# STAT 3400 - Homework #12

Alex Ojemann

Due May 4, 2023

## Problem 26.5.1

a. H_0: The the log odds of whether a student uses marijuana is the same whether or not their parents used it.

H_A: The the log odds of whether a student uses marijuana is not the same whether or not their parents used it.

b. The sample slope is 0.791 and the p value is <0.0001.

c. We reject the null hypothesis because the p value is less than our level of significance (0.05). This means that the log odds of a student using marijuana is not the same whether or not their parents used it.

## Problem 26.5.2

a. H_0: The the log odds of whether a patient died is the same whether they were in the control or treatment group.

H_A: The the log odds of whether a patient died is not the same whether they were in the control or treatment group.

b. The sample slope is 0.395 and the p value is 0.0594.

c. We fail to reject the null hypothesis because the p value is greater than our level of significance (0.05). This means that we don't have enough evidence to say that the log odds of whether a patient died isn't the same whether they were in the control or treatment group.

## Problem 26.5.3

a. There are 26 observations in fold 2. In the model with tail length as a predictor, eight were correctly predicted to be from Victoria and two were incorrectly predicted to be from Victoria.

b. 78 observations are used to build the model that predicts for fold 2 because each fold has 26 observations and folds 1, 3, and 4 are used for this.

c. One coefficient was estimated for the model using tail length as a predictor and two coefficients were estimated for the model using sex and length as predictors.

## Problem 26.5.4

a. 76 of 104 observations were correctly classified, so the proportion is 0.73.

b. 58 of 104 observations were correctly classified, so the proportion is 0.56.

c. I would choose the model with tail length because it has a superior classification accuracy.

d. The model with tail length is preferable given the predictions because it has a superior classification accuracy.

## Problem 26.5.5

a. There are 298 observations in fold 2. In the model using weight and mature as predictors, 10 were correctly predicted to be premature and five were incorrectly predicted to be premature.

b. 596 observations are used to build the model that predicts for fold 2 because each fold has 298 observations and folds 1 and 3 are used for this.

c. Most of the births in the original data set are full term. Only 109 of the 894 total observations are premature.

d. Six coefficients were estimated for the model which uses mage, weight, mature, visits, gained, and habit as predictors and two coefficients were estimated for the model which uses weight and mature as predictors.

## Problem 26.5.6

a. 816 of the 894 observations were predicted correctly in the model with six predictors, so the proportion is 0.91.

b. 815 of the 894 observations were predicted correctly in the model with two predictors, so the proportion is 0.91.

c. I would choose the model with two predictors because the accuracies are extremely close and the model with two predictors is less likely to have redundant features or multicollinearity.