Homework 4

ST557

Nick Sun

November 8, 2019

Question 1

Here we are doing a test on a sample of 66 students where each student takes 2 reading tests before and after a reading instruction program. Each student produces 4 scores, 2 scores for before and after the first exam and 2 scores for before and after the second exam.

Let μ_1 be the population mean vector for the scores before the training and let μ_2 be the population mean vector for the scores after the training.

Part a.

A paired test is appropriate for testing $H_0 : \mu_1 = \mu_2$ since each experimental unit is being measured before and after some treatment, in this case the reading instruction program.

The data can be thought of as being composed of pairs of observations where each pair is correlated somehow.

Part b.

Performing a level $\alpha = .05$ test of this null hypothesis means performing a paired Hotelling's T^2 test.

The test statistic will have the form $T^2 = n\bar{D}^T\Sigma_D^{-1}\bar{D}$.

```r
n <- nrow(reading.data)

D <- cbind(reading.data$PRE1-reading.data$POST1, reading.data$PRE2-reading.data$POST2)
p <- ncol(D)
D.bar <- colMeans(D)
D.sd <- cov(D)

reading.stat <- n*t(D.bar)%*%solve(D.sd)%*%D.bar
reading.crit <- p*(n-1)/(n-p)*qf(0.95, df1 = p, df2 = (n-p))

reading.stat > reading.crit
```

```
##      [,1]
## [1,] TRUE
```

Our test statistic is much larger than our critical scaled F-statistic so we have evidence to say that the reading instruction program produces a difference in the scores of the two exams.

Part c.

Performing simultaneous Bonferroni tests on the same data can be done with the `t.test` function or directly using linear algebra in R.

Here we are performing 2 simultaneous tests, one for the paired scores of exam 1 and another for the paired scores of exam 2.

Either method will produce the following:

```
## [1] 0.8629862 2.5612562
```

This is the simulatenous Bonferroni confidence interval for paired scores in the first exam.

