Homework 10

The purpose of this homework is to gain practice running analysis of variance hypothesis tests. Please fill in the appropriate code and write answers to all questions in the answer sections, then submit a compiled pdf with your answers through Gradescope by 11:59pm on Sunday November 29th.

As always, if you need help with any of the homework assignments, please attend the TA office hours which are listed on Canvas and/or ask questions on Piazza. Also, if you have completed the homework, please help others out by answering questions on Piazza.

Part 1: One-way Analysis of Variance (ANOVA) for testing more than two means

In class 9 we used dplyr to examine which genres of movies critics and regular audiences enjoy the most based on the ratings of over 500 movies from the Rotten Tomatoes website. In the analyses you did in class 9 you only created descriptive statistics to explore whether different genres of movies had different scores.

Let's now use a one-way ANOVA to test whether there are indeed differences in the mean of critics scores for different genres of movies. In particular, in this first set of questions you will use the movie data to assess whether there is a difference in mean critic scores for the genres "Action & Adventure", "Comedy" and "Drama".

Part 1.1 (3 points): Start your hypothesis test with step 1, by stating the null and alternative hypotheses in symbols and words. Also state the alpha level that is most commonly used.

Answers

In words

Null hypothesis: The mean of scores is the same score for all genres.

Alternative hypothesis: There is a difference between the mean scores of Action & Adventure, Comedy, and Drama movies.

In symbols

$$H_0 : \mu_{action} = \mu_{comedy} = \mu_{drama}$$
$$H_A : \mu_i \neq \mu_j$$

for some pair of i, j where i and j are {action, comedy, drama}

The significance level

$$\alpha = 0.05$$

Part 1.2a (5 points):

Now use dplyr to create a data frame called `films`. This data set should be derived from the movies data set and should have the following properties:

1. The type of movie (i.e., `title_type`) should only be “Feature Films”.
2. The only variables in the data frame should be: `title`, `genre`, `runtime`, `mpaa_rating`, `critics_score`, and `audience_score`.
3. The only genres should be “Action & Adventure”, “Comedy”, and “Drama”.
4. The only MPAA ratings should be: “PG”, “PG-13” and “R”.
5. Be sure to apply the `droplevels()` function to the final data frame to remove all levels of categorical variables that you are not using.

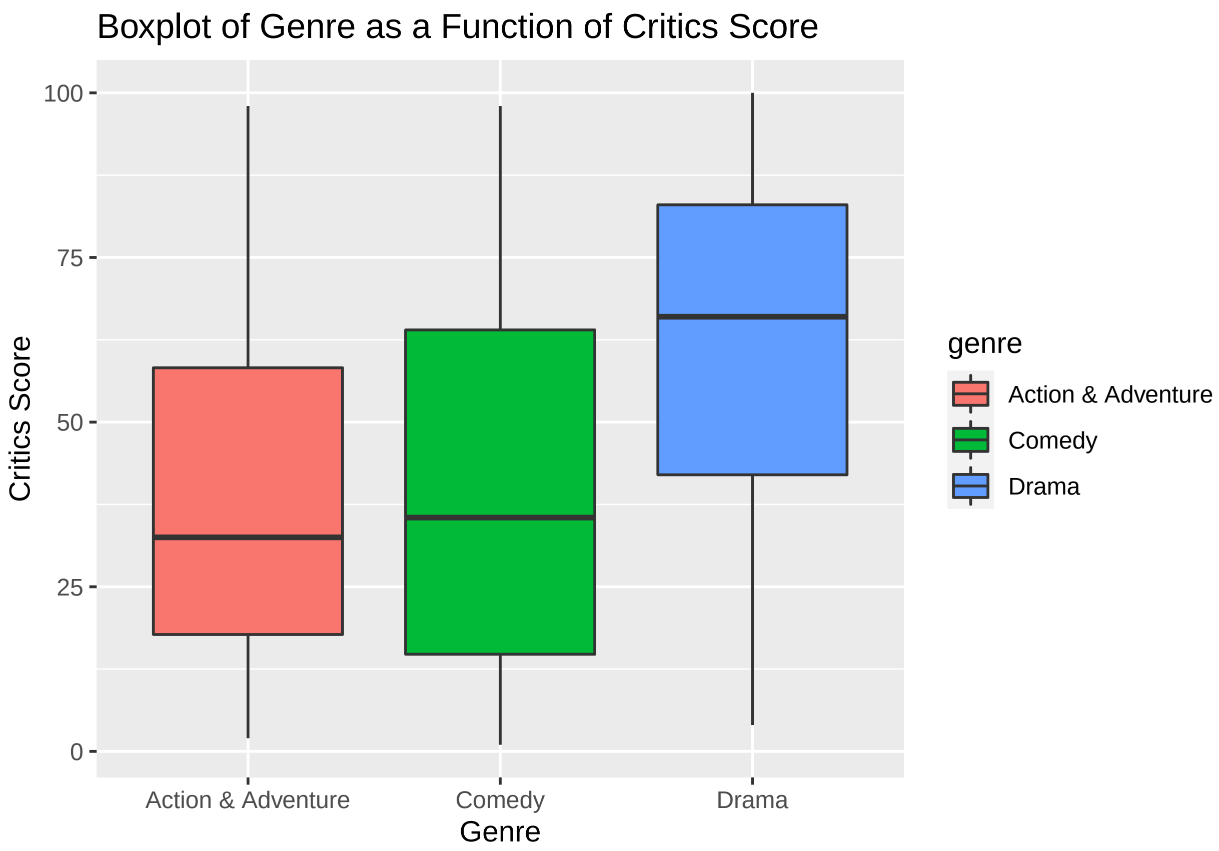
If you have created this data frame correctly, it should contain 433 rows and 6 columns.

