mean equivalent SAT score for admitted students Fraction of all undergraduates who received a federal Pell grant for tuition, and Fraction of all undergraduates receiving a federal student loan. Figure 1 shows histograms describing the distributions of each of these variables.

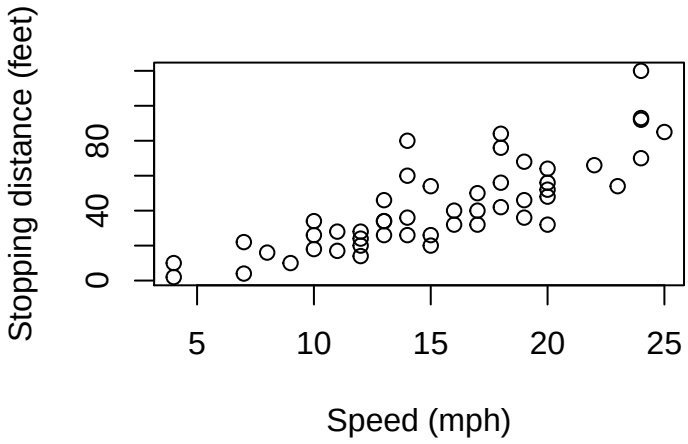

Figure 2: Distance it took cars at each speed to stop.

1. Explore the key variables you need to answer the Department of Education's questions. Describe them with any necessary univariate EDA.
2. Identify and specify your response variable and its distribution.
3. For the questions asked in Results below: Do multivariate EDA to explore the relationship between predictors and the response.
4. Describe any trends or interesting features that you see that suggest what you will find in the analysis.

Modeling & Diagnostics

1. Construct two models to answer the research questions, one linear and one additive (using your EDA to decide which terms should be nonlinear). For the additive model, use the `gam` package and `s()` with its default effective degrees of freedom.
2. Briefly explain what it means to control for the covariates you controlled for (such as SAT scores and prior economic status), and why you want to do so in this model. Briefly explain how you controlled for the covariates in your models.
3. Present model diagnostics. Discuss possible improvements and modifications to your model to address any violations of the model assumptions.
4. Use cross-validation to determine whether the linear or additive model fits best to the data, in terms of prediction error.
5. Comment on whether the difference between the two models appears significant, based on the uncertainty in your estimates of the prediction error; a formal test is not required here.
6. For your chosen model, examine the residual diagnostics to determine what type of bootstrap would be appropriate for this data.

Results

1. Using the selected model, determine whether students who attend more expensive school earn more money after graduation, as requested by the Department of Education. You can use appropriate model results or plots to answer this question.
2. Using the selected model, determine whether the relationship between price and earnings is the same at public, private, and for-profit solutions. Use an interaction term and a hypothesis test. Make sure you clearly state the null and alternative hypotheses, your test statistic, and the assumptions you made.
3. Carnegie Mellon University is listed in the data. Build a confidence interval for the mean earnings of students after graduation for a school just like Carnegie Mellon. Make this confidence interval by obtaining the standard error from `predict` (with the `se.fit` argument). State the assumptions this method makes.
4. Compare the confidence interval you got from `predict` to one calculated by bootstrapping, and comment on which you believe is more reliable by comparing the assumptions each method makes.

Conclusions

1. Summarize your main findings in the analysis. What conclusions can your client draw?
2. Discuss possible reasons for these findings.