Stat2170 Assignment

Zhi Wei Alphonsus Chua (46253009)

14/06/2023

Contents

Question 1

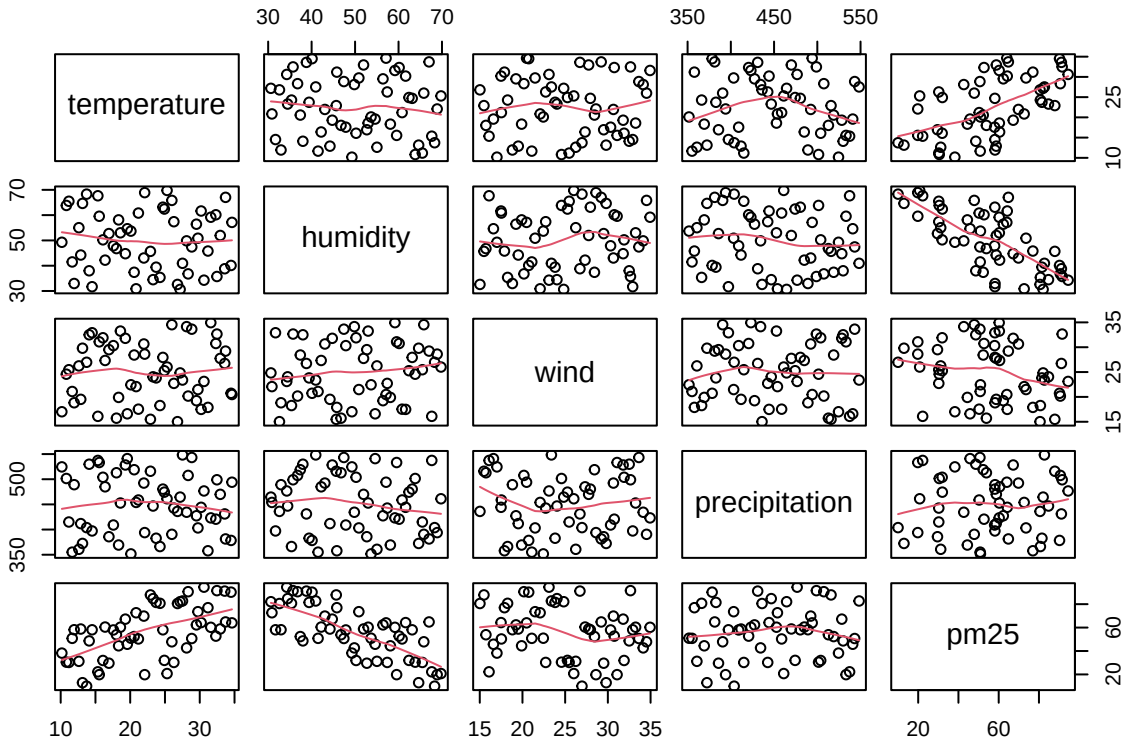
- **Loading the data**

```
pm25  = read.csv ("pm25.csv", header =TRUE)
head(pm25)
#>   temperature humidity  wind precipitation  pm25
#> 1       24.91    62.40 25.27        474.34 31.99
#> 2       26.28    57.40 22.72        442.67 30.24
#> 3       16.94    52.70 29.51        385.76 29.59
#> 4       34.50    40.10 20.73        378.40 90.23
#> 5       29.74    50.87 21.48        403.26 73.57
#> 6       12.89    44.22 19.51        411.81 58.77
```

Part A

- **Produce a plot and correlation matrix of the data. Comment on possible relationships**

```
pairs(pm25, panel = panel.smooth)
```




```
cor(pm25)
#>                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
```

The correlation between "pm25" and "temperature" is 0.571, signifying that they are positively correlated. A higher temperature corresponds to a higher PM2.5 concentration. There are also weak negative correlations between "humidity" and "temperature" as well as "precipitation" and "temperature" at -0.073 and -0.051. This means that a higher precipitation or humidity would translate to a very slight decrease in temperature on average.

