GR5291 Advanced Data Analysis Problem Set ANOVA 1

Francis Zhang

October 17, 2024

Question

Consider the ChickWeight data in R. The body weights of the chicks were measured at birth (i.e., time=0) and every second day thereafter until Day 20. They were also measured on Day 21. There were four groups of chicks on different protein diets.

1.Determine whether there is a significant difference in the mean weights of the four groups on Day 20:

a)Without adjusting for Birth Weight

b)Adjusting for Birth Weight. Give the LS Means (i.e., adjusted for Birth Weight).

c)For the model in part b), check the validity of your assumptions, including parallelism. Suggest measures that you would take if the assumptions are not satisfied.

2.For 1a), perform pairwise comparisons among the 4 groups using each of the following, and comment on the results

a)Bonferroni method

b)Tukey method

3.Repeat 1a) using the Kruskal-Wallis test

Solution

Question 1

Data Preparation

```r
# Load the ChickWeight dataset
data("ChickWeight")
head(ChickWeight)
```

```
##   weight Time Chick Diet
## 1     42    0     1    1
## 2     51    2     1    1
## 3     59    4     1    1
## 4     64    6     1    1
## 5     76    8     1    1
## 6     93   10     1    1
```

```r
# Filter data for Day 20 only
day20_data <- subset(ChickWeight, Time == 20)
day20_data
```

```
##    weight Time Chick Diet
## 11    199   20     1    1
```

```
## 23      209   20      2     1
## 35      198   20      3     1
## 47      160   20      4     1
## 59      220   20      5     1
## 71      160   20      6     1
## 83      288   20      7     1
## 95      125   20      8     1
## 106     100   20      9     1
## 118     120   20     10     1
## 130     181   20     11     1
## 142     195   20     12     1
## 154      91   20     13     1
## 166     259   20     14     1
## 193     133   20     17     1
## 207     144   20     19     1
## 219     115   20     20     1
## 231     318   20     21     2
## 243     164   20     22     2
## 255     170   20     23     2
## 267      76   20     24     2
## 279     259   20     25     2
## 291     236   20     26     2
## 303     185   20     27     2
## 315     212   20     28     2
## 327     279   20     29     2
## 339     157   20     30     2
## 351     235   20     31     3
## 363     291   20     32     3
## 375     156   20     33     3
## 387     327   20     34     3
## 399     361   20     35     3
## 411     225   20     36     3
## 423     169   20     37     3
## 435     280   20     38     3
## 447     250   20     39     3
## 459     295   20     40     3
## 471     199   20     41     4
## 483     269   20     42     4
## 495     199   20     43     4
## 517     197   20     45     4
## 529     231   20     46     4
## 541     210   20     47     4
## 553     303   20     48     4
## 565     233   20     49     4
## 577     264   20     50     4
```

Part (a) Without Adjusting for Birth Weight

```
# Perform ANOVA
anova_no_adjust <- aov(weight ~ Diet, data = day20_data)

# Summary of the ANOVA
summary(anova_no_adjust)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Diet          3  55881   18627   5.464 0.00291 **
## Residuals    42 143190    3409
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value from the ANOVA is far less than 0.05, even less than 0.01, closer to 0.001. So, it indicates that there is a significant difference in mean weights among the four diet groups on Day 20.

Part (b) Adjusting for Birth Weight

```r
# Get birth weight data for each chick at time = 0
birth_weight <- ChickWeight[ChickWeight$Time == 0, c("Chick", "weight")]
colnames(birth_weight)[2] <- "birth_weight"

# Merge birth weight with Day 20 data
day20_data <- merge(day20_data, birth_weight, by = "Chick")

# Perform ANCOVA
ancova_model <- aov(weight ~ Diet + birth_weight, data = day20_data)

# Summary of the ANCOVA
summary(ancova_model)
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## Diet           3  55881   18627   5.594 0.00261 **
## birth_weight   1   6672    6672   2.004 0.16447
## Residuals     41 136519    3330
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Load the emmeans package for LS Means
library(emmeans)
```

```
## Welcome to emmeans.
## Caution: You lose important information if you filter this package's results.
## See '? untidy'
```

```r
# Calculate LS Means for Diet groups adjusted for birth weight
lsmeans <- emmeans(ancova_model, ~ Diet)
lsmeans
```

