**HOMEWORK 8**

*Model Selection, 1 way ANOVA*

Notes:

* Round all numbers to 3 decimal places unless otherwise specified.

1. Major League Baseball winning percentage. (described in problem 3.56 on page 152)

The data are in MLB2007Standings\_new.jmp

Answer these questions, not the book’s question.

The goal is to predict winpct (the percent of games won by each team). Use all subsets model selection with the AICc criterion to identify the best set of variables to predict winpct. Only consider additive linear models (no interactions, no quadratic terms). All 17 variables (League, then BattingAvg to WHIP) are potential explanatory variables. You can leave League as a categorical variable; you do not need to create an indicator. Please set the maximum number of terms to 17 and leave the number of best models to show at its default of 56.

a. How many explanatory variables are in the “best” model?

4

b. BattingAvg is one of the variables in the “best” model (T/F)

T

c. Doubles is one of the variables in the “best” model (T/F)

F

d. How many **additional** models are considered reasonable alternative models?

14

e. My advice is that you shouldn’t use model selection as you have just done. Choose the best reason why you shouldn’t do model selection for this data set.

Choices:

17 variables is too many to use in a model selection

You shouldn’t do model selection when the response is a proportion

This: You need at least 6 observations per potential variable; this data set has 30 observations and 17 variables.

You need at least 100 observations to do model selection

1. The data in banksalary.jmp is a classic data set from a court case that alleged discrimination against women employees of a small bank. The variables are:

Sal77: annual salary in 1977

Log[Sal77]: natural log transformed 1977 salary

Sex: male or female

Female: indicator variable for sex, 1 if female, 0 if male

Senior: seniority, # months hired at the bank

Age: of the employee in months

Education: number of years of education (15 is “has BA/BS”)

Experience: prior to being hired by the bank, in months

All the individuals in this data set are hired as cashiers.

We are interested in the difference in log salary between men and women. The problem is that salary depends on many other characteristics of the individual, and we don’t know which ones. We will use model selection to choose appropriate variables to adjust, then add “female” to that model.

a. Use AICc to select appropriate variables. Remember female is not included in the list of potential variables. Select each variable that is selected to be in the best AICc model

Age

Education

b. If you used BIC do you get the same model (i.e., same selected variables) as when you used AICc?

Yes

c. Add female to the best AICc selected model. Report the estimated regression coefficient for female.

-0.114

d. Fill in the blank: the median predicted salary for female cashiers is 11.4 percent less than that for male cashiers

1. Exercise 5.27 from the book (Mouse serotonin) with new questions. We will consider only the female mice. These data are in MouseBrainF.jmp. This data set has all the Male mice hidden and excluded so any analysis only considers the female mice. The data has 3 genotype groups. The response variable is the number of social contacts each mouse had during an experiment.

Fit a 1 way ANOVA model that tests the null hypothesis that there are no differences among the 3 genotypes.

a. Report the degrees of freedom for genotype.

Note: You should be able to explain how this number relates to the number of genotypes.

2

3 categories - 1

b. Report the F statistic for this analysis

3.65

c. And the p-value

0.0429

d. Here are different conclusions that you might make. Choose the one(s) that are appropriate.

At least one genotype has a different mean # of contacts