36-402/608 Homework #7 Solutions 3/4

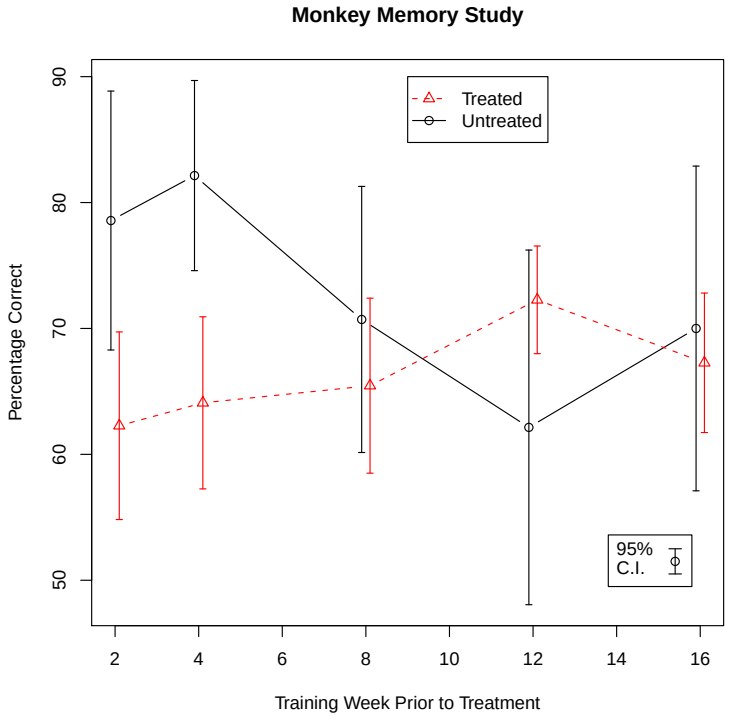
1. Monkey memory (40 points)

   This problem uses the monkey memory data of Sleuth Chapter 16, case 1 from case1601.csv. See page 463 for a description.

   (a) Read in the data, produce a plot similar to the one on page 463, but with full (up and down) confidence bars. Each CIs should be based on the 7 or 11 values that constitute the corresponding mean; in other words don't pool any variances. The "multiplier" for each CIs will be either `qt(0.975,6)` or `qt(0.975,10)`. Turn in your R code. One nice variant you could try is to add a small fixed amount to the x-values for one of the treatments so the error bars don't overlap. It is not necessary to add the horizontal lines at the ends of the error bars or the second (lower right) legend box.

```
mem = read.csv("case1601.csv")
names(mem) = casefold(names(mem))
# Group means and SEs as 2 by 5 matrices
means = as.matrix(aggregate(mem[,3:7], list(mem$treatment), mean)[,-1])
sds = as.matrix(aggregate(mem[,3:7], list(mem$treatment), sd)[,-1])
ns = as.matrix(aggregate(mem[,3:7], list(mem$treatment), length)[,-1])
SEs = sds/sqrt(ns-1)
multiplier = matrix(qt(0.975, ns-1), ncol=5, nrow=2)
upper = means + multiplier*SEs
lower = means - multiplier*SEs

# Start with an empty plot that will hold everything
lim = c(min(lower), max(upper))
plot(c(2,16), lim, type="n", xlab="Training Week Prior to Treatment",
     ylab="Percentage Correct",
     main="Monkey Memory Study")

# Means and CIs by treatment
x = c(2,4,8,12,16)
delta = 0.1
epsilon = 0.075
for (i in 1:2) {
  xa = x+c(-1,1)[i]*delta
  lines(xa, means[i,], col=i, lty=i, pch=i, type="b")
  segments(xa, lower[i,], xa, upper[i,], col=i)
  segments(rep(xa-epsilon,2), c(lower[i,],upper[i,]),
```

```
            rep(xa+epsilon,2), c(lower[i,],upper[i,]), col=i)
}

# Legend
legend(9, 90, c("Treated","Untreated"), lty=2:1, col=2:1, pch=2:1)

# Optional "CI" box:
text(c(14,14), c(52.5,51), c("95%","C.I."), adj=0)
segments(c(15.25,15.4,15.25), c(50.5,50.5,52.5), c(15.55,15.4,15.55), c(50.5,52
points(15.4,51.5)
rect(13.8, 49.5, 15.8, 53.6)
```



