STA 207: Assignment I

Greg DePaul (917835494)

Instructions You may adapt the code in the course materials or any sources (e.g., the Internet, classmates, friends). In fact, you can craft solutions for almost all questions from the course materials with minor modifications. However, you need to write up your own solutions and acknowledge all sources that you have cited in the Acknowledgement section.

Failing to acknowledge any non-original efforts will be counted as plagiarism. This incidence will be reported to the Student Judicial Affairs.

A consulting firm is investigating the relationship between wages and occupation. The file `Wage.csv` contains three columns, which are

- `wage`, the wage of the subject,
- `ethnicity`, the ethnicity of the subject,
- and `occupation`, the occupation of the subject.

We will only use `wage` and `occupation` in this assignment.

```
Wage=read.csv('Wage.csv');
library(gplots)
library(car)
library(DescTools)
attach(Wage)
```

---

(1) Write down a one-way ANOVA model for this data. For consistency, choose the letters from $\{Y, \alpha, \mu, \epsilon\}$ and use the factor-effect form.

The one-way ANOVA model will be of the form

$$Y_{i,j} = \mu + \alpha_i + \epsilon_{i,j}$$

for $1 \leq i \leq r, 1 \leq j \leq n_i$ and $\epsilon_{i,j} \sim N(0, \sigma^2)$. Further, we make the assumption that

$$\sum_{i=1}^{r} n_i \alpha = 0$$

---

(2) Write down the least squares estimator of α_i for all i. Find the expectation and variance of this estimate in terms of $\{n_i\}$ and the parameters of the model.

We know the least squares estimator will be (as stated in the notes): Let $n_T = \sum_{i=1}^{r} n_i$. Then we want to identify our estimators by taking the derivative of our least squares potential function:

$$\underset{\mu,\alpha_i}{\operatorname{argmin}}\{L_1(\mu, \alpha)\}$$

We can identify the estimators by analytically optimizing:

$$0 = \frac{L_1(\mu,\alpha)}{\partial \alpha_i} = -\sum_{j=1}^{n_i} 2(Y_{i,j} - (\mu + \alpha_i)) = -2\sum_{j=1}^{n_i} Y_{i,j} + 2\sum_{j=1}^{n_i}(\mu + \alpha_i)$$

$$\iff 2\sum_{j=1}^{n_i} Y_{i,j} = 2\sum_{j=1}^{n_i}(\mu + \alpha_i)$$

$$\iff \sum_{j=1}^{n_i} Y_{i,j} = \sum_{j=1}^{n_i}(\mu + \alpha_i) = n_i(\mu + \alpha_i)$$

$$\implies \mu + \alpha_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{i,j}$$

$$\implies \alpha_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{i,j} - \mu$$

$$0 = \frac{\partial L_1(\mu,\alpha)}{\partial \mu} = -\sum_{i=1}^{r}\sum_{j=1}^{n_i} 2(Y_{i,j} - (\mu + \alpha_i)) = -2\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{i,j} + 2\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\mu + \alpha_i)$$

$$= -2\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{i,j} + 2\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\mu + \alpha_i) = -2\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{i,j} + 2\sum_{i=1}^{r} n_i(\mu + \alpha_i) = -2\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{i,j} + 2n_T\mu + 2\sum_{i=1}^{r} n_i\alpha_i$$

$$\iff 2n_T\mu + 2\sum_{i=1}^{r} n_i\alpha_i = 2\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{i,j}$$

$$\iff n_T\mu + \sum_{i=1}^{r} n_i\alpha_i = \sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{i,j}$$

Applying the constraint $\sum_{i=1}^{r} n_i\alpha_i = 0$, then we see that:

$$\implies n_T\mu = \sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{i,j}$$

$$\implies \mu = \frac{1}{n_T}\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{i,j} = \sum_{i=1}^{r}\frac{1}{n_T}\frac{n_i}{n_i}\sum_{j=1}^{n_i} Y_{i,j} = \sum_{i=1}^{r}\frac{n_i}{n_T}\overline{Y_i}$$

