Assignment ZanCui 47384263

Question 1

a

The correlation matrix and plot are shown below.

Relationships between response and predictors:

- There is a moderate positive relationship between pm25 and temperature with correlation coefficient as 0.57.
- There is a strong negative relationship between pm25 and humidity with correlation coefficient as -0.72.
- There is a slightly negative relationship between pm25 and wind with correlation coefficient as -0.22.

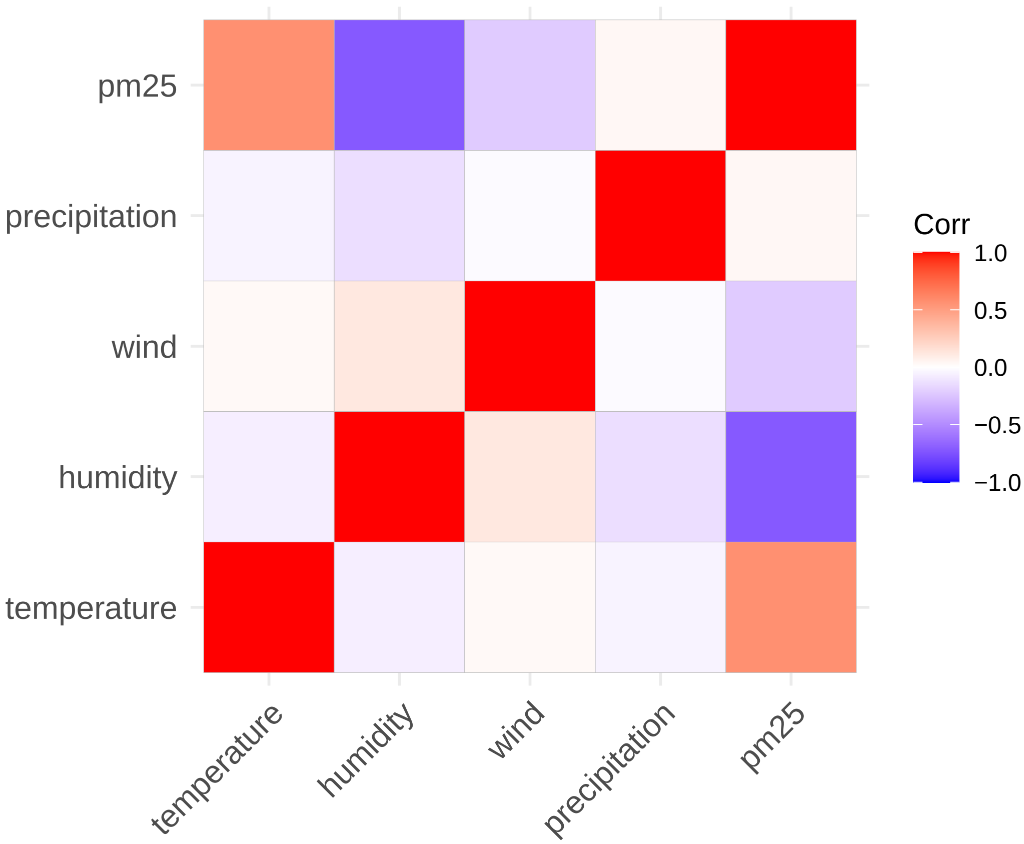
Relationships between the predictors themselves:

- There is a slightly positive relationship between humidity and wind with correlation coefficient as 0.12.
- There is a slightly negative relationship between humidity and precipitation with correlation coefficient as -0.14.

```r
library(ggplot2)
library(ggcorrplot)
setwd("/Users/selinayqi/Desktop/stat2170-zancui")
data <- read.csv("data/pm25.csv")
corr <- cor(data)
corr
```

```
##               temperature    humidity        wind precipitation        pm25
## temperature    1.00000000 -0.07264891  0.02861166   -0.05050014  0.57191961
## humidity      -0.07264891  1.00000000  0.12406351   -0.13550607 -0.71965591
## wind           0.02861166  0.12406351  1.00000000   -0.01525977 -0.21866823
## precipitation -0.05050014 -0.13550607 -0.01525977    1.00000000  0.03759033
## pm25           0.57191961 -0.71965591 -0.21866823    0.03759033  1.00000000
```

```r
ggcorrplot(corr)
```



b

The 95% confidence interval for the coefficient of humidity is (-1.515, -1.039).

