GR5291 Advanced Data Analysis Problem Set ANOVA 1

Francis Zhang

October 23, 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.Perform ANCOVA, adjusting for birthweight, to determine whether there is a significant difference in the mean weights of the four groups using the measurements separately at each time point: Day 10, 18, and 20.

2.Perform an appropriate repeated measures ANOVA, adjusting for birthweight, to determine whether there is a significant difference in the mean weights of the four diet groups using the measurements on Days 10, 18, and 20. (Make sure to treat the "Chick" and "Time" variables as unordered factors.)

- Do the analyses assuming compound symmetry and unstructured covariance structures and compare the results repeated measures.

3.Check the validity of your assumptions in each case, and comment on the approaches used in 1 and 2 above.

Solution

Question 1

```
# Load dataset
data("ChickWeight")

# Filter for each time point and merge birth weight data
timepoints <- c(10, 18, 20)

birth_weight <- ChickWeight[ChickWeight$Time == 0, c("Chick", "weight")]
colnames(birth_weight)[2] <- "birth_weight"

for (time in timepoints) {
  # Filter data for the specific time point
  day_data <- subset(ChickWeight, Time == time)
  day_data <- merge(day_data, birth_weight, by = "Chick")

  # Perform ANCOVA
  ancova_model <- aov(weight ~ Diet + birth_weight, data = day_data)
  print(paste("Results for Day", time))
  print(summary(ancova_model))
}
```

```
## [1] "Results for Day 10"
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Diet          3   8314  2771.4   6.336 0.00115 **
```

```
## birth_weight  1     57    57.3   0.131 0.71908
## Residuals    44  19247   437.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Results for Day 18"
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Diet          3  36690   12230   4.729 0.00623 **
## birth_weight  1   6229    6229   2.409 0.12818
## Residuals    42 108612    2586
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Results for Day 20"
##              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
```

- Day 10:

- The effect of the Diet is statistically significant ($p = 0.00115$). This suggests that there is a significant difference in the mean weights of the chicks across the four diet groups on Day 10, after adjusting for birth weight.

- Birth weight is not a significant covariate ($p = 0.71908$), meaning that on Day 10, the initial weight of the chicks does not significantly influence their weight after adjusting for diet.

- Day 18:

- The effect of the Diet remains statistically significant ($p = 0.00623$), indicating that there is still a significant difference in the mean weights across diet groups on Day 18.

- Birth weight is again not significant ($p = 0.12818$), suggesting that the initial weight is not a major determinant of the weight differences between diet groups at this time point.

- Day 20:

- The effect of Diet is also statistically significant ($p = 0.00261$), meaning that the diet groups continue to show significant differences in weight at this later time point.

- Birth weight is not significant ($p = 0.16447$), similar to the previous days, indicating that the initial weight does not significantly affect the final weight at Day 20 after adjusting for diet.

Overall Conclusion:

Across Days 10, 18, and 20, the diet consistently shows a statistically significant effect on the body weight of the chicks, meaning that the different protein diets lead to significant differences in weight. However, birth weight does not significantly affect the weight at any time point after adjusting for the diet. This suggests that diet is the main driver of weight differences between the groups, independent of the initial weight of the chicks.

