STA 207: Assignment I

(Sirapat Watakajaturaphon, Student ID: 920226951)

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('/Users/ploysirapat/Documents/WQ2023/STA207/HW207/Wage.csv')
library(gplots)
attach(Wage)

table(Wage$occupation)
N = nrow(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.

Let $\mu_i = \mu + \alpha_i$. The factor-effect form of the one-way ANOVA model can be written as:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad j = 1, ..., n_i, \quad i = 1, ..., 6,$$

where Y_{ij} is the wage of the jth sample in the ith occupation. Here, we define the occupations: $i = 1$ management, $i = 2$ office, $i = 3$ sales, $i = 4$ services, $i = 5$ technical, and $i = 6$ worker. The constant μ is defined as $\sum_{i=1}^{6}(n_i/N)\mu_i$ where μ_i is the cell mean when the occupation is i and $N = \sum_{i=1}^{r} n_i$ is the total sample size. The main effects plot in Part 3, shows the sample size for each occupation $n_1 = 55$, $n_2 = 97$, $n_3 = 38$, $n_4 = 83$, $n_5 = 105$, and $n_6 = 156$. The effect of the ith occupation is represented by α_i which satisfies $\sum_{i=1}^{6} n_i\alpha_i = 0$, and the random errors ϵ_{ij} are i.i.d. $N(0, \sigma^2)$.

(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.

For $i = 1, ..., 6$, the least squares estimator of α_i is

$$\hat{\alpha}_i = \bar{Y}_{i\cdot} - \hat{\mu}, \quad \text{where} \quad \hat{\mu} = \sum_{i=1}^{6} \frac{n_i}{N} \bar{Y}_{i\cdot}$$

First, we consider the expectation and variance of $\bar{Y}_{i\cdot}$ for all $i = 1, ..., 6$.

$$E\left(\bar{Y}_{i\cdot}\right) = E\left(\frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}\right) = \frac{1}{n_i}\sum_{j=1}^{n_i} E(Y_{ij}) = \frac{1}{n_i}\sum_{j=1}^{n_i} \mu_i = \mu_i = \mu + \alpha_i$$

$$\text{Var}\left(\bar{Y}_{i\cdot}\right) = \text{Var}\left(\frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}\right) = \frac{1}{n_i^2}\text{Var}\left(\sum_{j=1}^{n_i} Y_{ij}\right) = \frac{1}{n_i^2} n_i \sigma^2 = \frac{\sigma^2}{n_i}$$

Hence, we obtain the expectation of $\hat{\alpha}_i$ as follows.

$$\begin{aligned}
E(\hat{\alpha}_i) = E(\bar{Y}_{i\cdot} - \hat{\mu}) = E\left(\bar{Y}_{i\cdot}\right) - E(\hat{\mu}) &= E\left(\bar{Y}_{i\cdot}\right) - E\left(\sum_{i=1}^{6} \frac{n_i}{N} \bar{Y}_{i\cdot}\right) \\
&= E\left(\bar{Y}_{i\cdot}\right) - \sum_{i=1}^{6} \frac{n_i}{N} E\left(\bar{Y}_{i\cdot}\right) \\
&= (\mu + \alpha_i) - \sum_{i=1}^{6} \frac{n_i}{N} (\mu + \alpha_i) \\
&= (\mu + \alpha_i) - \mu \sum_{i=1}^{6} \frac{n_i}{N} - \frac{1}{N} \sum_{i=1}^{6} n_i \alpha_i \\
&= (\mu + \alpha_i) - \mu \cdot 1 - 0 \\
&= \alpha_i
\end{aligned}$$