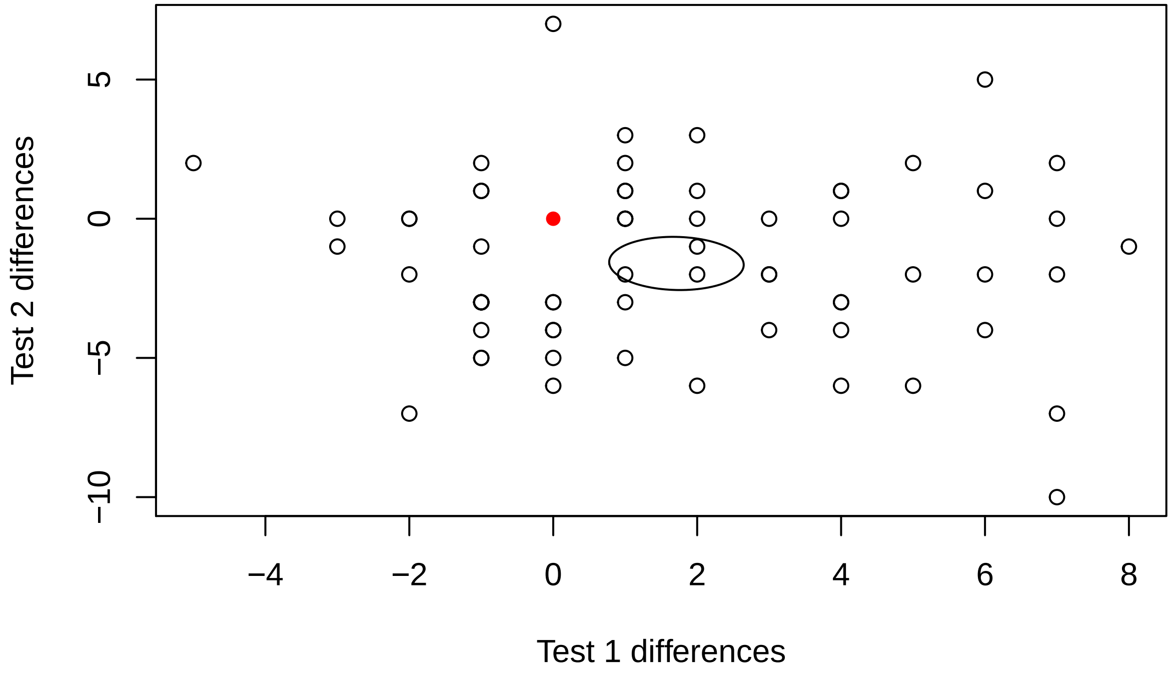
```
## [1] -2.4739413 -0.7381799
```

This is the simulatenous Bonferroni confidence interval for paired scores in the second exam.

Part d.

Constructing a Hotelling's T^2 confidence region is fairly straightforward with the code from the supplements.

```
plot(D, xlab = "Test 1 differences", ylab ="Test 2 differences")
points(0, 0, pch=16, col="red")
contour(muTest.test1, muTest.test2, reading.tstats, levels = (n-1)*p/(n-p)*
          qf(0.95,p,n-p), drawlabels = F, add = T)
```



The red dot on the contour plot represents the point (0,0). Notice that it is outside the confidence region, meaning that we reject Hotelling's T^2 with $\mu_0 = (0, 0)$.

Question 2

Here we are interested in monthly temperature data taken across multiple decades from 20 weather stations positioned around Corvallis.

Part a.

Our null hypothesis is $H_0 : \mu_1 = \mu_2 = \ldots = \mu_6$ where μ_i is the average temperature across all stations in decade i.

Our test here will be a repeated measures Hotelling's T^2. It is repeated measures since it is logical to assume that the recordings taken at the same weather station might be related and we should capture that covariance structure in our test.

Our test statistic will have the form $T^2 = n\bar{Y}^T S_Y^{-1}\bar{Y}$.

Our contrast matrix will be:

```
Temp.C <- cbind(rep(1,5), diag(x = -1, nrow = 5, ncol = 5))
Temp.C
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1   -1    0    0    0    0
## [2,]    1    0   -1    0    0    0
## [3,]    1    0    0   -1    0    0
## [4,]    1    0    0    0   -1    0
## [5,]    1    0    0    0    0   -1
```

```
Temp.tstat
```

```
##          [,1]
## [1,] 394.2733
```

```
Temp.crit
```

```
## [1] 18.37487
```

```
Temp.tstat > Temp.crit
```

```
##      [,1]
## [1,] TRUE
```

Our T^2 statistic is greater than our critcal scaled F-statistic so we have statistically significant evidence to say that the temperature means between the different decades are not equal.

Part b.

Let's construct some simultaneous Bonferroni confidence intervals for the differences in these μs.

We can use the `t.test` function to create these confidence intervals since we have a few intervals to make.

	Lower Bound	Upper Bound
mu2-mu1	0.0915605	0.2296395
mu3-mu1	-0.0698622	0.1172622
mu4-mu1	0.1690805	0.4350195
mu5-mu1	0.5837893	0.8764107
mu6-mu1	0.4298272	0.7907728

One of these intervals contains 0, the difference between $\mu_{1970s} - \mu_{1950s}$. This gives us evidence to say that these means are not signifcantly different from each other and this matches the outcome of the hypothesis test - most of the means between the decades are statistically different!

Question 3

For this question, we are interested in the performance of non-pitchers in the MLB during years just before free agency.

Part a.

No, a paired structure is not needed here. The experimental units are the players, however a players cannot be both a free agent and a non-free agent in this dataset so there is not covariance structure to capture here betweent the free agents and non-free agents.

Part b.