(b) Examine the diagonals of the two separate 5 by 5 covariance matrices for the two treatments. Which time has the larges ratio of the variances for the two treatments, and what is that ratio? Examine the two separate 5 by 5 correlation matrices. What pair of times has the largest difference between corresponding correlations, and what is the difference?

```
CTRL=mem$treatment=="CONTROL"
round(diag(cov(mem[CTRL,3:7])),3)
#   week2   week4   week8  week12  week16
# 105.952  57.143 111.905 198.810 166.667
round(diag(cov(mem[!CTRL,3:7])),3)
```

```
#   week2   week4   week8  week12  week16
# 111.818  94.091  97.273  36.818  61.818
## OR just sds^2 from above
apply(sds^2, 2, function(x) max(x)/min(x))
#    week2    week4    week8   week12   week16
# 1.055363 1.646591 1.150423 5.399765 2.696078

round(abs(cor(mem[!CTRL,3:7]) - cor(mem[CTRL,3:7])) ,3)
#        week2 week4 week8 week12 week16
# week2  0.000 0.809 0.227  0.352  0.203
# week4  0.809 0.000 0.260  0.352  0.844
# week8  0.227 0.260 0.000  0.231  0.525
# week12 0.352 0.352 0.231  0.000  0.325
# week16 0.203 0.844 0.525  0.325  0.000
```

Week 12 has a variance 5.4 times higher for controls than for treated subjects. The correlation values for week 4 and 16 differ by 0.84 for control vs. treated subjects. (Control: -0.68, treated: +0.16)

(c) Ignore the fact that we have fairly strong violations of the assumption of equal covariance across treatments, which will probably distort the results. Perform the MANOVA test of $\boldsymbol{\mu_C} = \boldsymbol{\mu_T}$ where each $\boldsymbol{\mu}$ represents all five individual times. Give the R code, the R result, and your interpretation. (Note that you need to run `aov()` with a matrix for the response, and then look at the `anova()` of that object. The `summary()` of the `aov()` object just gives the 5 one-way between-subjects ANOVAs.)

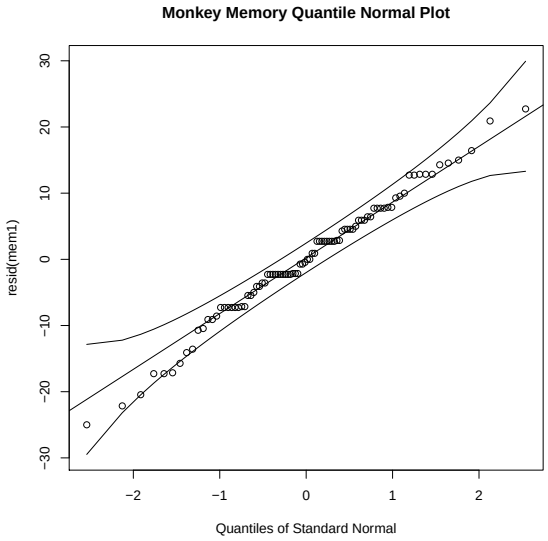
```
mem1 = aov(as.matrix(mem[,3:7])~treatment, data=mem)
anova(mem1, test="Hotelling")
# Analysis of Variance Table
#             Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
# (Intercept)  1          290.837   698.01      5     12 2.363e-14 ***
# treatment    1            2.331     5.59      5     12  0.006878 **
# Residuals   16

# Here is an alternative, which is better because it is more clear
# which columns you are looking at.  Also, the default test is
# "Pillai's trace" instead of "Hotelling's T^2".  Asymptotically
# they are the same.  For non-large samples, some suggest that Pillai
# is better because it is more robust to violation of model assumptions,
# but either is acceptable.
anova(aov(cbind(week2,week4,week8,week12,week16)~treatment, data=mem))
# Analysis of Variance Table
#             Df  Pillai approx F num Df den Df    Pr(>F)
```

```
# (Intercept)  1 0.99657   698.01      5     12 2.363e-14 ***
# treatment    1 0.69978     5.59      5     12  0.006878 **
# Residuals   16
```

I reject the null hypothesis (p=0.0069, F=5.59, df=5,12) that the 5 measurements have the same population means for control and treated monkeys.

