STA 207: Assignment II

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 some demographic factors. 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.

```
Wage=read.csv('Wage.csv');
library(gplots)
library(lme4)
attach(Wage)
```

---

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

The two-way ANOVA model is as follows:

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j} + \epsilon_{i,j,k}$$

over the cells enumerated as:

$$1 \leq i \leq a, \quad 1 \leq j \leq b, \quad 1 \leq i \leq n_{i,j}$$

subject to the constraints:

$$n_{\cdot,j} = \sum_{i=1}^{a} n_{i,j} \quad \forall j$$

$$n_{i,\cdot} = \sum_{j=1}^{b} n_{i,j} \quad \forall i$$

$$\sum_{i=1}^{a} n_{i,\cdot}\alpha_i = 0$$

$$\sum_{j=1}^{b} n_{\cdot,j}\beta_j = 0$$

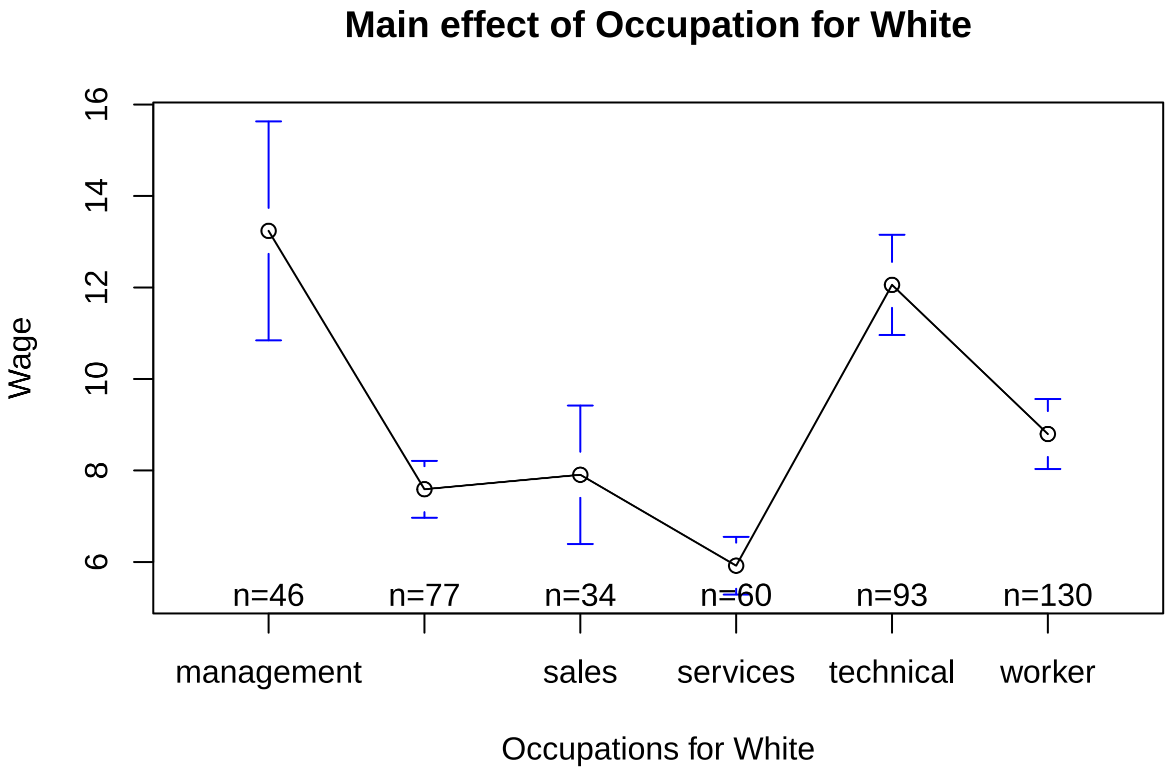
$$\sum_{i=1}^{a} n_{i,j}(\alpha\beta)_{i,j} = 0 \text{ for all } j$$