Once you have created this data frame, create a box plot of the critics scores as a function of genre. Based on this plot, do you believe there is a difference in the average scores critics give to these different movie genres?

```
# load the data
load(’movies.Rdata’)

films<-
  movies%>%
  filter(title_type == "Feature Film")%>%
  select(title, genre, runtime, mpaa_rating, critics_score, audience_score)%>%
  filter(genre %in% c(’Action & Adventure’, ’Comedy’, ’Drama’))%>%
  filter(mpaa_rating %in% c(’PG’, ’PG-13’, ’R’))%>%
  droplevels()



  dim(films)
```

```
## [1] 433   6
```

```
films%>%
  ggplot(aes(genre, critics_score, fill = genre))+
  geom_boxplot()+
  xlab("Genre")+
  ylab("Critics Score") + ggtitle("Boxplot of Genre as a Function of Critics Score")
```



Answer

Based on this plot, since the Drama average critic score (the median of the blue boxplot) is notably greater than the median critics score for Action & Adventure and Comedy, with strengthens our intuition that genre influences the mean critics score.

Part 1.2b (15 points): In order for the F-distribution to be a valid null distribution for our F-statistic, two conditions must hold. The two conditions that need to be met are: 1) the variances (or standard deviations) in each group must be approximately the same and 2) the data from each group must be relatively normal (also, as with almost all hypothesis tests, it needs to be the case that the data points are independent, but we will assume that is the case here). One can check these conditions either at the start of the analysis, or at the end before one draws a final conclusion. Let's check these conditions now!

We can check the first condition, as to whether the variances within each group are approximately the same, by comparing the standard deviations between the groups. As long as the largest standard deviation is not twice as big as the smallest standard deviation this condition is met. Please use dplyr to check this condition and in the answer section below describe whether this condition appears to be met.

We can check the second condition, as to whether the data is relatively normal in each group, by visually inspecting the data. This can This can be done in a several ways including: 1) creating a histogram of the residuals (differences between each point and its group mean), and 2) creating a Quantile-Quantile plot between the the residuals and a normal distribution.

To create these plots it will be useful to first create a data frame called `films2` that is derived from the `films` data frame but also has two additional variables. The first variable is called `mean_genre_scores` and should have the average critics score for each genre, and the second variable is called `critic_genre_residuals` and should contain the difference of each critics score from the mean genre score. Using the `group_by()` function in conjunction with the `mutate()` function will be helpful for doing this.

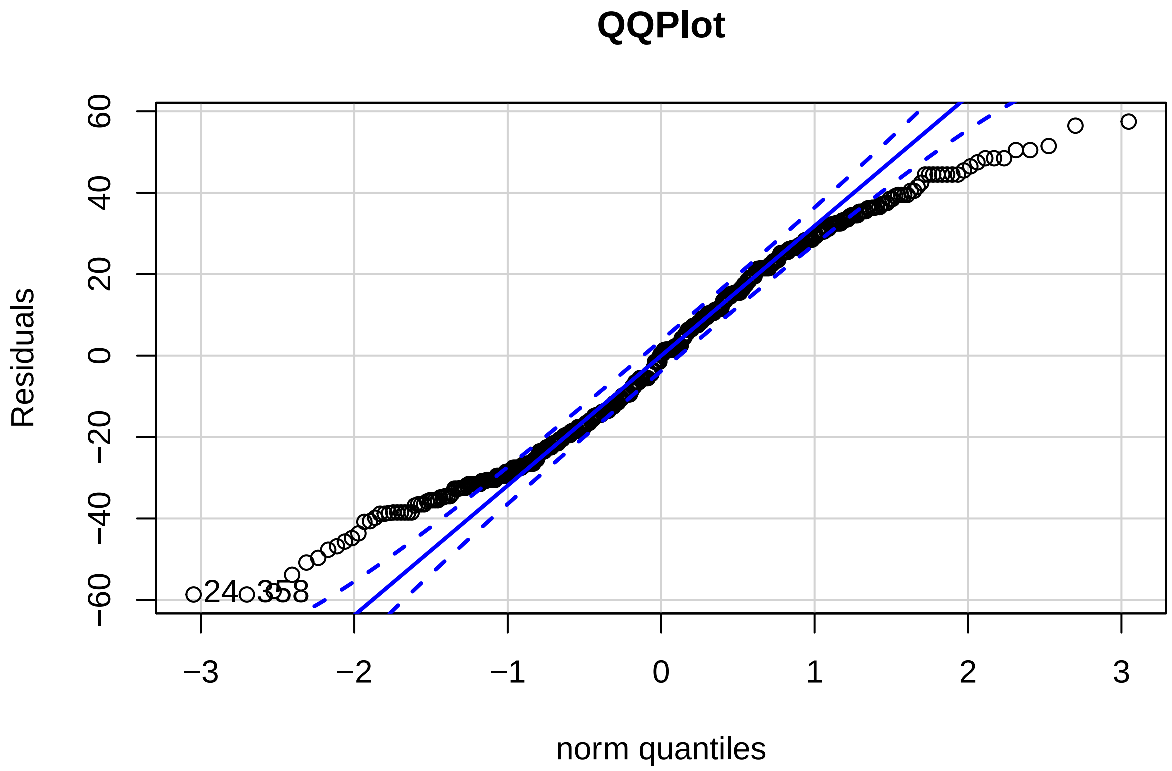
Once you have created the `films2` data frame, create the Q-Q plot using either the `car::qqPlot()` function or the `qqnorm()` function. Also use either ggplot or base R to create a histogram of the residuals (you could also plot a normal distribution on top of this histogram but this is not required). Based on this data, do the residuals appear normally distributed?