We have 95% confidence that the change in PM2.5 concentration for each extra percentage of relative humidity is between -1.515 and -1.039.

```r
model <- lm(pm25~temperature+humidity+wind+precipitation, data=data)
summary(model)
```

```
##
## Call:
## lm(formula = pm25 ~ temperature + humidity + wind + precipitation,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.759  -6.804  -1.649   6.857  20.975
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   102.72259   14.71953   6.979 5.88e-09 ***
## temperature     1.62142    0.18762   8.642 1.46e-11 ***
## humidity       -1.27742    0.11854 -10.776 9.49e-15 ***
## wind           -0.58016    0.23405  -2.479   0.0165 *
## precipitation  -0.01091    0.02350  -0.464   0.6444
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 51 degrees of freedom
## Multiple R-squared:  0.8127, Adjusted R-squared:  0.7981
## F-statistic: 55.34 on 4 and 51 DF,  p-value: < 2.2e-16
```

```r
b <- model$coefficients[3]
n <- length(data[,1])
t <- qt(0.975, n-5)
se <- 0.11854
ci.lower <- b-t*se
ci.upper <- b+t*se
c(ci.lower, ci.upper)
```

```
##  humidity  humidity
## -1.515402 -1.039444
```

c

The regression model is:

$pm25 = \beta_0 temperature + \beta_1 humidity + \beta_2 wind + \beta_3 precipitation + u.$

We conduct the overall ANOVA test for the above model.

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

$H_a : not all \beta_1, \beta_2 and \beta_3 are equal to 0$

The anova table are shown below.

Test statistic: $F_obs = \frac{Reg.M.S.}{Res.M.S} = 55.34.$

If H0 is true then F is distributed according to the F distribution with (k, n-k-1)= (4, 51) degrees of freedom.

P-value: ${P(F_{4,51}>=55.34)=0.0000<0.05}.

Reject at 5% level. There is a significant linear relationship between percentage response and at least one of the three predictor variables. The overall model is significant.

```r
anova(model)
```

```
## Analysis of Variance Table
##
## Response: pm25
##               Df  Sum Sq Mean Sq  F value    Pr(>F)
## temperature    1  9014.4  9014.4  89.0853 8.908e-13 ***
## humidity       1 12739.7 12739.7 125.9013 2.200e-15 ***
## wind           1   622.6   622.6   6.1533   0.01646 *
## precipitation  1    21.8    21.8   0.2156   0.64440
## Residuals     51  5160.6   101.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
k=4
ess = 9014.4+12739.7+622.6+21.8
rss = 5160.6
f=ess/k/(rss/(n-k-1))
f
```

```
## [1] 55.3387
```

```r
p_value = 1-pf(f, k, n-k-1)
p_value
```