$$\sum_{j=1}^{b} n_{i,j}(\alpha\beta)_{i,j} = 0 \text{ for all } i$$

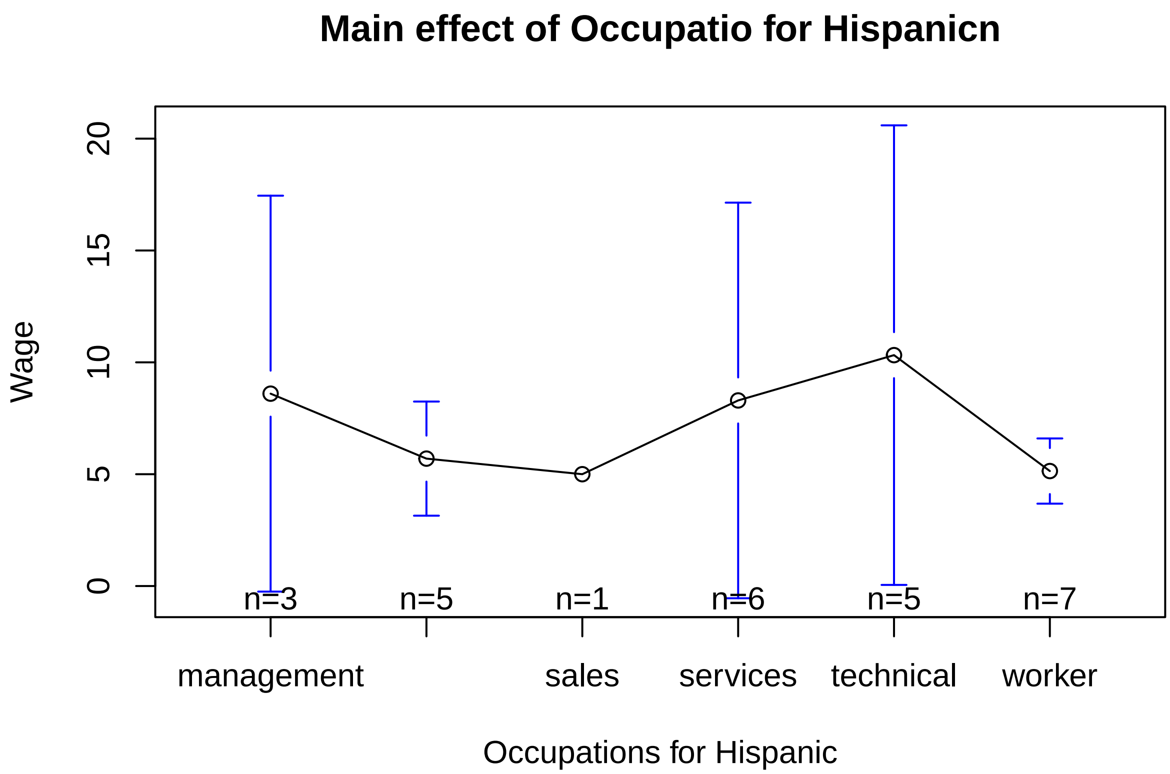
---

(2) Obtain the main effects plots and the interaction plot. Summarize your findings.

