R_ANCOVA, ANOVA involved numeric values

2023-10-14

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.

```
# import data
library(datasets)
data("ChickWeight")
summary(ChickWeight)
```

```
##      weight           Time           Chick     Diet
##  Min.   : 35.0   Min.   : 0.00   13     : 12   1:220
##  1st Qu.: 63.0   1st Qu.: 4.00   9      : 12   2:120
##  Median :103.0   Median :10.00   20     : 12   3:120
##  Mean   :121.8   Mean   :10.72   10     : 12   4:118
##  3rd Qu.:163.8   3rd Qu.:16.00   17     : 12
##  Max.   :373.0   Max.   :21.00   19     : 12
##                                  (Other):506
```

```
# get baseline
birth_weight <- ChickWeight[ChickWeight$Time == 0, c("Chick", "weight")]
colnames(birth_weight) <- c("Chick", "weight_initial")
chick_adjusted <- merge(ChickWeight, birth_weight, by = "Chick", all.x = TRUE)
```

Q1:

Perform ANCOVA adjusting for baseline to determine whether there is a significant difference in the mean weights of the four groups, separately at each timepoint: Day 16, Day 20 and Day 21, .

```
# ancova
print("Day16")
```

```
## [1] "Day16"
```

```
data_16=chick_adjusted[chick_adjusted$Time == 16, ]
summary(aov(weight ~ weight_initial+Diet, data_16))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)  
## weight_initial  1   9606    9606   5.239 0.0272 *
## Diet            3  14578    4859   2.650 0.0611 .
## Residuals      42  77015    1834                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print("Day20")
```

```
## [1] "Day20"
```

```
data_20=chick_adjusted[chick_adjusted$Time == 20, ]
summary(aov(weight ~ weight_initial+Diet, data_20))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)  
## weight_initial  1  20415   20415   6.131 0.0175 *
## Diet            3  42138   14046   4.218 0.0109 *
## Residuals      41 136519    3330                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print("Day21")
```

```
## [1] "Day21"
```

```
data_21=chick_adjusted[chick_adjusted$Time == 21, ]
summary(aov(weight ~ weight_initial+Diet, data_21))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)  
## weight_initial  1  20538   20538   5.112 0.0293 *
## Diet            3  43763   14588   3.631 0.0208 *
## Residuals      40 160703    4018                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q1_Answer:

P-values are 0.0109 and 0.0208 separately for Day 20 and Day 21. There is a significant difference in the mean weights of the four groups at Day 20 and Day 21. P-value is 0.0611 for Day 16. There is no significant difference in the mean weights of the four groups at Day 16.

Q2:

Perform an appropriate repeated measures ANOVA, adjusting for baseline, to determine whether there is a significant difference in the mean weights of the four groups using the measurements on Days 16, 20, and 21.

```r
data_repeated <- subset(ChickWeight, Time %in% c(16, 20, 21))
library(carData)
summary(aov(weight ~ Diet * Time + Error(Chick), data = data_repeated))
```

```
##
## Error: Chick
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Diet       3 126606   42202   4.725 0.00637 **
## Time       1   8076    8076   0.904 0.34722
## Diet:Time  1    366     366   0.041 0.84051
## Residuals 41 366195    8932
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Error: Within
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Time       1  62507   62507 240.578  < 2e-16 ***
## Diet:Time  3   6528    2176   8.375 5.94e-05 ***
## Residuals 87  22604     260
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q2_Answer:

From between-subject analysis, based on p-value as 0.00637, there are significant differences in the mean weights of different diet groups.