Note: as discussed in class, the ANOVA is fairly robust to departures from homoskedasticity and normality. However, if one is concerned about these conditions not being met, one could always run a permutation test instead (see class 15).

```
#first condition - variances -  check st. dev. of pts within each group
films%>%
  group_by(genre)%>%
  summarize(std = sd(critics_score), .groups = 'drop')
```

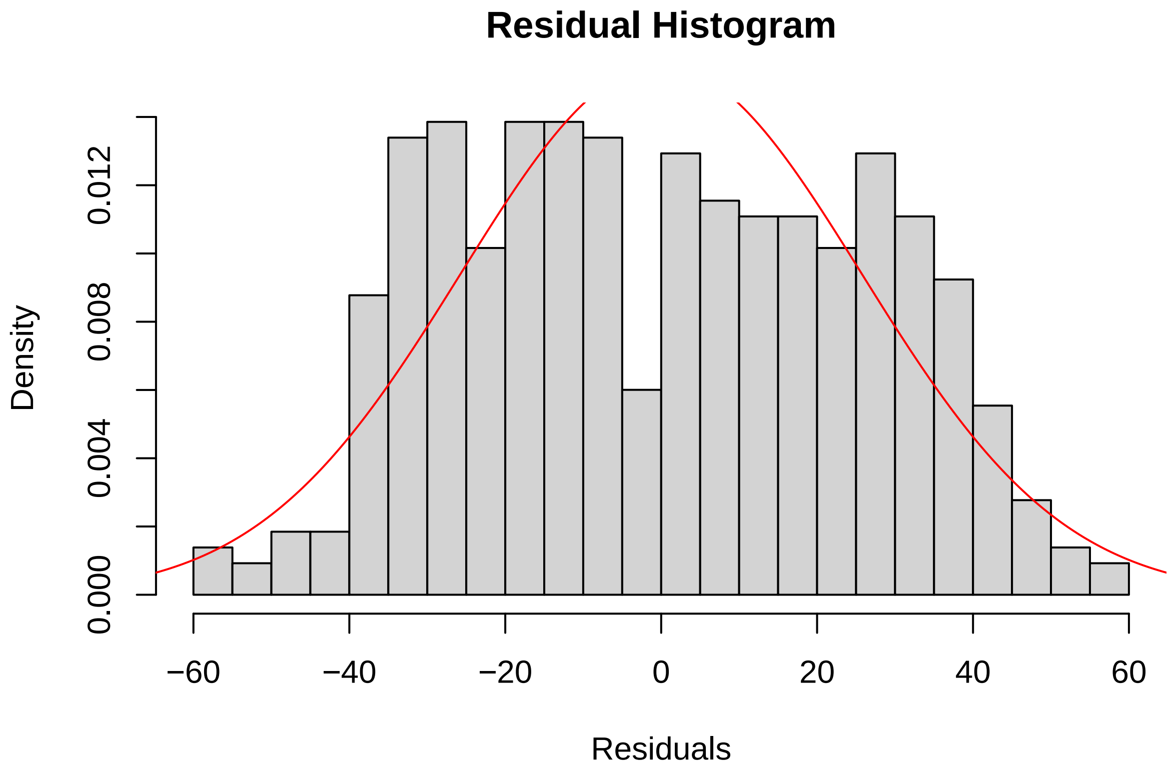
```
## # A tibble: 3 x 2
##   genre                std
##   <fct>              <dbl>
## 1 Action & Adventure  26.8
## 2 Comedy              27.3
## 3 Drama               25.1
```

```
  #second condition - normality

      #create a df that has the residuals = difference of critics score and mean critics score of that

films2<-
films%>%
  group_by(genre, .groups = 'drop')%>%
  mutate(mean_genre_score = mean(critics_score))%>%
  mutate(residuals = mean_genre_score - critics_score)


car:: qqPlot(films2$residuals,ylab = "Residuals", main = "QQPlot")
```



```
## [1]  24 358
```

```
hist(films2$residuals,
     xlab = "Residuals",
     ylab = "Density",
     main = "Residual Histogram",
     breaks = 40, freq = FALSE)
x<-seq(-100, 100, by = 0.1)
y<-dnorm(x, mean(films2$residuals), sd(films2$residuals))
points(x, y, type = 'l', col = 'red')
```



