Statistics 138

PROJECT

Winter Quarter 2021

This project is due on Thursday, March 18. You are highly encouraged to form a group of 3 registered students (including yourself) in the class. Only one report per group needs to be submitted and names of the group members must appear on the front page. It is important that the name of the project also appears on the front page of the report. Every student must be in only one group so that only one project per student is submitted.

Please note that you are not allowed to discuss the questions with anyone other than the instructor and the TA. This includes a tutor, your classmates (for example, comparing answers), and posting the questions online.

The project should be typed and the main body of the project should be no longer than 5-6 pages. You may use an appendix for the graphs and charts as needed. Please attach your R codes in the Appendix

Guidelines:

The project should include the following along with justifications and explanations in each step:

1. A description of the data and the goal of the analysis.
2. The type of statistical methods and models used in the analysis.
3. Transformations of variables (if needed) and interactions (if needed) in the analysis.
4. Initial analysis that includes all the variables and all the summary statistics such as parameter estimates, their standard errors, p-values etc.
5. All the relevant diagnostics (as appropriate) needed for the initial analysis (step 4).
6. Model selection.
7. The recommended final model along with the summary statistics for this model (as in part 4), and the relevant plots (as needed).
8. Summary of findings, conclusion and recommendations for further analysis (if any).

For each of the two problems, your report may include the following sections:

(i) Introduction: Statement of the problem
(ii) Materials and Methods: Description of the data and methods used in the analyses.
(iii) Results: Explanation of the results of your analyses. You can cut and paste the relevant parts of your computer outputs and refer to them in explaining your results.
(iv) Conclusion and Discussion: Highlight the main points and discuss them.

Format:

- The report should be typed and well formatted as a complete stand-alone document (not a list or bullet points, etc).

- The report should not contain code or raw R output. The R codes should be in an appendix.
- There should be a title page with names and student IDs of all group members.

Problem 1: Low birth rate or not. (file: baby)

The goal is to investigate if the probability of low birth weight of infant is related to information on mother such as age, weight smoking status etc. The response variable is birth with values 1 (low birth weight), 0 (no low birth weight). In addition to the first order terms, you may consider including the following interaction terms at the initial stage of your analysis: age and weight, weight and hypertension, and weight and pre. Use logistic regression to perform data summary, goodness-of-fit, and model selection. Use the final model to estimate the percentage of correct classification. [Note: if for any case if the estimated probability is larger than 1/2, the case is classified as birth=1.].

The file baby.xls has the following columns:

Column 1: age, the age of the mother,

Column 2: weight, the weight (in pounds) of the mother (before pregnancy),

Column 3: smoke, smoking status during pregnancy, with levels yes and no,

Column 4: pre, history of pre-mature labor, with levels yes and no,

Column 5: hyp, history of hypertension, with levels yes or no,

Column 6: visits, the number of visits during the first trimester (first three months),

Column 7: birth, if the birth weight of the infant was low or not (1=low, 0=not low).

2. **Ischemic heart disease. (file: ischemic)**

Data were collected by a health insurance company on its subscribers who had made claims resulting from ischemic (heart disease) for the time period of January 1, 1998 through December 31, 1999. The response is the number emergency room visits, and the goal is to model its mean as a function of 8 other variables. You may try models with all the predictor variables untransformed, and predictor variables transformed by square root (except gender). Use Poisson regression to perform data summary, goodness-of-fit and model selection.

The data are given in the file ischemic.xls. The columns are

Column 1: cost, total cost of claims made by subscriber (dollars),

Column 2: age, age of subscriber (years),

Column 3, gender of subscriber (1=male, 0=otherwise),

Column 4: inter, total number of interventions or procedures carried out,

Column 5: drugs, number of tracked drugs prescribed,

Column 6: complications, number of other complications that arose during the heart disease treatment,

Column 7: comorbidities, number of other diseases that the subscriber had during the period,

Column 8: duration, number of days of duration of treatment condition,

Column 9: visits, number of emergency room visits.