So our least-squares estimators will be:

$$\hat{\mu} = \sum_{i=1}^{r}\frac{n_i}{n_T}\overline{Y}_i$$

$$\implies \hat{\alpha}_i = \overline{Y}_i - \hat{\mu}$$

Since $\epsilon_{i,j} \sim N(0,\sigma^2)$, then it follows that

$$E\left[\hat{\mu}\right] = E\left[\sum_{i=1}^{r}\frac{n_i}{n_T}\overline{Y}_i\right] = \sum_{i=1}^{r}\frac{1}{n_T}E\left[n_i\overline{Y}_i\right] = \frac{1}{n_T}\sum_{i=1}^{r}\sum_{j=1}^{n_i} E\left[Y_{i,j}\right] = \frac{1}{n_T}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\mu+\alpha_i) = \mu + \underbrace{\frac{1}{n_T}\sum_{i=1}^{r} n_i\alpha_i}_{=0} = \mu$$

$$E[\alpha_i] = E[\overline{Y}_i - \hat{\mu}] = \mu_i - \mu = \alpha_i$$

Now, recall for every $\overline{Y_i} \sim N(\mu_i, \frac{\sigma^2}{n_i})$ and are independent for each i, then we see that

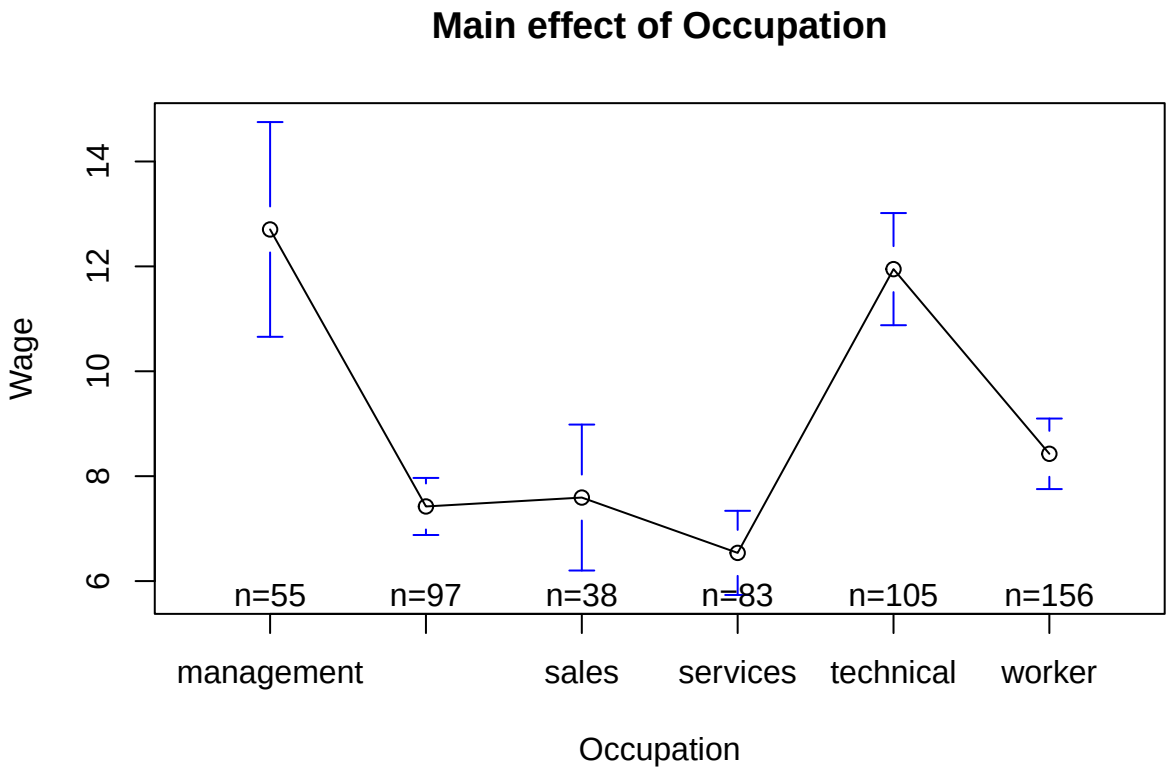
$$Var(\hat{\mu}) = Var\left(\frac{1}{n_T}\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{i,j}\right) = \frac{1}{n_T^2}\sum_{i=1}^{r}\left(\frac{\sigma^2}{n_i}\right) = \frac{\sigma^2}{n_T}$$

