STA 207: Assignment II

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 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)

Wage$ethnicity = as.factor(Wage$ethnicity)
Wage$occupation = as.factor(Wage$occupation)
```

```
knitr::kable(table(Wage$ethnicity, Wage$occupation),
             caption='Table1: The numbers of observations in each cell')
```

Table 1: Table1: The numbers of observations in each cell

	management	office	sales	services	technical	worker
cauc	46	77	34	60	93	130
hispanic	3	5	1	6	5	7
other	6	15	3	17	7	19

---

(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.

Solution: The factor-effect form of the two-way ANOVA model can be written as:

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \epsilon_{ijk}, \quad k = 1, ..., n_{ij}, \quad j = 1, ..., 6 \quad i = 1, 2, 3,$$

where the random errors ϵ_{ij} are i.i.d. $N(0, \sigma^2)$. Y_{ijk} is the wage of the kth sample from the ith ethnicity and jth occupation. Ethnicity has $a = 3$ levels $i = 1$ Caucasian, $i = 2$ Hispanic, and $i = 3$ Others. And Occupations has $b = 6$ levels: $j = 1$ Management, $j = 2$ Office, $j = 3$ Sales, $j = 4$ Services, $j = 5$ Technical, and $j = 6$ Worker. The total sample size is $n_T = 534$.

The overall mean $\mu_{..}$ is defined as $\sum_{i=1}^{3} \sum_{j=1}^{6} \mu_{ij}/18$ where μ_{ij} is the cell mean determined by one unique combination of two factors. Table 1 shows that the numbers of observations vary across the cells, so it is an unbalanced design. The effect of the i ethnicity is represented by α_i and that of the jth occupation by β_j. The constraints on these effects are:

$$\sum_{i=1}^{3} \alpha_i = \sum_{j=1}^{6} \beta_j = 0$$

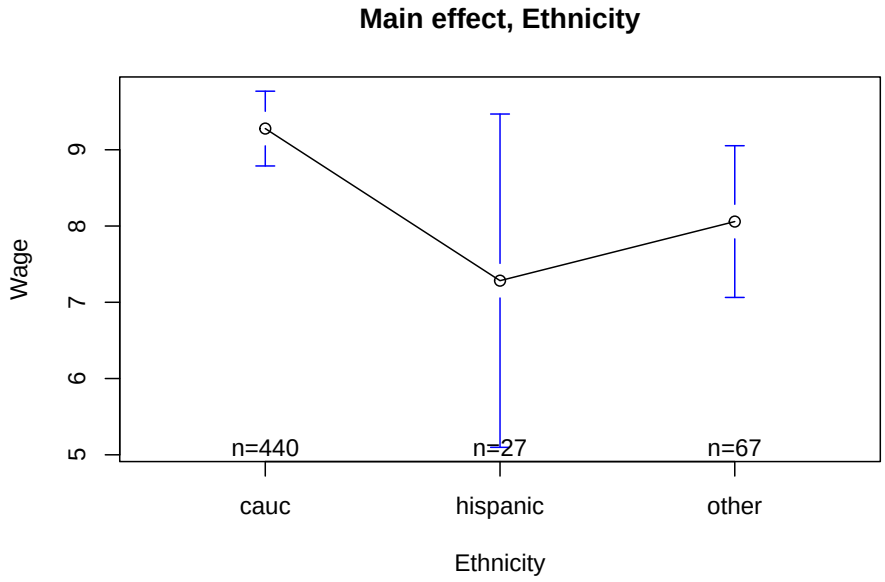
---

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

Solution:

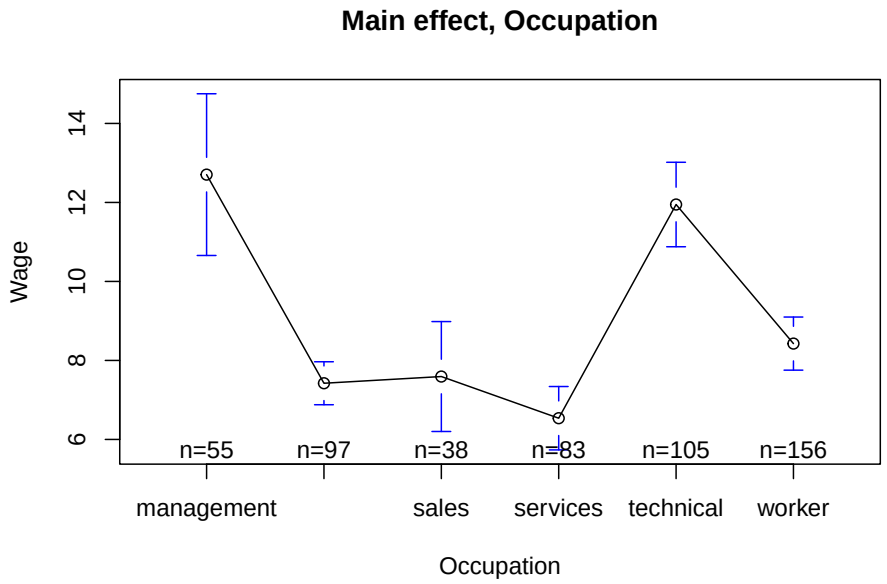
```
library(gplots)
par(mfrow=c(1,1))