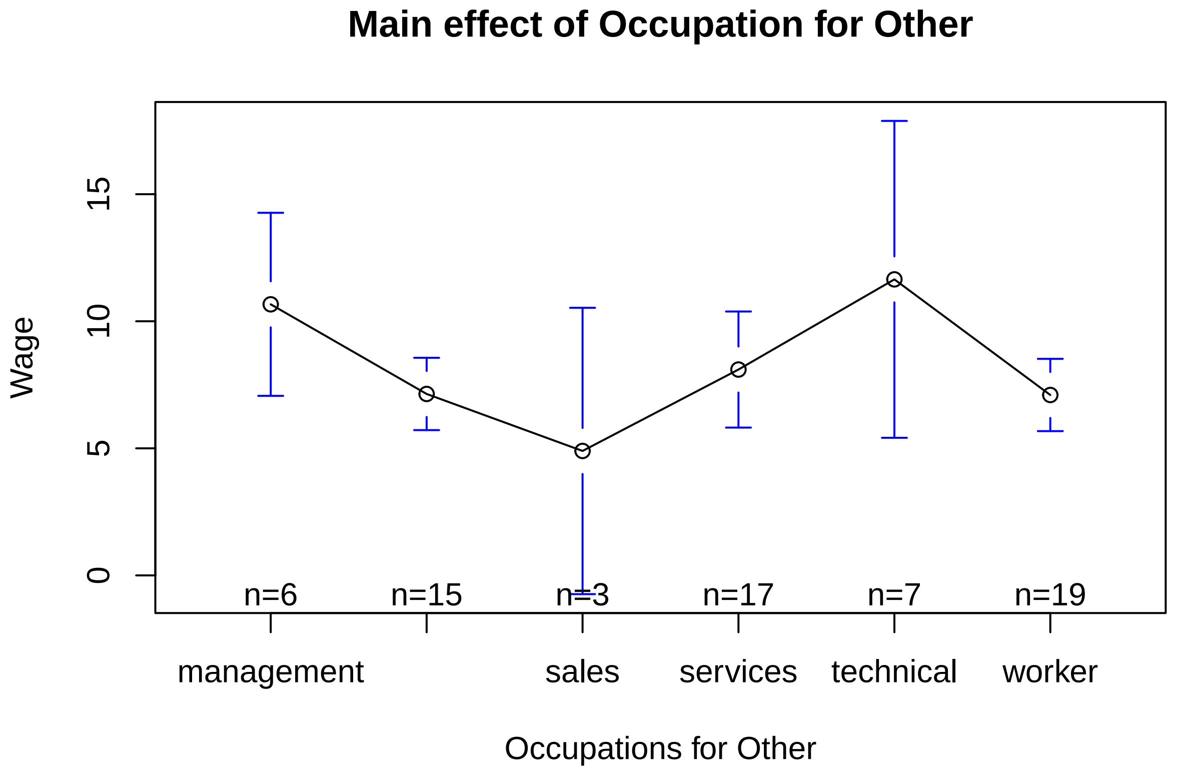
```
#par(mfrow=c(2,2))

white_wages = Wage$wage[Wage$ethnicity == "cauc"]
white_occupations = Wage$occupation[Wage$ethnicity == "cauc"]

plotmeans(white_wages ~ white_occupations,
          xlab = "Occupations for White", ylab = "Wage",
          main="Main effect of Occupation for White")
```



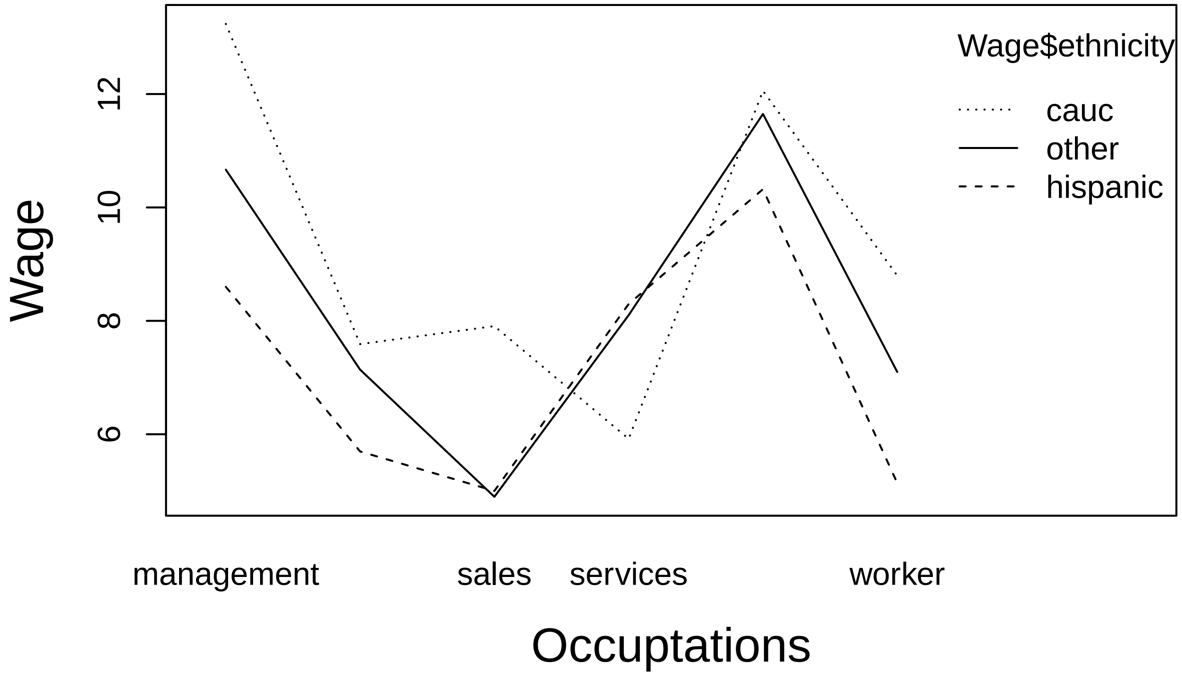
```
hisp_wages = Wage$wage[Wage$ethnicity == "hispanic"]
hisp_occupations = Wage$occupation[Wage$ethnicity == "hispanic"]

plotmeans(hisp_wages ~ hisp_occupations,
          xlab = "Occupations for Hispanic", ylab = "Wage",
          main="Main effect of Occupatio for Hispanicn")
```



```
other_wages = Wage$wage[Wage$ethnicity == "other"]
other_occupations = Wage$occupation[Wage$ethnicity == "other"]

plotmeans(other_wages ~ other_occupations,
          xlab = "Occupations for Other", ylab = "Wage",
          main="Main effect of Occupation for Other")
```



```
interaction.plot(Wage$occupation, Wage$ethnicity, Wage$wage
                ,cex.lab=1.5,ylab="Wage",xlab='Occuptations')
```



