STAT 425: Midterm 2

Instructor: A. Chronopoulou

Wednesday, November 17, 2021

<u>Instructions:</u>

– This is a closed-notes, closed-books exam.

– You can use your calculator, but cell-phones are not allowed to be used as calculators.

– The duration of the exam is 1h 20min.

– Read the questions carefully!

– Please mark your answers clearly and make sure that you *show your work.*

– Be as precise as possible in the interpretation-type answers you give.

Please write your name on top of your answer sheet!
In the end of the exam, scan **your exam** and upload it on *Gradescope.*
You do not need to scan your Cheat Sheets.

GOOD LUCK!

Multiple Choice Questions *(30 points - 5 points each)*
On your answer sheet, please write down the question number and the correct answer. *No justification needed.* There is only one correct answer:

1. The following output is the first step in a stepwise selection process applied to the `swiss` data set from the `datasets` **R** library. The goal of the analysis is to estimate *Fertility* in 47 French-speaking provinces of Switzerland from a set of predictors.

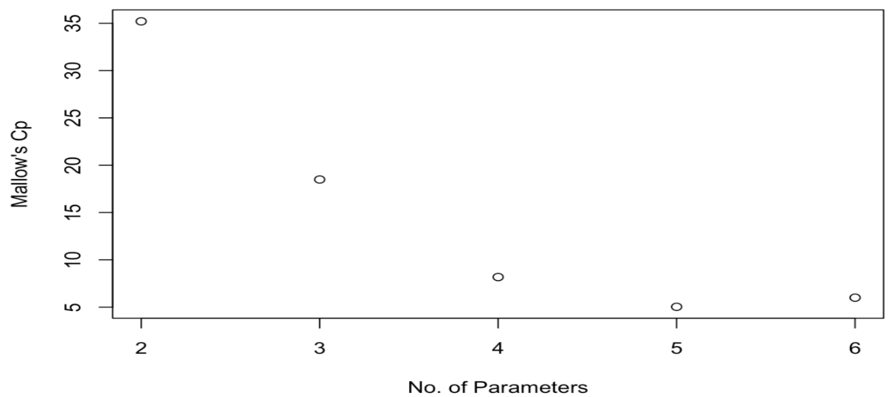
   According to the following output, the first variable to be removed from the model would be:

```
step(full.model, direction="both", k=log(n))
```

```
## Start:  AIC=201.79
## Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##
##                    Df Sum of Sq    RSS    AIC
## - Examination       1     53.03 2158.1 199.11
## <none>                          2105.0 201.79
## - Agriculture       1    307.72 2412.8 204.35
## - Infant.Mortality  1    408.75 2513.8 206.28
## - Catholic          1    447.71 2552.8 207.00
## - Education         1   1162.56 3267.6 218.61
##
```

   (a) Agriculture
   (b) Catholic
   (c) Education
   (d) Examination
   (e) Infant Mortality
   (f) We should not remove any of the variables.

2. In the same problem of predicting `fertility` using the `swiss` data, the following output corresponds to the Mallows C_p values were calculated while doing a backward regression in **R** .



According to the plot and the Mallow's C_p criterion, how many parameters does the optimal model have?

(a) 2

(b) 3

(c) 4

(d) 5

(e) 6

(f) None of the above.

3. In the same problem of predicting `fertility` using the `swiss` data, we applied the `pcrcomp` **R** function and obtained the following output:

```
X = swiss[,-1]
model.pcr = prcomp(X)
summary(model.pcr)
```

```
## Importance of components:
##                            PC1     PC2     PC3     PC4     PC5
## Standard deviation     43.360 21.4311 7.67001 3.72776 2.75052
## Proportion of Variance  0.777  0.1898 0.02431 0.00574 0.00313
## Cumulative Proportion   0.777  0.9668 0.99113 0.99687 1.00000
```

How many Principal Components should we keep in the model, if we want at least 97% of the variation of Y to be explained?

(a) 1
(b) 2
(c) 3
(d) 4
(e) 5
(f) None of the above.

4. What is the penalty term in the Ridge regression?
   (a) the square of the magnitude of the coefficients.
   (b) the square root of the magnitude of the coefficients.
   (c) the absolute sum of the coefficients.
   (d) the sum of the coefficients.