(d) Turn in a quantile normal plot (using the qqn() function to get confidence bands) and a brief statement on the degree of violation of the normality assumption.

```
qqn(resid(mem1), main="Monkey Memory Quantile Normal Plot")
```



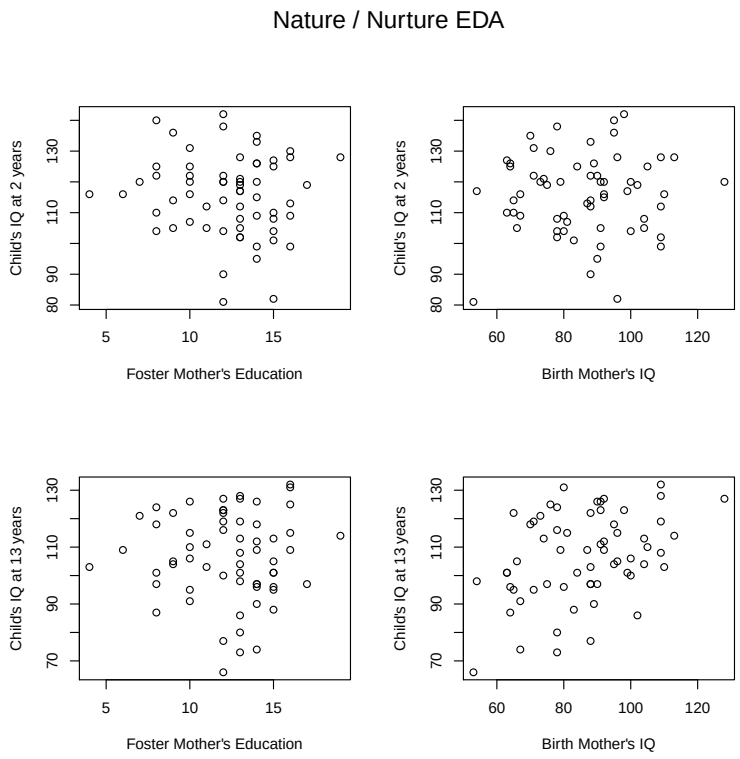
The normality assumption is reasonably well met.

2. Nature / Nurture IQ Data (30 points)

(a) Load the data from ex1605.csv, and make an EDA scatter plot that shows in four panels 1) the foster mothers education vs. age 2 IQ, 2) the birth mother's IQ vs. age 2 IQ, and the same two plots for age 13. Note: put the outcome on the y-axis. Use appropriate (non-default) labeling.

```
par(mfrow=c(2,2), oma=c(0,0,1.5,0))
with(nn, plot(fmed, age2iq, xlab="Foster Mother's Education",
              ylab="Child's IQ at 2 years"))
with(nn, plot(tmiq, age2iq, xlab="Birth Mother's IQ",
              ylab="Child's IQ at 2 years"))
with(nn, plot(fmed, age13iq, xlab="Foster Mother's Education",
              ylab="Child's IQ at 13 years"))
```

```
with(nn, plot(tmiq, age13iq, xlab="Birth Mother's IQ",
              ylab="Child's IQ at 13 years"))
mtext("Nature / Nurture EDA", outer=T, cex=1.3)
```



(b) Load the "car" package, and run the following code (which we need to get more useful results in this more complex problem):

```
library(car)
idata = data.frame(time = ordered(1:4, labels=paste("age",c(2,4,8,13),sep="")))
rslt = Manova(lm(cbind(age2iq,age4iq,age8iq,age13iq)~fmed+tmiq,nn),
       idata=idata, idesign=~time)
print(rslt)
```