We see that for four of the six occupations, there is a trend: caucasian is the highest paid, followed by other, and hispanic as last. The only positions that deviate are sales and services. But the bulk of the recorded datapoints are for white individuals.

(3) Fit the ANOVA model described in Part 1. Obtain the ANOVA table and state your conclusions. Are the findings here consistent with your initial assessments from Part 2?

```
full_model=lm(wage~as.factor(occupation)+as.factor(ethnicity)+as.factor(occupation)*as.factor(ethnicity)
anova(full_model)
```

```
## Analysis of Variance Table
##
## Response: wage
##                                              Df  Sum Sq Mean Sq F value Pr(>F)
## as.factor(occupation)                         5  2537.7  507.54 23.4347 <2e-16
## as.factor(ethnicity)                          2    93.5   46.76  2.1592 0.1165
## as.factor(occupation):as.factor(ethnicity)   10   270.2   27.02  1.2474 0.2579
## Residuals                                   516 11175.3   21.66
##
## as.factor(occupation)                       ***
## as.factor(ethnicity)
## as.factor(occupation):as.factor(ethnicity)
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
The relationship between occupation and wage is very strong. This relationship does a good job of hiding
```

---

(4) Carry out a test to decide if the effect of ethnicity is present on the full data set, at the significance level $\alpha = 0.01$.

We see by the summary table above that the p-value is 0.1165, which doesn't exceed our desired significance level, is not considered significant to indicate the effect of ethnicity. But this could indicate the conservative nature of the statistic. If we consider the first order reduced model, we see:

```
reduced_model=lm(wage~as.factor(occupation)+as.factor(ethnicity),data=Wage)
summary(reduced_model)
```

```
##
## Call:
## lm(formula = wage ~ as.factor(occupation) + as.factor(ethnicity),
##     data = Wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.863  -3.068  -0.851   2.217  31.637
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     12.8626     0.6351  20.253  < 2e-16 ***
## as.factor(occupation)office     -5.2617     0.7879  -6.678 6.16e-11 ***
## as.factor(occupation)sales      -5.1789     0.9846  -5.260 2.10e-07 ***
## as.factor(occupation)services   -6.0816     0.8135  -7.476 3.22e-13 ***
## as.factor(occupation)technical  -0.7925     0.7769  -1.020   0.3082
## as.factor(occupation)worker     -4.2879     0.7316  -5.861 8.12e-09 ***
## as.factor(ethnicity)hispanic    -1.8032     0.9266  -1.946   0.0522 .
## as.factor(ethnicity)other       -0.5525     0.6177  -0.894   0.3715
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.665 on 526 degrees of freedom
## Multiple R-squared:  0.1869, Adjusted R-squared:  0.1761
## F-statistic: 17.27 on 7 and 526 DF,  p-value: < 2.2e-16
```

We still see on the reduced model that we don't have a strong enoguh p-value to justify the significance of ethnicity. But this could be due to the severe lack of datapoints for ethnicity.