The correlation between "pm25" and "humidity" is -0.720, signifying that they are negatively correlated. A higher humidity corresponds to a lower PM2.5 concentration.

The correlation between "pm25" and "wind" is -0.219, this means that there is a weak negative correlation. A higher wind speed weakly corresponds to a lower PM2.5 concentration. However, there is also very little to no positive correlation between "temperature" and "wind" at 0.029, and weak positive correlation between "humidity" and "wind" at 0.124. Hence an increase in "temperature" or "humidity" may mean that there is small to no observable change to "wind".

There is a very weak positive correlation between "precipitation" and "pm25" at 0.038. Precipitation is only very slightly related to PM2.5 concentration levels.

Part B

- **Fit a model using all predictors to explain the pm25 response**

```
pm25.lm = lm(pm25~temperature+humidity+wind+precipitation, data = pm25)

summary(pm25.lm)
#>
#> Call:
#> lm(formula = pm25 ~ temperature + humidity + wind + precipitation,
#>     data = pm25)
#>
#> 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
```

Temperature and *humidity* are significant predictors as they have a P-Value of "*1.46e-11*" and "*9.49e-15*" respectively. *Wind* is a less significant predictor as its P-Value is *0.0165*. Precipitation is an insignificant predictor as its P-Value is *0.644* and greater than 0.05.

In regards to energy response to all predictors, an increase in *temperature* results in a *1.621* increase in *PM2.5* concentration. On the other hand, *PM2.5* concentration decreases by *1.277* and *0.580* for each incremental unit in *humidity* and *wind* respectively. On the other hand, with every increase in rooms, there is a reduction of 5.383 in energy. These observations were made while holding all other parameters constant.

- **The impact of humidity on PM2.5 concentration**

```
summary(pm25.lm)$coefficients
#>                  Estimate  Std. Error     t value     Pr(>|t|)
#> (Intercept)   102.72258771 14.71952825   6.9786603 5.881953e-09
#> temperature     1.62141831  0.18762464   8.6418198 1.463129e-11
#> humidity       -1.27742262  0.11854373 -10.7759612 9.490343e-15
#> wind           -0.58015926  0.23405331  -2.4787484 1.653279e-02
#> precipitation  -0.01090918  0.02349567  -0.4643059 6.444046e-01
```

```
qt(0.05/2, 51, lower.tail = FALSE)
#> [1] 2.007584
```

Terms of interest: * $\beta_{humidity} = 2331.116239$ * $s.e.(\beta_{humidity}) = 250.918960$ * $t_{51,1-0.05/2} = 2.007584$

For each percentage increase in mean relative humidity, we anticipate a change in PM2.5 concentration of,

$\beta_{humidity} \pm t_{51,1-0.05/2} s.e.(\beta_{humidity}) = -1.27742262 \pm 2.007584 \times 0.11854373 =$ (-1.039436, -1.515409)

We are 95% certain that every additional percentage increase in mean relative humidity will result in a negative change in PM2.5 concentration between *-1.039436* and *-1.515409*.

Part C

- **Conduct F-test for overall regression**
- **Full Mathematical multiple regression model**

```
coefficients(pm25.lm)
#>   (Intercept)   temperature      humidity          wind precipitation
#>  102.72258771    1.62141831   -1.27742262   -0.58015926   -0.01090918
```

$\hat{pm25} = \beta_0 + \beta_1$ temperature + β_2 humidity + β_3 wind + β_4 precipitation + ϵ

$\hat{pm25}$ = 102.72258771 + 1.62141831 temperature - 1.27742262 humidity - 0.58015926 wind - 0.01090918 precipitation

- **Hypotheses for the Overall ANOVA test of multiple regression**

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$
$H_1 : \beta_i \neq 0$ (at least one β_i parameter is not zero)

- **ANOVA table for the overall multiple regression model**

```
anova(pm25.lm)
#> 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
```

Full RegSS $= RegSS_{temperature} + RegSS_{humidity|temperature} + RegSS_{wind|temperature,humidity} + RegSS_{precipitation|temperature,hu}$

