STAT 3340 Assignment 2 Solutions, Fall 2024

(out of 45 points)

1. The length of a species of fish is to be represented as a function of the age and water temperature. The fish are kept in tanks at 25, 27, 29 and 31 degrees Celsius. The following reads some data on the age, water temperature, and length of fish, and fits a number of regression models.

```
data=read.csv("http://chase.mathstat.dal.ca/~bsmith/stat3340/Data/fish.csv",header=T)
age=data[,1]
temp=data[,2]
length=data[,3]
lm0=lm(length~1)
lm1=lm(length~age+temp)
lm2=lm(length~age+temp+age:temp)
lm3=lm(length~age+temp+I(age^2)+I(temp^2))
lm4=lm(length~age+temp+age:temp+I(age^2)+I(temp^2))
```

- a) compare models lm0 and lm2.
- a i) Write down the linear regression models associated with lm0 and lm2. (4 points - 2 points for each model)

For lm0 - $length = \beta_0 + \epsilon$

For lm2 - $length = \beta_0 + \beta_1 age + \beta_2 temp + \beta_3 age \times temp + \epsilon$.

Here, and elsewhere, the model can be stated as above, or more generic variables names can be used provided they are properly defined, and the models can include an index for the observation number. For example:

$y =$ length, $X_1 =$ age, $X_2 =$ temp

lm2: $y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$

- a ii) What are the associated null and alternative hypotheses when comparing the two models. (2 points - 1 point for each of the hypotheses)

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

H_A: at least one of of $\beta_1, \beta_2, \beta_3$ is non-zero.

- a iii) Use the anova command to compare the outputs lm0 and lm2.

```
anova(lm0,lm2)
```

```
## Analysis of Variance Table
##
## Model 1: length ~ 1
## Model 2: length ~ age + temp + age:temp
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     38 69117501
## 2     35 10335491  3  58782010 66.353 1.614e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- a iv) What are the observed value of F and the p-value. (2 points - 1 point for F and 1 point for the p-value)

From the output, the $F \approx 75.6$ and p-value ≈ 0 (4×10^{-16}).

- b) Compare two quadratic models, one which includes an interaction term, and the other which doesn't.
  - b i) Write down the *full* and *reduced* regression models. (4 points - 2 for each model)

This is comparing lm3 and lm4

full model: $length = \beta_0 + \beta_1 age + \beta_2 temp + \beta_3 age \times temp + \beta_4 temp^2 + \beta_5 age^2 + \epsilon$

reduced model: $length = \beta_0 + \beta_1 age + \beta_2 temp + + \beta_4 temp^2 + \beta_5 age^2 + \epsilon$

- b ii) What are the associated null and alternative hypotheses. (2 points - 1 for each hypothesis)

$H_0 : \beta_3 = 0$ $H_A : \beta_3 \neq 0$

- b iii) Use the anova command to compare the outputs for the full and reduced models.

```
anova(lm3,lm4)
```

```
## Analysis of Variance Table
##
## Model 1: length ~ age + temp + I(age^2) + I(temp^2)
## Model 2: length ~ age + temp + age:temp + I(age^2) + I(temp^2)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     34 2838064
## 2     33 1979329  1    858734 14.317 0.0006188 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- b iv) What are the observed value of F and the p-value. (2 points - 1 for each of F and p-value)

From the anova output, $F \approx 13.8$ and p-value $\approx 6.9 \times 10^{-4}$.

2. An experiment was carried out to assess the yield of 4 different crop types on yield. The data are entered into R as follows:

```
yield=c(123,128,166,151,156,150,178,125,112,174,187,117,100,116,153,155,
168,109,195,158,135,175,140,167,130,132,145,183,176,120,159,142,120,187,
131,167,155,184,126,168,156,186,185,175,180,138,206,173,147,178,188,154,
146,176,165,191,193,190,188,169)

crop=as.factor(rep(c("W","C","S","R"),15))
```

- 2a) Write down a multiple regression model corresponding to a one way analysis of variance of yield as a function of crop type.

(Hint: Define indicator variables for the different crop types, as done above for crop type W. Your multiple regression model should include three of the indicator variables as predictors.)