---

(5) For this part and the next, assume that the occupations have been selected randomly. Write down an appropriate ANOVA model that is additive in the factors and explain the terms in the model.

```
my_wage = Wage$wage
my_occupation = as.factor(Wage$occupation)
my_ethnicity = as.factor(Wage$ethnicity)

full_model=lmer(my_wage ~ my_ethnicity + (1 | my_occupation))
summary(full_model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: my_wage ~ my_ethnicity + (1 | my_occupation)
##
## REML criterion at convergence: 3171.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4974 -0.6629 -0.1867  0.4927  6.8278
##
## Random effects:
##  Groups        Name        Variance Std.Dev.
##  my_occupation (Intercept)  6.205   2.491
##  Residual                  21.760   4.665
## Number of obs: 534, groups:  my_occupation, 6
##
## Fixed effects:
##                        Estimate Std. Error t value
## (Intercept)              9.2654     1.0451   8.865
## my_ethnicityhispanic    -1.8077     0.9265  -1.951
## my_ethnicityother       -0.5760     0.6175  -0.933
##
## Correlation of Fixed Effects:
##             (Intr) my_thnctyh
## my_thnctyhs -0.051
## my_thnctyth -0.077  0.093
```

---

(6) Assuming that the model in Part 5 is appropriate, obtain an estimate of the proportion of variability that is due to variability in occupation.

$$\text{proportion of variability} = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma^2} = \frac{6.205}{6.205 + 21.760} = 0.2219 \approx 0.22$$

So we estimate the proportion of variability to be about 22%

---

(7) Consider a two-way ANOVA model with fixed effects

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \epsilon_{i,j,k}, \; i = 1, \dots, a, j = 1, \dots, b, k = 1, \dots, n \tag{1}$$

where $\{\alpha_i\}$ satisfies that $\sum_i^a \alpha_i = 0$, $\{\beta_j\}$ satisfies that $\sum_j^b \beta_j = 0$, and $\{\epsilon_{i,j,k}\}$ are i.i.d. $N(0, \sigma^2)$. Derive the least squares estimator from the above equation.

First we define the least squares potential function as:

$$L_1(\mu, \alpha, \beta) = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{i,j}} (Y_{i,j,k} - \mu - \alpha_i - \beta_j)^2$$

Then we want to identify our estimators by taking the derivative of our least squares potential function:

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

We can identify the estimators by analytically optimizing:

$$0 = \frac{L_1(\mu, \alpha, \beta)}{\partial \alpha_i} = -\sum_{j=1}^{b} \sum_{k=1}^{n_{i,j}} 2(Y_{i,j,k} - \mu - \alpha_i - \beta_j)$$

$$\iff 2\sum_{j=1}^{b} \sum_{k=1}^{n_{i,j}} Y_{i,j,k} = 2\sum_{j=1}^{b} \sum_{k=1}^{n_{i,j}} (\mu + \alpha_i + \beta_j)$$

$$\iff \sum_{j=1}^{b} \sum_{k=1}^{n_{i,j}} Y_{i,j,k} = \sum_{j=1}^{b} n_{i,j}(\mu + \alpha_i + \beta_j) = n_{i,\cdot}(\mu + \alpha_i) + \sum_{j=1}^{b} n_{i,j}\beta_j = n_{i,\cdot}(\mu + \alpha_i)$$

$$\mu + \alpha_i = \frac{1}{n_{i,\cdot}} \sum_{j=1}^{b} \sum_{k=1}^{n_{i,j}} Y_{i,j,k}$$