$$Var(\hat{\alpha}) = Var\left(\overline{Y_i} - \frac{1}{n_T}\sum_{i=1}^{r} n_i\overline{Y_i}\right) = \left(1 - \frac{n_i}{n_T}\right)^2 Var(\overline{Y_i}) + \sum_{k\neq i}\frac{n_k^2}{n_T^2}Var(\overline{Y_k})$$

$$= \left(1 - \frac{n_i}{n_T}\right)^2 \frac{\sigma^2}{n_i} + \sum_{k\neq i}\frac{n_k^2}{n_T^2}\frac{\sigma^2}{n_k} = \left(\frac{1}{n_i} - \frac{1}{n_T}\right)\sigma^2$$

(3) Obtain the main effects plots. Summarize your findings.

```
plotmeans(wage ~ occupation, data = Wage,
          xlab = "Occupation", ylab = "Wage",
          main="Main effect of Occupation")
```



(4) Set up the ANOVA table using R for your model. Briefly explain this table.

```
res.aov <- aov(wage ~ occupation, data = Wage)
summary(res.aov)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## occupation    5   2538   507.5   23.22 <2e-16 ***
## Residuals   528  11539    21.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(5) Test whether there is any association between `occupation` and `wage`. In particular, you need to (a) define the null and alternative hypotheses using the notation in Part 1, (b) conduct the test, and (c) explain your test result.

(a) Suppose that occupation has no effect on wage. Then it would follow that the effect terms would be zero. Therefore, we formulate the null and alternative hypotheses to be:

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_r = 0$$

$$H_A : \text{not all } a_i \text{ are zero}$$

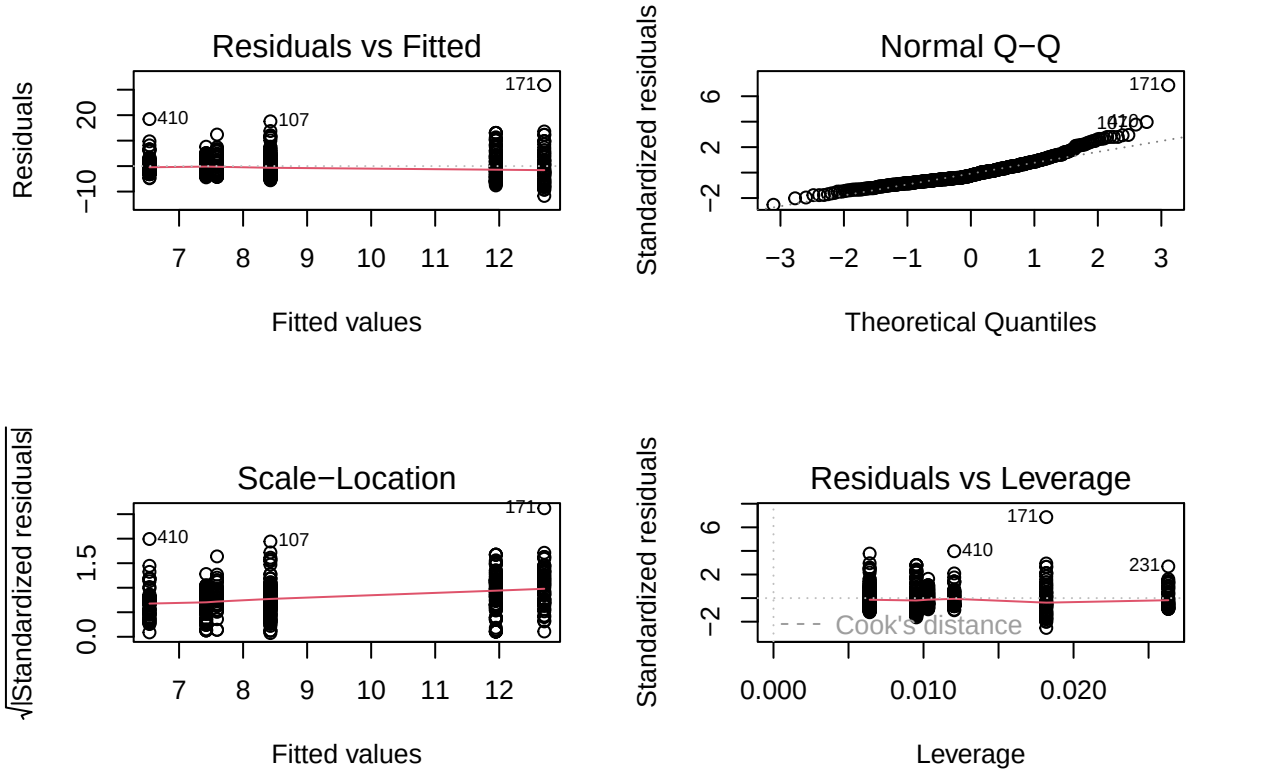
(b) As stated in problem 4, and our F-statistic is 23.22, which has a p-value of 2e-16. Therefore, we see that we satisfy for any reasonable critical value the ability to reject our null hypothesis.

(c) By rejecting the null hypothesis, we conclude that occupation does has an association on wage.

---

(6) For the model fitted in Part 4, carry out the necessary diagnostics to check if the model assumptions given in Part 1 are valid. What are your findings?