Answer

Our first assumption was the assumption that there are equal variances. This assumption is met because none of the standard deviations are double any of the other standard deviations. Our second assumption of normality is met because although the graph is not very normal, we are going to proceed with the ANOVA since ANOVA tends to be a really robust method. A good task later would be to run a permutation test later since a permutation test does not require normality.

Part 1.2c (15 points): Next let's use dplyr to create an ANOVA table. ANOVA tables have the following form:

Source	df	Sum of Sq.	Mean Square	F-stat	p-value
Groups	K - 1	SSG	MSG = SSG/(K-1)	MSG/MSE	
Error	N - K	SSE	MSE = SSE/(N-K)		
Total	N - 1	SSTotal			

Where:

$SSG = \sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2$ $SSE = \sum_{i=1}^{K} \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$ $SST = \sum_{i=1}^{K} \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{tot})^2$

Please use dplyr to fill in the values in the ANOVA table using the `films2` data frame (using the `films2` will be easier than using the `films` data frame). Be sure to print out all the values as you compute them in the R chunk below to "show your work" (i.e., print out the dfs, SSG, SSE, MSG, etc.). You will fill in the p-value in this table later during part 1.5 of this homework.

Note: You can use the `lm()` `aov` or `anova()` to check your answers, but **you are not allowed to use these functions** to compute the F-statistic; i.e., you need to use just the `films2` data frame and dplyr to fill in the ANOVA table and to compute the F-statistic.

```
#calculate stat - check the variance across groups / variance within groups = if this is large, we can

(N <-nrow(films))
```

```
## [1] 433
```

```
(K<-length(unique(films$genre)))
```

```
## [1] 3
```

```
df_groups<-K-1
df_error<-N-K
df_total<-N-1



(SSG<-sum((films2$mean_genre_score - mean(films2$critics_score))^2))
```

```
## [1] 45064.59
```

```
(SSE<-sum(films2$residuals^2))
```

```
## [1] 285699.7
```

```
(SSTotal<-sum((films2$critics_score - mean(films2$critics_score))^2))
```

```
## [1] 330764.3
```

```
(MSG<-SSG/df_groups)
```

```
## [1] 22532.29
```

```
(MSE<-SSE/df_error)
```

```
## [1] 664.4179
```

```
(F_stat<-MSG/MSE)
```

```
## [1] 33.91283
```

```
anova(lm(critics_score~genre, data = films))
```

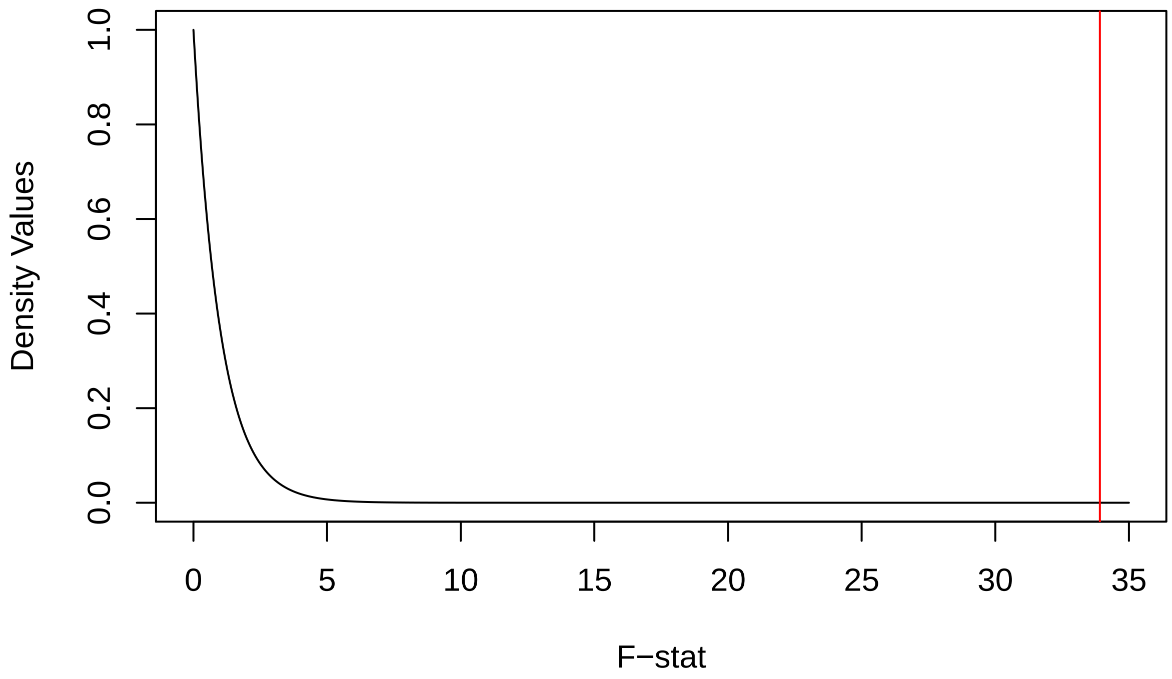
```
## Analysis of Variance Table
##
## Response: critics_score
##            Df Sum Sq Mean Sq F value              Pr(>F)
## genre       2  45065 22532.3  33.913 0.00000000000002109 ***
## Residuals 430 285700   664.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

Source	df	Sum of Sq.	Mean Square	F-stat	p-value
Groups	2	45064.59	22532.29	33.913	0.00000000000002109128
Error	430	285699.7	664.4179		
Total	432	330764.3			

Exercise 1.3 (6 points): Now it's time to do step 3 of the hypothesis test by plotting the null distribution. To create the appropriate F-distribution, use the degrees of freedom you calculated in part 1.2c. Then use either base R graphics or ggplot to plot the F-distribution density function, and add the observed statistic as a red vertical line to the plot. From looking at this null distribution, what do you think the p-value is?

