APPLIED STATS

HOMEWORK 5

In this homework, you will use the **ICU** data set, within the *Stat2Data* package. In class and in Homework 4, we investigated the relationship between survival status and various potential predictors. In this assignment, we will build multiple logistic regression models to predict “Survive”.

If you recall, in Homework 4 we found that:

* Age group and presence of infection were significant predictors of survival. Older individuals and those with infections were less likely to survive.
* Pulse was not a significant predictor of survival.

1. **Age group and presence of infection**

Since we already know these two variables are significant predictors of survival, it seems logical to start with them.

* 1. Fit a logistic regression model with age group and infection as the explanatory variables. Use the “AgeGroup” variable from the data set, and make sure you tell R that “AgeGroup” is a *categorical* variable. Call this “model1”.
  2. One of the p-values is 0.0134. Interpret this p-value – be specific about what, exactly, it’s telling you!
  3. The intercept of the model is 2.697. Interpret this quantity in context.
  4. For Age Group 2, e^(slope) = 0.376. Interpret this quantity in context.
  5. Calculate a 90% confidence interval for the coefficient on AgeGroup3, and interpret this interval in context.
  6. Is the linearity condition met in this case? Why or why not?

1. **Adding sex**
   1. EDA: Make a two-way table of survival status by Sex.
   2. Does there appear to be a relationship between these two variables? If so, what is the nature of that relationship?
   3. Fit a logistic regression model with age group, infection, and sex as the explanatory variables. Call this “model2”.
   4. Is sex a statistically useful addition to this model? Be specific about what leads you to conclude that it is or is not.
2. **Adding Emergency**
   1. EDA: Make a two-way table of survival status by Emergency.
   2. Does there appear to be a relationship between these two variables? If so, what is the nature of that relationship?
   3. Fit a logistic regression model with age group, infection, and emergency as the explanatory variables. Call this “model3”.
   4. Is emergency a statistically useful addition to this model? Be specific about what leads you to conclude that it is or is not.
   5. Interpret the coefficient on the Emergency variable in context.
3. **Deleting Infection**

With Emergency in the model, it appears that Infection is now superfluous.

* 1. Conduct a nested drop-in-deviance test to determine if infection is worth keeping in the model that contains age group and emergency. Make a conclusion in context.
  2. Fit a logistic regression model with age group and emergency as the explanatory variables. Call this “model4”.
  3. Write out the fitted model.
  4. Using this model, find the probability of survival for a middle-aged person who was admitted through the emergency room. Then find the probability of survival for a middle-aged person who had elective admission.

1. **Adding Systolic Blood Pressure (SysBP)** 
   1. EDA: Make a boxplot of survival status by SysBP.
   2. Does there appear to be a relationship between these two variables? If so, what is the nature of that relationship?
   3. Fit a logistic regression model with age group, emergency, and SysBP as the explanatory variables. Call this “model5”.
   4. Is SysBP a statistically useful addition to this model? Be specific about what leads you to conclude that it is or is not.
   5. Calculate a 90% confidence interval for the coefficient on SysBP, and interpret this interval in context.
   6. Is the linearity condition met in this case? Why or why not?
2. **Comparison**
   1. Calculate the misclassification rates for model1, model4, and model5.
   2. Compare model1, model4, and model5. Which do you think is the “better” model, and why? In your answer, you should discuss simplicity, deviance, misclassification, and conditions.