Full RegSS $= 9014.4 + 12739.7 + 622.6 + 21.8 = 22398.5$

RegM.S.$= \frac{Reg.S.S}{k} = \frac{22398.5}{4} = 5599.625$

- **Compute the F statistic**

$Fobs = \frac{RegM.S}{ResM.S} = \frac{5599.625}{101.2} = 55.33226$

- **Null distribution for the test statistic**

$\epsilon \sim N(0, \sigma^2)$

- **Compute the P-Value**

```
pf(55.33226 ,df1= 4,df2 = 51, lower.tail =FALSE)
#> [1] 6.096688e-18
```

$P(f_{4,51} \geq 55.33226) = 6.096688e - 18$

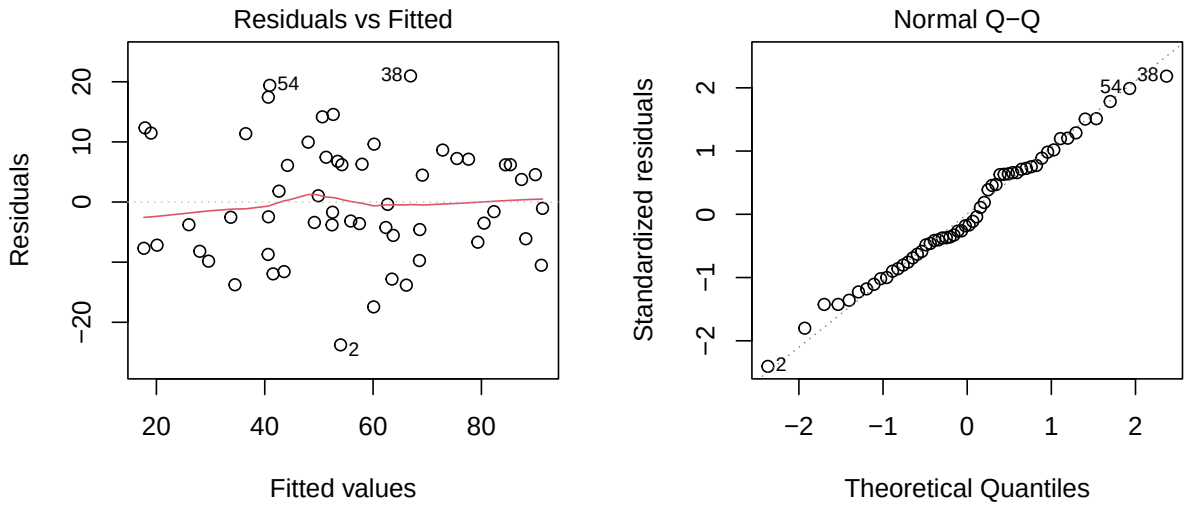
- **Conclusion**

At 5% significance level, reject the null hypothesis as P-Value is lesser than 0.05 or α. Therefore, we may conclude that there is a significant linear relationship between pm25 response and at least one of the four predictor variables.

