STA 712: HW 2

Due: Friday, September 15, 12:00pm (noon) on Canvas

Instructions: Submit your work as a single PDF, or as two separate PDFs (one for Parts 1 and 2, and another for Part 3). Parts 1 and 2 should be created using R Markdown or Quarto, so that all code needed to reproduce your results is included in the knitted document. Part 3 should be created using LaTeX; see the course website for a homework template file and instructions on getting started with LaTeX and Overleaf. See the Overleaf guide on mathematical expressions to get started writing math in LaTeX.

1 Reproducing the dengue results

Now that you've read the dengue paper by Tuan *et al.*, we will try to reproduce 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)

1. 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, and specificity for the rapid test.

2. 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, and specificity 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)
library(tidyverse)
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()
```

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

   (a) Use code from class 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 2?

   (c) Explain why the researchers preferred the model from question 2.

2 Data analysis

You are contacted by the US Small Business Administration (SBA), a government agency dedicated to helping support small businesses. The SBA provides loans to small businesses, but some businesses *default* on their loan (i.e., fail to pay it back). Researchers at the SBA are interested in predicting whether a business will default on the loan, and they have collected a random sample of 5000 different loans.

You can load the SBA data into R by

```
sba <- read.csv("https://sta712-f22.github.io/homework/sba_small.csv")
```

For each loan, we have the following variables:

- LoanNr_ChkDgt: Loan ID number that uniquely identifies each loan
- Name: Name of business receiving the loan
- City: City the business is based in
- State: State the business is based in (two-letter abbreviation)
- Zip: ZIP code the business is based in
- Bank: Name of bank making the loan
- BankState: State of the bank making the loan (two-letter abbreviation)
- NAICS: North American Industry Classification System code identifying the industry of the business receiving the loan
- ApprovalDate: Date of approval (YYYY-MM-DD) of the loan
- ApprovalFY: Fiscal year of approval of the loan
- Term: Length of the loan term (months)
- NoEmp: Number of employees of the business before receiving the loan
- NewExist: 1 if business already existed, 2 if business is new
- CreateJob: Number of jobs the business expects to create using the loan money
- RetainedJob: Number of jobs the business expects to retain because they received the loan
- FranchiseCode: For businesses that are franchises, a unique five-digit code identifying which brand they are a franchise of. 0 or 1 if the business is not a franchise.
- UrbanRural: 1 if business is in urban area, 2 if business is in rural area, 0 if unknown
- RevLineCr: Y if this is a revolving line of credit, N if not
- LowDoc: Y if loan was issued under the 'LowDoc Loan' program, which allows loans under $150,000 to be processed with a short one-page application. N if loan is issued with a standard application, which is much longer
- ChgOffDate: The date (YYYY-MM-DD) the loan was declared to be in default, if the borrower stopped paying it back

- DisbursementDate: Date (YYYY-MM-DD) the loan money was disbursed to the business
- DisbursementGross: The amount of money disbursed (loaned), in dollars
- BalanceGross: The amount of money remaining to be paid back, in dollars
- MIS_Status: Current loan status. CHGOFF = charged off, P I F = paid in full.
- ChgOffPrinGr: Amount of money charged off, if the borrower defaulted, in dollars
- GrAppv: Gross amount of loan approved by the bank, in dollars
- SBA_Appv: Amount of the loan guaranteed by the SBA, in dollars

Research question: Suppose that researchers at the SBA would like to predict which loans are likely to default; this will help them decide how to best allocate loans to the businesses applying. They ask you to build a model to predict loan default, taking into account the following preferences:

- Saving money is a top priority, so they want a model that will do a good job at predicting defaults for new loans