```
par(mfrow=c(2,2))
plot(res.aov)
```



It's very obvious that there are 3 outliers which should be removed if we are to get accurate results for our Bartlett and Levene Tests for homogeneity. Specifically:

```
outliers <- c(107, 410, 171, 231)
outliers_removed <- Wage[-outliers]
head(outliers_removed)
```

```
##    wage ethnicity occupation
## 1  5.10  hispanic     worker
## 2  4.95      cauc     worker
## 3  6.67      cauc     worker
## 4  4.00      cauc     worker
## 5  7.50      cauc     worker
## 6 13.07      cauc     worker
```

We perform the test for homogeneity of variances.

```
print(bartlett.test(wage ~ occupation, data = outliers_removed))
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  wage by occupation
## Bartlett's K-squared = 94.531, df = 5, p-value < 2.2e-16
```

```
print(leveneTest(wage ~ occupation, data = outliers_removed))
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   5  9.7025 7.043e-09 ***
##       528
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both p-values from Bartlett's and Levene's tests are less than the significance level of 0.05. This indicates that there is evidence of significantly unequal variances of error terms. This inspires doubt as to the treatment of data collection for these different groups of professions. It could be some professions may be more forthcoming with this wage information than others.

We also check for normality, we see that we have no scepticism about any departure from normality:

```
aov_residuals <- residuals(object = res.aov)
shapiro.test(x = aov_residuals )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.91749, p-value < 2.2e-16
```

(7) Assuming that the assumptions you made are true, can you statistically conclude if there is one occupation where the mean wage is the highest? Use the most appropriate method (use $\alpha = 0.05$) to support your statement.

$$H_0 : \alpha_1 = \alpha_2, \alpha_2 = \alpha_3, \alpha_1 = \alpha_3, \ldots, \alpha_5 = \alpha_6$$

