# STAT2401 Analysis of Experiments – Final Exam

## 2021 Electronic Answer Sheet

## Student Details:

| | |
|---|---|
| Family Name: | PARK |
| Given Name: | SUNGHYUN |
| Student Number: | 22209962 |
| Your Script Version: | 2 |

**Answers outside the answer boxs will not be marked. Please also PRINT this answer sheet as a new PDF file to avoid any data loss.**

## Question 1

### Question 1 (a)(i) [3 Marks]

>cor(watershd)

According to the correlation matrix, X5 has the highest correlation to variation in Q which is 0.2530.

### Question 1 (a)(ii) [3 Marks]

>cor(log(watershd)

According to the new correlation matrix with log variables, it appears to be X4 has the highest correlation to variation in Q which is -0.5012.

## Question 1 (b) [5 Marks]

```
> watershd.full = lm(Q~.,data=watershd)
> drop1(watershd.full, test="F")
> watershd.lm.up = update(watershd.lm.up, .~.-Xn)
```

This command should give summary of each variables F-test and p-value. To proceed backward selection you should eliminate lowest F-value(highest p-value) factor. Then repeat these steps until model has only significant contributor to Q.

For example, this watershd data should be performed order of elimination X2,X8,X7,X9,X6,X1,X3,X4,X5

Therefore there are no significant contributor for Q.

## Question 1 (c) [4 Marks]

```
> watershd.log = log(watershd)
> watershd.lm.log = lm(Q~.,data = watershd.log)
>summary(watershd.lm.log)
fitted model:
log(Q)(hat) = -14.3371 - 0.5380log(X1) + 0.0978log(X2) - 0.5037log(X3) - 0.8399log(X4) + 6.5237log(X5)
+ 0.1193log(X6) - 0.2492log(X7) + 0.2439log(X8) - 0.1448log(X9)

F-statistic = 7.037
p-value = 1.894e-06
```

## Question 1 (d) [3 Marks]

According to summary in part(c), it has p-value of 1.894e-06 which much less than 0.05, so it is possible to think there is at least some of statistically contributing variables to variation in log(Q).

## Question 1 (e) [5 Marks]

> watershd.lm.log.null = lm(Q~1,data=watershd.log)
>add1(watershd.lm.log.null,watershd.lm.log,test = "F")
>watershd.lm.log.up = update(watershd.lm.log.up,.~.+Xn)
>add1(watershd.lm.log.up,watershd.lm.log,test = "F")
repeat until all independent variables are not statistically significant.

Therefore, in 5% significance level from forward variable selection it gives equation of
fitted model = log(Q)(hat) = -13.2236 - 0.8287log(X4) - 0.5646log(X1) + 6.0401log(X5)

## Question 1 (f) [5 Marks]

```
> watershd.lm.log = lm(Q~.,data = watershd.log)
> drop1(watershd.lm.log, test="F")
> watershd.lm.log.up = update(watershd.lm.log.up, .~.-Xn)
repeat until all independent variables are statistically significant.
```

Therefore, in 5% significance level from backward variable selection it gives equation of fitted model = log(Q)(hat) = -13.2236 - 0.5646log(X1) - 0.8287log(X4) + 6.0401log(X5)

## Question 1 (g) [4 Marks]

Both final models from part(e) and (f) have X1,X4,X5 as common explanatory variables that means X1,X4,X5 statistically contributes to Q.

# Question 2:

## Question 2 (a) [3 Marks]

```
> M1 = Crab.lm1 = lm(log(Force)~log(Height),data=Crab)
> M2 = Crab.lm2 = lm(log(Force)~log(Height)+Species,data=Crab)
> M3 = Crab.lm3 = lm(log(Force)~log(Height)+Species+log(Height):Species,data=Crab)
```

## Question 2 (b) [6 Marks]

First, comparing Model 2 and Model 3
anova(Crab.lm2,Crab.lm3)
The p-value = 0.3199 > 0.05 = alpha, therefore we prefer Model 2, the parallel lines model

Then compare Model 1 and Model 2
> anova(Crab.lm1,Crab.lm2)
The p-value = 0.2883 > 0.05 = alpha, from these two we prefer Model 1 the simple linear regression model.