- However, the final model also has to be understandable by loan officers and the businesses applying. In particular, the SBA wants to be able to explain to businesses *why* an application has been denied.

Taking these preferences into account, build a model to predict loan defaults using the SBA data. Then answer the following questions. Make sure to provide all code and results. (Note: the questions do not ask you about exploratory data analysis, but EDA is always a good idea).

4. Summarize your final model, and describe how you chose that model.
5. Interpret some of the coefficients in your final model.
6. Assess and interpret the predictive performance of your final model.

3 Two views of AIC

In class, we discussed train/test splits, k-fold cross validation, and leave-one-out cross-validation (LOOCV) as methods for estimating model performance on new data, and thereby choosing between different models. We also discussed AIC and BIC as alternatives which are less computationally intensive. While AIC and BIC intuitively penalize the log-likelihood for the number of parameters in the model, it is not immediately clear how AIC or BIC is connected to model performance on new data. In this section, you will explore two different perspectives of AIC to better understand why AIC is a reasonable metric for comparing models: a derivation of AIC for estimating the KL divergence, and an asymptotic equivalence between AIC and LOOCV.

AIC and the KL divergence

Let Y denote the response variable of interest in a model, and let $f_0(y)$ denote the *true* probability function of the response (i.e., the true distribution that generated the data). For simplicity, assume here that Y is continuous, so $f_0(y)$ is a density.

We consider a family of distributions $\{f_\theta(y) : \boldsymbol{\theta} \in \Theta\}$, parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^p$. Our goal, when building a model, is to choose the parameters $\boldsymbol{\theta}$ which make $f_\theta(y)$ as close to the *true*

distribution $f_0(y)$ as possible.

One method of measuring how close $f_\theta(y)$ is to $f_0(y)$ is with the **Kullback-Leibler** (KL) divergence:

$$K(f_\theta, f_0) = \int \log\left(\frac{f_0(y)}{f_\theta(y)}\right) f_0(y)dy. \tag{1}$$

We fit our model via maximum likelihood estimation, producing the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ (which depends on the observed data). Then, $K(f_{\widehat{\theta}}, f_0)$ is one measure of how close our model is expected to fit a *new* set of data. And one way of choosing a model would be to choose the model which does best at predicting new data, i.e. which minimizes the KL divergence.

In the following questions, you will explore some basics of KL divergence, and show that choosing a model to minimize AIC is equivalent to minimizing an estimate of the KL divergence.

1. Show that $K(f_\theta, f_0) = 0$ if $f_\theta = f_0$.
2. Using the fact that $\log(x) \leq x - 1$ for all x, show that $K(f_\theta, f_0) \geq 0$ if $f_\theta \neq f_0$.
3. Conclude that, if we *could* calculate KL divergence for our models, and the true model f_0 was among the models considered in the model selection procedure, then minimizing KL divergence would find the true model.
4. Explain why we can't actually calculate KL divergence when modeling real data.

Since we can't calculate KL divergence, we to approximate it. Let $\boldsymbol{\theta}^*$ be the value of $\boldsymbol{\theta}$ which minimizes $K(f_\theta, f_0)$. A Taylor expansion gives

$$K(f_{\widehat{\theta}}, f_0) \approx K(f_{\theta^*}, f_0) + \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \mathcal{I}(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*), \tag{2}$$

where $\mathcal{I}(\boldsymbol{\theta}^*)$ is the information matrix at $\boldsymbol{\theta}^*$.

1. We know from STA 711 that, if the model is correct and the sample size is sufficiently large, then $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \mathcal{I}(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \approx \chi^2_p$. Conclude that under these assumptions,

$$\mathbb{E}[K(f_{\widehat{\theta}}, f_0)] \approx K(f_{\theta^*}, f_0) + p/2. \tag{3}$$

So far, our approximation to $K(f_{\widehat{\theta}}, f_0)$ is $K(f_{\theta^*}, f_0) + p/2$ (we can already see a penalty term that involves the number of parameters!). However, $K(f_{\theta^*}, f_0)$ still depends on f_0. To estimate $K(f_{\theta^*}, f_0)$, let $\boldsymbol{Y} = [Y_1, ..., Y_n]^T$ denote the observed data used to calculate $\widehat{\boldsymbol{\theta}}$, and consider the log-likelihood $\ell(\widehat{\boldsymbol{\theta}}) = -\sum_i \log f_{\widehat{\theta}}(Y_i)$.

