Stat 6021: Homework Set 11 & 12

Note that this is a homework that combines Modules 11 and 12, and is due Dec 5. You can work on questions 1a to 1g and 2 after Module 11.

1. For this question, we will revisit the `penguins` data set from the `palmerpenguins` package. The data set contains information regarding measurements of adult penguins near Palmer station, Antarctica. We will focus on using the four measurement variables (bill length, bill depth, flipper length, body mass) to model the gender of the penguins. Since there are three species involved, we also want to control for species in the logistic regression. We will not consider the island and year in this logistic regression.

   When you read the data in, notice that there are a number of penguins with missing values for gender. Remove these observations from the data frame. Also remove columns 2 and 8 since we are not considering island and year in this logistic regression. Then, randomly split your data into a training and test set (80-20 split respectively). For reproducibility, use `set.seed(1)` while performing the split. You can run the following block of code to carry out the needed steps.

```
library(palmerpenguins)

Data<-penguins
##remove penguins with gender missing
Data<-Data[complete.cases(Data[ , 7]),-c(2,8)]
##80-20 split
set.seed(1)
sample<-sample.int(nrow(Data), floor(.80*nrow(Data)), replace = F)
train<-Data[sample, ]
test<-Data[-sample, ]
```

   The questions below should be answered using the training set.

   (a) Create some data visualizations to explore the relationship between the various body measurements and the gender of penguins. Be sure to briefly comment on your data visualizations.

(b) Use R to fit the logistic regression model (first order only involving the 4 measurement variables and species). Based on the results of the Wald tests for the individual coefficients, which predictor(s) appears to be insignificant in the model?

(c) Based on your answer in part 1b, drop the predictor(s) and refit the logistic regression. Write out the estimated logistic regression equation. If you did not drop any predictor, write out the logistic regression equation from part 1b.

(d) Based on your estimated logistic regression equation in part 1c, how would you generalize the relationship between some of the body measurement predictors and the (log) odds of a penguin being male?

(e) Interpret the estimated coefficient for bill length contextually.

(f) Consider a Gentoo penguin with bill length of 49mm, bill depth of 15mm, flipper length of 220mm, and body mass of 5700g. What are the log odds, odds, and probability that this penguin is male?

(g) Conduct a relevant hypothesis test to assess if the logistic regression in part 1c is a useful model. Be sure to write out the null and alternative hypotheses, report the value of the test statistic, and write a relevant conclusion.

(h) Validate your model on the test data by creating an ROC curve. What does your ROC curve tell you?

(i) Find the AUC associated with your ROC curve. What does your AUC tell you?

(j) Create a confusion matrix using a threshold of 0.5. What is the false positive rate? What is the false negative rate? What is error rate?

(k) Discuss if the threshold should be changed. If it should be changed, explain why, and create another confusion matrix with a different threshold.

2. (You may only use R as a simple calculator or to find p-values or critical values) A health clinic sent fliers to its clients to encourage everyone to get a flu shot. In a follow-up study, 159 elderly clients were randomly selected and asked if they received a flu shot. A client who received a flu shot was coded $y = 1$, and a client who did not receive a flu shot was coded $y = 0$. Data were also collected on their age, x_1, health awareness rating on a 0-100 scale (higher values indicate greater awareness), x_2, and gender, x_3, where males were coded $x_3 = 1$ and females were coded $x_3 = 0$. A first order logistic regression model was fitted and the output is displayed below.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.17716    2.98242  -0.395  0.69307
age          0.07279    0.03038   2.396  0.01658 *
aware       -0.09899    0.03348  -2.957  0.00311 **
gender       0.43397    0.52179   _____  _______

    Null deviance: 134.94  on 158  degrees of freedom
Residual deviance: 105.09  on 155  degrees of freedom
```

(a) Interpret the estimated coefficient for x_3, gender, in context.

(b) Conduct the Wald test for β_3. State the null and alternative hypotheses, calculate the test statistic, and make a conclusion in context.

(c) Calculate a 95% confidence interval for β_3, and interpret the interval in context.

(d) Comment on whether your conclusions from parts 2b and 2c are consistent.

(e) Suppose you want to drop the coefficients for age and gender, β_1 and β_3. A logistic regression model for just awareness was fitted, and the output is shown below.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.91133    1.62651    3.02  0.00253 **
aware       -0.11931    0.03013   -3.96  7.5e-05 ***

    Null deviance: 134.94  on 158  degrees of freedom
Residual deviance: 113.20  on 157  degrees of freedom
```

Carry out the appropriate hypothesis test to see if the coefficients for age and gender can be dropped.

(f) Based on your conclusion in question 2e, what are the estimated odds of a client receiving the flu shot if the client is 70 years old, has a health awareness rating of 65, and is male? What is the estimated probability of this client receiving the flu shot?

3. Please remember to complete the Module 9 to 12 Guided Question Set Participation Self- and Peer-Evaluation Questions via Test & Quizzes on Collab.