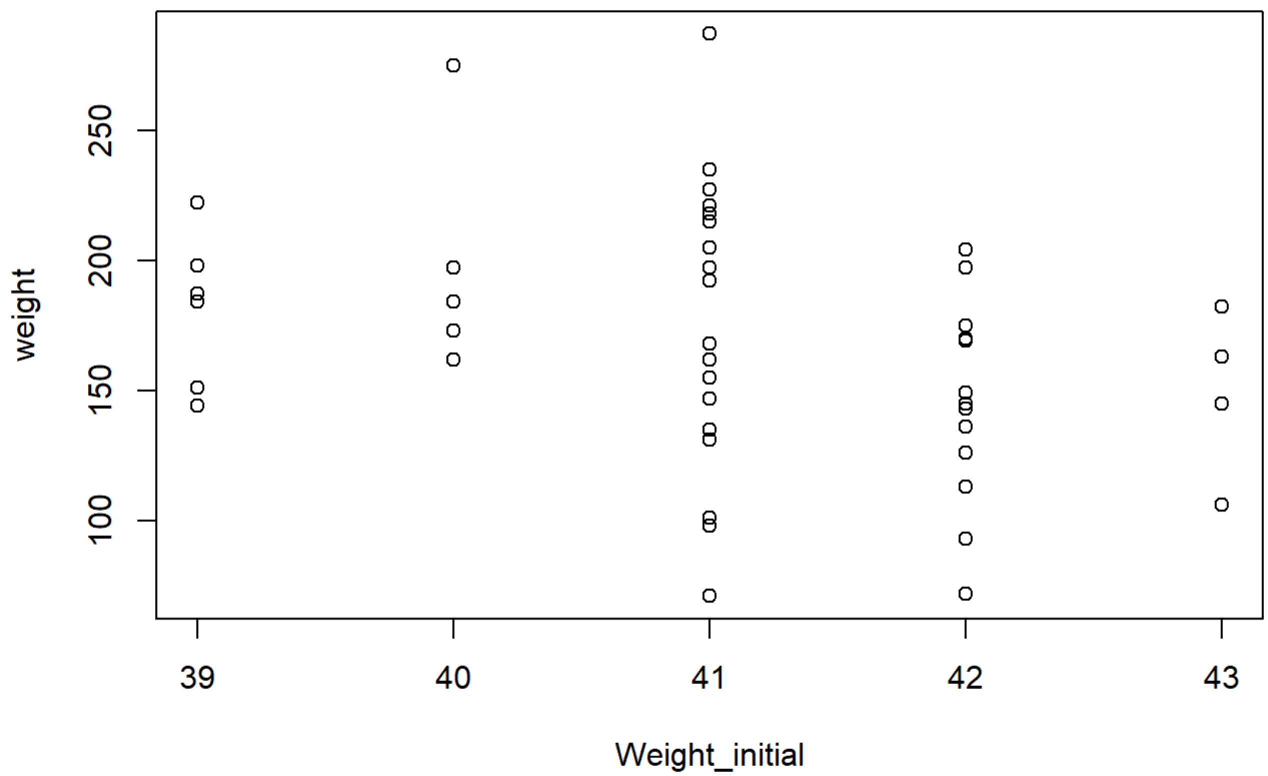
Because we measure chicks at multiple time points, we need to take considerations of repeated measures in within analysis. Based on p-value as $p < 2e-16$, different time points have a significant effect on weights. The interaction has $p = 5.94e-05$, showing significant interactions between the diet and time variables.

In sum, adjusting for baseline, there is a significant difference in the mean weights of the four groups using the measurements on Days 16, 20, and 21.

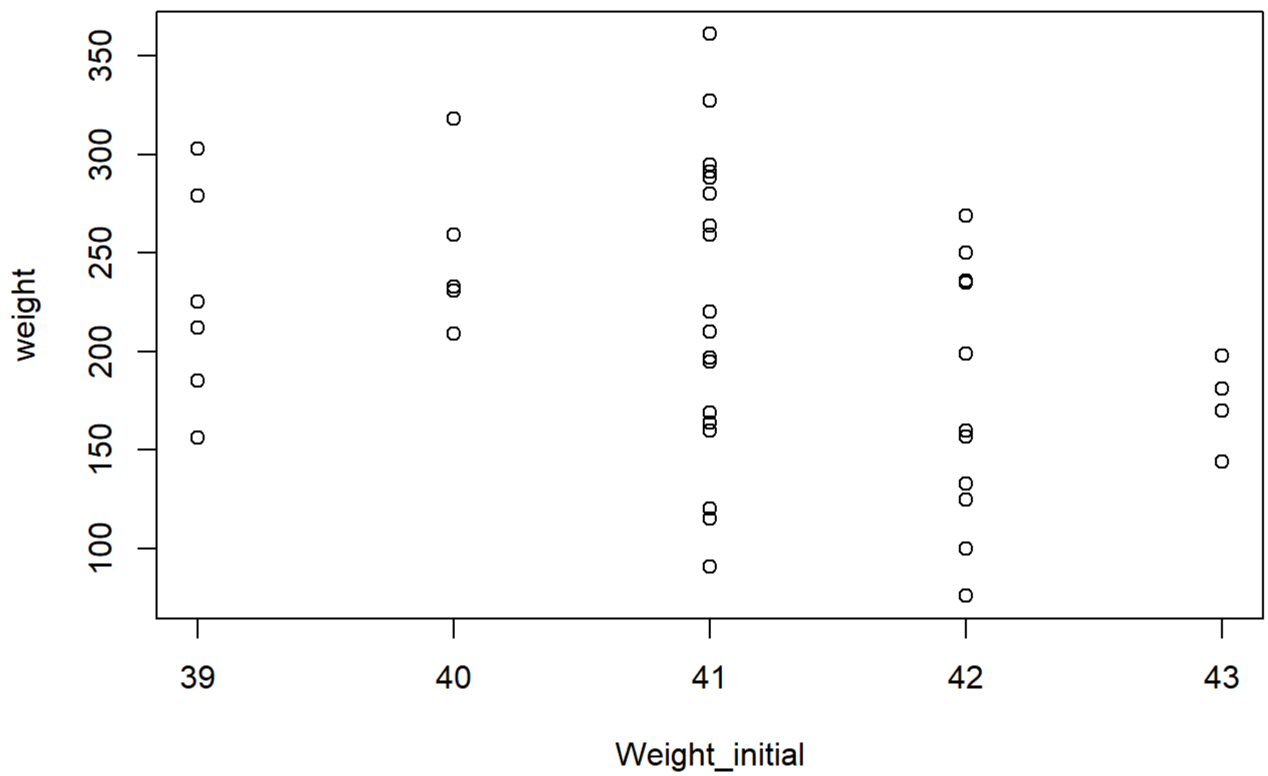
Q3:

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

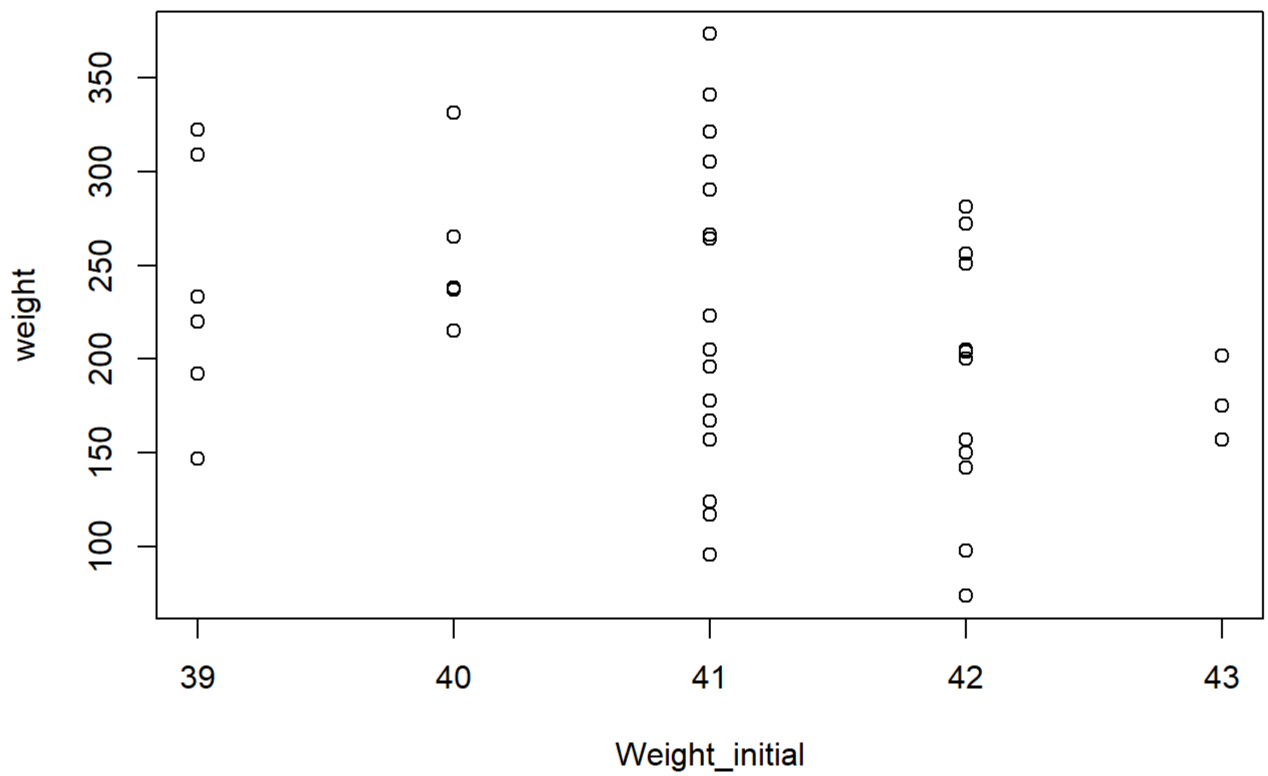
```r
#check for ANCOVA
# Linearity between the covariate and Y
plot(data_16$weight_initial, data_16$weight, xlab = "Weight_initial", ylab = "weight")
```




```
plot(data_20$weight_initial, data_20$weight, xlab = "Weight_initial", ylab = "weight")
```




```
plot(data_21$weight_initial, data_21$weight, xlab = "Weight_initial", ylab = "weight")
```




```
# Y normally distributed
shapiro.test(resid(aov(weight ~ Diet + weight_initial, data_16)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(aov(weight ~ Diet + weight_initial, data_16))
## W = 0.98261, p-value = 0.7019
```