5. Recall the regularization parameter λ in the Ridge regression. What does it mean if $\lambda = 0$?
   (a) Large coefficients are not penalized.
   (b) Overfitting problems are not accounted for.
   (c) The loss function is as same as the ordinary least square loss function.
   (d) All of the above.

6. Recall the regularization parameter t in the Lasso regression. Which of the following options is true, if the regularization parameter is very high?
   (a) Cannot be used to select important features of a data set.
   (b) Shrinks the coefficients of less important features to exactly 0.
   (c) The loss function is as same as the ordinary least square loss function.
   (d) The loss function is as same as the Ridge Regression loss function.

Problem 1 *(30 points - 5 points each)*
A manufacturer of felt tip markers conducted an experiment to investigate whether a proposed new display, featuring a picture of a physician, is *more effective in drug stores* than the present counter display, featuring a picture of an athlete and designed to be located in the stationary area. 15 drugstores of similar characteristics were chosen for the study. They were assigned at random in equal numbers to one of the following treatments:

- *Treatment 0:* present counter display in stationary area,
- *Treatment 1:* new display in stationary area,
- *Treatment 2:* new display in checkout area.

In order to determine whether the new display is more effective, the manufacturer also decided to include in their model *historical sales* for a 3-week period in all 15 stores, denoted by X. Sales, Y, were recorded for the next 3-week period in all 15 stores.

The experimenter fitted two ANCOVA models, with and without interactions, and obtained the output below (Figures 1, 2, and 3):

```
marker.fit1 = lm(Sales ~ Treatment*X, data=display)
anova(marker.fit1)
```

```
## Analysis of Variance Table
##
## Response: Sales
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## Treatment    2 2984.40 1492.20 209.4878 2.836e-08 ***
## X            1  125.95  125.95  17.6817   0.00229 **
## Treatment:X  2   21.14   10.57   1.4842   0.27729
## Residuals    9   64.11    7.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1: Problem 1 ANOVA Table for Interactions Model

```
marker.fit2 = lm(Sales ~ Treatment+X, data=display)
anova(marker.fit2)
```

```
## Analysis of Variance Table
##
## Response: Sales
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Treatment  2 2984.40 1492.20 192.537 2.753e-09 ***
## X          1  125.95  125.95  16.251  0.001978 **
## Residuals 11   85.25    7.75
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2: Problem 1 ANOVA Table for Additive Model

```
marker.fit2 = lm(Sales ~ Treatment+X, data=display)
summary(marker.fit2)
```

```
##
## Call:
## lm(formula = Sales ~ Treatment + X, data = display)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.140 -1.586  0.292  1.804  3.820
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.2280     8.8623   1.041  0.32010
## Treatmenttrt1  10.9320     5.4678   1.999  0.07088 .
## Treatmenttrt2  13.0960     4.0056   3.269  0.00747 **
## X               0.7400     0.1836   4.031  0.00198 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.784 on 11 degrees of freedom
## Multiple R-squared:  0.9733, Adjusted R-squared:  0.966
## F-statistic: 133.8 on 3 and 11 DF,  p-value: 6.158e-09
```

Figure 3: Problem 1 Summary Table for Additive Model

(a) Using the appropriate output, test whether or not the interaction term is statistically significant. State your hypothesis, decision rule and conclusion. Use significance level $\alpha = 0.05$.

(b) Using the appropriate output,

(i) state the regression line that corresponds to Treatment 0

(ii) state the regression line that corresponds to Treatment 1

(iii) state the regression line that corresponds to Treatment 2

(iv) Do the lines have the same slope and/or intercept terms? Explain.

(c) Does the sales vary according to the Treatment? Justify your answer.

Problem 2 *(40 points = 10+5+5+5+5+10 points)*
A university computer service conducted an experiment in which *three* identical coin-operated computer terminals were placed at each of *four* different locations on campus last semester during midterms week and again during final week of classes. The data recorded the number of hours (Y) each terminal was *not* in use during the week at the four `locations` (factor A) and for the two different `weeks` (factor B).

*All relevant **R** output is attached in the end of the Problem in pages 8, 9.*

(a) State the factor effects (full) model that corresponds to this experiment. Make sure you explain the notation you use and state all necessary assumptions **and constraints**. Make sure that the constraints you state are *compatible with the ANOVA output provided by **R***.