Question 2

Compound Symmetry Model: We will use the gls() function from the nlme package to fit the repeated measures model under the assumption of compound symmetry.

```r
library(nlme)

# Prepare data for repeated measures
```

```r
repeat_data <- subset(ChickWeight, Time %in% c(10, 18, 20))
repeat_data <- merge(repeat_data, birth_weight, by = "Chick")
repeat_data$Time <- as.factor(repeat_data$Time)

# Fit the model with compound symmetry
cs_model <- gls(weight ~ Diet + birth_weight,
                correlation = corCompSymm(form = ~ 1 | Chick),
                data = repeat_data)
summary(cs_model)
```

```
## Generalized least squares fit by REML
##   Model: weight ~ Diet + birth_weight
##   Data: repeat_data
##        AIC      BIC    logLik
##   1558.966 1579.406 -772.4831
##
## Correlation Structure: Compound symmetry
##  Formula: ~1 | Chick
##  Parameter estimate(s):
##       Rho
## 0.1283637
##
## Coefficients:
##                 Value Std.Error   t-value p-value
## (Intercept)  455.8115 239.96527  1.899490  0.0596
## Diet2         22.4534  17.14204  1.309842  0.1924
## Diet3         58.9842  16.98393  3.472939  0.0007
## Diet4         42.8567  16.88870  2.537599  0.0123
## birth_weight  -7.6412   5.76901 -1.324531  0.1875
##
##  Correlation:
##              (Intr) Diet2  Diet3  Diet4
## Diet2        -0.313
## Diet3        -0.282  0.407
## Diet4        -0.224  0.392  0.389
## birth_weight -0.999  0.290  0.258  0.201
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -1.84674225 -0.80540325 -0.02223785  0.75399132  2.48551353
##
## Residual standard error: 64.1698
## Degrees of freedom: 142 total; 137 residual
```

```r
# Perform ANOVA
anova(cs_model)
```

```
## Denom. DF: 137
##              numDF  F-value p-value
## (Intercept)      1 775.2744  <.0001
## Diet             3   6.0069  0.0007
## birth_weight     1   1.7544  0.1875
```

The compound symmetry model assumes that the correlations between measurements taken at different time points are equal and the variances across time points are the same.

- The diet variable shows a statistically significant effect on weight (F = 6.0069, p = 0.0007), indicating that the four diet groups differ in their effects on chick weight.
- Birth weight is not statistically significant (p = 0.1875), suggesting that after accounting for diet, birth weight does not significantly influence the weight at the three time points.
- The rho value (0.12836) from the compound symmetry correlation structure suggests a modest correlation between measurements within the same chicks.

From these results, we conclude that different diets lead to significant weight differences, while birth weight does not significantly affect the outcomes. The assumption of equal correlation between time points (compound symmetry) is used in this model, which may not fully capture the variability between time points.

Unstructured Covariance Model: Fit the repeated measures model assuming an unstructured covariance matrix, which allows for a different covariance for each pair of time points.

```
# Convert Time to numeric for correlation structure
repeat_data$Time <- as.numeric(repeat_data$Time)

# Fit the model with unstructured covariance
unstruct_model <- gls(weight ~ Diet + birth_weight,
                      correlation = corSymm(form = ~ Time | Chick),
                      weights = varIdent(form = ~ 1 | Time),
                      data = repeat_data)
summary(unstruct_model)
```

```
## Generalized least squares fit by REML
##   Model: weight ~ Diet + birth_weight
##   Data: repeat_data
##        AIC      BIC    logLik
##   1302.604 1334.724 -640.3022
##
## Correlation Structure: General
##  Formula: ~Time | Chick
##  Parameter estimate(s):
##  Correlation:
##   1     2
## 2 0.901
## 3 0.871 0.996
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | Time
##  Parameter estimates:
##        1        2        3
## 1.000000 3.893129 4.566131
##
## Coefficients:
##                 Value Std.Error  t-value p-value
## (Intercept)  37.35726  83.65020 0.446589  0.6559
## Diet2        10.03439   5.99585 1.673557  0.0965
## Diet3         8.09458   5.94016 1.362688  0.1752
## Diet4        32.04677   5.90151 5.430269  0.0000
## birth_weight  0.96610   2.01044 0.480541  0.6316
##
##  Correlation:
##              (Intr) Diet2  Diet3  Diet4
```

```
## Diet2        -0.316
## Diet3        -0.285  0.412
## Diet4        -0.227  0.398  0.394
## birth_weight -0.999  0.292  0.261  0.203
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -0.8749333  0.4815911  0.7575887  1.1659672  2.5932007
##
## Residual standard error: 29.67928
## Degrees of freedom: 142 total; 137 residual
```

```
# Perform ANOVA
anova(unstruct_model)
```

```
## Denom. DF: 137
##              numDF   F-value p-value
## (Intercept)      1 1730.5517  <.0001
## Diet             3   10.0809  <.0001
## birth_weight     1    0.2309  0.6316
```

The unstructured covariance model, which allows for a different covariance between each pair of time points, was also fitted:

- The diet variable remains statistically significant ($F = 10.0809$, $p < 0.0001$), confirming the significant effect of diet on weight. The diets differ significantly in their impact on chick weight.
- Birth weight is not statistically significant here either ($p = 0.6316$), consistent with the compound symmetry model, indicating that birth weight is not a major factor in weight differences across the three time points.
- The estimated variances for each time point are different, with greater variability at later time points (Day 20 has a variance estimate of 4.566). This suggests that weight differences between chicks grow larger over time.
- The log-likelihood for the unstructured model is higher (-640.3022) than for the compound symmetry model (-772.4831), which suggests that the unstructured model provides a better fit to the data.