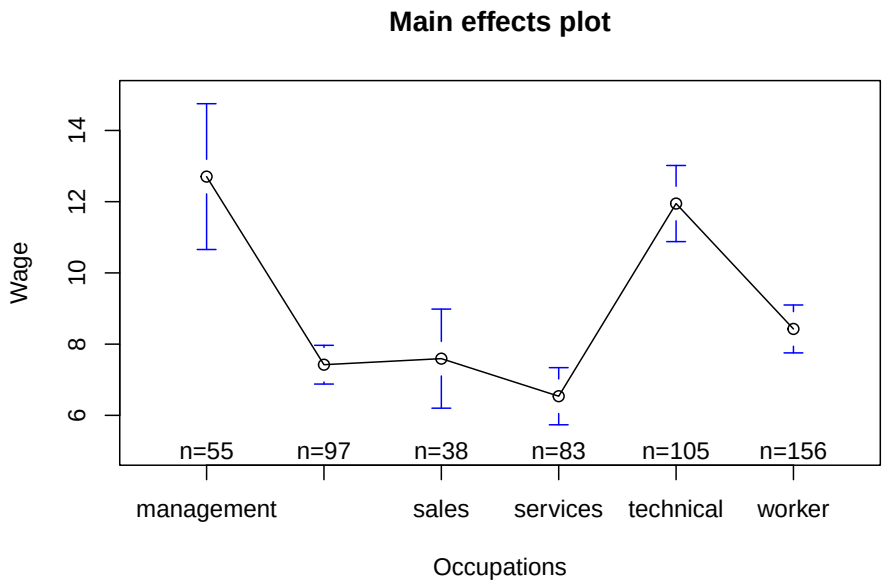
And the variance of $\hat{\alpha}_i$:

$$\begin{aligned}
\text{Var}(\hat{\alpha}_i) = \text{Var}\left[\bar{Y}_{i\cdot} - \sum_{i=1}^{6} \frac{n_i}{N}\bar{Y}_{i\cdot}\right] &= \text{Var}\left[\left(1-\frac{n_i}{N}\right)\bar{Y}_{i\cdot} - \sum_{k\neq i}\frac{n_k}{N}\bar{Y}_{k\cdot}\right] \\
&= \left(1-\frac{n_i}{N}\right)^2 \text{Var}(\bar{Y}_{i\cdot}) + \text{Var}\left(\sum_{k\neq i}\frac{n_k}{N}\bar{Y}_{k\cdot}\right) \\
&= \left(1-\frac{n_i}{N}\right)^2 \frac{\sigma^2}{n_i} + \sum_{k\neq i}\frac{n_k^2}{N^2}\frac{\sigma^2}{n_k} \\
&= \frac{(N-n_i)^2}{N^2 n_i}\sigma^2 + \sum_{k\neq i}\frac{n_k}{N^2}\sigma^2 \\
&= \frac{(N-n_i)^2}{N^2 n_i}\sigma^2 + \frac{N-n_i}{N^2}\sigma^2 \\
&= \frac{N^2 - Nn_i}{N^2 n_i}\sigma^2 \\
&= \left(\frac{1}{n_i} - \frac{1}{N}\right)\sigma^2
\end{aligned}$$

---

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

```
plotmeans(wage ~ occupation, data = Wage, ylim=c(5,15), xlab = "Occupations", ylab = "Wage",
          main= "Main effects plot")
```



- Management has the highest wage and Technical is a close second. The rest of the occupations have significantly lower wages.
- Management has the largest variability while Office has the lowest variability.
- The sample sizes vary across the occupations (imbalanced).

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

```
anova.fit = aov(wage ~ as.factor(occupation), data = Wage)
summary(anova.fit)
```

```
##                          Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(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
```

Treatment sum of squares (SSTR) is 2538 with degree of freedom 5. Residual sum of squares (SSE) is 11539 with degree of freedom 528. The mean square for treatments is MSTR=SSTR/df(SSTR)=507.5. The mean square for residuals is MSE=SSE/df(SSE)=21.9. The F test statistics is F^*=MSTR/MSE=23.22. The p-value is less than 2e-16, so it is obviously less than 0.05.