```
x_vals<-seq(0,35, by = 0.01)
y_vals<-df(x_vals, df_groups, df_error)
plot(x_vals, y_vals, type = "l", xlab = "F-stat", ylab = "Density Values")

abline(v = F_stat, col = "red")
```



Answer

The p-value seems to be less than 0.01 (fairly close to 0), at around 0.00000000000001 since the area to the right of the red line is close to 0. In other words, the F-distribution is asymptotic at y = 0 so the area to the right of the red line will be trivial so the p-value will be very small.

Part 1.4 (4 points): Now do step 4 of hypothesis testing by calculating the p-value using the `pf()` function. Report what the p-value is (and make sure you look at the correct tail). Is this close to what you estimated by looking at the null distribution above? Also, fill in the p-value in the ANOVA table in part 1.2c above.

```r
pf(F_stat, df_groups, df_error, lower.tail = FALSE)
```

```
## [1] 0.00000000000002109128
```

Answers

The p-value is 0.00000000000002109128. This is close to what my predicted p-value was (close to 0 /0.00000000000001).

Part 1.5 (3 points): Now complete step 5 of hypothesis testing by making a judgment. Are you able to reject the null hypothesis? What do you conclude?

Answer We reject the null hypothesis since the p value is 0.00000000000002109128, which is substantially less than

$$\alpha = 0.05$$

. We can conclude that the mean of scores is the same for all genres.

Exercise 1.6 (5 points): As we have discussed, we can use R's `lm()` function to run an ANOVA. Running an ANOVA and creating table requires two steps:

1. We must fit a model using the syntax: `fit <- lm(response_variable ~ categorical_predictor, data = my_data)`
2. We can then print an ANOVA table using `anova(fit)`

Please use the `lm()` function to fit a model that predicts critics score from the different genres of movies. Save the model that you fit to an object called `fit_genre`. Then use the `anova()` function to see whether the results match the results you had in parts 1.1 to 1.5 above. Report below whether the results match.

```
fit<-lm(critics_score~genre, data = films)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: critics_score
##            Df Sum Sq Mean Sq F value              Pr(>F)
## genre       2  45065 22532.3  33.913 0.00000000000002109 ***
## Residuals 430 285700   664.4
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