# main effect plot of ethnicity
plotmeans(wage~ethnicity, data=Wage, xlab="Ethnicity", ylab="Wage",
          main="Main effect, Ethnicity")
```



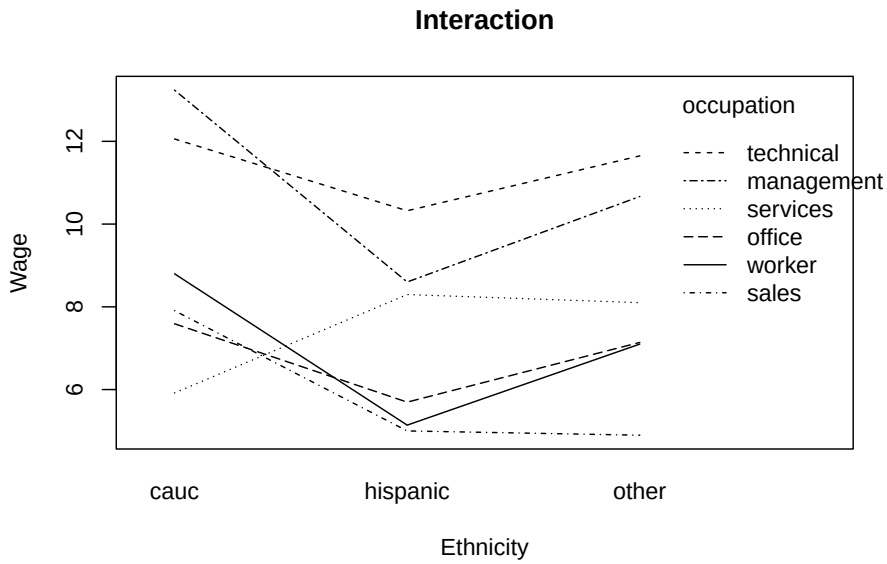
- Caucasian seems to have the overall highest wage. But there is no big difference in wages across the ethnicities.
- The variability in the Hispanic's wages is very large. Caucasian has the lowest variability
- The sample size for Caucasian is very large compared to that for other ethnicities.

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



- Management and Technical have high wages. 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.

```
# interaction plot
with(Wage, interaction.plot(ethnicity, occupation, wage,
                            xlab="Ethnicity",ylab="Wage", main="Interaction"))
```



- The interaction effect is not obviously present.

(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?

Solution: First, we test to see whether the interaction effects are present, $H_0 : (\alpha\beta)_{ij} = 0$ vs H_1: not all $(\alpha\beta)_{ij} = 0$, at the significance level $\alpha = 0.01$. The full model is the two-way ANOVA model with main effects and interaction $Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$ and the reduced model is the one in Part 1. Since the resulting p-value 0.26 is larger than 0.01, we conclude that the interaction terms are not significant at $\alpha = 0.01$.

```
# test for interactions
full.model    = aov(wage ~ ethnicity * occupation, data=Wage)  # with interactions
reduced.model = aov(wage ~ ethnicity +  occupation, data=Wage) # additive (no interactions)
anova(reduced.model, full.model)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ ethnicity + occupation
## Model 2: wage ~ ethnicity * occupation
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    526 11446
## 2    516 11175 10    270.16 1.2474 0.2579
```

Because it is an unbalanced design and there is no interaction effect, the type II ANOVA is preferred. We fit the model described in Part 1 and obtain the ANOVA table as follows.

