Data Analysis Assignment 2

Due: 11:55pm, Friday, Oct 7

Instructions

This assignment involves linear regression. The data can be found on Sakai: go to Resources → Data Analysis Assignment Datasets → Assignment 2. You are required to type your solutions using R Markdown. You will submit 1) a PDF produced from R Markdown with answers to the questions, and 2) your Rmd file. Submissions should be made on gradescope: go to Assignments → Data Analysis Assignment 2 and Assignments → Data Analysis Assignment 2 RMD CODE.

DO NOT INCLUDE R CODE OR OUTPUT IN YOUR SOLUTIONS/REPORTS *All R code must be included in an appendix, and R outputs should be converted to nicely formatted tables. Feel free to use R packages such as* `kable`*,* `xtable`*,* `stargazer`*, etc. Please consult the example report and corresponding Rmd file for guidance.*

Also, you should round up ALL numbers/estimates to 2 decimal places (4 decimal places at the most to avoid exact zeros when possible).

Reminder: You are allowed and even encouraged to talk to each other about general concepts, or to the instructor/TAs. However, the write-ups, solutions, and code MUST be entirely your own work.

Your answers for this assignment must be formatted as a report. You should write as if you are presenting results to a client with some basic knowledge of the methods you are using. Again, you should not have any R code or output in your report. You may include tables and figures that are formatted nicely (raw regression output does not count as formatted nicely). Pay careful attention to the page limits for each question.

Questions

Question 1 was taken and adapted from Chapter 7 of Ramsey, F.L. and Schafer, D.W. (2013), "The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)."

1. Old Faithful Geyser in Yellowstone National Park, Wyoming, derives its name and its considerable fame from the regularity (and beauty) of its eruptions. As they do with most geysers in the park, rangers post the predicted times of eruptions on signs nearby, and people gather beforehand to witness the show. R.A. Hutchinson, a park geologist, collected measurements of the eruption durations (in minutes) and the subsequent intervals before the next eruption (also in minutes) over an 8-day period. *The data for this question can be found in the file "OldFaithful.csv" on Sakai. Your answer to this question should be at most 2 pages.*

    - Fit a regression model for predicting the `interval` between eruptions from the `duration` of the previous one, to the data, and interpret your results.
    - Include the 95% confidence interval for the slope, and explain what the confidence interval reveals about the relationship between duration and waiting time.
    - Describe in a few sentences whether or not you think the regression assumptions are plausible based on residual plots (do not include any plots).
    - Fit another regression model for predicting `interval` from `duration` and `day`. Treat `day` as a categorical/factor variable. Is there a significant difference in mean intervals for any of the days (compared to the first day)? Interpret the effects of controlling for the days (do so only for the days with significant effects, if any).
    - Using k-fold cross validation (with k=10), compare the average RMSE for this model and the average RMSE for the previous model excluding `day`. Which model appears to have higher predictive accuracy based on the average RMSE values?

2. These days, it is widely understood that mothers who smoke during pregnancy risk exposing their babies to many health problems. This was not common knowledge fifty years ago. One of the first studies that addressed the issue of pregnancy and smoking was the Child Health and Development Studies, a comprehensive study of all babies born between 1960 and 1967 at the Kaiser Foundation Hospital in Oakland, CA. The original reference for the study is Yerushalmy (1964, *American Journal of Obstetrics and Gynecology*, pp. 505-518). The data and a summary of the study are in Nolan and Speed (2000, *Stat Labs*, Chapter 10) and can be found at the book's website.

   The data for this question can be found in the file "smoking.csv" on Sakai.

There were about 15,000 families in the study. You will only analyze a subset of the data. The researchers interviewed mothers early in their pregnancy to collect information on socioeconomic and demographic characteristics, including an indicator of whether the mother smoked during pregnancy. The variables in the dataset are described in the code book at the end of this document.

Note that this is an observational study, because mothers decided whether or not to smoke during pregnancy; there was no random assignment to smoke or not to smoke. Thus, we cannot make causal inference statements from the results of a standard regression model.

In 1989, the Surgeon General asserted that mothers who smoke have increased rates of premature delivery (before 270 days) and low birth weights. We will analyze the data to see if there is an association between smoking and birth weight. To simplify analyses, we'll compare babies whose mothers smoke to babies whose mothers have never smoked. The data file you have access to has only these people, although there were other types of smokers in the original dataset.

Our questions of interest include the following:

- Do mothers who smoke tend to give birth to babies with lower weights than mothers who do not smoke?
- What is a likely range for the difference in birth weights for smokers and non-smokers?
- Is there any evidence that the association between smoking and birth weight differs by mother's race? If so, characterize those differences.
- Are there other interesting associations with birth weight that are worth mentioning?

First build your model, then do model assessment and validation. You should only proceed to answer the questions when you are satisfied with your final model; you should answer all the questions using that final model.

Analyze the data and investigate these questions using a linear model.

