Stat 302, Assignment 3, Due Thursday March 10, 2016 at 4:30pm

There are 5 multi-part questions and a set of questions based on 7 pages of reading.
Please see the Stats Workshop for help, or see me in office hours (Tue 1-2, Thur 3:30-4:30), or e-mail me at jackd@sfu.ca

Explanations and relevant computer output should be included, but plots do not need to be.
Necessary R code is included for every question, after the reading questions.
There are also four practice problems, mostly surrounding review material.

Total	/ 67	Name ______________________________
Q1	/ 11	
Q2	/ 11	
Q3	/ 6	Student Number _____________________
Q4	/ 16	
Q5	/ 6	
Reading	/ 17	

1) Consider the dataset airquality, which is embedded into R. This real dataset has daily measurements of temperature (in Fahrenheit), ozone concentration, and average wind speed (in miles per hour) from 111 of 153 days over a single summer. (The remaining 42 days are removed due to missing data)

A 2pts) Construct a regression using Temp as a response to Wind, Ozone, and Solar.R. Give the regression equation.

B 1pt) What proportion of the variation in daily temperature is explained by wind and ozone concentration.

C 1pt) What is the Akaike information criterion (AIC) for this model?

D 2pts) Would the AIC improve if you dropped a variable from this model? Which variable(s) could be dropped? Are these improvements large enough that they couldn't be due to randomness. Comment on how well this does (or doesn't) line up with the p-values from the model summary.

E 2pts) Do the variance inflation factors indicate that there is a co-linearity problem in this model?

F 3pts) Construct a regression using Temp as a response to only Wind and Ozone. Give the regression equation and the proportion of variance in temperature that is explained by this new model. Compare your answers to those in 1A and 1B. Is the removal of Solar.R justified?

2) Consider again the dataset airquality. This time we're interested in the factors behind

A 2pts) Create a scatterplot with Ozone in the Y and Temp as X. Does there appear to be a relationship between Ozone and Temperature? Is it a linear relationship?

B 1 pt) Create a regression model with Ozone as a response and Temperature and Wind as explanatory variables. Report the two variance inflation factors (VIFs) from the model and the AIC.

C 3pts) Add the interaction between Temperature and Wind to the model from 2B. Report the three VIFs and the AIC. Does the AIC indicate that the model with the interaction term is better? Why are these VIFs so large?

D 3pts) Replace the interaction term from the model in 2C with a Wind-squared term. Report the three VIFs and the AIC. Does the AIC indicate that including the squared term is better than not including it? Why is the VIF for Temperature still small?

E 2pts) Construct a model that includes both the Wind-Temperature interaction term and the Wind-squared term. Report the FOUR VIFs, and the AIC. Does the AIC indicate that the model including both the squared term and the interaction term is best of all?

.

---

3) Consider the dataset airquality one last time.
Also consider the set of explanatory variables (Ozone, Wind, Solar.R), the squares of these variables, and the interactions between these variables (9 variables in all) .

A 3pts) Starting with this set of explanatory variables, find the model with the AIC using the stepwise method. Give the regression equation of the final model and the proportion of variance explained.

b 3pts) Repeat 3a, but use BIC instead of AIC as your optimality criterion. Mention any differences and explain them by using the practical difference between AIC and BIC. Reminder for explanation: There are n = 111 observations.

---

4) Consider the dataset farms.csv

a 2pts) Construct a crosstab of the variables 'fertilizer' and 'land'. Which category within each variable will be considered the baseline.

B 3pts) Construct a regression with Yield as the response and 'fertilizer' and 'land' as the explanatory variables. Interpret the values of the intercept and each of the four dummy variables.

c 4pts) Do a hypothesis test on each of the three dummy variables for fertilizer at alpha = 0.05. Construct a TukeyHSD of the regression in 4b. Do the relevant hypothesis tests in the TukeyHSD analysis agree with those from the dummy variables?

D 1pt) Why are the p-values for (nature touch vs. none) and for (skotz vs. none) larger in the TukeyHSD analysis than the equivalent dummy variables in the regression?

E 2pts) Construct a regression with Yield as the response and 'fertilizer' and 'land' and the fertilizer:land interaction as explanatory terms. Do a hypothesis test at alpha = 0.05 for each of the interaction dummy variables.

F 2pt) Run an ANOVA on this model with interactions. Do the ANOVA results agree with hypothesis tests in 4e?

G 2pts) Compare the AICs of the model with and without the interaction term. Which model is better according to the AIC? Do you agree? Briefly justify your answer.

5) Consider the dataset gapminder.csv and the model of birth rates in response to `agri_in_gdp` , `co2_emit` , `female_work` , `GINI` , `HDI` , and `health_spending`

A 3pts) Starting with this set of explanatory variables, find the model with the AIC using the stepwise method. Give the regression equation of the final model and the proportion of variance explained.

