Final Project

VERSION: 2018-11-13 · 00:17:59

1. Instructions

- Due Monday, December 3 at 11:59pm.
- You may work in teams of up to 3 students from either section.
- Each project consists of a computer-typed report strictly between 8-10 pages including figures, but excluding a mandatory Appendix containing (but not limited to) all **R** code.
- Reports must be submitted as a single PDF document. There are numerous free programs available online to convert proprietary formats such as DOC into PDF, such as this. However, given the presentation requirements detailed in the marking scheme below, it is very likely that a better grade can be obtained by writing the report with **R** Markdown.
- Reports must be submitted online via LEARN and/or Crowdmark. Specific instructions will be provided at a later time.

2. Project Description

The file **`chds_births.csv`** contains information on 1236 healthy male single-fetus births collected as part of the Child Health and Development Studies (CHDS) – a comprehensive survey of all pregnancies among women enrolled in the Kaiser Foundation Health Plan in the San Francisco/East Bay area between 1960-1967. The variables in the dataset are:

- `wt`: The birth weight (ounces).
- `gestation`: The length of the gestation period (days).
- `parity`: The total number of previous pregnancies (including fetal deaths and still births).
- `meth`: The self-reported ethnicity of the mother: 0-5 = Caucasian, 6 = Mexican, 7 = African-American, 8 = Asian, 9 = Mixed, 10 = Other.
- `mage`: The mother's age at termination of pregnancy (years).
- `med`: The mother's education: 0 = elementary school, 1 = middle school, 2 = high school, 3 = high school + trade school, 4 = high school + some college, 5 = college graduate, 6 = trade school, 7 = high school unclear.
- `mht`: The mother's height (inches).
- `mwt`: The mother's pregnancy weight (pounds).
- `feth`: The father's ethnicity (same coding as `meth`).
- `fage`: The father's age at end of pregnancy (years).
- `fed`: The father's education (same coding as `med`).

- `fht`: The father's height (inches).
- `fwt`: The father's weight (pounds).
- `marital`: The mother's marital status: 1 = married, 2 = legally separated, 3 = divorced, 4 = widowed, 5 = never married.
- `income`: The family yearly income in $2500 (USD) increments: 0 = under 2500, 1 = 2500-4999, 2 = 5000-7499, 3 = 7500-9999, 4 = 10000-12499, 5 = 12500-14999, 6 = 15000-17499, 7 = 17500-19999, 8 = 20000-22499, 9 = over 22500.
- `smoke`: Does the mother smoke at time of pregnancy? 0 = never, 1 = smokes now, 2 = until pregnancy, 3 = used to, not anymore.
- `time`: Time since the mother quit smoking before pregnancy: 0 = never smoked, 1 = still smokes, 2 = during pregnancy, 3 = less than a year, 4 = 1-2 years, 5 = 2-3 years, 6 = 3-4 years, 7 = 5-9 years, 8 = more than 10 years, 9 = quit but don't know when.
- `number`: Number of cigarettes smoked per day by mother when she was smoking: 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 = more than 60, 9 = smoked but don't know how much.

The goal of this project is to explore the relation healthy male single-fetus birth weight and some explanatory variables. To do this, write a report containing the following sections:

1. Summary

A maximum of 200 words describing the objective of the report, an overview of the statistical analysis, and summary of the main results.

2. Model Selection

Using linear regression methods, consider several models for birth weight as a function of the other variables in the dataset.

- Your analysis must contain at least one pre-fitting data diagnostic, and at least one automated model selection.
- Proper treatment of `NA`'s in the dataset is required. In general, it's acceptable to throw out observations containing missing covariates if they account for less than about 5-10% of your dataset. Throwing out more than 10% of observations can seriously bias your analysis. In this case, the simplest solution[1] is to instead discard covariates having too many `NA`'s (with a proper explanation).
- Discuss any issues encountered during model fitting, e.g., `NA`'s, `±Inf`'s, and how you addressed them.
- Narrow your search down to two candidate models for closer inspection.

[1] While beyond the scope of this course, much better approaches to the problem of `NA`'s consist of "imputing" the missing data, as briefly discussed here.

3. Model Diagnostics

Perform an in-depth comparison of the two candidate models you have proposed by examining the following diagnostics:

- Different types of residual plots. For assessing normality, please use the residuals that would look most normal if the model is correct.
- Leverage and influence measures.
- Cross-validation. Produce boxplots for root mean square prediction error (rPMSE),

$$\text{rPMSE} = \sqrt{\frac{1}{N_{\text{test}}} \sum_{i \in \mathcal{S}_{\text{test}}} (y_i - \hat{y}_i^{\text{train}})^2},$$

  where $\hat{y}_i^{\text{train}} = x_i' \hat{\beta}_{\text{train}}$ is the predicted value[2] for testing observation i based on the training set $\mathcal{S}_{\text{train}}$.

Analyze these diagnostics, and retain one final model. Display its parameter estimates, standard errors, and p-values in a clear and compact table.

4. Discussion

Report what this analysis has taught you about the factors associated/influencing with birth weight of healthy single-fetus males in the CHDS sample. For example:

- What are the most important factors associated with/influencing birth weight?
- Low birth weight is considered to be 88 ounces or less. Based on this analysis, would you be able to recommend behavioral changes to parents in order to avoid low birth weight? If so, please carefully formulate your recommendation.
- Are there any coefficients with high p-values retained in the final model? If so, why?
- Are there any outlying observations that might be appropriate to remove?
- Are any of the regression assumptions of the final model violated? If so, which ones? What are the possible deficiencies of the final model? How do these deficiencies nuance your conclusions/recommendations above?

[2]This formula only applies for regressions with the response fit on the original scale. For log-additive models with the response fit on the log scale, a different formula must be used.

3. Marking Scheme

Content: 50%

- *Correct and efficient programming.*

  Categorical variables should be encoded as factors with meaningful names.

- *Correct and insightful interpretation of results.*

- *Justification of subjective decisions.*

  Data can only be thrown out for valid reasons. A lesser fitting model can be selected if it has a more meaningful interpretation.

- *Originality.*

  Anything that goes beyond an uninspired "all interactions included" model selection and copy-pasting the plots out of the class notes.

Presentation: 50%

- *Organization of information, overall legibility.*

  Present only the most relevant models and output, optionally including further analyses in the Appendix.

- *Clarity of explanations.*

  Use full sentences. Avoid using abbreviations such as `meth` when giving explanations.

- *Properly commented **R** code.*

  A suggestion is to divide the Appendix into clearly labeled blocks of code, each starting with a description of what it does and where to find it in the report.

- *Properly labeled figures, succinctly formatted regression output.*

  Include figure captions or titles. Do not waste space by displaying 100 rows of a matrix 1/3 of the page width at 6 decimal places.