	BatAvg	OBP	Runs	Hits	Doubles	Triples	HRs	
BatAvg	0.0013847	0.0013284	0.6062737	1.222733	0.2119166	0.0322690	0.0833346	0.4641
OBP	0.0013284	0.0019948	0.7503599	1.101477	0.1869919	0.0255388	0.1257932	0.5029
Runs	0.6062737	0.7503599	755.9731231	1109.447368	205.1219280	30.8140501	183.9135338	601.7707
Hits	1.2227331	1.1014774	1109.4473684	2057.424812	377.8157895	47.7142857	261.8233083	1007.8646
Doubles	0.2119166	0.1869919	205.1219280	377.815789	97.8272360	7.1415105	58.8984962	211.4084
Triples	0.0322690	0.0255388	30.8140501	47.714286	7.1415105	5.9806980	1.3233083	11.6719
HRs	0.0833346	0.1257932	183.9135338	261.823308	58.8984962	1.3233083	106.3120301	262.9398
RBI	0.4641271	0.5029809	601.7707328	1007.864662	211.4084839	11.6719785	262.9398496	824.1815
Walks	0.2831319	0.7180618	510.8568062	622.917293	111.4513523	10.2764000	130.8646617	410.1511
StrikeOuts	0.0186002	0.2720562	504.1301201	646.537594	126.0598698	11.3481091	252.3796992	609.2866
SB	0.1240902	0.1621579	158.3909774	175.037594	15.3007519	9.7142857	-1.0601504	6.8947
Errors	0.0450634	0.0287639	28.0742902	67.233083	8.3514757	0.8874425	6.4436090	29.2827

	BatAvg	OBP	Runs	Hits	Doubles	Triples	HRs	
BatAvg	0.0016813	0.0015939	0.4020978	0.871891	0.1622746	0.0223295	0.0648282	0.3790
OBP	0.0015939	0.0022721	0.5567157	0.927463	0.1762044	0.0207519	0.1064637	0.4715
Runs	0.4020978	0.5567157	776.7007267	1361.935326	250.7841779	44.9676877	144.1820221	661.0559
Hits	0.8718910	0.9274630	1361.9353265	2741.478808	494.2825684	84.2943472	250.0661854	1279.4401
Doubles	0.1622746	0.1762044	250.7841779	494.282568	108.1292982	13.1296396	52.9903673	249.4052
Triples	0.0223295	0.0207519	44.9676877	84.294347	13.1296396	6.7844706	3.3391211	31.4605
HRs	0.0648282	0.1064637	144.1820221	250.066185	52.9903673	3.3391211	60.8097352	185.7181
RBI	0.3790779	0.4715535	661.0559186	1279.440155	249.4052822	31.4605911	185.7181388	772.8763
Walks	0.2384808	0.5593518	530.4561040	919.795347	169.0643077	23.4000878	113.5597961	485.8329
StrikeOuts	0.1290740	0.2703929	676.0398966	1269.080110	240.5384822	37.7119202	190.1970931	723.4059
SB	0.0701659	0.0953606	183.0638931	301.258060	46.8771887	19.5448227	11.9329854	104.2990
Errors	0.0273579	0.0266983	80.7989806	170.895454	29.8011023	4.4045506	10.1456372	67.5430

One thing that I notice is that the covariance of the number of Errors seems much different between the eligible and noneligible players. It appears that the noneligible players have a much higher covariance on average in this row.

Let's see how this affects the determinants.

```
det.eligible <- det(eligible.cov)
det.noneligible <- det(noneligible.cov)

det.eligible/det.noneligible
```

```
## [1] 2.303601
```

The determinant of the covariance matrix for eligible players is over twice as large as the determinant for the noneligible players. I think that this is evidence to say these covariance matrices are different.

Part c.