Comparison of Models:

- Better Fit: The unstructured covariance model fits the data better than the compound symmetry model, as evidenced by its higher log-likelihood and its ability to account for different variances and covariances between time points.
- Diet Effect: In both models, the diet has a significant effect on chick weight, but the unstructured model suggests a stronger effect (with a larger F-value) than the compound symmetry model.
- Birth Weight: In both models, birth weight is not a significant predictor of weight, indicating that diet is the primary driver of weight differences.
- Variability over Time: The unstructured model reveals that the variance of weight measurements increases over time, which is important for understanding how weight variability grows as chicks age.

Question 3

For ANCOVA (Question 1):

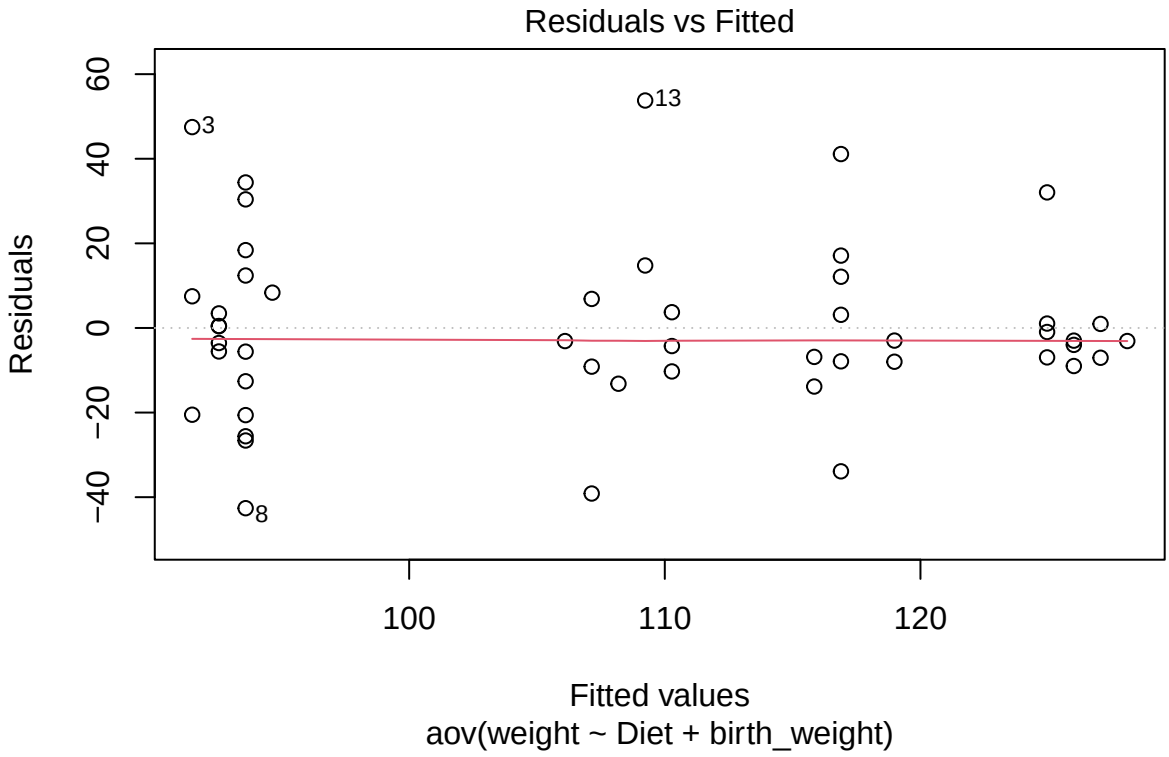
Linearity: • Assumption: The relationship between the covariate (birth weight) and the dependent variable (weight at different time points) should be linear within each diet group.

