Module 4 R Activity

Zahlen Zbinden

1. Read in SeedsData.csv. Save this data as "seeds" and display the first six rows.

```
# A tibble: 6 x 8
   Area Perimeter Compactness Length Width Asymmetry GrooveLength Variety
  <dbl>     <dbl>       <dbl>  <dbl> <dbl>     <dbl>        <dbl>   <dbl>
1  15.3      14.8       0.871   5.76  3.31      2.22         5.22       1
2  14.9      14.6       0.881   5.55  3.33      1.02         4.96       1
3  14.3      14.1       0.905   5.29  3.34      2.70         4.82       1
4  13.8      13.9       0.896   5.32  3.38      2.26         4.80       1
5  16.1      15.0       0.903   5.66  3.56      1.36         5.18       1
6  14.4      14.2       0.895   5.39  3.31      2.46         4.96       1
```

2. Read in "PollutionData.csv". Save this data as "pollut" and display the first six rows.

```
# A tibble: 6 x 4
   Wind SolarRad   NO2    O3
  <dbl>    <dbl> <dbl> <dbl>
1     8       98    12     8
2     7      107     9     5
3     7      103     5     6
4    10       88     8    15
5     6       91     8    10
6     8       90    12    12
```

3. Using the seeds data, calculate the equal-variance version of the two-sample Hotellings T^2 test statistic to test that seed Variety 1 and seed Variety 2 have the same population mean vector. What is the value of the scaled version of the resulting T^2 test statistic?

```
[1] "The test statistic is 780.67 and the scaled test statistic is 106.68"
```

4. What is the p-value corresponding to the equal-variance version of the two sample Hotelling's T^2 test?

```
[1] "The p-value of the Hotelling's T2 test is 0"
```

5. Based on the results of the equal-variance two sample Hotelling's T^2 test, what is your hypothesis test decision at level .05?

My decision is to reject the null hypothesis that the population mean vectors for both samples is the same, as the p_value « .05. i.e. Reject $H_o : \mu_1 = \mu_2$

6. For the seeds data use the HotellingsT2() function to confirm your calculations for the equal-variance version of the two-sample Hotelling's T^2 test statistic.

```
	Hotelling's two sample T2-test

data:  seeds_1[1:7] and seeds_2[1:7]
T.2 = 106.68, df1 = 7, df2 = 132, p-value < 2.2e-16
alternative hypothesis: true location difference is not equal to c(0,0,0,0,0,0,0)
```

7. For the seeds data, use the T2.test() function to confirm your calculations for the equal-variance version of the two-sample Hotelling's T^2 test statistic.

```
	Two-sample Hotelling test

data:  seeds_1[1:7] and seeds_2[1:7]
T2 = 780.67, F = 106.68, df1 = 7, df2 = 132, p-value < 2.2e-16
alternative hypothesis: true difference in mean vectors is not equal to (0,0,0,0,0,0,0)
sample estimates:
                  Area Perimeter Compactness   Length    Width Asymmetry
mean x-vector 14.33443  14.29429   0.8800700 5.508057 3.244629  2.667403
mean y-vector 18.33429  16.13571   0.8835171 6.148029 3.677414  3.644800
              GrooveLength
mean x-vector     5.087214
mean y-vector     6.020600
```

8. Using the seeds data, calculate the unequal-variance version of the two-sample Hotelling's T^2 test statistic.

```
[1] "The unequal T2 test variance is 780.674774617226"
```

9. What is the p-value corresponding to the unequal-variance version of two-sample Hotelling's T^2 test?

```
[1] "The p-value of the unequal variance test statistic is 0"
```

10. Based on the results of the unequal-variance two-sample Hotelling's T^2 test, what is your hypothesis test decision at level 0.05?

I reject the null $H_o : \mu_1 = \mu_2$ as the p-value « 0.05.

11. Using the seeds data, perform MANOVA by hand to test the hypothesis that the population mean vectors for the three varieties are equal. $H_o : \mu_1 = \mu_2 = \mu_2$

```
[1] "Wilk's lambda test statistic is 0.0352871816793775"
```

12. What is the value of the scaled version of the Wilk's lambda statistic, that we would compare to the chi-squared distribution?

```
[1] "The scaled version of Wilk's lambda is 682.224043268659"
```

13. What is the p-value from the chi-squared approximation to the distribution of teh scaled Wilk's lambda test statistic?

```
[1] "The p-value from the chi-squared approximation is 0"
```

14. What is your conclusion based on this result: at a 0.05 confidence level, would you conclude that it is plausible that the three different wheat varieties have the same population mean vectors?