```
# fit model in Part 1
library(car)
options(knitr.kable.NA = '')
knitr::kable(Anova(lm(wage ~ ethnicity + occupation, data=Wage), type = 'II'),
             caption = 'Table 2: ANOVA Table of model in Part 1')
```

Table 2: Table 2: ANOVA Table of model in Part 1

	Sum Sq	Df	F value	Pr(>F)
ethnicity	93.52643	2	2.149098	0.1176119
occupation	2458.56892	5	22.597704	0.0000000
Residuals	11445.47501	526		

- Based on the test for interactions, the interaction effects between ethnicity and occupation are not present, which corresponds to what we observed from the interaction plot in Part 2.
- Based on the output p-values in Table 2, occupation is very significant to wages while ethnicity is not as much. This is similar to what we noticed from the main effects plots in Part 2.

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

Solution: The null and alternative hypotheses are:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \quad \text{vs} \quad H_1 : \alpha_1 \neq 0 \text{ or } \alpha_2 \neq 0 \text{ or } \alpha_3 \neq 0.$$

We use an F test where the test statistic is F^* =MSA/MSE which follows an F-distribution with $df = (a-1, (n_T - 1)ab)$. From Table 2, the output p-value for ethnicity is calculated to be 0.12 which is larger than 0.01. Hence, we fail to reject the null, indicating the effect of ethnicity is not present at $\alpha = 0.01$.

---

(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.

Solution: The additive two-way ANOVA with random effects for both factors can be written as:

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \epsilon_{ijk}, \quad k = 1, ..., n_{ij}, \quad j = 1, ..., 6 \quad i = 1, 2, 3,$$

where the random errors ϵ_{ij} are i.i.d. $N(0, \sigma^2)$, the random main effects of the ith ethnicity α_i are i.i.d. $N(0, \sigma_\alpha^2)$, the random main effects of the jth occupation β_j are i.i.d. $N(0, \sigma_\beta^2)$, and all these random variables are mutually independent.

In this model (random effect model), $\mu_{..}$ represents population mean across *all possible* factor levels. And the factor levels are selected randomly from a larger pool of levels.

---

(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.

Solution:

```
# random effect model
library(lme4)
fit2 = lmer(wage ~ (1 | ethnicity) + (1 | occupation), data = Wage)
summary(fit2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: wage ~ (1 | ethnicity) + (1 | occupation)
##    Data: Wage
##
## REML criterion at convergence: 3178
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4827 -0.6618 -0.2216  0.4895  6.8410
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  occupation (Intercept)  6.2280  2.4956
##  ethnicity  (Intercept)  0.3239  0.5691
##  Residual               21.7670  4.6655
## Number of obs: 534, groups:  occupation, 6; ethnicity, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)    8.781      1.115   7.874
```

From the output, we have the estimates $\hat{\sigma}^2_\alpha = 0.32$ (ethnicity), $\hat{\sigma}^2_\beta = 6.23$ (occupation), and $\hat{\sigma}^2 = 21.77$ (error term). Thus, the total variation is $0.32 + 6.23 + 21.77 = 28.32$. The estimate of the proportion of variability that is due to variability in occupation is $6.23/28.32 = 0.22$.

---

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

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

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.

Solution: The least squares (LS) estimators $\hat{\mu}_{LS}$, $\hat{\alpha}_i$, and $\hat{\beta}_j$ are the minimizers of

$$Q(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} \left[Y_{ijk} - (\mu + \alpha_i + \beta_j)\right]^2.$$

Taking

$$\begin{aligned} 0 = \frac{\partial Q}{\partial \mu} &= -2\sum_i\sum_j\sum_k [Y_{ijk} - (\mu + \alpha_i + \beta_j)] \\ &= -2\left\{\sum_i\sum_j\sum_k Y_{ijk} - abn\mu - nb\sum_i \alpha_i - na\sum_j \beta_j\right\} \\ &= -2\left\{\sum_i\sum_j\sum_k Y_{ijk} - abn\mu - 0 - 0\right\} \end{aligned}$$

yields the LS estimator of μ

$$\hat{\mu}_{LS} = \frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk}.$$

For $i = 1, ..., a$, consider

$$\begin{aligned} 0 = \frac{\partial Q}{\partial \alpha_i} &= -2\sum_j\sum_k [Y_{ijk} - (\mu + \alpha_i + \beta_j)] \\ &= -2\left\{\sum_j\sum_k Y_{ijk} - nb\mu - nb\alpha_i - n\sum_j \beta_j\right\} \\ &= -2\left\{\sum_j\sum_k Y_{ijk} - nb\mu - nb\alpha_i - 0\right\} \end{aligned}$$

Thus, the LS estimator of α_i is

$$\hat{\alpha}_i = \frac{1}{nb}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk} - \hat{\mu}_{LS}, \quad \text{for } i = 1, ..., a.$$

For $j = 1, ..., b$, consider

$$\begin{aligned} 0 = \frac{\partial Q}{\partial \beta_j} &= -2 \sum_i \sum_k [Y_{ijk} - (\mu + \alpha_i + \beta_j)] \\ &= -2 \left\{ \sum_i \sum_k Y_{ijk} - na\mu - n \sum_i \alpha_i - na\beta_j \right\} \\ &= -2 \left\{ \sum_i \sum_k Y_{ijk} - na\mu - 0 - na\beta_j \right\} \end{aligned}$$

Thus, the LS estimator of β_j is

$$\hat{\beta}_j = \frac{1}{na} \sum_{i=1}^{a} \sum_{k=1}^{n} Y_{ijk} - \hat{\mu}_{LS}, \quad \text{for } j = 1, ..., b.$$

---

(8) Consider the following models

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

and

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

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 $\{\beta_{l,m} : l = 1, \ldots, a; m = 1, \ldots, b\}$ using $\{\mu_{i,j} : i = 1, \ldots, a; j = 1, \ldots, b\}$.

Solution: For $1 \leq i \leq a$ and $1 \leq j \leq b$, consider

$$\begin{aligned} \mu_{i,j} &= \sum_{l=1}^{a} \sum_{m=1}^{b} \beta_{l,m} X_{l,m;i,j,k} \\ &= \sum_{l=1}^{a} [\beta_{l,1} X_{l,1;i,j,k} + ... + \beta_{l,j} X_{l,j;i,j,k} + ... + \beta_{l,b} X_{l,b;i,j,k}] \\ &= \sum_{l=1}^{a} [\beta_{l,j} X_{l,j;i,j,k}] = [\beta_{1,j} X_{1,j;i,j,k} + ... + \beta_{i,j} X_{i,j;i,j,k} + ... + \beta_{a,j} X_{a,j;i,j,k}] = \beta_{i,j} X_{i,j;i,j,k} = \beta_{i,j} \end{aligned}$$

Hence, for $1 \leq l \leq a$ and $1 \leq m \leq b$,

$$\beta_{l,m} = \mu_{l,m} = \mu_{l,m} X_{l,m;l,m,k} = \sum_{i=1}^{a} [\mu_{i,m} X_{l,m;i,m,k}] = \sum_{i=1}^{a} \left[\sum_{j=1}^{b} \mu_{i,j} X_{l,m;i,j,k} \right]$$

Therefore, $\{\beta_{l,m} : l = 1, \ldots, a; m = 1, \ldots, b\}$ can be expressed using $\{\mu_{i,j} : i = 1, \ldots, a; j = 1, \ldots, b\}$ as follows:

$$\beta_{l,m} = \sum_{i=1}^{a} \sum_{j=1}^{b} \mu_{i,j} X_{l,m;i,j,k}$$

---