```
##  Diet emmean   SE df lower.CL upper.CL
##  1       176 14.6 41      147      206
##  2       201 18.5 41      164      239
##  3       256 18.4 41      218      293
##  4       232 19.3 41      193      271
##
## Confidence level used: 0.95
```

The ANCOVA results indicate that the type of diet has a significant effect on chick weight on Day 20 ($p = 0.00261$), suggesting that mean weights differ across the four diet groups. However, birth weight does not have a statistically significant effect on Day 20 weight ($p = 0.16447$), implying that initial weight differences among chicks do not significantly impact the weight outcomes once diet is considered in the model.

The LS Means (Least Squares Means), adjusted for birth weight, reveal the following estimated mean weights for each diet group: Diet 1 has a mean weight of 176 grams (95% CI: 147, 206), Diet 2 has a mean weight of

201 grams (95% CI: 164, 239), Diet 3 has the highest mean weight at 256 grams (95% CI: 218, 293), and Diet 4 has a mean weight of 232 grams (95% CI: 193, 271). This adjustment controls for any variation in initial birth weight, providing a clearer picture of the effect of diet on final weight.

In summary, Diet 3 yields the highest average weight on Day 20, suggesting it may be the most effective for promoting weight gain. Conversely, Diet 1 results in the lowest adjusted mean weight. The confidence intervals for each group indicate the range within which we expect the true mean weight to fall, at a 95% confidence level. Further pairwise comparisons between diet groups could be performed to determine whether the differences in adjusted means between specific diets are statistically significant.

```r
# Get birth weight data for each chick at time = 0
birth_weight <- ChickWeight[ChickWeight$Time == 0, c("Chick", "weight")]
colnames(birth_weight)[2] <- birth_weight
```

```
## Warning in colnames(birth_weight)[2] <- birth_weight: number of items to
## replace is not a multiple of replacement length
```

```r
# Merge birth weight with Day 20 data
day20_data <- merge(day20_data, birth_weight, by = "Chick")

# Perform ANCOVA
ancova_model <- aov(weight ~ Diet + birth_weight, data = day20_data)

# Summary of the ANCOVA
summary(ancova_model)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Diet          3  55881   18627   5.594 0.00261 **
## birth_weight  1   6672    6672   2.004 0.16447
## Residuals    41 136519    3330
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Load the emmeans package for LS Means
library(emmeans)

# Calculate LS Means for Diet groups adjusted for birth weight
lsmeans <- emmeans(ancova_model, ~ Diet)
lsmeans
```

```
##  Diet emmean   SE df lower.CL upper.CL
##  1       176 14.6 41      147      206
##  2       201 18.5 41      164      239
##  3       256 18.4 41      218      293
##  4       232 19.3 41      193      271
##
## Confidence level used: 0.95
```

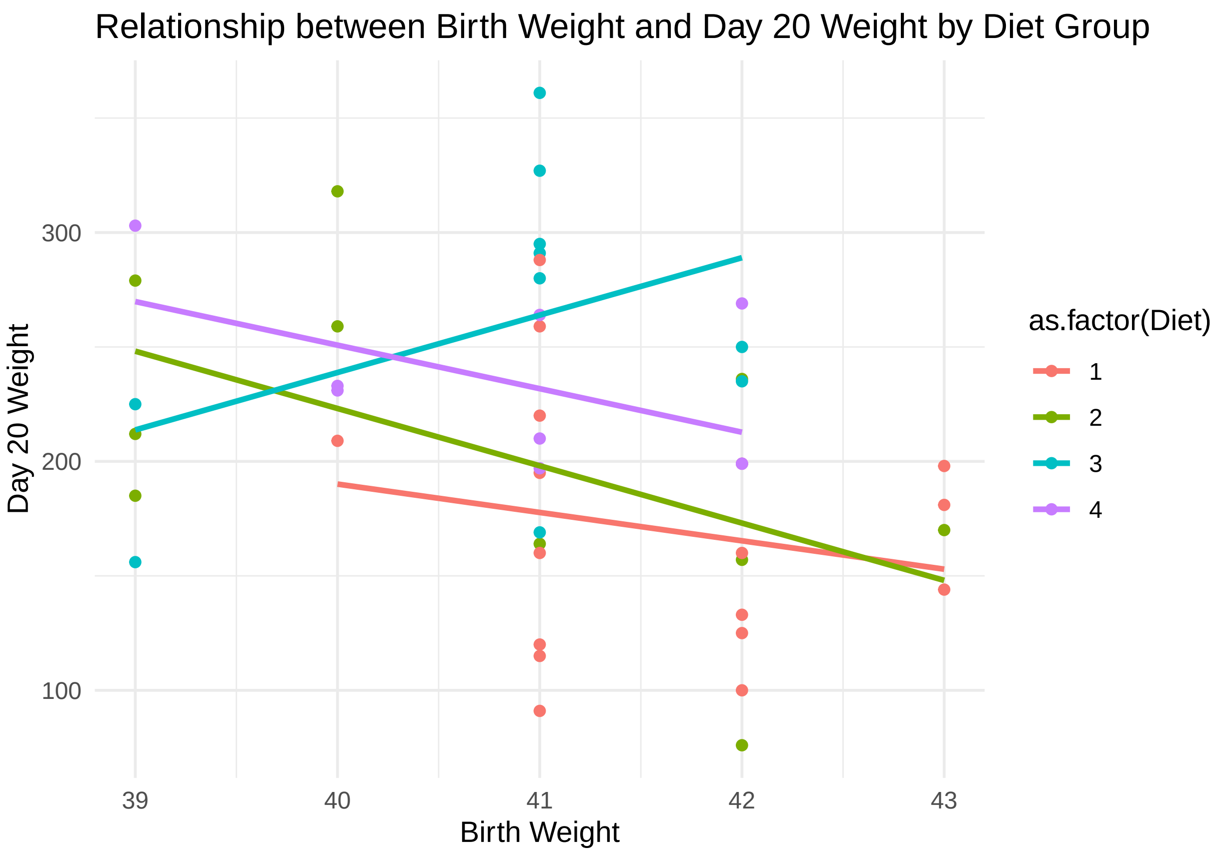
Part (c) Check Validity of Assumptions for the ANCOVA Model

- Linearity: Check if the relationship between birth weight and Day 20 weight is linear within each diet group.
- Parallelism: The effect of the covariate (birth weight) should be consistent across the diet groups.
- Normality: Residuals of the model should be normally distributed.
- Homoscedasticity: Residuals should have constant variance across groups.

- Independence: Observations should be independent of one another, meaning that the weight of one chick should not influence the weight of another.

