BIOS 6612 Final Project 2021

Guidelines

Basic rules:

1. You should not discuss this exam with anyone else.
2. You may use any resources (books, literature, internet) except another individual.
3. If you have questions about the exam, please post on the discussion board so that all students can benefit from responses.

General guidelines for answering questions:

1. Answer each question completely, but be concise.
2. For any models you build, clearly write out the form of the model and any relevant modeling assumptions (i.e. independence, homoscedasticity, data distribution, etc.)
3. Organize your answers so that they are easy to follow and easy to read. You should type your answers.
4. Do not submit unnecessary computer output. The output that you submit should be referenced in your answer, and the output should be organized and annotated so that we know how you are interpreting the results.
5. Summarize or interpret an analysis for a non-statistical (e.g., clinical) investigator. When answering these kinds of questions, you should use statistical terminology that would be understood by such an investigator.
6. Append any extra R code used for analysis to the end of your document

The Data

> The data are from a prospective cohort study of 609 white males in Evans County, Georgia who were followed for 7 years. Investigators are interested in the association between coronary heart disease (CHD) and stress, as measured by catecholamine level in the blood (adrenal glands send catecholamines into your blood when you are physically or emotionally stressed). Other potential confounding variables were measured as well.

The raw data are stored in an external file: `evanscounty.csv`. Each row of the data set contains the following variables:

- `id`: the subject identifier
- `chd`: the outcome, coronary heart disease status (0 = no chd, 1 = chd)
- `cat`: catecholamine level, the main exposure of interest (0 = normal, 1 = high)
- `age`: age in years
- `chl`: cholesterol level (mg/dL)
- `smk`: smoking status (0 = nonsmoker, 1 = smoker)
- `ecg`: ECG status (0 = normal, 1 = abnormal)
- `dbp`: diastolic blood pressure
- `sbp`: systolic blood pressure
- `hpt`: hypertension status (0 = normal, 1 = high blood pressure)

Part 1: EDA

Perform exploratory data analysis to better understand the association between the outcome and exposure of interest. Explore relationships between CHD status and potential confounders as well. Present your results as a table one of descriptive statistics in both the CHD and no CHD groups. Also include any informative plots that highlight these relationships.

Part 2: Regression Modeling

Investigators want to to assess the relationship between CHD and catecholamine level. Build a model to capture this relationship, controlling for any potential confounders or effect modifiers. Defend your modeling choices using findings from your exploratory analysis, test(s) of model fit and model diagnostic plot(s). Present the results of your final model in a neat table.

Part 3: Model Interpretation

Summarize and interpret the findings from your model for your collaborators in language they will understand. Include relevant effect size(s), p-value(s), and confidence interval(s).

Part 4: Prediction

As a secondary analysis, you decide to assess the predictive capacity of your model. Using a threshold of 0.5 on the predicted probabilities of your model, generate a confusion matrix for your model. What are the false positive and false negative rates? Construct and interpret an ROC curve for your model. What is the AUC? Is the threshold of 0.5 reasonable?