Stat 302, Assignment 3, KEY

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
Q1	/ 11
Q2	/ 11
Q3	/ 6
Q4	/ 16
Q5	/ 6
Reading	/ 17

Name ________________________________

Student Number ______________________

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.

Temperature = 72.42 – 0.32*Wind + 0.17*Ozone + 0.007*Solar.R+error

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

There was a typo. The question should have also mentioned Solar.R.

If you interpreted the question to include Solar.R, ***0.4999, or 49.99%*** *of variance is explained.*
If you interpreted the question as it is, ***0.4957, or 49.57%*** *of variance is explained.*

Either answer is acceptable.

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

Using the AIC() function, 747.58,
Using the drop1() function 430.58,
Using the extractAIC() function 430.58,

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.

Dropping Solar.R would improve the AIC to 746.50 (or 429.50), an improvement of 1.08
Dropping Wind would improve the AIC to 747.55 (or 430.55), an improvement of 0.03

Any model within 2 points of AIC is statistically equal at alpha = 0.05, so the improvements to the model from dropping either Wind or Solar.R could feasibly be due to randomness.

Notice that the improvement is the same regardless of the function used to compute AIC.

E 2pts) Do the variance inflation factors indicate that there is a co-linearity problem in this model?
Variable – VIF
Ozone – 1.817
Wind – 1.622
Solar.R – 1.154

None of these VIFs are greater than 5, so there is no indication of substantial co-linearity.

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?

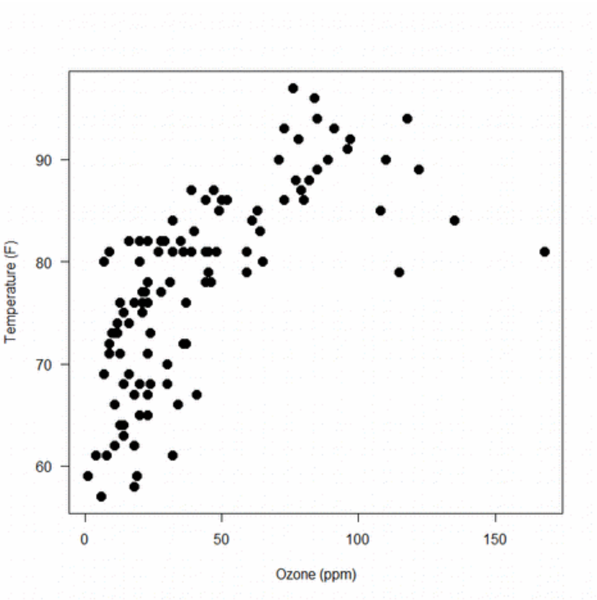
*Temperature = 73.14 – 0.30*Wind + 0.18*Ozone + error*

Variance explained: 0.4957, or 49.57%

The model and its R-squared are nearly the same as they were when Solar.R was included. This means Solar.R has a trivial impact on the model, and for the sake of parsimony (simpleness), we should remove it.

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?



There is a pattern between Ozone and Temperature, but it has a definite curve to it. There is a non-linear relationship between these variables.

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.

AIC: XXXXX

Variable – VIF
Temp – 1.328
Wind – 1.328

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?

AIC:XXXXXXX

Variable – VIF

Temp – 8.481
Wind – 61.572
Temp:Wind – 47.75

The AIC for this model is 15 points lower than the AIC in 2B, so the model with an interaction term is significantly better than simpler linear model.
*The VIFs are all large because both Temperature and Wind speed are included in two different variables. Temperature has to have some correlation with Temp*Wind by definition.*

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?

AIC: XXXXXXX

Variable – VIF
Temp – 1.37
Wind – 20.48
Wind^{2}*– 19.32*

The AIC for this model is 25 points lower than the AIC in 2B, so the model Wind-squared is significantly better than both the simpler linear model AND the model with an interaction term.

The VIF for temperature is small because only one variable uses temperature.

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?

AIC: XXXXXX

Variable – VIF
Temp – 17.73
Wind – 265.34
Wind^{2}*– 39.81*
Temp:Wind – 98.40

The AIC for this model is 1.5 points higher than that of the squared-term-only model in 2D. This means that including the interaction term doesn't improve the model if it already contains Wind-squared.

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.

*Temp = 71.86 + 0.33*Ozone – 0.73*Wind + 0.02*Solar.R – 0.0023*Ozone^2 – 0.00022*Solar.R^2 + 0.00069*Ozone*Solar.R + 0.0031*Wind*Solar.R*

This model explains 64.21% of the variance in temperature.

