Applied Statistical Modelling and Inference

Short Project, due Friday, May 1, 5:00 PM

1 Introduction

Birth weight is the body weight of a baby at its birth. It is viewed as an important health indicator. Low birth weight, defined as birth weight less than 2500 grams, is an outcome that has been of concern to physicians for years. This is due to the fact that infant mortality rates and birth defect rates are very high for low birth weight babies. A woman's health and health behavior during pregnancy (such as smoking habits and whether receiving prenatal care) can greatly alter the chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight.

In this project, you are asked to prepare a short data analysis on low birth weight. You will use the data set `birthwt.csv`. This dataset is part of a larger study conducted at Bay State Medical Center in Springfield, Massachusetts. The dataset contains a short list of variables that are believed to be associated with low birth weight in the obstetrical literatures.

Dataset Description The explanatory variables collected in this study include: Mother's age in years (`MOTHAGE`), Mother's weight in pounds (`MOTHWT`), Mother's Race (`RACE`, 1 = White, 2 = Black, 3 = Other), Mother's history of premature labour (`PREM`, number of times), Mother's history of hypertension (`HYPE`, 1= Yes, 0 = No), Mother's history of urinary irritation (`URINIRR`, 1 = Yes, 0 = No), number of physician visits (`PHYSVIS`), Mother's smoking status during pregnancy (`SMOKE`, 0 = No, 1 = Yes). The outcome variable newborn's birthweight is recorded in grams (`BIRTHWT`).

The goal of the current study is two-fold:

1. To provide an explanatory model for the factors that may be associated with the weight of newborns (in grams) in the clinical population being served by Bay State Medical Center;

2. To build a second explanatory model with a specific focus on whether a mother's smoking status during pregnancy is associated with the chance that a baby is born of low birth weight, controlling for other variables.

2 Requirements

Your task is to conduct a data analysis and write a report with the following specifications. The report should be written to a general scientific audience and will be graded on the aspects of sophistication of the analyses, correctness of the interpretations and procedures used as well as spelling, grammar and readability of the report. To keep your report readable,

use appropriate section headings to help guide the reader through the information presented. Your report must not exceed **10 pages** double-spaced including any necessary figures and tables.

Make sure that `R` handles any factor or categorical variables appropriately after importing the dataset. The report should have the following sections:

1. Start with a short paragraph summary of the problems associated with being born at low birthweight (10pts). You may choose to conduct a small internet search to learn more about this problem.

2. A section on data and descriptives (25pts). This section should include the following information

   - Table(s) containing descriptive statistics of the people in the sample. For continuous variables, present the following five measures: sample mean, sample standard deviation, sample median, min and max. For categorical variables, present the number of individuals in each category as well as the corresponding sample proportion.
   - Figures pertaining to exploratory data analysis that shed light about some of the relationships in the data. Only include plots that are relevant to the questions of interest. Some examples of graphs are
     - Histogram of the outcome variable birth weight.
     - Scatterplots between birth weight and quantitative variables
     - Boxplots between birth weight and categorical valued variables.

     All plots need appropriate captions and should be explained in the text so that a general scientific audience may understand what is being presented (i.e. you may assume that you reader knows what a mean or median is and how to read a box plot, but not how to interpret patterns seen over different box plots or histograms).
   - Take a close examination of the data and the descriptive statistics.
     - Pay attention to some values that are considered as "error" such as any category values that are not included in the above dataset description or negative birth weight values. Code those values as missing values `NA` in `R` and report patterns of missing value.
     - Adjust the tables and figures, if necessary, after data cleaning.
   - Test the hypothesis that the birth weight differs by smoking status (without controlling for any other variables) using an appropriate statistical test. Write down the null and alternative hypotheses, report the test results. Comment on the results.

3. Prepare a section on a linear regression analysis using birthweight as a continuous outcome (30pts).

- When fitting this model, think about the choices of treating a covariate as quantitative or categorical. In particular, pay attention to variables `PREM` and `PHYSVIS`. Both variables record quantitative count values, however, the counts are low in both cases. Comments on whether it is appropriate to model a linear relationship between birth weights and variables such as `PREM` and `PHYSVIS`. In addition, if you decide to treat `PREM` and `PHYSVIS` as categorical values, pay attention to the decision whether you choose to collapse some of the categories, both from a statistical and a substantive considerations. [For this part, you can work with an additive model that include all variables in the dataset to specify a desired way to model `PREM` and `PHYSVIS`.]
- Use appropriate model selection criterions (for examples, you may consider F statistics, R^2, adjusted R^2, AIC and BIC among other options) to choose an optimal model among candidate models that include additive relationships and two way interactions between smoking status and any other variables. In addition, you should also use the substantive knowledge you learned in this section to aid you with the decisions. Articulate your model selection strategies clearly and briefly explain why.
- Include a table summarizing your final model, that includes the estimates of the regression coefficients and their standard errors, the relevant test statistics and p-values (or 95% confidence intervals of the estimates). Interpret the coefficients in the table.
- Conduct residual diagnostics for your choice of model and report your findings.

4. In this section, using the physicians definition of low birthweight constituting any weight less than 2500g, create a new, binary variable that categorizes the newborns as either low birthweight($< 2500g$) or not($\geq 2500g$). Conduct a logistic regression analysis and find the best explanatory model for being low birthweight (25pts).

    - When fitting the best explanatory model, the candidate models should include both additive models and two-way interaction models with Smoking Status.
    - Explain your model selection strategies, and comment on the choices of criterion you use.
    - Present the odds ratio and 95% confidence interval for all parameter estimates in the model you select. Interpret the effects in terms of odds ratio and pay particular attention to the effect of smoking status as well as relevant interaction effects on being low birthweight.
    - For the model of your choice, present the classification table, and perform an ROC analysis and estimate the area under the curve using step function approximation.

5. In a conclusion section, recap the substantive problem of interest, methods used, and results found (10pts). Do your results from the first hypothesis test, regression analysis

and logistic regression analysis agree or disagree? Why or why not? In this section, make sure to answer both questions of interest posed in the study goals.

6. Bonus(5pts): For the logistic regression, conduct an extended additional model selection task that include all two-way interactions between available variables. Report the results of the optimal model and compare it with the one you determined in the previous logistic regression analysis. The bonus part should be included as an Appendix, and should not exceed 2 pages long (in addition to the 10-page limit of the main report).

Submit your report via NYU Classes. You may use any functions in `R` of your choosing, but must justify all modeling decisions and tests used in your report. Submit the code used in your analysis as a supplementary, organized and well-commented, `.R` file.