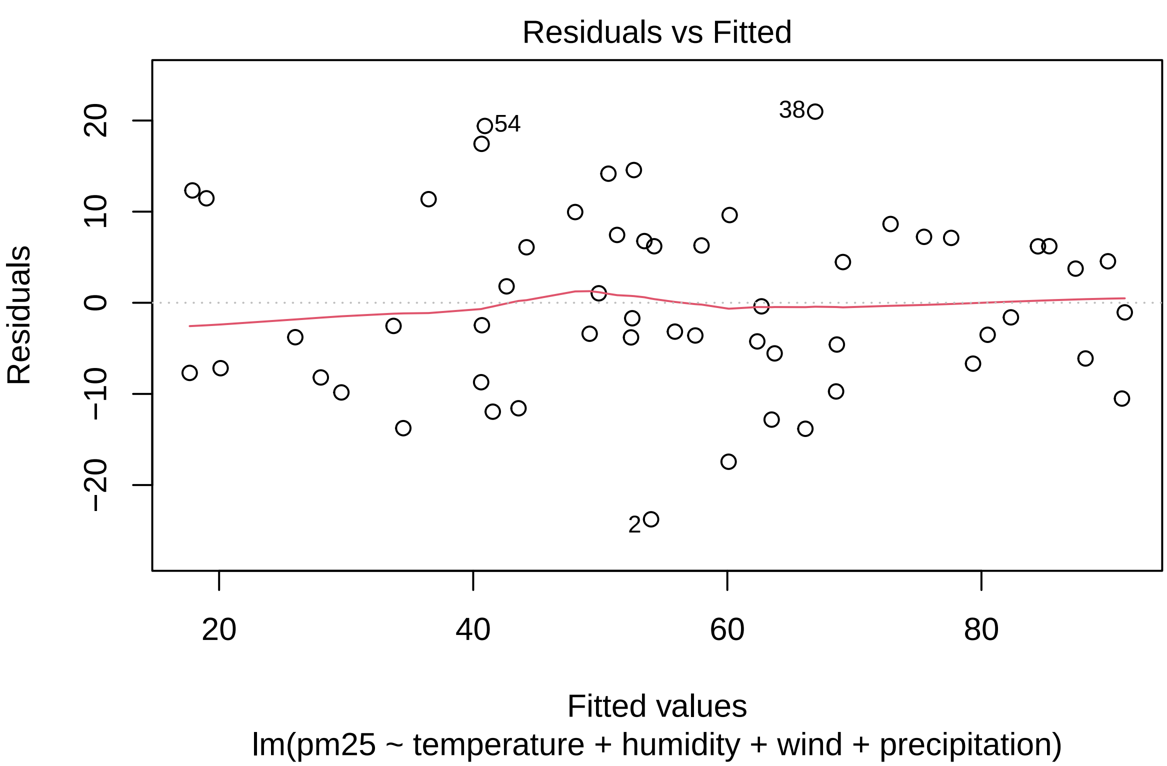
```
## [1] 0
```

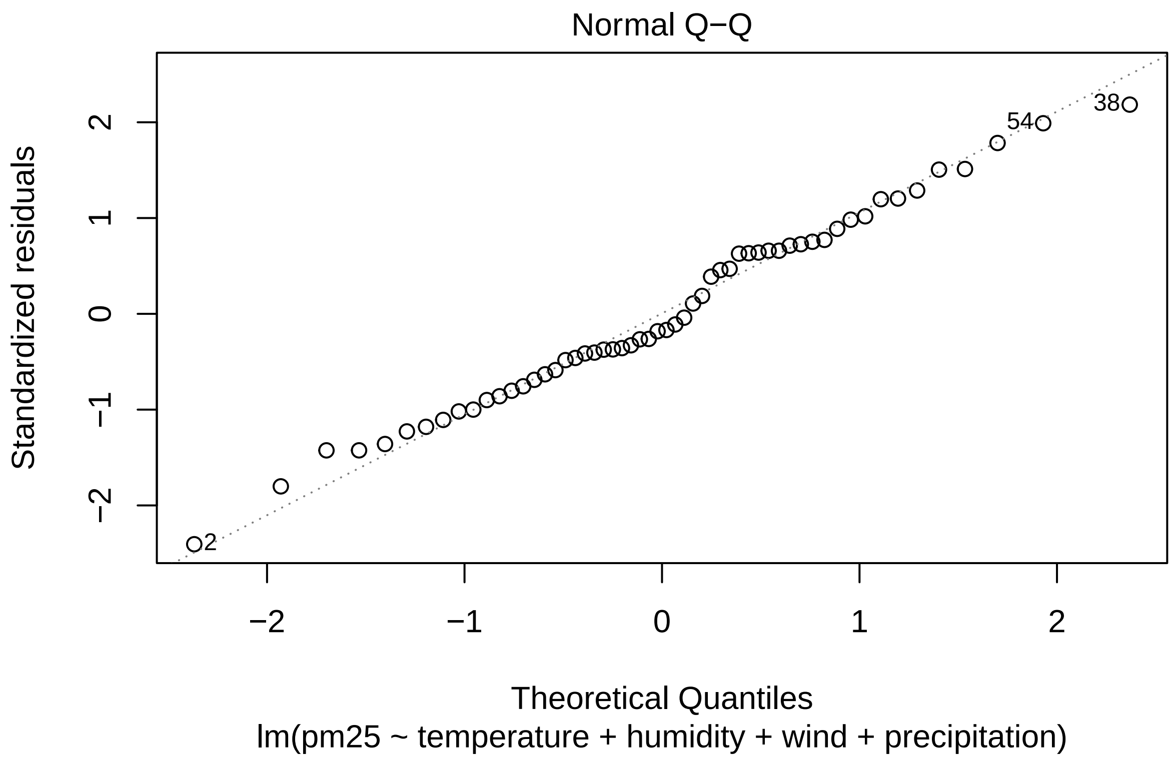
d

We check diagnostics, finding that there is no sign of heteroskedasticity. The normality assumption is also met. We plot the residuals against predictors and there is no sign of curvature.

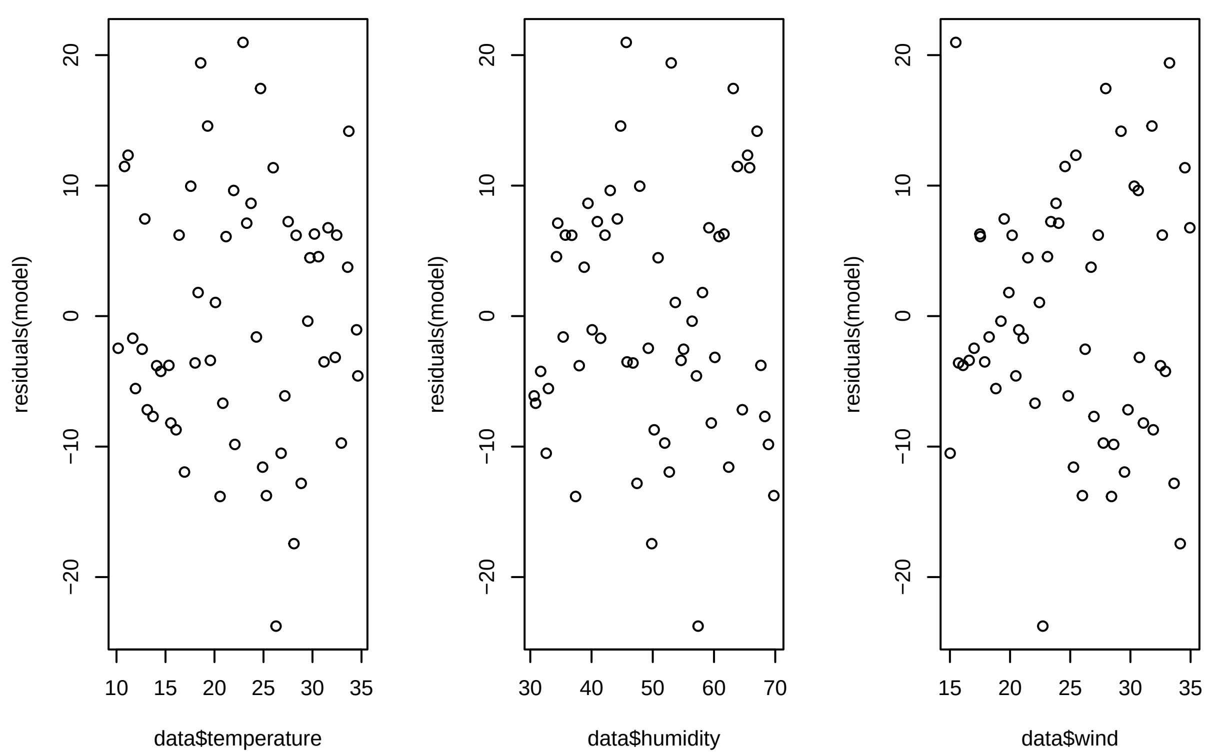
The overall model significance F-test indicates that the model is significant. The regression coefficient for temperature, humidity and wind are significant at 0.01 significance level, indicating that the above variables can be used to explain the PM2.5 concentration. As long as the location's temperature, humidity and wind are within the range of the sample variables, this model is appropriate to explain the PM2.5 concentration at various test locations.

```r
#check diagnostics
plot(model, which = 1:2)
```






```r
# check residuals against predictors
par(mfrow = c(1, 3))
plot(data$temperature, residuals(model))
plot(data$humidity, residuals(model))
plot(data$wind, residuals(model))
```



e

R^2 is 0.8127, the regression model explains 81.27% of the variation in pm25.

f

Use stepwise backward selection.