1. Under the same assumptions as the previous question, $2(\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}^*)) \approx \chi^2_p$. Use this to argue that

$$K(f_{\theta^*}, f_0) \approx \mathbb{E}[-\ell(\widehat{\boldsymbol{\theta}})] + p/2 + \int \log(f_0(y)) f_0(y)dy. \tag{4}$$

2. Combine the previous questions to argue that an estimate of $K(f_{\widehat{\theta}}, f_0)$ is

$$\widehat{K(f_{\widehat{\theta}}, f_0)} = -\ell(\widehat{\boldsymbol{\theta}}) + p + \int \log(f_0(y)) f_0(y)dy. \tag{5}$$

3. Conclude that minimizing $\widehat{K(f_{\widehat{\theta}}, f_0)}$ is equivalent to minimizing

$$AIC = -2\ell(\widehat{\boldsymbol{\theta}}) + 2p.$$

This derivation isn't super rigorous, but it captures the main steps and the important intuition: AIC is a reasonable metric for model selection because minimizing AIC is equivalent to minimizing an estimate of the KL divergence.

AIC and LOOCV

Another way of viewing AIC is as a computationally efficient approximation of leave-one-out cross-validation (LOOCV). Indeed, Stone (1977) showed that (under appropriate assumptions), *AIC and LOOCV are asymptotically equivalent.* Stone's paper is quite short (indeed, it is actually published in the journals *Notes, Queries, and Comments* rather than as a full-length research article), and this part of the assignment will guide you through reading the original paper.

> Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 44-47.

Begin by reading sections 1 – 3 of the paper, which set up the background and problem the author is trying to solve. (Because this is a short research note, rather than a full paper, the structure of the article is a bit different than usual).

1. In this paper, what is the LOOCV metric that we want to minimize? That is, what value is calculated for each held-out observation?
2. The goal of the article is to show that model selection by AIC and LOOCV are asymptotically equivalent. In particular, which two quantities (describe the quantities and give the equation numbers) does the author plan to show are equivalent?

Now read section 4 (you may need to read it several times). For your first read-through, I suggest skimming over details like regularity conditions, and focusing on the general outline of the mathematical results. In particular:

1. What technique (which we have used many times in class!) does the author use to characterize the large-sample behavior of A?
2. What is the key assumption that makes A asymptotically equivalent to AIC?

Finally, read through section 4 again, keeping the outline and direction of the derivation in mind as you read. This time, pay more attention to the mathematical details and see if you can follow the steps. The key equations are (4.4), (4.5), and (4.6).

1. What asymptotic results are used to move from equation (4.4) to equation (4.5)?
2. How does the "key assumption" allow us to write equation (4.5) as equation (4.6)? In particular, what is required for $L_1 = -L_2$? (We have seen this equivalence before in STA 711!)

Reflecting

In this assignment, we have seen two justifications of AIC. First, using AIC is reasonable because minimizing AIC is the same as minimizing an estimate of the KL divergence. And minimizing an estimate of the KL divergence is reasonable because *if* the true model is contained in our search, and *if* we could minimize the actual KL divergence (not the estimate) then we would choose the true model. Second, minimizing AIC is reasonable because it is asymptotically equivalent to minimizing LOOCV error. As long as we think LOOCV error is a good method of assessing model performance, then AIC should be too.

Let's finish the assignment by comparing these two different perspectives of the AIC.

1. What are the similarities between the two perspectives of AIC (KL divergence and LOOCV)? Think about:

   - What KL divergence and LOOCV are trying to measure
   - Similar methods used in both perspectives

   Provide an intuitive explanation for why the LOOCV error in the Stone (1977) paper (the log likelihood calculated with each training observation held out in turn) would be similar to the KL divergence (which involves expected log-likelihoods...)