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.

*Temp = 63.66 + 0.50*Ozone – 0.0024*Ozone^2*

This model explains 61.11% of the variance in temperature.

Compared to the AIC-optimal result, this model is much simpler. It has two explanatory variables instead of seven. Log(111) = 4.71 > 2, therefore the BIC imposes a greater penalty per explanatory variable than the AIC does.

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.

“A-None” will be the baseline for fertilizer. “Flat” will be the baseline for land. These categories are the first alphabetically in their respective variables.

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.

Intercept: The average yield of a farm with no fertilizer on flat ground is 66.8

The average yield from using sloped ground instead of flat is 13.4 less
The average yield from using Greeno fertilizer instead of none is 21.6 more
The average yield from using Nature Touch fertilizer instead of none is 12.1 more
The average yield from using Skotz fertilizer instead of none is 5.9 more

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?

The p-values for the dummy variables for Greeno, Nature Touch, and Skotz are 3.65x10-8, 0.000382, and 0.0635 respectively.

We reject the nulls that average yield is different using Greeno and using Nature Touch.
We fail to reject the null that average yield is different using Skotz.

The relevant TukeyHSD comparisons are (nature touch vs none), (greeno vs none), and (skotz vs none) in the fertilizer section. They all agree with their dummy variable counterparts.

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?

Tukey's HSD adjusts for multiple comparisons, and increases the p-values accordingly. The regression does not.

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.

The p-values for the interaction dummy variables are 0.799, 0.506, and 0.506 respectively.

We fail to reject all three interaction null hypotheses.

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

The p-value for the interaction in the ANOVA is 0.5989. We fail to reject the null that the mean yield changes with any of the interactions. This agrees with our findings 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.

The AIC of the model without the interaction term is 274.5
The AIC of the model with the interaction term is 278.2

The simpler model is better. It has a lower AIC, and the interaction terms were found not to be significant.

---

5) Consider the dataset gapminder again.
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.

Birth Rate = 58.82 + 0.161(GINI) – 68.35(HDI) + error

81.22% 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.

Birth Rate = 58.82 + 0.161(GINI) – 68.35(HDI) + error

81.22% Variance explained

This is the same outcome model as when we used stepwise to find the best AIC. This similarity implies that both of the terms in the AIC-based model are contributing enough to the model to overcome the increased penalty that BIC imposes.

The penalty per term for BIC is log(n) = log(150) > 2
If you use the n after incomplete observations are removed, log(33) > 2 as well.
The output model is the same by either penalty.

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.

Model selection is used to select a single best model from a set of many candidates.
(Not needed) It computes a criterion, such as AIC, for each candidate and selects the model with the best value for this criterion.

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

With a null hypothesis test, you can only draw inferences from one hypothesis at a time. "Model selection offers a way to draw inferences from a set of multiple competing hypotheses."

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

Users are not restricted to evaluating a single model.
Models can be ranked and weighted.
Model averaging can be used if multiple models are supported equally by the data.

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

AIC (Akaike information criterion) and SC / BIC (Schwarz criterion / Bayesian information criterion).

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?

Model Averaging, which reduces model selection bias and accounts for selection uncertainty.

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?
Model selection is used to identify selective pressures that shape adaptations in the wild.
There are often several mechanisms that could explain evolutionary change, and since model selection compares several models, it can compare the viability of these mechanisms.

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.

Models should predict known patterns, and they should not generate implausible estimates.

For interest only, 0 pts), . What framework does model selection offer to ecosystem science?
It offers a framework for empirical support for a set of food-web models.

```
############ PREAMBLE CODE
## Load the car package so we can use VIF
install.packages("car") ### Only needed once.
Install.packages("MASS")
library(car) ## Needed every time you open R
library(MASS)

########### 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 model with a squared term, get AIC and VIFs
mod = lm(Ozone ~ Temp + Wind + Temp:Wind + I(Temp^2), 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 for question 5

Q5 = read.csv("gapminder.csv")
Q5 = subset(Q5, !is.na(GDPpercap) & !is.na(co2_emit) & !
is.na(agri_in_gdp) & !is.na(female_work) & !is.na(GINI) & !is.na(HDI)
& !is.na(health_spending))


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

mod_end = stepAIC(mod_start) ### Or use step(mod_full)
summary(mod_end)

## Now for BIC
n = nrow(Q5)

mod_end = stepAIC(mod_start, k = log(n))
summary(mod_end)
```