Any distinct names are reasonable. Only 3 of the indicators should be used in the "lm" command. (3 points for any 3 correctly defined indicator variables)

```
IW=ifelse(crop=="W",1,0)
IC=ifelse(crop=="C", 1,0)
IS=ifelse(crop=="S", 1,0)
IR=ifelse(crop=="R", 1,0)
```

- 2b) Fit the regression model in R using the "lm" command, and show the summary output. (4 points for a model which uses 3 of the indicator variables. subtract 1 point for each error.)

```
lm.out=lm(yield~IW+IC+IS)  #any 3 of the indicators can be used
summary(lm.out)
```

```
##
## Call:
## lm(formula = yield ~ IW + IC + IS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.200 -19.267   3.933  19.983  46.533
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  159.667      6.567  24.312   <2e-16 ***
## IW           -13.200      9.288  -1.421    0.161
## IC            -3.467      9.288  -0.373    0.710
## IS             7.800      9.288   0.840    0.405
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.44 on 56 degrees of freedom
## Multiple R-squared:  0.08606,    Adjusted R-squared:  0.0371
## F-statistic: 1.758 on 3 and 56 DF,  p-value: 0.1657
```

- 2c) What is the observed value of the F statistic? (2 points - 1 for each of F and p-value)

From the observed $F \approx 1.758$ and the p-value $\approx .166$.

3. An experiment was carried out to assess the effect of diet on weight loss.

Five mice were put on each of three diets. At the beginning of the experiment, each animal's weight was measured, and recorded recorded as the variable x. After 3 months on diet, the animal's weight was measured again, and recorded as y.

Write down a single multiple regression model which allows for different slopes and different intercepts between y and x for each of the three diets. That is, the one multiple regression model should allow for 3 different linear regressions of y on x, one regression for each diet, and allowing for the 3 regression lines to have different slopes and intercepts.

Carefully define each variable to be used in the regression model.

(Hint: you'll need to define appropriate indicator variables to code for the different diets.) (8 points - 2 for correct definition of 2 of the indicators, and 6 points for a correct model. Subtract 1 point for each incorrect or missing term in the model.)

Define 2 or 3 indicator variables for diet. Any names are OK, and only 2 of the indicators should be used in the model statement.

$Z_1 = 1$ for diet A, 0 otherwise.

$Z_2 = 1$ for diet B, 0 otherwise.

$$y = \beta_0 + \beta_1 x + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 Z_1\, x + \beta_5 Z_2\, x + \epsilon$$

.

In terms of your model parameters, state the null and alternative hypotheses to be used when testing that the slopes of the 3 regression lines are the same, but allowing for the intercepts to be different. (2 points - 1 for each of the hypotheses.)

$H_0 : \beta_4 = \beta_5 = 0$

H_A : at least one of β_4, β_5 is non-zero.

4. An experiment was carried out to assess the effect of sex (Male and Female), and diet type (I, II or III) on weight loss. Five mice were randomly assigned to each each combination of gender and diet. The outcome variable y was the individual's change in weight after 3 months on the diet.

Write down a single linear regression model that can be used to fit a two way analysis of variance model for weight change, which allows for an interaction between sex and diet type.

Carefully define each variable to be used in the regression.

(Hint: you'll need to define appropriate indicator variables to code for sex and diet.)

Include 1 indicator variable for sex (here defined as X, and 2 indicator variables for diet (here defined as Z). (8 points - 3 points for a correct set of indicator variables and 5 points for a correct model. Subtract 1 point for each incorrect or missing term in the model.)

$X = 1$ for females, 0 for males

$Z_1 = 1$ for diet I, 0 otherwise.

$Z_2 = 1$ for diet II, 0 otherwise.

$$y = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 X Z_1 + \beta_5 X Z_2 + \epsilon$$

In terms of your model parameters, state the null and alternative hypotheses used when testing for the presence of an interaction. (2 points - 1 for each hypothesis.)

$H_0 : \beta_4 = \beta_5 = 0$

H_A : at least one of β_4, β_5 is non-zero.