From the regression in b, we find insignificant variable *precipitation*. Remove this variable and re-estimate the model.

The coefficient in the re-estimated model2 are all significant.

The final fitted model is: $\hat{pm}25 = 97.3224 + 1.6267 temperature - 1.2698 humidity - 0.5806 wind$.

```
model.2 <- lm(pm25~temperature+humidity+wind, data=data)
summary(model.2)
```

```
##
## Call:
## lm(formula = pm25 ~ temperature + humidity + wind, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.7588  -6.4368  -0.5659   6.4006  20.2813
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  97.3234     8.9561  10.867 5.45e-15 ***
## temperature   1.6267     0.1859   8.753 8.39e-12 ***
## humidity     -1.2698     0.1165 -10.899 4.89e-15 ***
## wind         -0.5806     0.2323  -2.500   0.0156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.983 on 52 degrees of freedom
## Multiple R-squared:  0.812,  Adjusted R-squared:  0.8011
## F-statistic: 74.84 on 3 and 52 DF,  p-value: < 2.2e-16
```

g

The R^2 for final model is 0.812 and the adjusted R^2 is 0.801.

Compared with the full model, the R^2 for final model is smaller but the adjusted R^2 is larger. Because when removing a predictor, the interpretation ability of the model tend to decrease (SSExplained decrease), leading to a smaller R^2.

The adjusted R^2 value takes into account the number of predictors in the model, and penalizes models with more predictors. Therefore, after removing the insignificant variable, the model become less complex and more parsimonious, leading to a larger ajusted R^2.

Question 2

a

The design is unbalanced because group sizes are not equal.

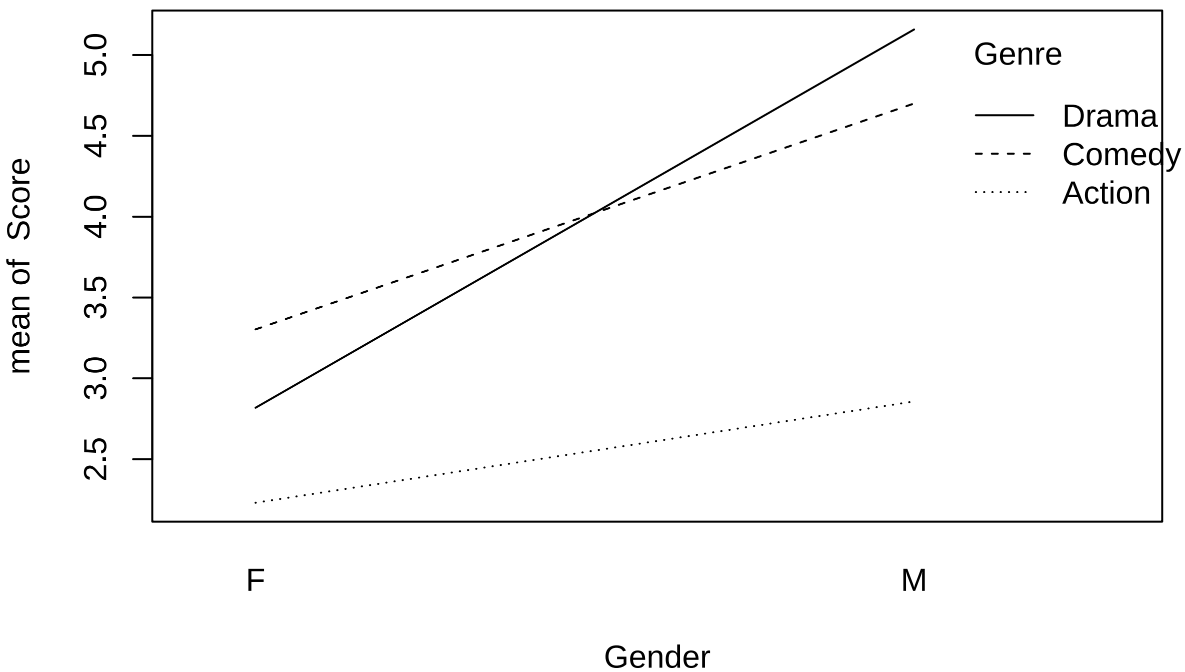
```
movie <- read.csv("data/movie.csv")
table(movie[, c("Gender", "Genre")])
```

```
##        Genre
## Gender Action Comedy Drama
##      F     39     33    22
##      M     14     10    19
```