(b) Test whether or not the interaction term is statistically significant. State the hypothesis test, decision rule and conclusion. Use significance level $\alpha = 0.05$.

The following questions (c), (d), (e) are based on the **additive** model and focus on the factor `location`.

(c) Test whether or not the factor location is statistically significant. State the hypothesis, decision rule and conclusion. Use significance level $\alpha = 0.05$. *Interpret your result in the context of the problem.*

(d) What are your conclusions based on Tukey's multiple comparisons for *all pairwise differences*? Make sure that your explanations are given in the context of the problem.

(e) We know that locations 1 and 2 are on the *North* side of campus, while locations 3 and 4 on the *South* side of campus. The experimenter wanted to estimate mean difference in hours each terminal was not used between South and North campus locations.

  (i) State the contrast that corresponds to this question.

  (ii) Estimate the contrast in (e (i)) using a 95% confidence interval. The multiplier here is equal to 2.093.

R Output for Problem 2

```
coin.fit1 = lm(hours ~ location*week, data=coins)
anova(coin.fit1)
```

```
## Analysis of Variance Table
##
## Response: hours
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## location       3 156.048  52.016  14.746 7.228e-05 ***
## week           1  47.884  47.884  13.574  0.002008 **
## location:week  3   7.948   2.649   0.751  0.537593
## Residuals     16  56.440   3.528
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4: Problem 2 ANOVA Table for Interactions Model

```
contrasts(coins$location)=contr.sum(levels(coins$location))
contrasts(coins$week)=contr.sum(levels(coins$week))
summary(coin.fit1)
```

```
##
## Call:
## lm(formula = hours ~ location * week, data = coins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3333 -1.0250  0.1667  1.3750  2.6667
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      15.3667     1.0844  14.171 1.79e-10 ***
## location2        -3.4333     1.5335  -2.239   0.0397 *
## location3        -3.5667     1.5335  -2.326   0.0335 *
## location4         3.7000     1.5335   2.413   0.0282 *
## week2             1.3667     1.5335   0.891   0.3860
## location2:week2   2.6667     2.1687   1.230   0.2366
## location3:week2   2.5000     2.1687   1.153   0.2659
## location4:week2   0.6667     2.1687   0.307   0.7625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.878 on 16 degrees of freedom
## Multiple R-squared:  0.7897, Adjusted R-squared:  0.6976
## F-statistic: 8.581 on 7 and 16 DF,  p-value: 0.0002031
```

Figure 5: Problem 2 Summary for Interactions Model

```r
coin.fit2 = lm(hours ~ location+week, data=coins)
anova(coin.fit2)
```

```
## Analysis of Variance Table
##
## Response: hours
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## location   3 156.048  52.016  15.349 2.592e-05 ***
## week       1  47.884  47.884  14.130  0.001329 **
## Residuals 19  64.388   3.389
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 6: Problem 2 ANOVA Table for Additive Model

```r
TukeyHSD(aov(hours ~ location+week, data=coins), "location")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = hours ~ location + week, data = coins)
##
## $location
##           diff       lwr       upr     p adj
## 2-1 -2.1000000 -5.088521 0.8885209 0.2316482
## 3-1 -2.3166667 -5.305188 0.6718542 0.1648147
## 4-1  4.0333333  1.044812 7.0218542 0.0061532
## 3-2 -0.2166667 -3.205188 2.7718542 0.9968952
## 4-2  6.1333333  3.144812 9.1218542 0.0000801
## 4-3  6.3500000  3.361479 9.3385209 0.0000520
```

Figure 7: Problem 2 Tukey HSD Output

		Week 1	Week 2
Location	1	$\bar{Y}_{11.} = 15.37$	$\bar{Y}_{12.} = 16.73$
	2	$\bar{Y}_{21.} = 11.93$	$\bar{Y}_{22.} = 15.97$
	3	$\bar{Y}_{31.} = 11.8$	$\bar{Y}_{32.} = 15.67$
	4	$\bar{Y}_{41.} = 19.07$	$\bar{Y}_{42.} = 21.1$

Figure 8: Problem 2 Treatment Averages