---

(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) The null and alternative hypotheses are:

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_6 = 0 \quad \text{vs} \quad H_1 : \text{ not all } \alpha_i \text{ are zero.}$$

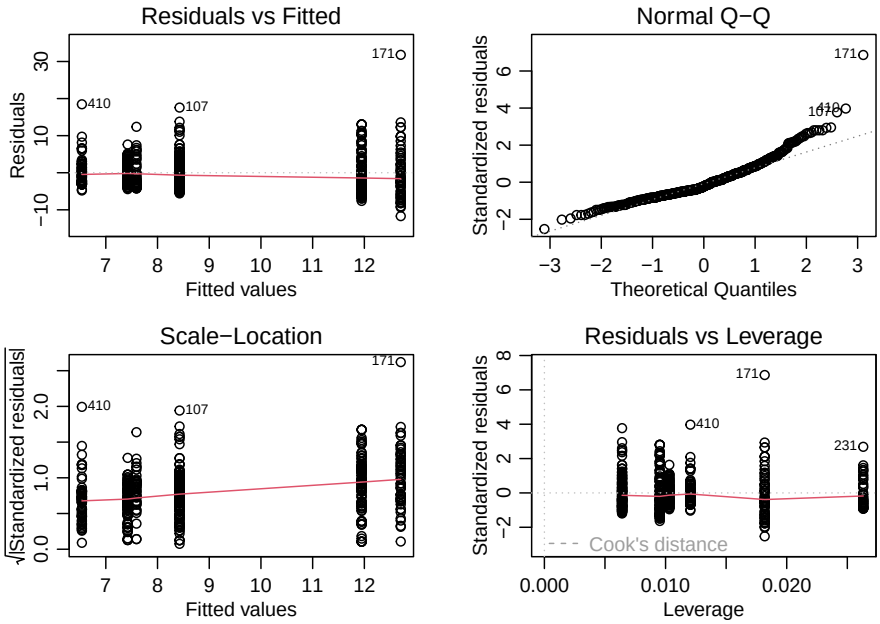
b) Conduct the F-test. From the ANOVA table in Part 4, the F test statistic is F^*=MSTR/MSE=23.22. Under H_0, the F^* follows an F-distribution with $(r-1, N-r) = (6-1, 534-6) = (5, 528)$ degrees of freedom. We then calculate the p-value $P(F_{5,528} \geq F^* = 23.22)$ which, based on the ANOVA table in Part 4, is less than 0.05.

c) Since the p-value is less than 0.05, we reject the null and conclude that there is a significant association between `occupation` and `wage` at the significance level of 0.05.

---

(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), mar=c(3,3,2,2), mgp=c(1.7,.7,0))
plot(anova.fit)
```



- There is a presence of outliers. We may need to remove those points 171, 410, and 107.
- From the Residuals vs Fitted plot, the points do not have an equal spread along the X-axis. So, there is an indication of non-constant variance. Moreover, from the Levene's test below, the p-value is much smaller than 0.05, hence we conclude that the equal variance assumption does not hold at $\alpha = 0.05$.

```
library(car)
leveneTest(wage ~ as.factor(occupation), data = Wage) # Levene's test
```

```
## 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
```

- From the Normal QQ plot, the distribution appears to be right-skewed. So, the normality assumption is violated. We can also check by Shapiro-Wilk normality test. Because of a very small p-value, we reach the same conclusion that the errors are not normally distributed.

```
shapiro.test(anova.fit$residuals) # Shapiro-Wilk normality test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  anova.fit$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.

Tukey-Kramer method is usually selected when performing pairwise comparisons. Since we have access to the data, we can apply all three methods and choose the one with the smallest multiplier (in order to get the narrowest confidence intervals). The table below confirms that the most appropriate method is Tukey-Kramer.

```
m = choose(6,2)
alpha = 0.05
r = length(anova.fit$coefficients)
B.stat = qt(1-alpha/(2*m), N-r)                    # Bonfferoni
T.stat = qtukey(1-alpha, nmeans=r, df=N-r)/sqrt(2) # Tukey-Kramer
S.stat = sqrt( (r-1)*qf(1-alpha, r-1, N-r) )       # Scheffe

knitr::kable(data.frame(Bonferroni=B.stat, Tukey=T.stat, Scheffe=S.stat),
             caption = paste("Comparison of multipliers"), digits = 2)
```