Part D

- **Validate the full model**

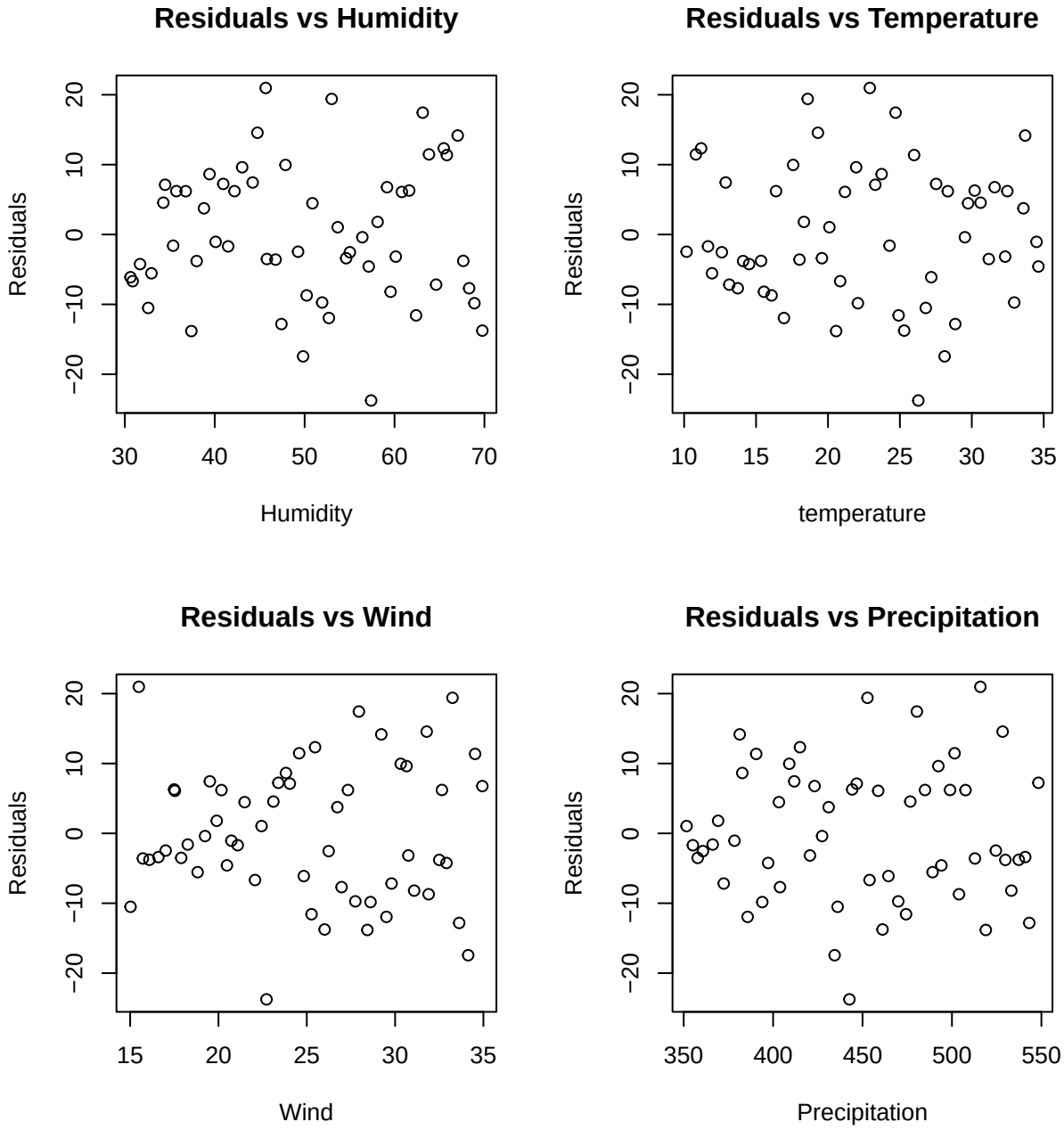
```
par(mfrow = c(1,2))
plot(pm25.lm , which = 1:2)
```



1. The Residual Normal Q-Q plot is close to linear with slight curvature, validating the assumption that any unexplained variation follows a normal distribution.

2. Data is relatively evenly distributed with only some data distrbuted towards the upper half of the graph and hence, meets the linearity assumption.

```
par(mfrow = c(2,2))
plot(pm25$humidity, pm25.lm$residuals, main = "Residuals vs Humidity",
     xlab = "Humidity", ylab = "Residuals")
plot(pm25$temperature, pm25.lm$residuals, main = "Residuals vs Temperature",
     xlab = "temperature", ylab = "Residuals")
plot(pm25$wind, pm25.lm$residuals, main = "Residuals vs Wind",
     xlab = "Wind", ylab = "Residuals")
plot(pm25$precipitation, pm25.lm$residuals, main = "Residuals vs Precipitation",
     xlab = "Precipitation", ylab = "Residuals")
```



The majority of residual vs predictor plots show even and random distribution of points along the horizontal axis. This suggests that the linear model is acceptable. This may suggest that all predictors have a relationship with energy.

