Final Project

VERSION: 2020-04-03 · 16:56:01

1. Instructions

- Due **Thursday, April 16 at 11:59pm**.
- Each group consists of **exactly 2 students** (see below). The group enrolment deadline is **Monday, March 30**. Students who have not enrolled in a group by then will be randomly assigned to a group.
- 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 created with R Markdown and submitted online via LEARN and/or Crowdmark. Specific instructions will be provided at a later time.
- **Lateness Penalty:** 10% per day. Projects turned in after April 19 at 11:59PM will not be graded.

Group Enrolment

1. Login to LEARN and join a Group: At the top of the screen, click

   `Connect > Groups > View Available Groups`

   Agree on a Group number (say N) between 1-89 with your other team members, and click `Join Group` beside `Project > Group` X.

2. Submit files: At the top of the screen, click `Assessments > Dropbox`, then `Group` N`: Project`, then `Add a File`.

The names of all collaborators must be written on your report.

2. Project Description

The file **`fhsd.csv`** contains information on 2306 individuals participating in the Framingham Heart Study. The dataset contains the following variables:

- `chdrisk`: A risk measure for coronary heart disease (CHD) (a probability between 0 and 1).
- `sex`: The sex of the individual.
- `totchol`: Serum total cholesterol (mg/dL).
- `age`: Age of individual (years).
- `sysbp`: Systolic blood pressure (mmHg).
- `diabp`: Diastolic blood pressure (mmHg).

- `cursmoke`: Currently a cigarette smoker.
- `cigpday`: Number of cigarettes smoked each day.
- `bmi`: Body mass index (kg/m^2).
- `diabetes`: Whether or not the individual is diabetic.
- `bpmeds`: Whether or not the individual is on anti-hypertensive medication.
- `heartrte`: Heart rate (beats/min).
- `glucose`: Casual serum glucose (mg/dL).
- `prevmi`: Whether or not the individual has had a myocardial infarction.
- `prevstrk`: Whether or not the individual has had a stroke.
- `prevhyp`: Whether or not the individual has hypertension.
- `hdlc`: High density lipoprotein cholesterol (mg/dL).
- `ldlc`: Low density lipoprotein cholesterol (mg/dL).

The goal of this project is to explore the relation between the risk score for CHD 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. Descriptive Statistics

Display summary statistics, pair plots, and calculate the VIF for the explanatory variables. Comment on your findings.

3. Candidate Models

Using

$$\text{logit}(\texttt{chdrisk}) = \log(\texttt{chdrisk}) - \log(1 - \texttt{chdrisk})$$

as the response variable, create two candidate models:

1. The first candidate model must be obtained using automated model selection. Justify your choice of inputs to the automated selection procedure(s)[1].
2. The second candidate model should be manually constructed using at least one F-test, with a priority given to interpretability of the model. Justify your decision process in arriving at this model.

4. 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.

[1] You are welcome to compare several automated selections, and explain how you pick just one.

- Leverage and influence measures.

Comment on your findings.

5. Model Selection

Pick one of the two candidate models using your judgment to balance predictive vs explanatory power. Perform a cross-validation analysis to assess the former. Produce boxplots for root mean square prediction error (rMSPE),

$$\mathrm{rMSPE} = \sqrt{\frac{1}{N_{\mathrm{test}}} \sum_{i \in \mathcal{S}_{\mathrm{test}}} (\mathtt{chdrisk}_i - E[\mathtt{chdrisk} \mid x = x_i, \hat{\beta}_{\mathrm{train}}, \hat{\sigma}_{\mathrm{train}}])^2}\,,$$

where $\hat{\beta}_{\mathrm{train}}$ and $\hat{\sigma}_{\mathrm{train}}$ are the parameter estimates based on the training data $\mathcal{S}_{\mathrm{train}}$.

For a regression model of the form $\mathrm{logit}(\mathtt{chdrisk}) \mid x \sim \mathcal{N}(x'\beta, \sigma^2)$, there is no closed-form solution for the conditional mean $E[\mathtt{chdrisk} \mid x]$. However, it can be approximated fast and accurately by the calculation described in Appendix A. Write an R function

```
logitnorm_mean <- function(mu, sigma)
```

to calculate $E[Y]$ for $\mathrm{logit}(Y) \sim \mathcal{N}(\mu, \sigma^2)$, and fully document this function as we have seen in class. Then, check that your function is correctly implemented by running the following code[2]:

```
mu <- c(0.7, 3.2, -1.1)
sigma <- c(.8, .1, 2.3)
# logitnorm_mean only accepts one (mu, sigma) pair at a time
sapply(1:3, function(ii) logitnorm_mean(mu[ii], sigma[ii]))
```

```
[1] 0.6491002 0.9606606 0.3530580
```

Based on the predictive cross-validation assessment, your judgment regarding interpretability of the models, and the model diagnostics of Section 4, retain one final model. Display its parameter estimates, standard errors, and p-values in a clear and compact table.

6. Discussion

Report what this analysis has taught you about the factors associated with CHD risk. For example:

- What are the most important factors associated with high CHD risk? With low CHD risk?