Using the equal covariance assumption for the Hotelling's T^2 test, we get the following test statistic and critical scaled F statistic:

```
Baseball.eq.tstat
```

```
##          [,1]
## [1,] 82.92731
```

```
Baseball.eq.crit
```

```
## [1] 22.11154
```

```
Baseball.eq.tstat > Baseball.eq.crit
```

```
##      [,1]
## [1,] TRUE
```

Our calculated Hotelling's T^2 is much larger than our critical F statistic so we have significant evidence to say that there is a difference between the performance of eligible and non eligible MLB players.

Part d.

Let's retry this hypothesis test except with an unequal covariance assumption and a χ^2 critical statistic instead of the scaled F statistic.

Notice that here instead of computing a pooled covariance estimate we simply calculate a weighted average of the two covariance estimates.

```
Baseball.neq.tstat
```

```
##          [,1]
## [1,] 85.04061
```

```
Baseball.neq.crit
```

```
## [1] 21.02607
```

```
Baseball.neq.tstat>Baseball.neq.crit
```

```
##      [,1]
## [1,] TRUE
```

Similar with the equal covariance Hotelling's T^2, we have a calculated Hotelling statistic that is greater than our critical statistic so we have statistically significant evidence to reject the null hypothesis. There is some difference between the eligible and noneligible MLB players.

Part e.

No, running both tests without considering the covariance structure beforehand is not good statistical practice. We are basically increasing our probability of Type I error and also likely trying to find a significant result where one might not exist.

This is a reprehensible statistical practice and pretty unprofessional. We should analyze the covariance beforehand and then decide which of the two tests to do.

Part f.

Let's take a look at the scaled F statistics vs the χ^2 statistics. After generating a bunch of different scaled F-statistics for different sample sizes, we get the following minimum and maximum critical F statistics (all with the same p, α):

```
v <- seq(min(eligible.n, noneligible.n), eligible.n + noneligible.n, 0.1)
crits <- rep(0, length(v))

for (i in 1:length(v)) {
  crits[i] <- v[i]*Baseball.p/(v[i]-Baseball.p+1)*qf(0.95, Baseball.p, v[i]-Baseball.p+1)
}

min(crits)
```

```
## [1] 22.1048
```

```
max(crits)
```

```
## [1] 23.94589
```

Compare these values with the value from the χ^2 distribution: 21.026. So it will be slightly easier to reject the test if we use the χ^2 test statistic, but the difference is not especially large.

Question 4

For this question we are analyzing skull data between different periods of ancient Egyptian history. Each skull has 4 distinct measurements taken and we are interested in seeing if the measurements change over time.

Part a.

Let's compare the covariance matrices between different time periods and see if they are similar. Doing this visually between 5 covariance matrices is fairly difficult, so we will just analyze the determinants of these matrices.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   35212   61828   74729   91573   96431  189667
```

We can see that there is a surprisingly large range between the determinants. The largest determinant is almost 6 times larger than the smallest determinant.

I would say this is evidence that the covariance matrices are not similar between these time periods.

Part b.

Here we want to perform a MANOVA test using Wilks Λ statistic. Following the supplemental code, we get the following statistic and critical values:

```
## [1] 59.25903
```

```
## [1] 26.29623
```

```
## [1] TRUE
```

Here our reference distribution is the χ^2 distribution and we see that our test statistic which comes from the Wilks $\Lambda = \frac{|W|}{|T|}$ is larger than our critical test statistic.

Therefore, we have statistically significant evidence to say that there is a difference in the population mean vectors between the different time periods.

Part c.

Let's perform a bunch of individual α^* ANOVA tests.

	stats	greater_than_crit
MB	5.954613	TRUE
BH	2.447420	FALSE
BL	8.305665	TRUE
NH	1.506997	FALSE

In the table above, the second column indicates whether or not the individual ANOVA statistic was larger than our critical F statistic with our adjusted α^*. The statistics for MB and BL are both larger than the critical F, indicating that on an individual variable level, the MB and BL measurements are different between the different time periods.

Now we are curious to see if we replace our MANOVA test with these individual simulataneous ANOVA tests where we reject the overall null hypothesis that the skull measurements are different if we reject **any** of the ANOVAs.

Doing some quick algebra, we see that

$$\begin{aligned}
P(\text{Reject any of the ANOVAs}) &= 1 - P(\text{Reject none of the ANOVAs}) \\
&= 1 - (1 - \cup_i P(\text{Reject the ith ANOVA})) \\
&\leq 1 - (1 - \sum_i P(\text{Reject the ith ANOVA})) \\
&= 1 - (1 - p\frac{\alpha}{p}) = \alpha
\end{aligned}$$

Note that we had to make use of the Bonferroni inequality here. But this shows that we are at least controlling the probability of type I error at α, although if there is any covariance between the ANOVA tests, we will probably not have a Type I error of exactly α.

Question 5

For this last question, we are interesting in modelling a multivariate response (NO_2 and O_3) air pollution with two covariates (wind and solar radiation).

Part a.

Let's fit a multivariate linear regression.

We'll fit a full model and a reduced model with the constraint $\beta_2 = 0$ enforced and perform an ANOVA test to see if the model's performance suffers significantly.

```r
mod.full <-lm(cbind(NO2, O3) ~ Wind + SolarRad, data = pollution.data)
mod_wind <- lm(cbind(NO2, O3) ~ Wind, data = pollution.data)
anova(mod.full, mod_wind)
```

```
## Analysis of Variance Table
##
## Model 1: cbind(NO2, O3) ~ Wind + SolarRad
## Model 2: cbind(NO2, O3) ~ Wind
##   Res.Df Df Gen.var.   Pillai approx F num Df den Df Pr(>F)
## 1     39      17.834
## 2     40  1   18.297 0.096851   2.0375      2     38 0.1444
```

The high p-value of .144 means that we do not have significant evidence to say that the full model, that is $\beta_2 \neq 0$ performs significantly better than the reduced model.

Part b.

Now we want to test whether $\beta_1 = 0$, aka the only covariate is Solar radiation. We can do this using the ANOVA test again.

```
mod_solar <- lm(cbind(NO2, O3) ~ SolarRad, data = pollution.data)
anova(mod.full, mod_solar)
```

```
## Analysis of Variance Table
##
## Model 1: cbind(NO2, O3) ~ Wind + SolarRad
## Model 2: cbind(NO2, O3) ~ SolarRad
##   Res.Df Df Gen.var.  Pillai approx F num Df den Df Pr(>F)
## 1     39      17.834
## 2     40  1   17.931 0.05962   1.2046      2     38  0.311
```

Here we have another large p-value of .311 which tells us that the full model does not perform signficantly better than the reduced model with only solar radiation. Interesting result, considering we also had the same result for the wind variable.

Part c.

Now we want to test the null hypothesis that $\beta_1 = \beta_2 = 0$. This is the same thing as fitting a model with only an intercept term.