I would reject the null hypothesis of $H_o : \mu_1 = \mu_2 = \mu_3$ as the p-value « 0.05, as well as Wilk's lamba not being "small".

15. Uint the seeds data, perform a MANOVA using the Wilk.test() function to confirm the results you got by hand.

```
    One-way MANOVA (Bartlett Chi2)

data:  x
Wilks' Lambda = 0.035287, Chi2-Value = 682.22, DF = 14.00, p-value <
2.2e-16
sample estimates:
      Area Perimeter Compactness   Length    Width Asymmetry GrooveLength
1 14.33443  14.29429   0.8800700 5.508057 3.244629  2.667403     5.087214
2 18.33429  16.13571   0.8835171 6.148029 3.677414  3.644800     6.020600
3 11.87386  13.24786   0.8494086 5.229514 2.853771  4.788400     5.116400
```

16. Using the pollut data, fit a multivariate multiple regression model with the polutant levels "NO2" and "O3" as response variables, "Wind" and "SolarRad" as predictor variables. Are either of the predictors significant at level 0.1 in either of the univariate response models?

We can see in NO2 ~ pollut_x that only the intercept is significant, and in O3 ~ pollut_x that SolarRad is significant at the 0.1 level.

```
Response NO2 :

Call:
lm(formula = NO2 ~ pollut_x)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7521 -2.2053 -0.5917  1.6852 10.4623 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)   
(Intercept)      10.11454    3.62607   2.789  0.00813 **
pollut_xWind     -0.21129    0.33917  -0.623  0.53694   
pollut_xSolarRad  0.02055    0.03094   0.664  0.51042   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.416 on 39 degrees of freedom
Multiple R-squared:  0.02311,   Adjusted R-squared:  -0.02698 
F-statistic: 0.4614 on 2 and 39 DF,  p-value: 0.6338


Response O3 :

Call:
lm(formula = O3 ~ pollut_x)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.9527 -3.5053 -0.2998  1.4703 14.7123 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)  
(Intercept)       8.27619    5.58044   1.483   0.1461  
pollut_xWind     -0.78682    0.52198  -1.507   0.1398  
```

```
pollut_xSolarRad  0.09518    0.04761   1.999   0.0526 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.257 on 39 degrees of freedom
Multiple R-squared:  0.1513,    Adjusted R-squared:  0.1078
F-statistic: 3.476 on 2 and 39 DF,  p-value: 0.04082
```

17. Fit a reduced multivariate regression model that has just "Wind" as a predictor variable (leave out "SolarRad") and use the "anova()" function to compare this reduced model to the full model. What is your conclusion based on this result?

We fail to reject the null hypothesis as the p-val = 0.1444 in the ANOVA test is > 0.05. There is no difference between the reduced model and the full model, which says that SolRad is not significant to determining the response variable. This is the opposite conclusion that we drew from the two single univariate tests, which showed that SolarRad was the only predictor variable that had significance in determining the response variable.

```
Response NO2 :

Call:
lm(formula = NO2 ~ pollut_x)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.6330 -2.4464 -0.5476  2.0109 10.3670 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  11.8037     2.5669   4.598 4.22e-05 ***
pollut_x     -0.2341     0.3351  -0.699    0.489    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.392 on 40 degrees of freedom
Multiple R-squared:  0.01206,   Adjusted R-squared:  -0.01264 
F-statistic: 0.4884 on 1 and 40 DF,  p-value: 0.4887


Response O3 :

Call:
lm(formula = O3 ~ pollut_x)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-9.6365 -3.6011 -0.4584  2.0148 15.1489 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  16.0999     4.1246   3.903 0.000355 ***
pollut_x     -0.8927     0.5384  -1.658 0.105128    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.451 on 40 degrees of freedom
Multiple R-squared:  0.06431,   Adjusted R-squared:  0.04092 
F-statistic: 2.749 on 1 and 40 DF,  p-value: 0.1051


Analysis of Variance Table

Model 1: pollut_y ~ pollut_x
Model 2: pollut_y ~ pollut_x
  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
```