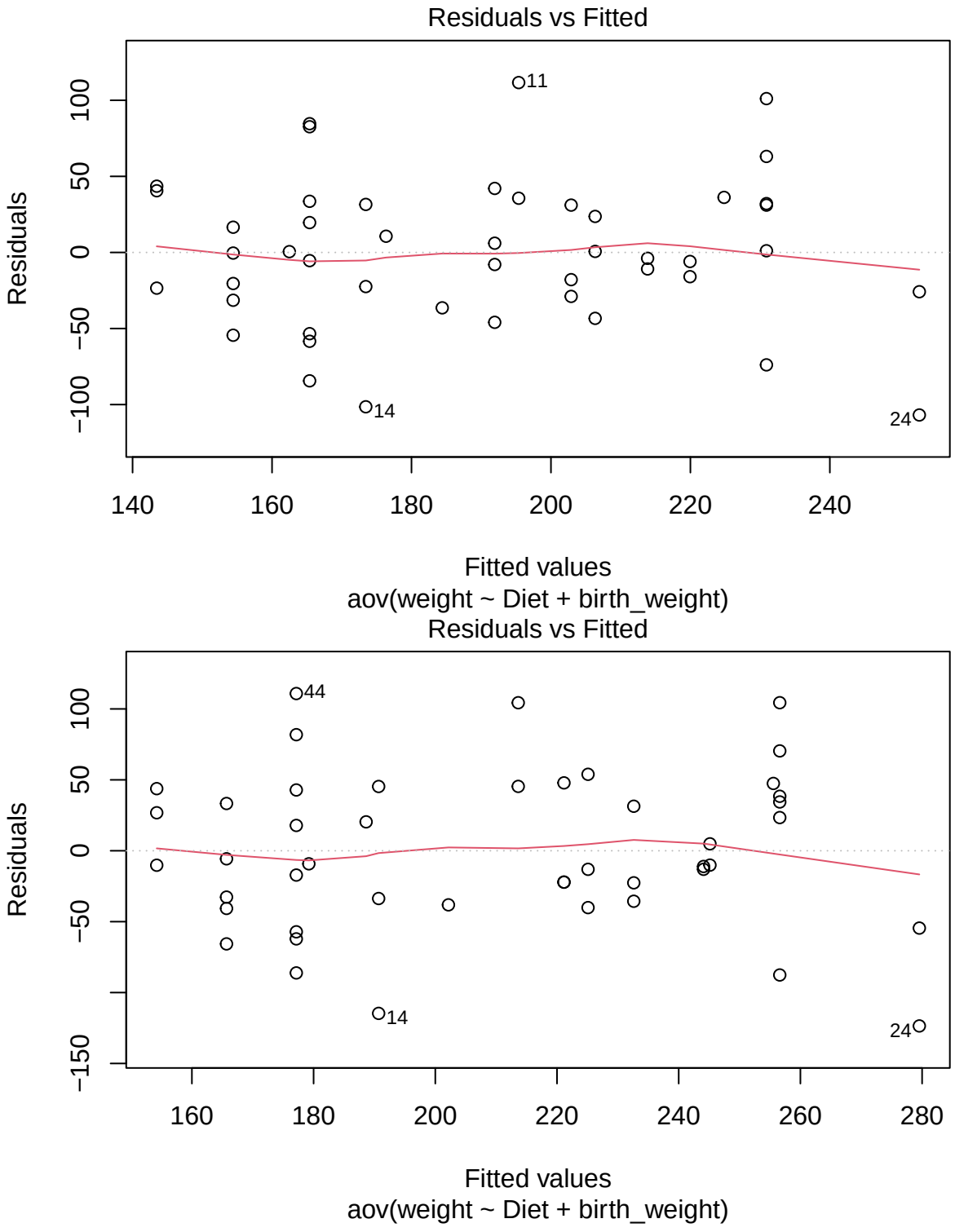
- How to Check: Use Residuals vs Fitted plots to ensure there are no clear patterns indicating non-linearity.

```r
# Residual plots for ANCOVA at each time point
for (time in timepoints) {
  day_data <- subset(ChickWeight, Time == time)
  day_data <- merge(day_data, birth_weight, by = "Chick")

  ancova_model <- aov(weight ~ Diet + birth_weight, data = day_data)

  # Residuals vs Fitted plot
  plot(ancova_model, which = 1)  # Residuals vs Fitted
}
```