- Write a report (maximum of 5 pages) describing your findings. Code and additional plots should be placed in an appendix (not included in the 5 pages). You should start this report on a new page after Question 1.
- Make sure to provide direct answers to each question using your model.
- Be sure to also include the following in your report:
  - some exploratory data analysis (e.g., a "table 1" that shows mean (SD) of continuous variables and N (%) of categorical variables stratified by mother's smoking status). Include comments about anything particularly notable in EDA,
  - the model you ultimately decided to use,
  - clear model building, that is, justification for the final model (e.g., model assessment and model selection procedures, though intermediate models do not need to be included),
  - the relevant regression output (includes: a table with coefficients and SEs, p-values, confidence intervals; and somewhere in the text or table the estimated regression standard deviation and R-squared),
  - and any potential limitations of the analysis.

To help organize your thoughts, you should organize your report into sections as follows.

- **Summary**: a few sentences describing the inferential question(s), the method used and the most important results.
- **Introduction**: a short but more in-depth introduction to the inferential question(s) of interest. Here, you are basically writing the experiment and questions of interest given in your own words.
- **Data**: your EDA, interesting features of the data, and how missing data were dealt with. Include one or two plots/tables of your most interesting EDA findings and describe those findings.
- **Model**: a detailed description of the model used, how you selected the model, how you selected the variables, model assessment, model validation, and presentation of the model results. What are your overall conclusions in context of the inferential problem(s)? Try to include one or two plots that can help drive your point home.
- **Conclusion**: the importance of your findings and potential limitations of the study.

There are some complexities in the original dataset to be aware of. Some variables have missing values, but the provided dataset includes complete observations only. In particular, the height and weight of the father are missing quite frequently, so those variables were removed. This is typical in data on births, as it is often difficult to get data about the fathers.

The data files also contain two outcome variables: gestational age and birth weight. Both of these could be affected by smoking, so both are outcomes rather than predictors. It does not make sense scientifically to include one as a predictor of the other; the two variables happen simultaneously and hence are a bivariate outcome. For this analysis, we exclude gestational age from the modeling. Of course, one could do a separate regression for gestational age to see if smoking has an effect on gestational ages. The file also contains an indicator variable for Premature (gestational age < 270 days), which is just a recoding of gestational age; ***we won't use that****.*

The main file also includes information on the number of cigarettes smoked and about timing for mothers who quit smoking. For this analysis you do not have to use those variables, as we just compare smokers and non-smokers. Also, for this analysis, you can ignore the birth date variable, you can collapse education categories from 6-7 into one category for education = trade school, and you can also collapse race categories from 0 - 5 into one category for race = white.

Code Book

Variable	Description
Id	id number
birth	birth date where 1096 = January1, 1961
gestation	length of gestation in days
bwt	birth weight in ounces (999 = unknown) *Response/outcome variable*
parity	total number of previous pregnancies, including fetal deaths and still births. (99=unknown)
mrace	mother's race or ethnicity 0-5=white 6=mexican 7=black 8=asian 9=mix 99=unknown
mage	mother's age in years at termination of pregnancy
med	mother's education 0 = less than 8th grade 1 = 8th to 12th grade. did not graduate high school 2 = high school graduate, no other schooling 3 = high school graduate + trade school 4 = high school graduate + some college 5 = college graduate 6,7 = trade school but unclear if graduated from high school 9 = unknown
mht	mother's height in inches
mpregwt	mother's pre-pregnancy weight in pounds

Variable	Description
drace	father's race or ethnicity 0-5 = white 6 = mexican 7 = black 8 = asian 9 = mix
dage	father's age in years at termination of pregnancy
ded	father's education 0 = less than 8th grade 1 = 8th to 12th grade. did not graduate high school 2 = high school graduate, no other schooling 3 = high school graduate + trade school 4 = high school graduate + some college 5 = college graduate 6,7 = trade school but unclear if graduated from high school 9 = unknown
dht	father's height
dwt	father's pre-pregnancy weight in pounds
marital	marital status of mother 1 = married 2 = legally separated 3 = divorced 4 = widowed 5 = never married
income	family yearly income in 2500 increments. 0 = under 2500, 1 = 2500-4999, ...
smoke	does mother smoke? 0 = never 1 = smokes now 2 = until preg 3 = once did, not now
time	If mother quit, how long ago did she quit? 0 = never smoked, 1 = still smokes, 2 = quit during pregnancy, 3 = up to 1 yr ago, 4 = up to 2 yr ago, 5 = up to 3 yr ago, 6 = up to 4 yr ago, 7 = 5 to 9yr ago, 8 = 10+yr ago, 9 = quit and don't know, 98 = unknown
number	number of cigs smoked a day for past and current smokers 0 = never smoked 1 = 1-4 2 = 5-9 3 = 10-14 4 = 15-19 5 = 20-29 6 = 30-39 7 = 40-60 8 = 60+, 9 = smoke but don't know
Premature	1 = baby born before gestational age of 270, and 0 = otherwise. *Ignore this for this assignment since it is just a dichotomized version of the gestational age.*

Grading

30 points: 10 points for question 1, 20 points for question 2.