STA 712 Project 1: Logistic regression

Due: Friday, September 29 at 12:00pm (noon) on Canvas

Data

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey of 350,000 people in the United States. As its name implies, the BRFSS is designed to identify risk factors in the adult population and report emerging health trends. For example, respondents are asked about their diet and weekly physical activity, their HIV/AIDS status, possible tobacco use, and even their level of healthcare coverage. The BRFSS Web site contains a complete description of the survey, including the research questions that motivate the study and many interesting results derived from the data.

We will focus on a random sample of 20,000 people from the BRFSS survey. While there are over 200 variables in this data set, we will work with a small subset for this project.

Variables

The dataset provided for this project contains the following columns:

- `genhlth`: respondents were asked to evaluate their general health, responding either excellent, very good, good, fair or poor.
- `exerany`: indicates whether the respondent exercised in the past month (1) or did not (0).
- `hlthplan`: indicates whether the respondent had some form of health coverage (1) or did not (0).
- `smoke100`: indicates whether the respondent had smoked at least 100 cigarettes in their lifetime.
- `height`: in inches
- `weight`: in pounds
- `wtdesire`: desired weight in pounds
- `age`: in years
- `gender`: biological sex, limited to male/female.

Downloading the data

To load the data, use the code below. It will import a data set called `cdc` into R.

```
source("http://www.openintro.org/stat/data/cdc.R")
```

LaTeX template

A LaTeX template for this project is available in a ZIP file on the course website.

- Download and extract the ZIP file
- Create a new project on Overleaf (e.g. "STA 712 Project 1")
- Upload all files and folders from the extracted template folder
- Compile
- Make edits to the manuscript, references, and images as needed

Research questions

You are asked by a team of researchers to investigate respondents' exercise habits. In particular, they are interested in the following two questions:

1. Is there a relationship between how much weight someone wants to lose, and the probability that they exercise regularly, after accounting for their age, general health, and health coverage?
2. How well can we predict whether a patient exercises regularly, using other variables in the data?

To address these research questions, you will perform exploratory data analysis, fit one or more models, and analyze the results.

Assignment

In this assignment, you will use logistic regression to address the research questions above. You will then communicate your findings in a written report. Your written report will be formatted similar to a submission to the academic journal *PLOS One*. Please read the *PLOS One* submission guidelines for full details, paying particular attention to the "Manuscript Organization", "Parts of a Submission", and "Statistical reporting" sections.

In particular, your manuscript must be organized with the following sections:

- **Abstract**: an overview of the main objectives and results
- **Introduction**: Background, motivation, a brief review of related literature, a summary of the research questions, and a summary of the main results
  - A brief literature survey of previous work related to the research questions (ballpark: 5–10 references) is required. All citations must be properly formatted.
  - State the research questions in your own words. Do not copy my wording from the prompt.
- **Methods**: Description of the data, an explanation of how the research questions will be addressed, and a summary of the statistical methods used. Enough detail should be provided that a statistician could reproduce your analysis without access to the source code.
- **Results**: The main results of your analysis, including the fitted regression models, diagnostics and assessments of model performance, and hypothesis tests.

- **Discussion**: A discussion of the results in context of the research questions, and a conclusion summarizing the main findings of the analysis.
- **Supplementary information**:
  - A link to a GitHub repository containing the code you used in your analysis
  - Any supplementary figures or tables (e.g., additional EDA) not included in the main text
- **Acknowledgments**: Acknowledge any students with whom you discussed the analysis.
- **References**: A properly formatted bibliography for all references in the manuscript.