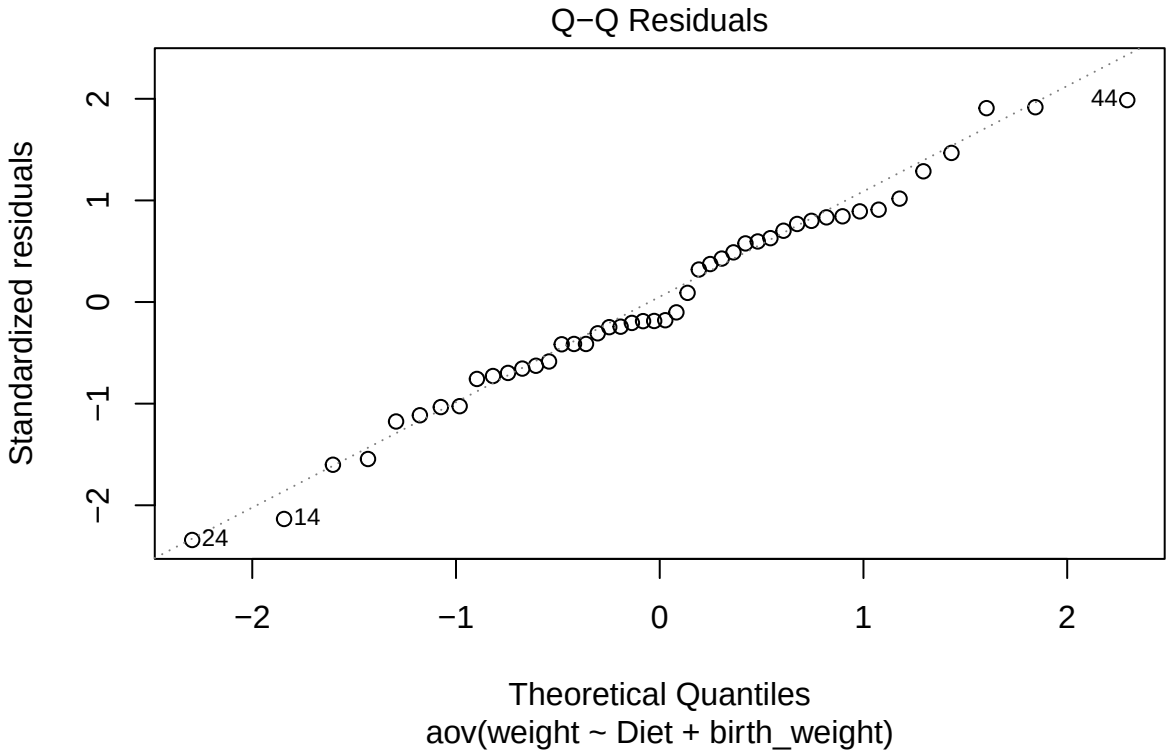
Homoscedasticity: • Assumption: Residuals should have constant variance across groups (homoscedasticity).

• How to Check: Use Residuals vs Fitted plots. If the plot shows a funnel shape, this indicates heteroscedasticity.

Normality of Residuals: • Assumption: Residuals should be normally distributed.

• How to Check: Use a Q-Q plot and the Shapiro-Wilk test to check for normality.

```
# Q-Q Plot for Normality
plot(ancova_model, which = 2)  # Q-Q Plot
```




```
# Shapiro-Wilk test for normality of residuals
shapiro.test(residuals(ancova_model))
```

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

Independence: • Assumption: Observations (weights) should be independent within each group.

• How to Check: Independence is typically ensured by the experimental design.

For Repeated Measures ANOVA (Question 2):

Sphericity: • Assumption: Sphericity assumes that the variances of the differences between all pairs of time points are equal.