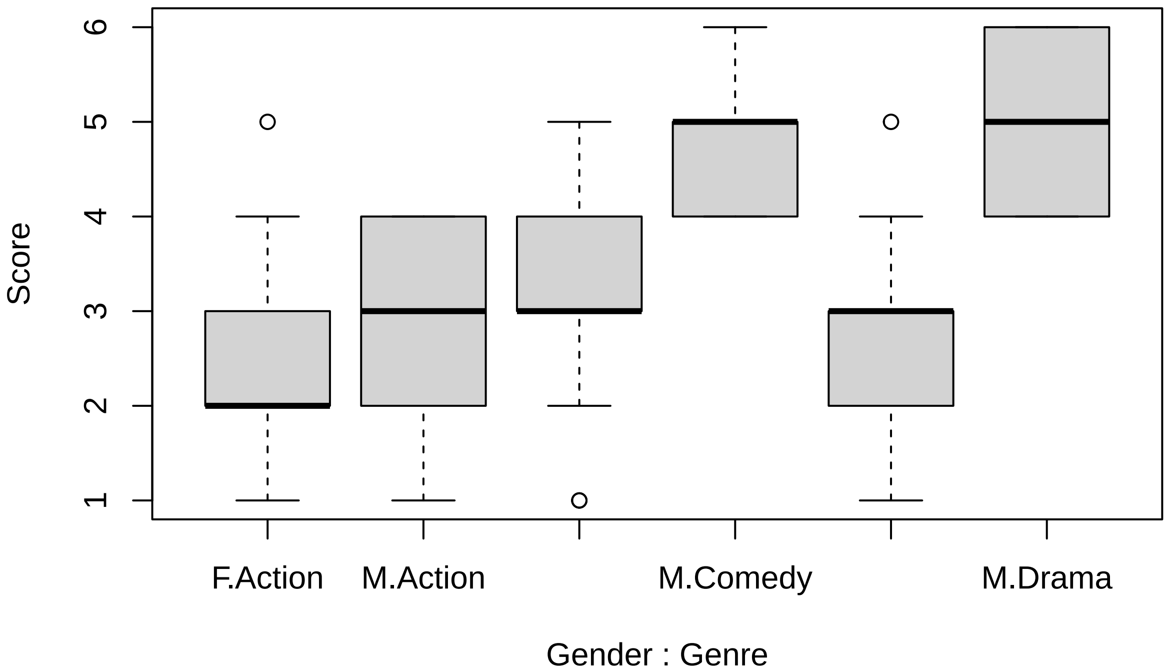
b

The interaction line plot shows that lines are not parallel, suggesting interaction. The boxplot shows variability, and possible outliers in the female movie score

```
par(mfrow = c(1,1))
with(movie, interaction.plot(Gender, Genre, Score))
```




```
boxplot(Score ~ Gender + Genre, data= movie)
```



c

The full mathematical model is:

$Score_{ijk} = \mu + \alpha_2 Gender_{i2} + \beta_2 Genre_{j2} + \beta_2 Genre_{j3} + \gamma_{22} Gender_{i2} Genre_{j2} + \gamma_{23} Gender_{i2} Genre_{j3} + \epsilon_{ijk}, \epsilon_{ijk} \stackrel{i.i.d.}{~} N(0, \sigma^2)$.