You will follow the *PLOS One* guidelines for manuscript style and format, with a few exceptions to make the manuscript easier to read:

- Figures and tables *should* be included in the manuscript
- You may use BibTex and a .bib file to format references
- Supplementary figures and tables may be included at the end of the manuscript, in the same file (you do not need to submit separate files)

Requirements

Content

Your report must contain:

- A brief literature survey of previous work related to the research questions (ballpark: 5–10 references). All citations should be properly formatted
- A description of the data, including
  - The source of the data
  - The size of the data, what a row in the data represents, what variables are available, and whether there is any missing data
- Any exploratory data analysis needed to address the research questions, including
  - A description of any transformations or manipulations (e.g., creating a new variable, removing missing observations, etc.)
  - A (well formatted) table summarizing the relevant variables (as an example, see Table 1 in the dengue paper from class)
  - Any essential figures informing your data analysis. **Note:** do not include all figures you create during your exploration of the data! Pick the most important 1–3 figures. Other figures can be included in the supplementary materials, if necessary.
- An explanation of how you chose the final model(s) to answer the research questions
- A (well formatted) table of coefficients for your final logistic regression model(s) (see, for example, Table 3 in the dengue paper from class)
- Model diagnostics and assessments for your final logistic regression model(s), such as:
  - Quantile residual plots for key quantitative explanatory variables

  - Assess influential points with Cook's distance, and multicollinearity with VIFs
  - ROC curves and performance metrics (AUC, sensitivity, specificity, etc.) if the goal is prediction

- When testing a hypothesis to answer a research question, clearly state the null and alternative hypotheses, report both a test statistic and p-value, and use the p-value to make a conclusion about the research question

- In the supplementary information, a link to a GitHub repository containing all code needed to reproduce the analysis in your report

Format and style

- The report should be written like an article or research paper: in full sentences and paragraphs, with headings for each section. You should not write your report with question numbers or as a list of bullet points. Scientific articles are generally written in third person, though "we" can also be acceptable ("we can see from Figure 2.1...").

- No R code or raw R output may be included in your report. Only properly formatted figures and tables may be included.

- Your document must be prepared using the LaTeX template provided, and submitted on Canvas as a pdf.

- Figures should have labeled axes, and should be clear and easy to read. Figures should also be captioned and numbered. Captions should provide enough information to understand what is being plotted, but interpretation can be left to the main text. Refer to figures by their number in the text. Make sure that any figures you include are discussed in the text.

- Tables should be nicely formatted, and have a number and caption. As with figures, refer to tables by their number in the text.

Grading

Your report will be graded on a mastered/not yet mastered basis. Mastery requires meeting all of the Criteria for Publication for *PLOS One* (see the link for full details on the criteria).

I will grade your manuscript as a reviewer for *PLOS One*. I will make one of the following decisions, and provide feedback on your initial submission:

- *Accept*: the manuscript can be accepted as-is; no revisions are necessary.

- *Minor revisions:* the manuscript is in good shape, but some minor changes (formatting, some additional explanation, etc.) need to be made before acceptance

- *Major revisions:* the work has potential promise, but major changes (e.g., a revised or expanded analysis) need to be made

You will have one opportunity to re-submit the manuscript if you do not receive an "Accept" on your first submission. To master the project, you must either:

- Receive an "Accept" on your initial submission, OR

- Receive either an "Accept" or "Minor revisions" on your second submission

Collaboration

You are welcome to work with other students on this project, but everyone needs to submit their own report. Acknowledge any collaboration in the "Acknowledgments" section of your report.

Resources

GitHub

You are required to host your code for this project in a GitHub repository. See the GitHub docs for guidance on getting started.

LaTeX

Your written report must be created using LaTeX. See the course website for resources on getting started with LaTeX. Overleaf also has a lot of good tutorials to help you get started.

Google Scholar is useful for getting BibTex references. In Google Scholar, search for an article, then click on "Cite", then "BibTex". Copy the resulting BibTex citation into your .bib file.

Figures and tables

- Examples are provided in the LaTeX template
- See the Overleaf tutorial on including figures and tables in LaTeX documents
- The `xtable` R package is useful for creating LaTeX tables
- The online LaTeX table generator is also helpful
- Consider the `gridExtra` and `patchwork` R packages for arranging multiple R figures into a single image