• How to Check: For mixed models with gls() (as used in Question 2), sphericity is handled by the model, but we can compare models with different covariance structures (e.g., compound symmetry vs. unstructured covariance).

```
# Compare models using likelihood ratio tests (ANOVA)
anova(cs_model, unstruct_model)
```

```
##                Model df      AIC      BIC    logLik   Test  L.Ratio p-value
```

```
## cs_model           1  7 1558.966 1579.406 -772.4831
## unstruct_model     2 11 1302.604 1334.724 -640.3022 1 vs 2 264.3618  <.0001
```

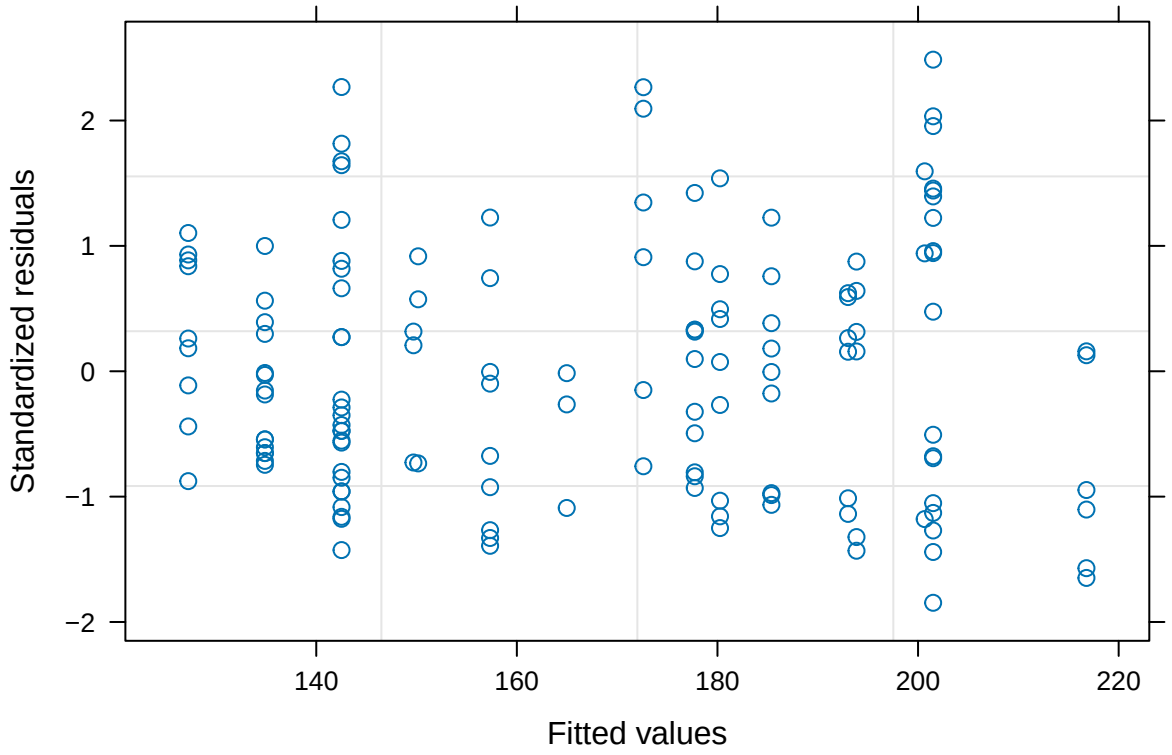
```
# Compare using AIC
AIC(cs_model, unstruct_model)
```

```
##                 df      AIC
## cs_model         7 1558.966
## unstruct_model  11 1302.604
```

Normality of Residuals: • Assumption: Residuals from the repeated measures model should be normally distributed.

• How to Check: Use a Q-Q plot and the Shapiro-Wilk test for the residuals from the repeated measures ANOVA.

```
# Q-Q Plot for normality in repeated measures model
plot(cs_model, which = 2)  # Assuming cs_model is fitted with gls()
```




```
# Shapiro-Wilk test for normality of residuals
shapiro.test(residuals(cs_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(cs_model)
## W = 0.97122, p-value = 0.004319
```

Model Fit and Covariance Structure: • Assumption: The covariance structure should be appropriate for the data. Compare the compound symmetry and unstructured covariance models using AIC and likelihood ratio tests.

• How to Check: If the unstructured model has a significantly lower AIC and better fit (based on likelihood ratio tests), it may be the better model for the data.