Problem Set 7 Solution

Shane T. Mueller shanem@mtu.edu

October 23, 2018

For each of the following questions, provide R commands used to solve the problem.

Linear Regression and Variable Selection

Suppose you are a biologist interested in developing metrics to determine the gender of fossilized kangaroo specimen based on skull measurements. Load the kanga data set from the faraway package. A number of the elements have missing data, and so select out the subset that have all observed measures:

```
library(faraway)
data(kanga)
kanga2 <- kanga[!is.na(rowSums(kanga[,3:20])),]
```

1. Predicting a categorical variable

Transform sex to a numeric variable, and predict the outcome Sex based on all of the other predictors, other than species. Make a plot examining how good the fit is. To predict Sex, a categorical value, you must transform it into a number. Doing as.numeric(kanga2$sex) will code Female=1 and Male=2. You are attempting to predict an integer- coded categorical value based on a sum of linear predictors, and your outcome is not an integer. So, use whether the outcome prediction value is greater than or less than 1.5 to decide whether the specimen is male or female (you can use round()). Create a table showing the accuracy and error rate for the sex and the model's guess at sex. Discuss the problems you think might occur when trying to predict a categorical value with a continuous linear model, and how accurate you end up being with this model.

2. Selecting variables

Use t-values or an Anova/F test to select a reduced set of predictors. You can use either t-values, or F-values if you compare models using anova or drop1(model,test="F"), as these produce the same results. Discuss howw few predictors can you use without impacting the goodness of fit appreciably? Then, use an AIC criterion to select the best model. Finally, use a BayesFactor regression. When appropriate, you can use the step function to automatically find the best model.

Identify the smallest, most predictive model using each method. Describe the resulting models, discuss whether they differ, and how good the final model is at predicting (again, show a table examining actual by predicted sex for each model).

3. Predicting missing data

First, Build a model predicting palate.width based on as many of the remaining variables you think are reasonable. Try to find the simplest model that predicts the variable well. Examine how well you predict palate.width in the data set, and examine and interpret the R^2 and multiple R^2 values to justify your

model. REMEMBER–YOU ARE NOT PREDICTING SEX IN THIS MODEL. Show a plot and discuss whether the prediction seems good.

Once you have a good model, consider that there were 24 missing palate.width values in the original data. Select the data with missing values along that dimension:

```
missing <- kanga[is.na(kanga$palate.width),]
```

For any model, you can use the predict function to produce the model's best estimates for a given observed data set, even if it was not part of the original model (as with the missing data).

```
newpred <- round(predict(lm.palate,missing))
```

Note that other variables are also sometimes missing in the set, so you should try to create a model that does not use the variables with missing values. Now, put these predicted values back into your kanga data set, like this:

```
kanga$palate.width[is.na(kanga$palate.width)] <- newpred
```

This is called imputing data. Finally, if your best model from the previous questions did not contain palate.width as a predictor, add it to that model. Otherwise use your best model that already contains palate.width as a predictor. With this model, predict sex of the missing cases. Make a table showing how well the model is at predicting the sex of the kangaroos.