Overall, the full model is adequate as it satisfies all assumptions. However, it can be advised to remove certain features, that are minor and may not affect on the response variable significantly so that a parsimonious model can be constructed.

Part E

- **Find the** R^2

```
summary(pm25.lm)$r.squared
#> [1] 0.8127448
```

$R^2 = \frac{TotalS.S - ResidualsS.S}{TotalS.S.} = \frac{(22398.5+5160.6)-5160.6}{22398.5+5160.6} = 0.8127448$

81.2% variance can explain the variation in response variable(PM2.5) around its predictor variables.

Part F

- **The best multiple regression model that describes the data**

Utilising Stepwise Backward Selection model, remove "precipitation" as it has the lowest probability of having a significant relationship with the response variable(PM2.5) because it has the highest P-Value in the t-test. Removing "precipitation" will result in changing the P-Value.

```
pm25.lm2 = lm (pm25~temperature+humidity+wind, data = pm25)
summary(pm25.lm2)
#>
#> Call:
#> lm(formula = pm25 ~ temperature + humidity + wind, data = pm25)
#>
#> 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
```

- **Change in P-Value**

```
pf(74.84 ,df1= 3,df2 = 52, lower.tail =FALSE)
#> [1] 7.143951e-19
```

There is a decrease in P-Value from 6.079192e-18 to 7.143951e-19 after removing "Precipitation" prediction variable from the model.

- **The final fitted regression model**
  $\hat{pm25} = 97.323 + 1.627temperature - 1.270humidity - 0.581wind$

Part G

Based on the final model, the R^2 value decreased from 0.8127 in the full model to 0.812. This is because the number of predictor variables was only decreased by one from four to three. However, introducing new predictor variables will lead to a greater R^2 value. This is because of its characteristic, which suggests that adding more variables will raise the value even if only marginally, regardless of significance to the model. In conclusion, R^2 does not contribute to the indication of an insignificant independent variable to the regression model. By looking at high R^2 predictors, this may lead to a large amount of insignificant predictors being included in the model.

In contrast to R^2, Adjusted R^2 will decrease when insignificant variables are included in the regression. The result is a more accurate and reliable analysis. In this scenario, the adjusted R-squared rose from 0.7981 initially to 0.8011 in the final fitted regression model. This indicates that the new model is parsimonious after omitting the "Precipitation" variable which does not match the model.

Finally, the changes in R^2 and adjusted R^2 differ since R^2 increases as the number of predictors increases, while adjusted R^2 decreases as the predictors are less significant. Hence, adjusted R^2 is a more reliable and robust model evaluator for determining the predictor variable's relevance to the base variable.

Question 2

- **Loading the data**

```
movie  = read.csv ("movie.csv", header =TRUE)
head(movie)
#>   Gender  Genre Score
#> 1      F Action     3
#> 2      F Action     3
#> 3      F Action     2
#> 4      F Action     1
#> 5      F Action     4
#> 6      F Action     2
```