.

b 3pts) Repeat 5a, but use BIC instead of AIC as your optimality criterion. Mention any differences and explain them by using the practical difference between AIC and BIC. Reminder for explanation: There are n = 150 observations.

Reading questions)
These questions pertain to “Model selection in ecology and evolution” by Jerald B. Johnson and Kristian S. Omland.
Trends in Ecology and Evolution
Vol.19 No.2 February 2004

The answer to R1 appears first in the text, and so on. Every question can be answered in 25 words or fewer. For this reading question, ignore the boxes and tables and focus on the main article. The answers are quite short; they are given a lot of marks to reward you for the effort it takes the read the article and find the answers.

R1, 2 pts) From the abstract (the first paragraph in bold), how is model selection used.

R2, 2 pts). What does model selection offer in contrast to a single null hypothesis test?

R3, 3 pts) What are three primary advantages of model selection?

R4, 2 pts) Name two commonly used criteria for model selection?

R5, 3 pts) Name a method to address the problem of several models all being equally (or nearly equally) viable? (In other words, if more than one model has equal support from the data). What are two advantages of using this model?

R6, 3 pts) What is a more recent application of model selection in evolutionary biology? What about model selection makes it well suited to this application?

R7, 2 pts. The authors suggest a requirement of the model being selected. This is in order to ensure the parameter estimates are biologically plausible. Describe this requirement.

For interest only, 0 pts), . What framework does model selection offer to ecosystem science?

```
############ PREAMBLE CODE
## Load the car package so we can use VIF
install.packages("car") ### Only needed once.

library(car) ## Needed every time you open R

########### EXAMPLE CODE QUESTION 1
### Load the air quality data
Q1 = airquality
## Remove the rows of data with a missing value
Q1 = Q1[!is.na(Q1$Ozone) & !is.na(Q1$Wind) & !is.na(Q1$Solar.R),]


### Create a model of temp in response to Ozone, Wind, and Solar.R
mod = lm( Temp ~ Ozone + Wind + Solar.R, data=Q1)
summary(mod)


### Get the Akaike information criteria values for this model,
### and the models with 1 variable removed
drop1(mod)

### Get the Variance Inflation Factors of variables in this model
vif(mod)

### Create a model of temp in response to Ozone and Wind only
mod = lm(Temp ~ Ozone + Wind, data=Q1)
summary(mod)

########### EXAMPLE CODE QUESTION 2
Q2 = Q1
plot(Q2$Ozone, Q2$Temp)

### Make a linear model and get the AIC and VIFs
mod = lm(Ozone ~ Temp + Wind, data=Q2)
AIC(mod)
vif(mod)

### Make a linear model with an interaction term, get AIC and VIFs
mod = lm(Ozone ~ Temp + Wind + Temp:Wind, data=Q2)
AIC(mod)
vif(mod)


### Make a model with a squared term, get AIC and VIFs
```

```
mod = lm(Ozone ~ Temp + Wind + I(Temp^2), data=Q2)
AIC(mod)
vif(mod)


### Make a dummy variable that is 1 when Ozone is over 50.
Q2$Temp80 = 0
Q2$Temp80[Q2$Temp > 80] = 1

### Make a model with a dummy variable, get AIC and VIFs
mod = lm(Ozone ~ Temp + Wind + Temp80, data=Q2)
AIC(mod)
vif(mod)

########### EXAMPLE CODE QUESTION 3
Q3 = Q1

### Use the stepwise method to find the best model by AIC
mod_start = lm(Temp ~ Ozone*Wind*Solar.R + I(Ozone^2) + I(Wind^2)
                                       + I(Solar.R^2), data=Q3)
mod_end = stepAIC(mod_start)
summary(mod_end)

### Use the stepwise method to find the best model by BIC instead
n = nrow(Q3)
mod_end = stepAIC(mod_start, k=log(n))
summary(mod_end)


########### EXAMPLE CODE QUESTION 4
Q4 = read.csv("farms.csv")

### Get a crosstab
table(Q4$fertilizer, Q4$land)

### Get a regression
mod = lm(yield ~ land + fertilizer, data=Q4)
summary(mod)

## Get a Tukey HSD for comparison
mod2 = aov(yield ~ land + fertilizer, data=Q4)
TukeyHSD(mod2)

## Get a regression with interactions
```

```
mod3 = lm(yield ~ land + fertilizer + land:fertilizer, data=Q4)
summary(mod3)

### Get the ANOVA
anova(mod3)

### Compare AICs
AIC(mod)
AIC(mod3)

########### EXAMPLE CODE QUESTION 5
### Very similar to question 3.
Q5 = read.csv("gapminder.csv")

mod_start = lm(birth_rate ~ agri_in_gdp + co2_emit + female_work +
GINI + HDI + health_spending, data=Q5)

### AIC
mod_end = stepAIC(mod_start)
summary(mod_end)

### BIC
n = nrow(Q5)
mod_end = stepAIC(mod_start, k=log(n))
summary(mod_end)
```