STA 214 Homework 6

Due: Friday, March 3, 12:00pm (noon) on Canvas.

Instructions: There are two parts to this assignment. In the first part, you will read a research paper about predicting dengue fever in hospital patients. In the second part, you will assess the predictive ability of logistic regression models on the dengue data, and experiment with variable selection.

Getting started: Begin by downloading the HW6 template from the course website:

`https://sta214-s23.github.io/homework/hw_06_template.Rmd`

Save this template file to your computer, then open it in RStudio. As you complete the assignment, you will write down your answers to all questions in the R Markdown file, and include all R code in code chunks. *If a question requires code, you will not receive credit if no code is provided.* Refer to the R Markdown instructions on the course website (`https://sta214-s23.github.io/resources/rmarkdown_instructions/`) if you have issues getting started.

Submission: When you have completed the assignment, knit your homework to HTML and submit on Canvas.

1 Reading a research paper

Statistics is an important research tool used in many fields. In this section of the assignment, you will read the original research article analyzing the dengue data which we have used in class. The purpose of this section is to help you learn how to read a research paper, and extract key details about the study and results. Later in this assignment, we will try to replicate the authors' results ourselves.

Overview

Dengue is a mosquito-borne viral disease which affects hundreds of millions of people each year. Early diagnosis is crucial for patients to have the best prognosis, but relies on a variety of laboratory tests. To enhance practitioners' ability to diagnose dengue, a 2015 paper by Tuan *et al.*[1] investigated the possibility of detecting dengue using a variety of clinical measurements like white blood cell count and platelet count. This activity will help you read the paper by Tuan *et al.*. The paper is available at

`https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0003638`

Outline of a research paper

Research papers in many fields, particularly the sciences, often contain the following main sections:

- **Abstract:** A short overview of the full paper, giving highlights of the motivation and background, the research question, the data, and the results.

[1] Tuan, Nguyen Minh, et al. (2015) "Sensitivity and specificity of a novel classifier for the early diagnosis of dengue." *PLoS neglected tropical diseases* 9.4 (2015): e0003638.

- **Introduction:** A broad overview of the research question the authors want to study, motivation for studying this question, and the authors' approach to answering their question. The introduction often starts very general, then narrows to the specific question addressed in this paper. More detail is provided in the introduction than in the abstract, and more time is spent on motivation and related literature.
- **Methods:** The data and analysis techniques used to answer the research question. This typically describes the what the data looks like, how and where it was collected, and any statistical tools (e.g. visualizations, regression, hypothesis testing) that were used when analyzing the data.
- **Results:** A summary of the analysis results, such as figures showing regression fits, and tables of regression coefficients and p-values.
- **Discussion:** A discussion of the analysis results, in context of the original research question. In this section, explanations for *why* particular results were observed may be proposed.
- **Conclusion:** A short summary of the paper and its key results, and their connections to broader scientific questions. The conclusion is often the reverse of the introduction: it starts with the specific question addressed by this paper, then discusses the implications of this research for science in general.

Reading a research paper

Reading a research paper, particularly in a field in which you are not an expert, can be challenging. The trick is to skim the paper for the most relevant information, and skip over technical details that are not essential to understanding the key take-aways. The questions below will guide you to the most important sections in the paper by Tuan *et al.*

The Abstract and Introduction

A good place to start is often with the Abstract and Introduction, which allow you get an overview of the paper, and usually don't contain too many technical details. The Abstract is more succinct than the Introduction, but it also provides less motivation. When the Introduction is long, you may want to skim for key details.

Read the Abstract and introduction. Then answer the following questions.

1. Why is it important for the researchers to build a model to detect dengue in hospital patients?
2. What is the specific purpose of the research study?

So far, we know what question the researchers are trying to answer, and we know that they are going to build some kind of model to predict dengue. Our goal for the rest of the paper is to understand how the authors conducted this analysis. In particular, we want to answer the following questions:

- Who participated in the study?
- What did the researchers record about each participant?
- What statistical methods did the researchers use?
- What did the researchers conclude from their study?

This information is provided in the Methods and Results sections of the paper. These sections also contain lots of other details which is valuable, but not crucial to understand on a first reading, so we will focus on the most important parts of the Methods and Results.

Study Participants

Read the *Patient Enrolment* subsection of the Methods, and then answer the following questions.

3. How many patients participated in the study?

Potential research subjects must meet certain criteria, defined by the researchers, to participate in a study. *Inclusion* criteria define requirements for inclusion in the study (e.g., a target age range or social group), while *exclusion* criteria are reasons a subject would be asked not to participate (e.g., certain medical conditions).