Part A

- **Is the design balanced or unbalanced?**

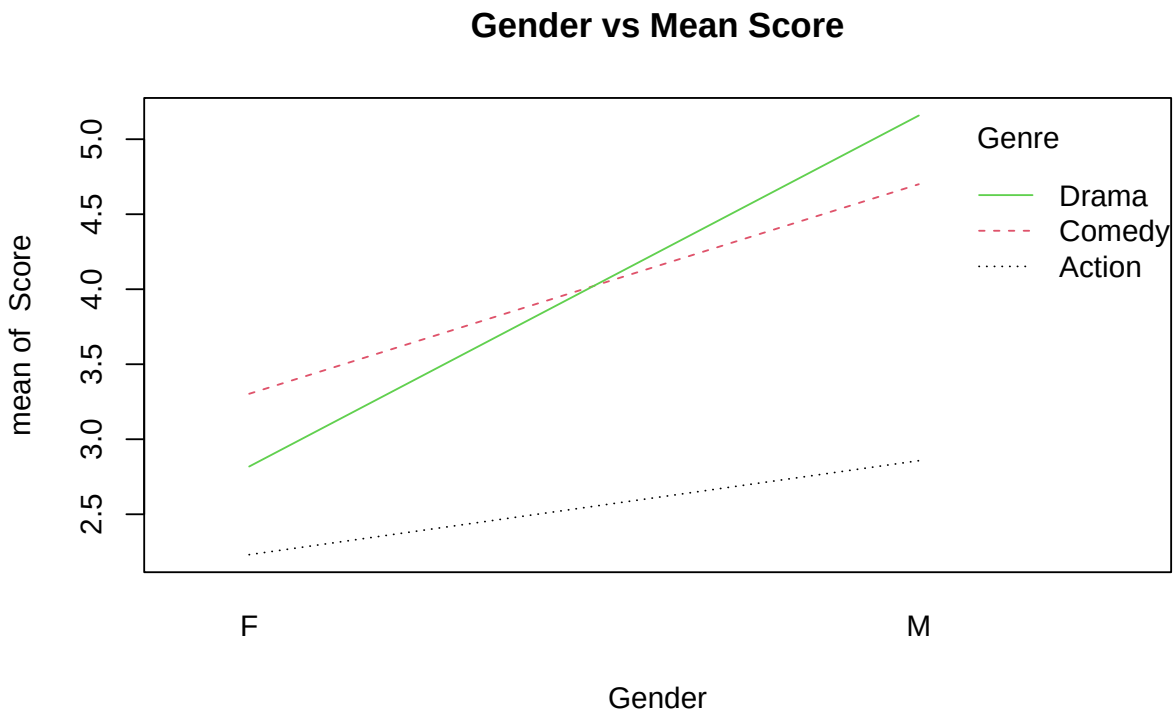
```
table(movie[1:2])
#>        Genre
#> Gender Action Comedy Drama
#>      F     39     33    22
#>      M     14     10    19
```

The design is unbalanced as the factor combinations of Gender and Genre are unequal.

Part B

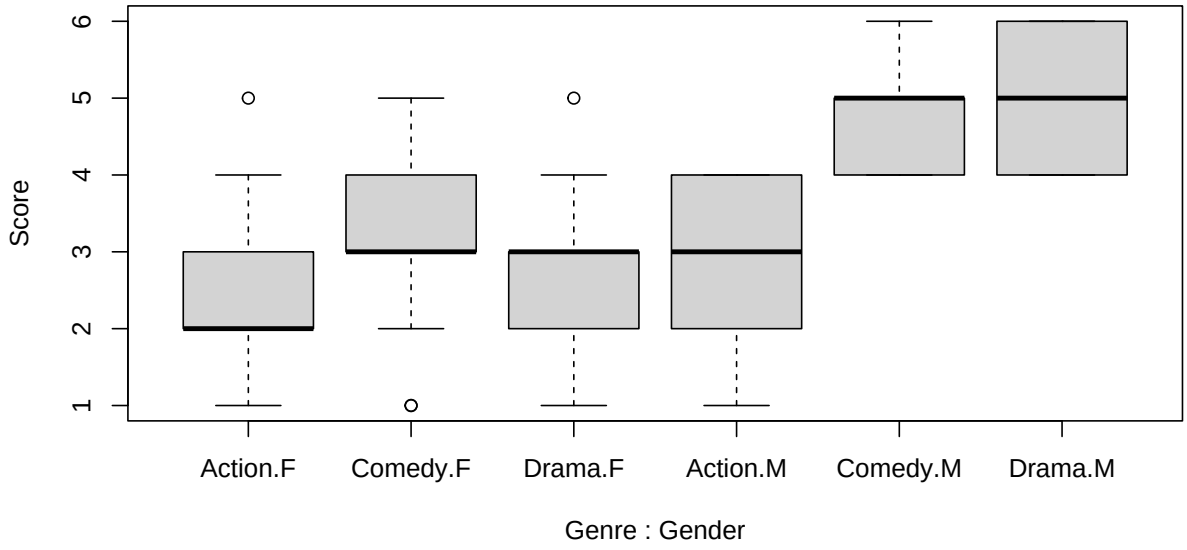
- **Construct 2 different preliminary graphs**

```
with(movie,interaction.plot(Gender,Genre,Score, col = 1:3, main = "Gender vs Mean Score"))
```



The plot shows possible interaction between response and independent variables as there is intersection between Drama and Comedy graphs. However the sample sizes for some of the factor combinations are small and so it may not be accurate and cannot be relied on to produce a valid conclusion from the graphs.