Table 1: Comparison of multipliers

Bonferroni	Tukey	Scheffe
2.95	2.86	3.34

By the main effects plot in Part 3, the two largest cell means belong to Management and Technical. So, we will focus on the difference of these two largest means. The forth row shows the lower bound (lwr=-2.98) and the upper bound (upr=1.47) of the difference between Technical and Management. We can see that zero is covered in the interval, so we cannot conclude that there is ONE occupation with the highest mean wage.

```
TukeyHSD(anova.fit) # Tukey-Kramer
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = wage ~ as.factor(occupation), data = Wage)
##
## $`as.factor(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
```

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

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

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 $L_2(\mu, \alpha)$.

The loss function associated with least squares is:

$$L_1(\mu, \alpha) = \sum_{i=1}^{6} \sum_{j=1}^{n_i} (Y_{ij} - (\mu + \alpha_i))^2 \tag{2}$$

The likelihood function is

$$\begin{aligned} L(\mu, \alpha) = \prod_{i,j} f(Y_{ij}|\mu, \alpha) &= \prod_{i,j} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(Y_{ij} - (\mu + \alpha_i))^2}{2\sigma^2} \right\} \\ &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{6} \sum_{j=1}^{n_i} (Y_{ij} - (\mu + \alpha_i))^2 \right\} \end{aligned}$$

Hence, the log-likelihood is:

$$L_2(\mu, \alpha) = \log L(\mu, \alpha) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{6} \sum_{j=1}^{n_i} (Y_{ij} - (\mu + \alpha_i))^2 \tag{3}$$

---

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

To obtain the MLE of μ, we will take the derivative of $L_2(\mu, \alpha)$ with respect to μ, set the derivative to zero, and solve for μ.

$$\begin{aligned} 0 = \frac{\partial}{\partial \mu} L_2(\mu, \alpha) &= \frac{1}{\sigma^2} \sum_{i=1}^{6} \sum_{j=1}^{n_i} (Y_{ij} - (\mu + \alpha_i)) \\ &= \frac{1}{\sigma^2} \left[\sum_{i=1}^{6} \sum_{j=1}^{n_i} Y_{ij} - \sum_{i=1}^{6} \sum_{j=1}^{n_i} \mu - \sum_{i=1}^{6} \sum_{j=1}^{n_i} \alpha_i \right] \\ &= \frac{1}{\sigma^2} \left[\sum_{i=1}^{6} \sum_{j=1}^{n_i} Y_{ij} - N\mu - \sum_{i=1}^{6} n_i \alpha_i \right] \\ &= \frac{1}{\sigma^2} \left[\sum_{i=1}^{6} \sum_{j=1}^{n_i} Y_{ij} - N\mu - 0 \right] \end{aligned}$$

Hence, the MLE of μ is

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{6} \sum_{j=1}^{n_i} Y_{ij} = \bar{Y}_{\cdot\cdot}$$

Next, we will find the MLE of α_i, for $i = 1, ..., 6$. Again take the derivative of $L_2(\mu, \alpha)$ with respect to α_i, set it tozero, and then solce for α_i.

$$\begin{aligned} 0 = \frac{\partial}{\partial \alpha_i} L_2(\mu, \alpha) &= \frac{1}{\sigma^2} \sum_{j=1}^{n_i} (Y_{ij} - (\mu + \alpha_i)) \\ &= \frac{1}{\sigma^2} \left[\sum_{j=1}^{n_i} Y_{ij} - \sum_{j=1}^{n_i} \mu - \sum_{j=1}^{n_i} \alpha_i \right] \\ &= \frac{1}{\sigma^2} \left[\sum_{j=1}^{n_i} Y_{ij} - n_i \mu - n_i \alpha_i \right] \end{aligned}$$

Thus, for $i = 1, ..., 6$, the MLE of α_i is

$$\hat{\alpha}_i = \left(\frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \right) - \hat{\mu} = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}$$

Acknowledgement

1. Course notes from STA 207 (both lecture and discussion)
2. Course notes from STA 106 (discussion)

Session information

```
# sessionInfo()
```