```
# Linearity: Scatter plot for each diet group with a linear regression line
library(ggplot2)
ggplot(day20_data, aes(x = birth_weight, y = weight, color = as.factor(Diet))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +  # Linear regression line
  labs(title =
         "Relationship between Birth Weight and Day 20 Weight by Diet Group",
       x = "Birth Weight", y = "Day 20 Weight") +
  theme_minimal()
```

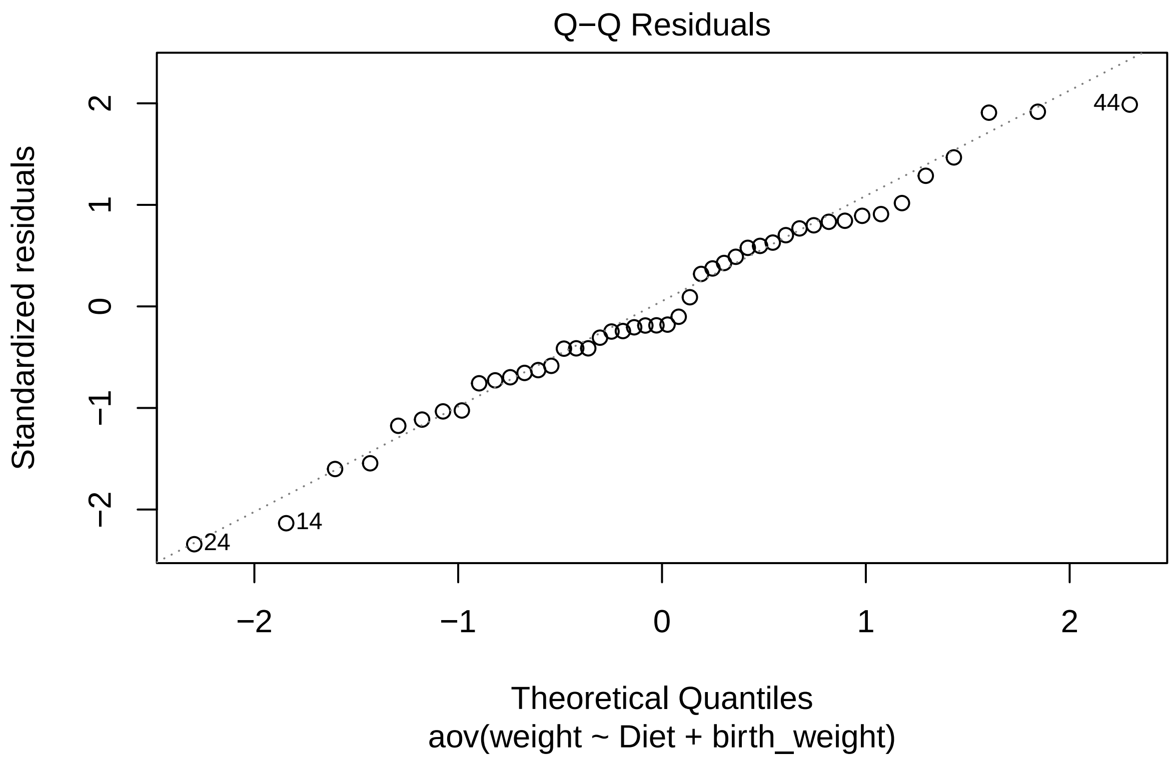
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Parallelism: Test by including interaction term