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



All groups seem to have a normal spread, distribution among the first, second and third groups is comparable, indicating that variation across these three are similar or equal.

Distribution of the fourth, fifth and sixth groups are also comparable, however this may not be accurate due to the small sample sizes of each effect.

Part C

- **Write down the full mathematical model**

$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \epsilon_{ijj} \sim N(0, \sigma^2)$

The terms and coefficients above represent:
* Y_{ijk} : Score response
* μ : Overall population mean
* α_i : Gender effect * β_j : Genre effect * γ_{ij} : Interaction effect between Gender and Genre
* ϵ_{ijk} : Unexplained variation for each replicated observation
* $\epsilon_{ijk} \sim N(0, \sigma^2)$

Part D

- **Null and alternative hypothesis**
- H_0: no interaction | $\gamma_{ij} = 0$ for all i, j.
- Gender has the same influence regardless of Genre, and vice versa.
- H_1: there is interaction | not all $\gamma_{ij} = 0$
- Gender has varying effects depending on the level of Genre, and vice versa
- **Fitting the model**

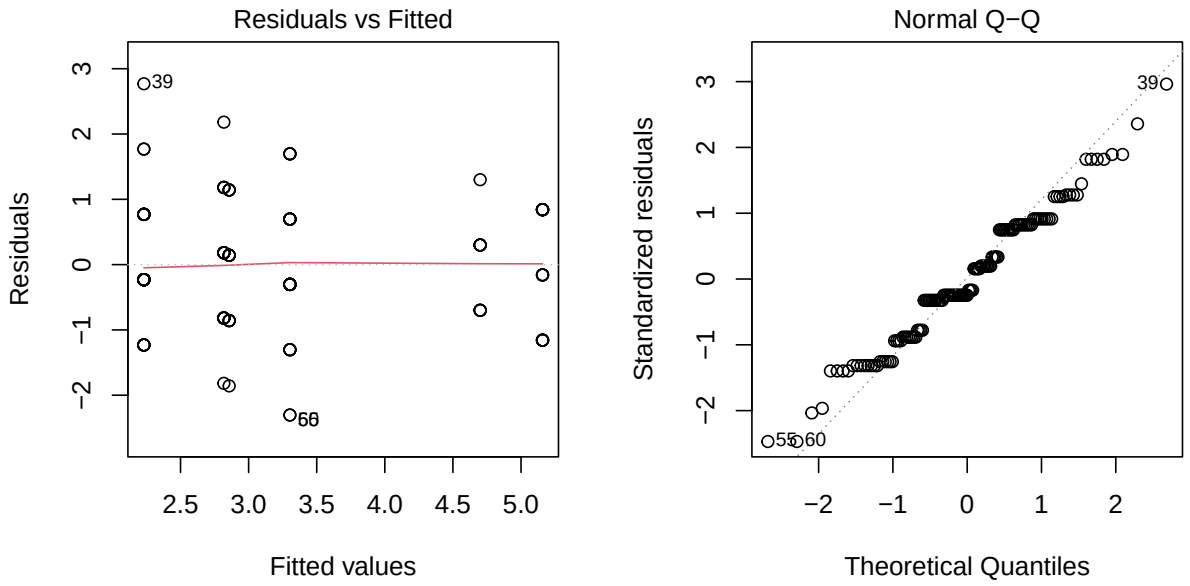
```
movie.lm = lm (Score ~ Gender * Genre ,  data = movie)
anova(movie.lm)
#> Analysis of Variance Table
#>
#> Response: Score
#>               Df  Sum Sq Mean Sq F value    Pr(>F)
#> Gender         1  71.583  71.583 79.8038 3.277e-15 ***
#> Genre          2  50.357  25.178 28.0698 7.152e-11 ***
#> Gender:Genre   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
```

The order needs to be reversed as the design is unbalanced.