To test simultaneous inference, we have access to several tests.

```
print(pairwise.t.test(Wage$wage, Wage$occupation, p.adjust.method="bonferroni"))
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Wage$wage and Wage$occupation
##
##           management office  sales   services technical
## office    8.4e-10    -       -       -        -
## sales     4.7e-06    1.000   -       -        -
## services  2.2e-12    1.000   1.000   -        -
## technical 1.000      2.7e-10 1.7e-05 2.8e-13  -
## worker    1.4e-07    1.000   1.000   0.046    6.7e-08
##
## P value adjustment method: bonferroni
```

```
TukeyHSD(res.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = wage ~ occupation, data = Wage)
##
## $occupation
##                            diff         lwr       upr     p adj
## office-management    -5.2814227 -7.53836588 -3.024479 0.0000000
## sales-management     -5.1113684 -7.93192217 -2.290815 0.0000046
## services-management  -6.1665301 -8.49132800 -3.841732 0.0000000
## technical-management -0.7565714 -2.98218646  1.469044 0.9265714
## worker-management    -4.2775256 -6.37435762 -2.180694 0.0000001
## sales-office          0.1700543 -2.38885725  2.728966 0.9999657
## services-office      -0.8851074 -2.88440478  1.114190 0.8032931
## technical-office      4.5248513  2.64180401  6.407898 0.0000000
## worker-office         1.0038970 -0.72503588  2.732830 0.5584749
## services-sales       -1.0551617 -3.67411581  1.563792 0.8589942
## technical-sales       4.3547970  1.82347368  6.886120 0.0000170
## worker-sales          0.8338428 -1.58502882  3.252714 0.9223496
## technical-services    5.4099587  3.44609529  7.373822 0.0000000
## worker-services       1.8890045  0.07238631  3.705623 0.0361288
## worker-technical     -3.5209542 -5.20878674 -1.833122 0.0000001
```

```
ScheffeTest(res.aov)
```

```
##
##   Posthoc multiple comparisons of means: Scheffe Test
##     95% family-wise confidence level
##
## $occupation
##                            diff     lwr.ci    upr.ci    pval
## office-management    -5.2814227 -7.9169347 -2.645911 3.5e-08 ***
## sales-management     -5.1113684 -8.4050284 -1.817708 7.9e-05 ***
## services-management  -6.1665301 -8.8812785 -3.451782 1.4e-10 ***
## technical-management -0.7565714 -3.3555005  1.842358 0.96671
## worker-management    -4.2775256 -6.7260702 -1.828981 3.6e-06 ***
## sales-office          0.1700543 -2.8180768  3.158185 0.99999
```

```
## services-office      -0.8851074 -3.2197573  1.449542 0.90059
## technical-office      4.5248513  2.3259507  6.723752 1.2e-08 ***
## worker-office         1.0038970 -1.0150388  3.022833 0.73711
## services-sales       -1.0551617 -4.1134066  2.003083 0.93181
## technical-sales       4.3547970  1.3988816  7.310712 0.00024 ***
## worker-sales          0.8338428 -1.9907588  3.658444 0.96463
## technical-services    5.4099587  3.1166862  7.703231 2.1e-11 ***
## worker-services       1.8890045 -0.2323245  4.010333 0.11742
## worker-technical     -3.5209542 -5.4918957 -1.550013 1.9e-06 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Acknowledging these three tests, we see that there is no conclusive rejection of the notion that technical versus managerial occupations have different mean values. Therefore, we cannot conclude that there is one occupation where the mean wage is the highest.

---

(8) Consider a one-way ANOVA model with fixed effects

$$Y_{i,j} = \mu + \alpha_i + \epsilon_{i,j},\ j = 1, \dots, n_i, i = 1, \dots, r, \tag{1}$$

where $\{\alpha_i\}$ satisfies that $\sum_i n_i \alpha_i = 0$ and $\{\epsilon_{i,j}\}$ are i.i.d. $N(0, \sigma^2)$. For the above model, write down the loss function associated with least squares, denoted as $L_1(\mu, \alpha)$, and write down the log-likelihood, denoted as $L_2(\mu, \alpha)$.