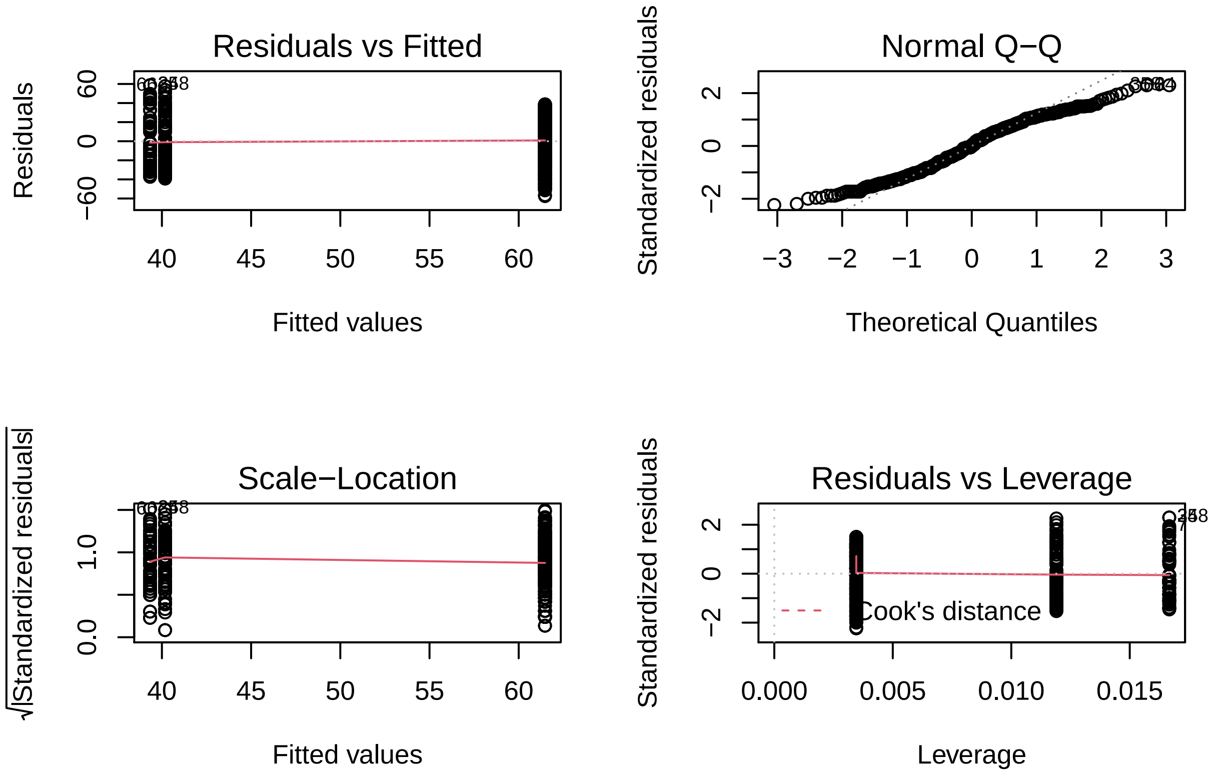
Answer:

The results that the ANOVA function generates are the same as the results from part 1.1 to 1.5.

Exercise 1.7 (5 points):

When we use the `lm()` function to run an ANOVA, we can use the `plot()` function on the model we have fit to get diagnostic plots which we can use to assess whether assumptions underlying the ANOVA were met (i.e., whether the residuals are normally distributed with equal variance). Please use the `plot()` function on the `fit_genre` model you created above to create diagnostic plots. Briefly discuss if you think these plots are more or less informative relative to the plots you created in part 1.2b. Also describe what is determining the x-axis location of the points in the residuals vs. fitted values plot (upper left plot), and why all the points are clustered together on the x-axis.

```
# create 4 subplot with diagnostic plots

par(mfrow = c(2,2))
plot(fit)
```



Answer:

The graph we made was better and more informative in QqPlot since it shows the same variables and has the intervals that we would expect for the graph to be normal. For the other graphs, the x values represent the means of each genre that we calculated (39, 40, and 61). All of the points are clustered together because each line of points clustered together represents one genre (3 categories, each category has the same mean group).

Part 2: Two-way analysis of variance

Let's now run a two-way ANOVA to assess whether both genre and the MPAA rating a movie gets affects critics' scores.

Part 2.1 (12 points): Start your hypothesis test with step 1 by stating the null and alternative hypotheses in symbols and words. State these null and alternative hypotheses for the main effect of genre, the main effect of MPAA rating, and also for the interaction effect of genre and MPAA rating.

Main effect for genre **In words** Null hypothesis: Genre has no effect on the mean critics score. Alternative hypothesis: Some genres have different mean critics scores compared to others.

In symbols

$H_0 : \mu_{action} = \mu_{comedy} = \mu_{drama}$

$H_A : \mu_i \neq \mu_j$ for some pair of i, j where i and j are {action, comedy, drama}

Main effect for MPAA rating **In words** Null hypothesis: MPAA rating has no effect on the mean critics score. Alternative hypothesis: Some MPAA ratings have different mean critics scores compared to others.

In symbols $H_0 : \mu_G = \mu_{PG} = \mu_{PG-13} = \mu_R$

$H_A : \mu_i \neq \mu_j$ for some $i \neq j$ where i and j are {G, PG, PG-13, R}

Interaction effect for genre and MPAA rating **In words** Null hypothesis: Score can be explained linearly by genre; scores can be predicted by expected effects between scores and genre alone. Alternative hypothesis: There is some combination of genre and scores that cannot be explained by genre and scores alone.

In symbols

$H_0 : Y_{ij} = 0$ where i represents genre and j represents critic score $H_A : Y_{ij} \neq 0$ where i represents genre and j represents critic score

Part 2.2 (5 points):

Now fit a model **that only has main effects** for genre and MPAA rating (i.e., that does not contain an interaction term). Use the `Anova()` function in the `car` package to create an ANOVA table using a type III sum of squares. Are the main effects statistically significant?