ancova_parallel <- aov(weight ~ Diet * birth_weight, data = day20_data)
summary(ancova_parallel)
```

```
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## Diet               3  55881   18627   5.924 0.00203 **
## birth_weight       1   6672    6672   2.122 0.15342
## Diet:birth_weight  3  17043    5681   1.807 0.16235
## Residuals         38 119476    3144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

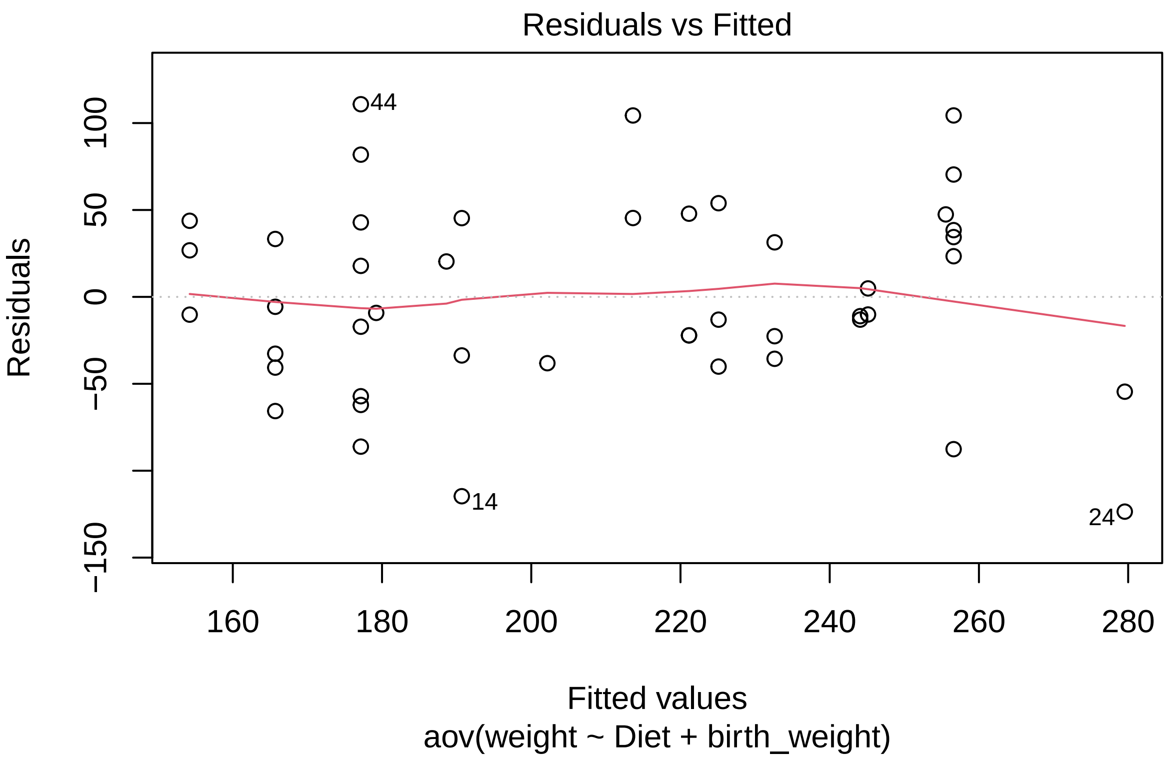
```
# Normality: Q-Q plot and Shapiro-Wilk test for residuals
plot(ancova_model, which = 2)  # Q-Q plot
```