Response: Score_{ijk} = kth replicate of the treatment at ith level in Gender and jth level in Genre.

μ = overall population mean.

$Gender_{i2}$: main effect of Gender.

$Genre_{j2}, Genre_{j3}$: main effect of Genre.

$Gender_{i2} Genre_{j2}, Gender_{i2} Genre_{j3}$: interaction effects.

ϵ_{ijk}: unexplained variation for each replicated observation.

d

There are three types of tests

1. Interaction

$H_0 : \gamma_{22} = \gamma_{23} = 0$

$H_A : not both \gamma_{22} \gamma_{23} are equal to 0$

Test statistic: $F_obs = 8.4054$. P-value is 0.0004<0.05. Reject at 5% level. There is a significant interaction effect.

2. Main effect Gender $H_0 : \alpha_2 = 0$

$H_A : \alpha_2 \neq 0$

Test statistic: $F_obs = 71.807$. P-value is 0.0000<0.05. Reject at 5% level. There is a significant gender main effect.

3. Main effect Genre $H_0 : \beta_2 = \beta_3 = 0$

$H_A : not both \beta_2 and \beta_3 are equal to 0$

Test statistic: $F_obs = 25.257$. P-value is 0.0000<0.05. Reject at 5% level. There is a significant genre main effect.

```
movie.1 = lm(Score ~ Gender * Genre, data=movie)
summary(movie.1)
```

```
##
## Call:
## lm(formula = Score ~ Gender * Genre, data = movie)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3030 -0.7000 -0.2308  0.7692  2.7692
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.2308     0.1517  14.709  < 2e-16 ***
## GenderM                 0.6264     0.2951   2.123   0.0357 *
## GenreComedy             1.0723     0.2240   4.787 4.51e-06 ***
## GenreDrama              0.5874     0.2525   2.326   0.0215 *
## GenderM:GenreComedy     0.7706     0.4516   1.706   0.0903 .
## GenderM:GenreDrama      1.7133     0.4184   4.095 7.34e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9471 on 131 degrees of freedom
## Multiple R-squared:  0.5383, Adjusted R-squared:  0.5207
## F-statistic: 30.55 on 5 and 131 DF,  p-value: < 2.2e-16
```

```
anova(movie.1)
```

```
## Analysis of Variance Table
##
## Response: Score
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## Gender         1  71.583  71.583 79.8038 3.277e-15 ***
## Genre          2  50.357  25.178 28.0698 7.152e-11 ***
## Gender:Genre   2  15.079   7.540  8.4054 0.0003677 ***
## Residuals    131 117.506   0.897
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
movie.2 = lm(Score ~ Gender * Genre, data=movie)
movie.2 = update(movie.1, . ~ . - Gender:Genre) # OR update by removing
summary(movie.2)
```

```
##
## Call:
## lm(formula = Score ~ Gender + Genre, data = movie)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.46788 -0.68389 -0.01153  0.71078  2.98847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0115     0.1459  13.790  < 2e-16 ***
## GenderM       1.4563     0.1881   7.742 2.20e-12 ***
## GenreComedy   1.2777     0.2050   6.232 5.64e-09 ***
## GenreDrama    1.2160     0.2110   5.763 5.49e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9984 on 133 degrees of freedom
## Multiple R-squared:  0.4791, Adjusted R-squared:  0.4673
## F-statistic: 40.77 on 3 and 133 DF,  p-value: < 2.2e-16
```

```r
anova(movie.2)
```

```
## Analysis of Variance Table
##
## Response: Score
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Gender      1  71.583  71.583  71.807 3.914e-14 ***
## Genre       2  50.357  25.178  25.257 5.036e-10 ***
## Residuals 133 132.585   0.997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

e

From the *movie.1* regression result in d), the slope coefficient for gender, genre and their interaction term all have a significant positive effect on the movie score.

The coefficient for GenderM:GenreDrama is 1.71, indicating that the drama movie score of female viewers are 1.71 higher than other combination of gender and genre. To maximise the brand recognition, they should place more drama to female viewers.