$$L_1(\mu, \alpha) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{i,j} - \mu - \alpha_i)_2^2$$

To derive the likelihood function, we first need to consider the distribution:

$$Y_{i,j} = \mu + \alpha_i + \epsilon_{i,j} \sim N(\mu + \alpha_i, \sigma^2)$$

Then, for every instance, we want to maximize the likelihood:

$$\text{Likelihood} = \prod_{i=1}^{r} \prod_{j=1}^{n_i} f(Y_{i,j}; \mu, \alpha) = \prod_{i=1}^{r} \prod_{j=1}^{n_i} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Y_{i,j} - (\mu + \alpha_i)}{\sigma}\right)^2}$$

$$\implies \log \text{Likelihood} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} -\frac{1}{2} \left(\frac{Y_{i,j} - (\mu + \alpha_i)}{\sigma}\right)^2 - \log(\sigma\sqrt{2\pi})$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(Y_{i,j} - (\mu + \alpha_i)\right)^2 - \sum_{i=1}^{r} \sum_{j=1}^{n_i} \log(\sigma\sqrt{2\pi})$$

---

(9) Find the maximum likelihood estimator of μ and α using the log-likelihood $L_2(\mu, \alpha)$ in Question 8.

Observe, we are identifying the estimators μ, α_i subject to the maximization problem:

$$\underset{\mu,\alpha}{\text{argmax}} \{L_2(\mu, \alpha)\}$$

$$\iff \underset{\mu,\alpha}{\text{argmax}} \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(Y_{i,j} - (\mu + \alpha_i)\right)^2 - \sum_{i=1}^{r} \sum_{j=1}^{n_i} \log(\sigma\sqrt{2\pi}) \right\}$$

Clearly linear transformations can be applied and yield the same estimators. Firstly, the translate can be removed:

$$\iff \underset{\mu,\alpha}{\text{argmax}} \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(Y_{i,j} - (\mu + \alpha_i)\right)^2 \right\}$$

And the scale parameter can be removed as well since α_i and μ do not rely on σ:

$$\iff \underset{\mu,\alpha}{\text{argmax}} \left\{ -\sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(Y_{i,j} - (\mu + \alpha_i)\right)^2 \right\}$$

Lastly, taking the dual with respect to the negative changes the problem to a minimization problem:

$$\iff \underset{\mu,\alpha}{\text{argmin}} \left\{ \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(Y_{i,j} - (\mu + \alpha_i)\right)^2 \right\}$$

Clearly, we see that this is exactly:

$$\iff \underset{\mu,\alpha_i}{\text{argmin}} \{L_1(\mu, \alpha)\}$$

Since these are equivalent, we realize that the maximum likelihood estimators are equivalent to those derived in question 2! So:

$$\hat{\mu}_{MLE} = \sum_{i=1}^{r} \frac{n_i}{n_T} \overline{Y}_i$$

$$\hat{\alpha}_{i,MLE} = \overline{Y}_i - \hat{\mu}$$

Acknowledgement

Extensive usage of the course notes (Chapter4ANOVA) as well as the discussion notes. Otherwise, all work presented is my own.

Session information

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Ventura 13.2
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
```

```
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] DescTools_0.99.47 car_3.1-1         carData_3.0-5     gplots_3.1.3
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.10        highr_0.10         compiler_4.2.2     cellranger_1.1.0
##  [5] bitops_1.0-7       class_7.3-20       tools_4.2.2        boot_1.3-28
##  [9] digest_0.6.31      evaluate_0.20      rootSolve_1.8.2.3  lattice_0.20-45
## [13] rlang_1.0.6        Matrix_1.5-1       cli_3.6.0          rstudioapi_0.14
## [17] yaml_2.3.7         mvtnorm_1.1-3      expm_0.999-7       xfun_0.37
## [21] fastmap_1.1.0      e1071_1.7-13       httr_1.4.4         knitr_1.42
## [25] gtools_3.9.4       caTools_1.18.2     gld_2.6.6          grid_4.2.2
## [29] data.table_1.14.6  R6_2.5.1           readxl_1.4.1       lmom_2.9
## [33] rmarkdown_2.20     htmltools_0.5.4    MASS_7.3-58.1      Exact_3.2
## [37] abind_1.4-5        KernSmooth_2.23-20 proxy_0.4-27
```