4. What are the inclusion/exclusion criteria for this study?

Data Collection and Analysis

Now that we know who was studied, we want to know what data was collected about each participant, and how it was analyzed. Read the *Clinical and laboratory investigations on the day of enrolment* and *Statistical methods* subsections of the Methods. Then answer the following questions.

5. Which variables were recorded for the patients in the study?
6. Which types of statistical methods were used to model the relation between the explanatory variables and whether the patient had dengue? (It is ok if you're not familiar with all these methods!)
7. How did the researchers choose which variables to include in their logistic regression model?
8. Which threshold did the researchers use when converting their predicted probabilities into binary predictions?

Results

Finally, let's see what the researchers concluded from their statistical models. Read the *Early Dengue Classifier* subsection of the Results, then answer the following questions.

9. Which variables did the researchers include in their final logistic regression model?
10. What sensitivity and specificity values did the researchers observe using their final model?
11. What Area Under the ROC Curve (AUC) did the researchers observe?

2 Replicating results

Now that you've read the dengue paper by Tuan *et al.*, we will try to replicate their results. I have downloaded their data, and performed some initial data cleaning for you. The prepared data can be loaded into R using the following command:

```
dengue <- read.csv("https://sta214-s23.github.io/homework/dengue.csv")
```

The prepared data contains 5720 patients, with the following variables:

- SiteNumber: The hospital at which the data was recorded
- Sex: patient's sex (female or male)

- Age: patient's age (in years)
- DiseaseDay: how long the patient has been ill
- Vomiting: whether the patient has experienced vomiting (0 = no, 1 = yes)
- Abdominal: whether the patient has abdominal pain (0 = no, 1 = yes)
- Temperature: patient's body temperature (in Celsius)
- BMI: the patient's body mass index (BMI)
- WBC: the patient's white blood cell count
- HCT: the patient's hematocrit
- PLT: the patient's platelet count
- RapidTest: predicted disease status from a rapid test (positive or negative)
- Dengue: whether the patient actually has dengue fever, based on a lab test (0 = no, 1 = yes)

12. First, let's look at the rapid test.

    (a) Create a confusion matrix for the predictions from the rapid test. Note that you will not need to threshold these predictions, as the rapid test already makes binary predictions!

    (b) Calculate the accuracy, sensitivity, specificity, and positive predictive value for the rapid test.

13. Next, let's look at the final model chosen by the researchers. Their Early Dengue Classifier uses age, white blood cell count, and platelet count to predict dengue status.

    (a) Fit a logistic regression model to predict dengue status using age, white blood cell count, and platelet count.

    (b) Create a confusion matrix for the predictions from your fitted model. Use the same threshold as in the paper (0.333).

    (c) Calculate the accuracy, sensitivity, specificity, and positive predictive value for your logistic regression model. Are the values close to the values reported in the original paper? (It is ok if they don't match exactly)

    (d) How does the logistic regression model perform, compared to the rapid test?

    (e) Now let's create an ROC curve for the logistic regression model, so we can assess predictive performance across different thresholds. If your logistic regression model is named `m1`, the following code will create the ROC curve and calculate the AUC. Run the code below to calculate the AUC and make the plot; is the AUC similar to the value reported in the original paper?

```
library(ROCR)
pred <- prediction(m1$fitted.values, dengue$Dengue)
perf <- performance(pred,"tpr","fpr")

performance(pred, "auc")@y.values

data.frame(fpr = perf@x.values[[1]],
           tpr = perf@y.values[[1]]) %>%
  ggplot(aes(x = fpr, y = tpr)) +
  geom_line(lwd=1.5) +
  geom_abline(slope = 1, intercept = 0, lty = 2,
              lwd = 1.5) +
  labs(x = "False positive rate (1 - Specificity)",
       y = "True positive rate (Sensitivity)") +
  theme_classic()
```

14. Finally, let's experiment with model selection to see if we get a different model than the one selected by the researchers.

    (a) Adapt the code from the class activity on February 13 to perform forward stepwise selection with AIC on the dengue data. Your response variable should be `Dengue`, and your full model (the `scope` in the `stepAIC` function) should contain all explanatory variables **except** for RapidTest and SiteNumber. Which variables are chosen in forward stepwise selection?

    (b) Calculate an AUC for the model chosen by forward stepwise selection. Is it very different from the AUC of the model in question 13?

    (c) Explain why the researchers preferred the model from question 13.