## Question 2 (c) [4 Marks]

Since I chose simple linear regression model all three species fitted model will be same which is log(Force)(hat) = 0.4981 + 0.7691log(Height)

**Question 2 (d)(i) [3 Marks]**

**Question 2 (d)(ii) [3 Marks]**

**Question 2 (d)(iii) [4 Marks]**

**Question 2 (d)(iv) [4 Marks]**

## Question 3:

### Question 3 (a) [6 Marks]

```
> scores.lm = lm(Final~.,data=scores)
> summary(scores.lm)
Model 1 = Final = 16.9821 + 0.4518Pre1 + 0.3450Pre2
> scores.lm2 = lm(Final~Pre1,data=scores)
> summary(scores.lm2)
Model 2 = Final = 42.1989 + 0.4929Pre1
> scores.lm3 = lm(Final~Pre2,data=scores)
> summary(scores.lm3)
Model 3 = Final = 49.1605 + 0.4079Pre2
```

### Question 3 (b) [6 Marks]

```
> predict(scores.lm,newdata=data.frame(Pre1 = 78, Pre2 = 85),interval="predict",level=0.95)
First model has 95% prediction interval as 59.6141 ~ 103.4849 (range is 43.87079)
> predict(scores.lm2,newdata=data.frame(Pre1 = 78, Pre2 = 85),interval="predict",level=0.95)
Second model has 95% prediction interval as 58.8059~ 102.4888 (range is 43.68288)
> predict(scores.lm3,newdata=data.frame(Pre1 = 78, Pre2 = 85),interval="predict",level=0.95)
Third model has 95% prediction interval as 61.5713 ~ 106.0975 (range is 44.52621)

Therefore, shortest 95% prediction interval is Second model which is 43.68288
```

### Question 3 (c) [5 Marks]

```
Model 4 has fitted model of Pre2 = 73.0846 + 0.1192Pre1
To verify Model2 coefficient Pre1 = Model1 coefficient Pre1 + Model1 coefficientPre2*Model4 coefficient
Pre1,
Model2 coefficient Pre1 = 0.4929
Model1 coefficient Pre1 = 0.4518
Model1 coefficient Pre2 = 0.3450
Model4 coefficient Pre1 = 0.1192
0.4929 = 0.4518 + 0.3450*0.1192
Therefore, it is numerically verified that Model2 coefficient Pre1 = Model1 coefficient Pre1 + Model1
coefficientPre2*Model4 coefficient Pre1.
```

### Question 3 (d) [4 Marks]

```
Model5 coefficient Pre2 will be (Model3 coefficient Pre2 - Model1 coefficient Pre2) / Model1 coefficient
Pre1
Model3 coefficient Pre2 = 0.4079
Model1 coefficient Pre2 = 0.3450
Model1 coefficient Pre1 = 0.4518
Then, (0.4079-0.3450)/0.4518
which will be 0.1392
Therefore, Model5 coefficient Pre2 is 0.1392
```

# Question 4:

## Question 4 (a) [4 Marks]

```
election= read.table(file="Election-Version-2.txt",header=T)
> election.lm = lm(V~I+D+W+G*I+P+N, data=election)
> summary(election.lm)
```

According to the summary, p-value for the F-test is 0.0768 which is larger than confidence level 0.05.
Therefore, the regression is not statistically significant.

## Question 4 (b) [6 Marks]

```
election.StanRes = rstandard(election.lm)
election.leverage = hatvalues(election.lm)
(1:21)[election.leverage >(2*8/21)]
```

it gives 4, 16 which means they are high leverage

```
(1:21)[abs(election.StanRes)>2]
```

it gives non.
Therefore, There are 0 observations identified as outliers.
There is 2 observation identified as high leverage point 4,16
There is 0 observation identified as influential point.

## Question 4 (c) [4 Marks]

```
> election.cd = cooks.distance(election.lm)
> (1:21)[election.cd>1]
```

There are 2 observation identified as influential point 4, 16