```
mod_intercept <- lm(cbind(NO2, O3) ~ 1, data = pollution.data)
anova(mod.full, mod_intercept)
```

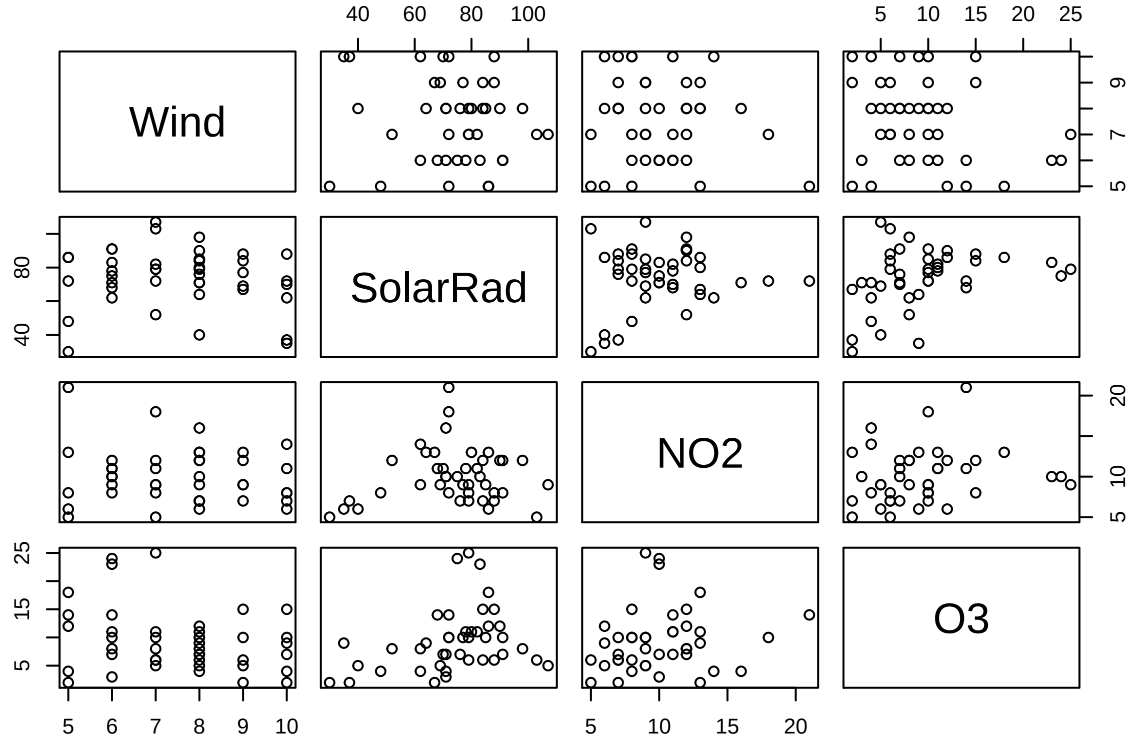
```
## Analysis of Variance Table
##
## Model 1: cbind(NO2, O3) ~ Wind + SolarRad
## Model 2: cbind(NO2, O3) ~ 1
##   Res.Df Df Gen.var.  Pillai approx F num Df den Df Pr(>F)
## 1     39      17.834
## 2     41  2   18.500 0.15921   1.6865      4     78 0.1615
```

Again we get a p-value that is not significant, indicating that the reduced model i.e. the intercept model sufficiently describes the data. This is unexpected, considering the other ANOVA tests maybe suggested that the individual covariates could describe the data adequately.

In fact, if we compare the intercept model to the models with just the individual covariates, we see that we get high p-values yet again. Therefore, perhaps neither solar radiation nor wind are actually useful in modelling air pollution.

The reason we got high p-values in the previous two tests may be that while neither covariate produces a decent model individually, when we add the covariates into the same model, the result is a model that performs even worse. This would explain the results from (a) and (b), given what we have seen from (c).

A matrix scatterplot further confirms our suspicions - there appears to be no linear relationship between wind, solar radiation, and either of the air pollution indicators.



```
anova(mod_solar, mod_intercept)
```

```
## Analysis of Variance Table
##
## Model 1: cbind(NO2, O3) ~ SolarRad
## Model 2: cbind(NO2, O3) ~ 1
##   Res.Df Df Gen.var.  Pillai approx F num Df den Df Pr(>F)
## 1     40      17.931
## 2     41  1   18.500 0.10587   2.3088      2     39 0.1128
```

```
anova(mod_wind, mod_intercept)
```

```
## Analysis of Variance Table
##
## Model 1: cbind(NO2, O3) ~ Wind
## Model 2: cbind(NO2, O3) ~ 1
##   Res.Df Df Gen.var.   Pillai approx F num Df den Df Pr(>F)
## 1     40      18.297
## 2     41  1   18.500 0.069005   1.4453      2     39  0.248
```