For this part, just turn in the R results for the above commands.

```
# Type II Repeated Measures MANOVA Tests: Pillai test statistic
#           Df test stat approx F num Df den Df   Pr(>F)
# fmed       1   0.00191   0.1131      1     59 0.737878
# tmiq       1   0.11226   7.4612      1     59 0.008302 **
# time       1   0.41590  13.5287      3     57 8.91e-07 ***
# fmed:time  1   0.01411   0.2718      3     57 0.845454
# tmiq:time  1   0.10846   2.3114      3     57 0.085792 .

# Simpler code is less useful:
```

```
anova(aov(cbind(age2iq,age4iq,age8iq,age13iq)~fmed+tmiq,nn))
# Analysis of Variance Table
#             Df  Pillai approx F num Df den Df  Pr(>F)
# (Intercept)  1 0.99173  1678.40      4     56 < 2e-16 ***
# fmed         1 0.01991     0.28      4     56 0.88693
# tmiq         1 0.17759     3.02      4     56 0.02504 *
```

My interpretation: First we look at the interaction p-values, and conclude that we do not have sufficient evidence to reject the null hypothesis that the pattern of children's IQ across time stays the same (although possibly with changing overall level) as the foster mother's education and/or the birth mother's IQ changes.

The time p-value of < 0.0001 indicates that we have good evidence to reject the null hypothesis that the IQ of the children does not change over time. (Note that this is based on so-called "type 2" sum of squares so it is OK to interpret this as if the interaction were not in the model.)

(c) Give an interpretation of the fmed and tmiq p-values. Use your EDA plot and/or some calculated means to state the direction of the significant effect.

We retain the null hypotheses that the IQs do not change with the foster mothers education (F=0.113, df=1,59, p=0.74).

We reject the null hypothesis that the IQs do not change with the birth mothers IQ (F=7.47, df=1,59, p=0.0083). From the EDA plot, at least at age 13, there is a positive correlation between birth mother's IQ and child's IQ. The effect is much less clear at age 2, although the p-value for the interaction that indicates a different pattern at different children's ages is (barely) not statistically significant.

3. Fake beetles (10 points)

Load "fakebeet.csv". You might want to try:

```
with(fakebeet, plot(P, Q, pch=as.numeric(species),
                    main="Fake Beetle Study"))
legend("topleft",levels(fakebeet$species),pch=1:2)
```

as EDA. Turn in the separate t-tests for comparing the species (one vs. two) in terms of measurement P and in terms of measurement Q. Also turn in a T^2 test of the multivariate null hypothesis that

$$\begin{bmatrix} \mu_{P1} - \mu_{P2} \\ \mu_{Q1} - \mu_{Q2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Briefly comment on what it all means.

```
fakebeet=read.csv("fakebeet.csv")

with(fakebeet, t.test(P[species=="one"], P[species=="two"]))
# t = -0.8064, df = 31.684, p-value = 0.4260

with(fakebeet, t.test(Q[species=="one"], Q[species=="two"]))
# t = 1.1734, df = 33.969, p-value = 0.2488

anova(aov(cbind(P,Q)~species, data=fakebeet), test="Hotelling")
# Analysis of Variance Table
#             Df Hotelling-Lawley approx F num Df den Df  Pr(>F)
# (Intercept)  1           357.11   5892.3      2     33 < 2e-16 ***
# species      1             0.23      3.8      2     33 0.03386 *
# Residuals   34
```

Neither measurement alone is sufficient to make a claim that the beetle species differ according to what is measured. The power of multivariate analysis is shown in the MANOVA which detects the (bivariate) difference in means between the species.

4. Writing Assignment (20 points) As in homework 4, you will write about the Algal Regrowth analysis. This time you are a newly hired junior statistician, and you must write a report for the chief statistician so that she can present your results to the CEO. You need to get your boss to understand what you did and why, but your boss is a busy person responsible for many projects, so you can only write a two-sided one page report.