(9) With some abuse of notation, we rewrite the regression model from Question 8 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 $\{\epsilon\} \sim \text{MVN}(0, \sigma^2\text{I})$, i.e., multivariate normal with covariance matrix $\sigma^2\text{I}$. 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.

Solution: The LS estimators vector:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

The fitted values vector:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat matrix. Note that $\mathbf{H}$ is a projection matrix, i.e., $\mathbf{H}' = \mathbf{H}$ and $\mathbf{H}^2 = \mathbf{H}\mathbf{H} = \mathbf{H}$.

The residuals vector:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Let $\mathbf{1} = (1, ..., 1)'_{n_T}$ be a $n_T \times 1$ vector of ones. Then define the $n_T \times n_T$ matrix

$$\mathbf{J} = \mathbf{1}\mathbf{1}' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

We can see that

$$\mathbf{H}\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{X},$$

then using the fact that some column of design matrix is $\mathbf{1} = (1, ..., 1)'_{n_T}$ (since every cell has at least one observations), we have $\mathbf{H}\mathbf{1} = \mathbf{1}$. This leads to

$$\mathbf{H}\mathbf{J} = \mathbf{H}\mathbf{1}\mathbf{1}' = \mathbf{1}\mathbf{1}' = \mathbf{J} \quad \text{and} \quad \mathbf{J}\mathbf{H} = \mathbf{1}\mathbf{1}'\mathbf{H} = \mathbf{1}\mathbf{1}' = \mathbf{J}.$$

Thus,

- $(\mathbf{I} - \mathbf{H})$ is a projection matrix because $(\mathbf{I} - \mathbf{H})' = (\mathbf{I} - \mathbf{H})$ and $(\mathbf{I} - \mathbf{H})^2 = \mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{H} = (\mathbf{I} - \mathbf{H})$.
- $(\mathbf{H} - \frac{1}{n_T}\mathbf{J})$ is a projection matrix because $(\mathbf{H} - \frac{1}{n_T}\mathbf{J})' = (\mathbf{H} - \frac{1}{n_T}\mathbf{J})$ and

$$(\mathbf{H} - \frac{1}{n_T}\mathbf{J})^2 = \mathbf{H} - \frac{1}{n_T}\mathbf{H}\mathbf{J} - \frac{1}{n_T}\mathbf{J}\mathbf{H} + \frac{1}{n_T}\mathbf{J} = \mathbf{H} - \frac{1}{n_T}\mathbf{J} - \frac{1}{n_T}\mathbf{J} + \frac{1}{n_T}\mathbf{J} = (\mathbf{H} - \frac{1}{n_T}\mathbf{J}).$$

Therefore, the residual sum of squares can be written as

$$\begin{aligned} \text{SSE} = \mathbf{e}'\mathbf{e} &= ((\mathbf{I} - \mathbf{H})\mathbf{Y})'((\mathbf{I} - \mathbf{H})\mathbf{Y}) \\ &= \mathbf{Y}'(\mathbf{I} - \mathbf{H})'(\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} \end{aligned}$$

and the explained sum of squares can be written as

$$\begin{aligned}\text{SSR} = (\hat{\mathbf{Y}} - \bar{\mathbf{Y}})'(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}) &= \left((\mathbf{H} - \frac{1}{n_T}\mathbf{J})\mathbf{Y}\right)' \left((\mathbf{H} - \frac{1}{n_T}\mathbf{J})\mathbf{Y}\right) \\ &= \mathbf{Y}'(\mathbf{H} - \frac{1}{n_T}\mathbf{J})'(\mathbf{H} - \frac{1}{n_T}\mathbf{J})\mathbf{Y} \\ &= \mathbf{Y}'(\mathbf{H} - \frac{1}{n_T}\mathbf{J})\mathbf{Y}\end{aligned}$$

We show that $(\mathbf{I} - \mathbf{H})\mathbf{Y}$ and $(\mathbf{H} - \frac{1}{n_T}\mathbf{J})\mathbf{Y}$ are uncorrelated as follows:

$$\begin{aligned}\text{Cov}\left((\mathbf{I} - \mathbf{H})\mathbf{Y},\ \left(\mathbf{H} - \frac{1}{n_T}\mathbf{J}\right)\mathbf{Y}\right) &= (\mathbf{I} - \mathbf{H})\text{Var}\{\mathbf{Y}\}\left(\mathbf{H} - \frac{1}{n_T}\mathbf{J}\right) \\ &= (\mathbf{I} - \mathbf{H})(\sigma^2\mathbf{I})\left(\mathbf{H} - \frac{1}{n_T}\mathbf{J}\right) \\ &= \sigma^2\left[\mathbf{H} - \frac{1}{n_T}\mathbf{J} - \mathbf{H}\mathbf{H} + \frac{1}{n_T}\mathbf{H}\mathbf{J}\right] \\ &= \sigma^2\left[\mathbf{H} - \frac{1}{n_T}\mathbf{J} - \mathbf{H} + \frac{1}{n_T}\mathbf{J}\right] \\ &= \mathbf{0}\end{aligned}$$

With Normality assumption on the error terms, $(\mathbf{I} - \mathbf{H})\mathbf{Y}$ and $(\mathbf{H} - \frac{1}{n_T}\mathbf{J})\mathbf{Y}$ are independent. Since SSE is a function of $(\mathbf{I} - \mathbf{H})\mathbf{Y}$ and SSR is a function of $(\mathbf{H} - \frac{1}{n_T}\mathbf{J})\mathbf{Y}$, we can conclude that SSE and SSR are independent.

Acknowledgement

1. Course notes from STA 207 (both lecture and discussion)
2. Course notes from STA 206
3. https://stat.ethz.ch/~meier/teaching/anova/random-and-mixed-effects-models.html#eq:cell-means-random

Session information

```
# sessionInfo()
```