Similarly, differentiating with respect of β_j:

$$0 = \frac{L_1(\mu, \alpha, \beta)}{\partial \beta_j} = -\sum_{i=1}^{a} \sum_{k=1}^{n_{i,j}} 2(Y_{i,j,k} - \mu - \alpha_i - \beta_j)$$

$$\iff 2\sum_{i=1}^{a} \sum_{k=1}^{n_{i,j}} Y_{i,j,k} = 2\sum_{1=1}^{a} \sum_{k=1}^{n_{i,j}} (\mu + \alpha_i + \beta_j)$$

$$\iff \sum_{i=1}^{a} \sum_{k=1}^{n_{i,j}} Y_{i,j,k} = \sum_{i=1}^{a} n_{i,j}(\mu + \alpha_i + \beta_j) = n_{\cdot,j}(\mu + \beta_j) + \sum_{i=1}^{a} n_{i,j}\alpha_i = n_{\cdot,j}(\mu + \beta_j)$$

$$\mu + \beta_j = \frac{1}{n_{\cdot,j}} \sum_{i=1}^{a} \sum_{k=1}^{n_{i,j}} Y_{i,j,k}$$

Lastly, differentiating with respect of μ:

$$0 = \frac{L_1(\mu, \alpha, \beta)}{\partial \beta_j} = -\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{i,j}} 2(Y_{i,j,k} - \mu - \alpha_i - \beta_j)$$

$$\iff \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{i,j}} Y_{i,j,k} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{i,j}} (\mu + \alpha_i + \beta_j) = n_T\mu + \sum_{i=1}^{a} n_{i,\cdot}\alpha_i + \sum_{j=1}^{b} n_{\cdot,j}\beta_j = n_T\mu$$

This gives us the final estimators:

$$\implies \hat{\mu} = \frac{1}{n_T}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{i,j}} Y_{i,j,k}$$

$$\implies \hat{\alpha}_i = \frac{1}{n_{i,\cdot}}\sum_{j=1}^{b}\sum_{k=1}^{n_{i,j}} Y_{i,j,k} - \hat{\mu}$$

$$\implies \hat{\beta}_j = \frac{1}{n_{\cdot,j}}\sum_{i=1}^{a}\sum_{k=1}^{n_{i,j}} Y_{i,j,k} - \hat{\mu}$$

---

(8) Consider the following models

$$Y_{i,j,k} = \mu_{i,j} + \epsilon_{i,j,k},\ \ k = 1, \dots, n, i = 1, \dots, a, j = 1, \dots, b, \tag{2}$$

and

$$Y_{i,j,k} = \sum_{l=1}^{a}\sum_{m=1}^{b} \beta_{l,m} X_{l,m;i,j,k} + \epsilon_{i,j,k},\ \ k = 1, \dots, n, i = 1, \dots, a, j = 1, \dots, b, \tag{3}$$

where $\{\epsilon_{i,j,k}\}$ are i.i.d. $N(0, \sigma^2)$ and $X_{l,m;i,j,k} = 1$ when $(l, m) = (i, j)$ and $X_{l,m;i,j,k} = 0$ otherwise. Express $\{\beta_{l,m} : l = 1, \dots, a; m = 1, \dots, b\}$ using $\{\mu_{i,j} : i = 1, \dots, a; j = 1, \dots, b\}$.

Observe:

$$\begin{aligned}
\mu_{i,j} &= \sum_{l=1}^{a}\sum_{m=1}^{b} \beta_{l,m} X_{l,m;i,j,k} \\
&= \beta_{i,j} X_{i,j;i,j,k} + \underbrace{\sum\sum \beta_{l,m} X_{l,m;i,j,k}}_{(l,m)\neq(i,j)} \\
&= \beta_{i,j} X_{i,j;i,j,k} + 0 \\
&= \beta_{i,j}
\end{aligned}$$

---

(9) With some abuse of notation, we rewrite the regression model as

$$Y = X\beta + \epsilon, \tag{4}$$

where Y is a n_T-dimensional vector, X is an $n_T \times p$ matrix, β is a p-dimensional vector, and $\{\epsilon\} \sim$ MVN$(0, \sigma^2\mathrm{I})$, i.e., multivariate normal with covariance matrix $\sigma^2\mathrm{I}$. Express the residual sum of squares and explained sum of squares in Y and X, and then show that these two sum of squares are independent.