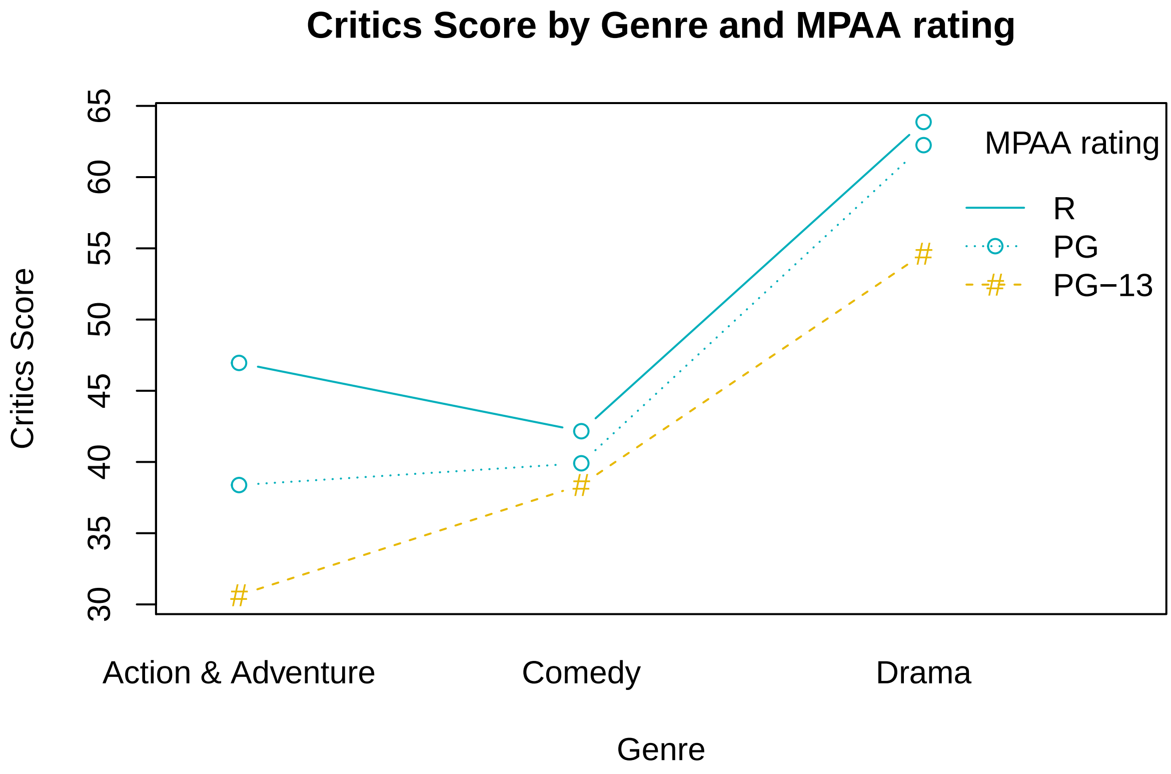
```r
fit_maineffects <- lm(critics_score~ genre+mpaa_rating, data = films)
car::Anova(fit_maineffects, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: critics_score
##             Sum Sq  Df  F value                Pr(>F)
## (Intercept)  69272   1 105.9400 < 0.00000000000000022 ***
## genre        37479   2  28.6589     0.000000000002086 ***
## mpaa_rating   5838   2   4.4637               0.01206 *
## Residuals   279862 428
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer Since the p-value is far less than 0.05. we can reject the null hypothesis and determine that the main effects are statistically significant.

Part 2.3 (7 points): Interaction effects Now let's visualize the data to see if there is an interaction between genre and MPAA rating. Use the `interaction.plot()` function to visualize the film's genre on the x-axis with different lines for the different MPAA rating levels. Try to make the plot look nice by making sure the labels are meaningful and choose a decent color scheme. Based on the this visualization, does there seem to be an interaction between genre and MPAA rating?

```
interaction.plot(x.factor = films$genre, trace.factor = films$mpaa_rating,
                 response = films$critics_score, fun = mean,
                 type = "b", legend = TRUE,
                 xlab = "Genre", ylab="Critics Score",
                 main = "Critics Score by Genre and MPAA rating",
                 trace.label = "MPAA rating",
                 pch=c(1,35), col = c("#00AFBB", "#E7B800"))
```



Answer: Based on the plot, there does not seem to be much interaction between movie genre and MPA rating based on how the plotted lines follow a roughly similar pathway. Still, an important note is that the difference in critics scores based on MPAA rating differs more within the Action & Adventure Genre. I recommend running a test to see if the differences are statistically significant between movie genre and MPAA rating.

Part 2.4 (points): Testing interactions effects Now fit a model has both the main effects and an interaction effect for genre and MPAA rating. Again use the `Anova()` function in the `car` package to create an ANOVA table using a type III sum of squares. Are the effects for all factor levels statistically significant?