[2] If you cannot get the `logitnorm_mean()` function to work, then you can conduct the analysis using the response variable $\log(\mathtt{chdrisk})$ instead (with some penalty). If you do this, make sure you calculate $E[\mathtt{chdrisk} \mid x, \hat{\beta}, \hat{\sigma}]$ properly as we have seen in class.

- Based on this analysis, would you be able to recommend behavioral changes to lower the risk of CHD? 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?

A. Appendix

Include **all R code** here, and any additional analyses that couldn't make it into the main body of the report.

3. Grading

In addition to submitting the PDF report, you must submit the R Markdown (`Rmd`) file used to generate the report, along with all accompanying helper files. Such helper files include:

- (Mandatory) Saved results of any calculation that takes more than a few seconds, as the TAs should not have to wait around for several minutes to compile your report. Please see the document *R Markdown: Time-Consuming Calculations* on LEARN for instructions on how to include such calculations in your report.
- (Optional) An external R script containing your code. While it is possible have all the R code directly in the R Markdown document, in a lengthy report such as this it can make the code difficult to manage as it spans across multiple blocks of text. Please see the updated document *R Markdown: Formatting Tips* on LEARN for instructions on how to include external R scripts in your report.

The grading of the report will consider the following elements:

- The report is well-written:
  - ⋆ Ideas are clearly expressed.
  - ⋆ All required elements are provided.
  - ⋆ Report is well organized with proper sections and subsections.
  - ⋆ Use complete sentences.
  - ⋆ Avoid abbreviations when providing explanations e.g., "`hdlc` is an important predictor".
  - ⋆ The page limit is *strict*: present only the most relevant models and output, optionally including further analyses in the Appendix.
  - ⋆ Correct and insightful interpretation of results.
  - ⋆ Justification of subjective decisions.
- The report is well-presented:

  - Figures and Tables have captions.
  - Sections, Figures, Tables, equations, etc., always referred to with hyperlinks.
  - Figures have proper axis labels (not e.g., `predict(M5)`), titles, and legends if more than one thing is being plotted.
  - All elements of Figures are properly sized (points, titles, axis labels, margins, legends, etc.).
  - Tables (and numbers in general) contain the appropriate amount of significant digits.
  - Tables do not waste space, e.g., by displaying 100 rows of a matrix at one quarter of the page width. Please take a look at the **kableExtra** package to help with this.
  - Equations are numbered and properly formatted with LaTeX.

- The R code is correct and efficient:
  - The submitted `Rmd` file (and helper files) generating the PDF report compiles without errors.
  - Use built-in R commands whenever possible.
  - Avoid inefficient for-loops.

- The R code is easy to read and assess for correctness:
  - Code is organized into clearly labelled sections.
  - Variables have informative names.
  - Functions are documented as we have seen in class and on HW1.
  - Code is extensively commented throughout.

A. Mean of the Logit-Normal Distribution

Suppose that Y is a random variable such that $\operatorname{logit}(Y) \sim \mathcal{N}(\mu, \sigma^2)$, and let $\nu = 1/(1+e^{-\mu})$. Then we have

$$E[Y] \approx \sum_{i=1}^{10} w_i \cdot \exp\{g(x_i)\},$$

where

$$\begin{aligned} g(x) = &\texttt{dnorm(}\operatorname{logit}(x)\texttt{, mean = }\mu\texttt{, sd = }\sigma\texttt{, log = TRUE)} - \log(1-x) \\ &- \texttt{dbeta(}x\texttt{, shape1 = }\alpha_1\texttt{, shape2 = }\alpha_2\texttt{, log = TRUE),} \end{aligned}$$

$$\alpha_1 = \frac{1}{\sigma^2(1-\nu)}, \qquad \alpha_2 = \frac{1}{\nu\sigma^2},$$

and $x = (x_1, \ldots, x_{10})$ and $w = (w_1, \ldots, w_{10})$ correspond to list elements `nodes` and `weights` as returned by the `gauss.quad.prob()` function in the R package **statmod**, namely:

$$\texttt{gauss.quad.prob(n = 10, dist = "beta", alpha = }\alpha_1\texttt{, beta = }\alpha_2\texttt{)}$$

Thus for example we have

```r
require(statmod) # load the statmod package (after having installed it)
gqp <- gauss.quad.prob(n = 10, dist = "beta", alpha = 1.5, beta = 6.8)
gqp$nodes # (x_1, ..., x_10)
```

```
[1] 0.01380234 0.05448303 0.11990180 0.20661845 0.31007705 0.42485193
[7] 0.54494738 0.66414978 0.77647526 0.87708843
```

```
gqp$weights # (w_1, ..., w_10)
```

```
[1] 6.281845e-02 1.903633e-01 2.671645e-01 2.402066e-01 1.501551e-01
[6] 6.583977e-02 1.952705e-02 3.580580e-03 3.349187e-04 9.738733e-06
```

Note that the calculation of $g(x)$ is on the log scale. While we could have calculated $h(x) = \exp\{g(x)\}$ directly using the ratio of `dnorm()` and `dbeta()`, this usually leads to considerable roundoff errors when numerator and denominator are both very big or very small.