We can rewrite the residual sum of squares and explained sum of squares as the following functions:

$$SSE = Y^T(I - X(X^TX)^{-1}X^T)Y = e^Te$$

$$SSR = Y^T(X(X^TX)^{-1}X^T - \frac{1}{n_T}1 \cdot 1^T)Y = g(\hat{\beta}, \hat{Y})$$

Observe,

$$\begin{aligned}
Cov(e, \hat{Y}) &= Cov((I - X(X^TX)^{-1}X^T)Y, X(X^TX)^{-1}X^TY) \\
&= (I - X(X^TX)^{-1}X^T)\sigma^2(Y)(X(X^TX)^{-1}X^T)^T \\
&= (I - X(X^TX)^{-1}X^T)\sigma^2(Y)X(X^TX)^{-1}X^T \\
&= \sigma^2(Y)(I - X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T & \text{since } \sigma^2(Y) = \sigma^2 I \\
&= \sigma^2(Y)(X(X^TX)^{-1}X^T - (X(X^TX)^{-1}X^T)^2) \\
&= \sigma^2(Y)(X(X^TX)^{-1}X^T - X(X^TX)^{-1}X^T) \\
&= 0
\end{aligned}$$

$$\begin{aligned}
Cov(e, \hat{\beta}) &= Cov((I - X(X^TX)^{-1}X^T)Y, (X^TX)^{-1}X^TY) \\
&= (I - X(X^TX)^{-1}X^T)\sigma^2(Y)X(X^TX)^{-T} & \text{since } \sigma^2(Y) = \sigma^2 I \\
&= \sigma^2(Y)(I - X(X^TX)^{-1}X^T)X(X^TX)^{-T} \\
&= \sigma^2(Y)(X(X^TX)^{-T} - X(X^TX)^{-T}) \\
&= 0
\end{aligned}$$

Now, we define the function:

$$SSE = f(e) = e^Te = \left\|(I - X(X^TX)^{-1}X^T)Y\right\|_2^2$$

Therefore, if we let $g = id : \mathbb{R}^n \to \mathbb{R}$, then we see that by the fact that if two sets of random variables, say $(Z_1, \dots, Z_s)$ and $(W_1, \dots, W_t)$, are independent with each other, then their functions, say $f(Z_1, \dots, Z_s)$ and $g(W_1, \dots, W_t)$, are independent, then it must follow that

$$e \perp\!\!\!\perp \hat{\beta} \implies f(e) \perp\!\!\!\perp g(\hat{\beta}) \implies SSE \perp\!\!\!\perp \hat{\beta}$$

Similarly, consider the function:

$$SSR = g(\hat{\beta}, \hat{Y}) = \left\|(X(X^TX)^{-1}X^T - \frac{1}{n_T}1 \cdot 1^T)Y\right\|_2^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

Then it must follow that

$$e \perp\!\!\!\perp \hat{\beta} \text{ and } \hat{Y} \implies f(e) \perp\!\!\!\perp g(\hat{\beta}, \hat{Y}) \implies SSE \perp\!\!\!\perp SSR$$

Acknowledgement

I appreciate the Piazza comments that helped me complete this homework assignment.

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] lme4_1.1-31  Matrix_1.5-1 gplots_3.1.3
## 
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.10        rstudioapi_0.14    knitr_1.42         splines_4.2.2     
##  [5] MASS_7.3-58.1      lattice_0.20-45    rlang_1.0.6        fastmap_1.1.0     
##  [9] minqa_1.2.5        highr_0.10         caTools_1.18.2     tools_4.2.2       
## [13] grid_4.2.2         nlme_3.1-160       xfun_0.37          KernSmooth_2.23-20
## [17] cli_3.6.0          htmltools_0.5.4    gtools_3.9.4       yaml_2.3.7        
## [21] digest_0.6.31      nloptr_2.0.3       bitops_1.0-7       evaluate_0.20     
## [25] rmarkdown_2.20     compiler_4.2.2     boot_1.3-28
```