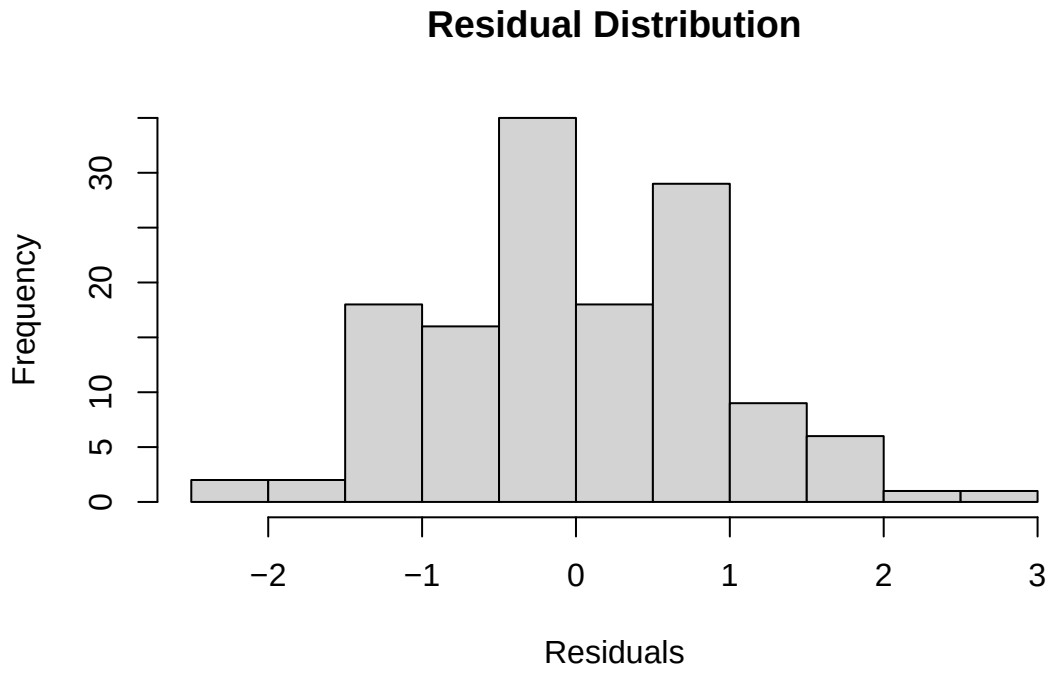
```
movie.lm = lm (Score ~  Genre *Gender ,  data = movie)
anova(movie.lm)
#> Analysis of Variance Table
#>
#> Response: Score
#>               Df  Sum Sq Mean Sq F value    Pr(>F)
#> Genre          2  62.190  31.095 34.6658 8.254e-13 ***
#> Gender         1  59.750  59.750 66.6117 2.388e-13 ***
#> Genre:Gender   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
```

- **Checking assumptions**

```
par(mfrow = c(1,2))
plot(movie.lm, which = 1:2, main = "")
```




```
hist(resid(movie.lm), main = "Residual Distribution", xlab = "Residuals")
```



The diagnostic graphs appear to support the model. The residual plot against fitted values exhibits no pattern or trends however there seems to be a gap in data in the center right section of the graph and the variability between effects can be observed to be constant.

The normal Q-Q plot does not look to be very linear as there is deviation from normality near the ends points and middle of the graph. A log transformation is suggested.

The histogram graph indicates that residuals are very slightly skewed to the right.

Part E

- **Discuss practical implications of your findings for the business**

Based on the results in the above, a business that aims to maximise brand recognition from product placement can expect to have significant impact on the response variable score by altering the Gender and Genre effects to optimise their brand recognition. The brand recall score for males in the effect drama genre may be higher than for females.