```
fit_interactioneffects <- lm(critics_score~ genre*mpaa_rating, data = films)
car::Anova(fit_interactioneffects, type = "III")
```

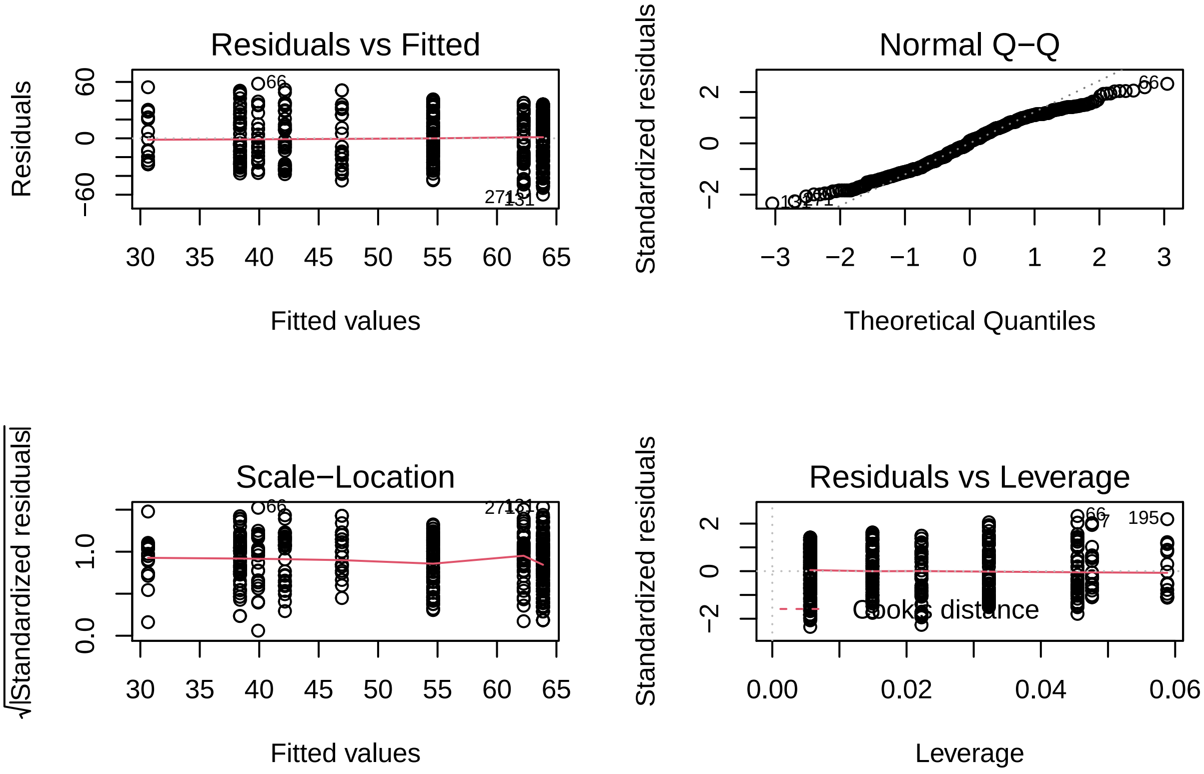
```
## Anova Table (Type III tests)
```

```
##
## Response: critics_score
##                   Sum Sq  Df F value           Pr(>F)
## (Intercept)        30935   1 47.0608 0.00000000002447 ***
## genre              11740   2  8.9298         0.000159 ***
## mpaa_rating         2580   2  1.9621         0.141837
## genre:mpaa_rating   1149   4  0.4370         0.781882
## Residuals         278713 424
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Answer: Only the main effect for genre is statistically significant at $\alpha = 0.05$, as its p-value, 0.00000000002447, is far less than the alpha level. The interaction effect between genre and MPAA rating is not statistically significant as its p-value of 0.781882 is far greater than the alpha value of $\alpha = 0.05$. Furthermore, the main effect of MPAA is not statistically significant since its p-value of 0.141837 is greater than $\alpha = 0.05$.

Part 2.5 (7 points): Checking ANOVA assumptions In order for our inferences to be valid, the assumptions underlying the ANOVA should be met. Please check these assumptions now and report whether they appear to be met.

```r
par(mfrow = c(2, 2))
plot(fit_interactioneffects)
```



```
(critics_score_summmary<-films%>%
  group_by(genre, mpaa_rating)%>%
  summarize(sd = sd(critics_score),
            .groups = "drop"))
```

```
## # A tibble: 9 x 3
##   genre              mpaa_rating    sd
##   <fct>              <fct>       <dbl>
## 1 Action & Adventure PG           22.2
## 2 Action & Adventure PG-13        26.1
## 3 Action & Adventure R            30.2
## 4 Comedy             PG           26.2
## 5 Comedy             PG-13        28.0
## 6 Comedy             R            28.1
## 7 Drama              PG           27.5
## 8 Drama              PG-13        23.9
## 9 Drama              R            24.5
```

```
max(critics_score_summmary$sd)/min(critics_score_summmary$sd)
```

```
## [1] 1.363499
```

Answer

The Q-Q line seems approximately normal so the first condition of normality seems to be met. Since the largest standard deviation is not twice as big (only 1.363499 times as big) as the smallest standard deviation the second condition of equal variances is met.

Reflection (3 points)

Please reflect on how the homework went by going to Canvas, going to the Quizzes link, and clicking on Reflection on homework 10.