```r
shapiro.test(residuals(ancova_model))  # Shapiro-Wilk test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(ancova_model)
## W = 0.98247, p-value = 0.7082
```

```r
# Homoscedasticity: Residuals vs Fitted plot and Breusch-Pagan test
plot(ancova_model, which = 1)  # Residuals vs Fitted plot
```




```r
library(lmtest)
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
bptest(ancova_model)  # Breusch-Pagan test for homoscedasticity
```

```
##
##  studentized Breusch-Pagan test
##
## data:  ancova_model
## BP = 6.247, df = 4, p-value = 0.1814
```

```r
# Independence: Durbin-Watson test for autocorrelation in residuals
library(car)
```

```
## Loading required package: carData
```

```r
durbinWatsonTest(ancova_model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.3209161      2.602097    0.08
##  Alternative hypothesis: rho != 0
```

- Linearity: The relationship between birth weight and Day 20 weight within each diet group was evaluated using scatter plots with linear regression lines. The plots show that some groups have a linear relationship, while others do not exhibit a clear linear trend. This might suggest that the linearity assumption is not fully met for all groups. Further investigation or potential transformations of the covariate (birth weight) could be considered if the non-linearity is significant.

- Parallelism: The interaction term between diet and birth weight in the ANCOVA model was tested. The p-value for the interaction term is 0.16235, which is greater than 0.05, indicating that the interaction is not significant. Therefore, the parallelism assumption is met, meaning the effect of birth weight on weight gain is consistent across the different diet groups.

- Normality: The Q-Q plot of the residuals shows that the residuals are approximately normally distributed, as most points lie along the diagonal line. The Shapiro-Wilk test for normality also supports this conclusion, with a p-value of 0.7082, which is greater than 0.05. Therefore, the normality assumption is satisfied.

- Homoscedasticity: The Residuals vs. Fitted plot shows no clear pattern, suggesting that the variance of residuals is relatively constant across fitted values. The Breusch-Pagan test yields a p-value of 0.1814, which is greater than 0.05, further supporting the assumption of homoscedasticity. Thus, the homoscedasticity assumption is met.

- Independence: The Durbin-Watson test for autocorrelation in the residuals produces a p-value of 0.08, which is slightly above the 0.05 threshold. This suggests that there is no strong evidence of autocorrelation in the residuals, and the independence assumption is reasonably satisfied. However, given that the p-value is close to the significance level, you may want to monitor this assumption closely, especially if additional data are collected.

Overall, the key assumptions for ANCOVA (linearity, parallelism, normality, homoscedasticity, and independence) are generally satisfied, with some minor concerns regarding linearity in some diet groups and the independence assumption. Based on these results, the ANCOVA model is valid, and the results can be interpreted with confidence.

Question 2

Part (a) Bonferroni Method

```r
pairwise.t.test(day20_data$weight,
                day20_data$Diet, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  day20_data$weight and day20_data$Diet
##
##   1      2      3
## 2 0.8278 -      -
## 3 0.0027 0.2853 -
## 4 0.0700 1.0000 1.0000
##
## P value adjustment method: bonferroni
```

Using the Bonferroni method to adjust for multiple comparisons, we find that there is only a significant difference between Diet 1 and Diet 3 ($p = 0.0027$). No other pairwise comparisons show a statistically significant difference after the Bonferroni correction. The Bonferroni adjustment is a conservative method, and the fact that only one comparison is significant after adjustment suggests that other differences are either small or not robust enough to survive this correction for multiple testing.

Part (b) Tukey Method

```r
# Tukey's HSD test
tukey_result <- TukeyHSD(anova_no_adjust)
tukey_result
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight ~ Diet, data = day20_data)
##
## $Diet
##          diff         lwr       upr     p adj
## 2-1  35.18824 -27.0570451  97.43352 0.4394508
## 3-1  88.48824  26.2429549 150.73352 0.0024869
## 4-1  63.47712  -0.9086564 127.86290 0.0545918
## 3-2  53.30000 -16.5496130 123.14961 0.1895019
## 4-2  28.28889 -43.4747665 100.05254 0.7186387
## 4-3 -25.01111 -96.7747665  46.75254 0.7877511
```

The Tukey method shows that there is a statistically significant difference between Diet 3 and Diet 1 (p = 0.0025). No other pairwise comparisons show statistically significant differences. Similar to the Bonferroni method, the significant result involves Diet 3 vs Diet 1, and Diet 4 vs Diet 1 is close to significance, but does not reach the 0.05 threshold.

Question 3

Part (a) Kruskal-Wallis test

```r
kruskal.test(weight ~ Diet, data = day20_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  weight by Diet
## Kruskal-Wallis chi-squared = 12.852, df = 3, p-value = 0.004969
```

Since the p-value (0.004969) is less than the standard significance level of 0.05, we reject the null hypothesis that all diet groups have the same distribution of weights. This indicates that there are significant differences in weight among the diet groups. The Kruskal-Wallis test is a non-parametric alternative to ANOVA and is used when the assumption of normality may not hold. Therefore, this result suggests that diet does indeed affect the weight, and further pairwise comparisons could be conducted to determine which specific groups differ.