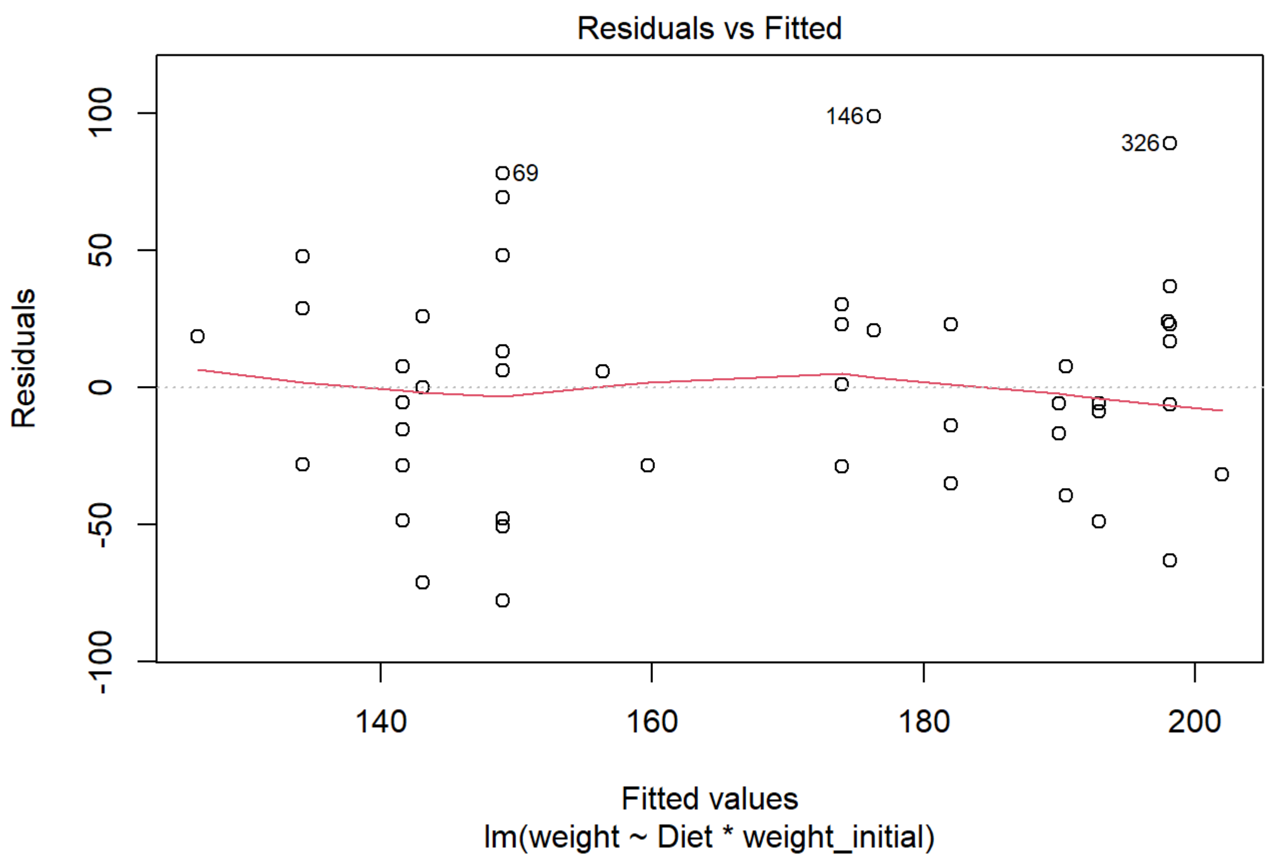
```
shapiro.test(resid(aov(weight ~ Diet + weight_initial, data_20)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(aov(weight ~ Diet + weight_initial, data_20))
## W = 0.98247, p-value = 0.7082
```

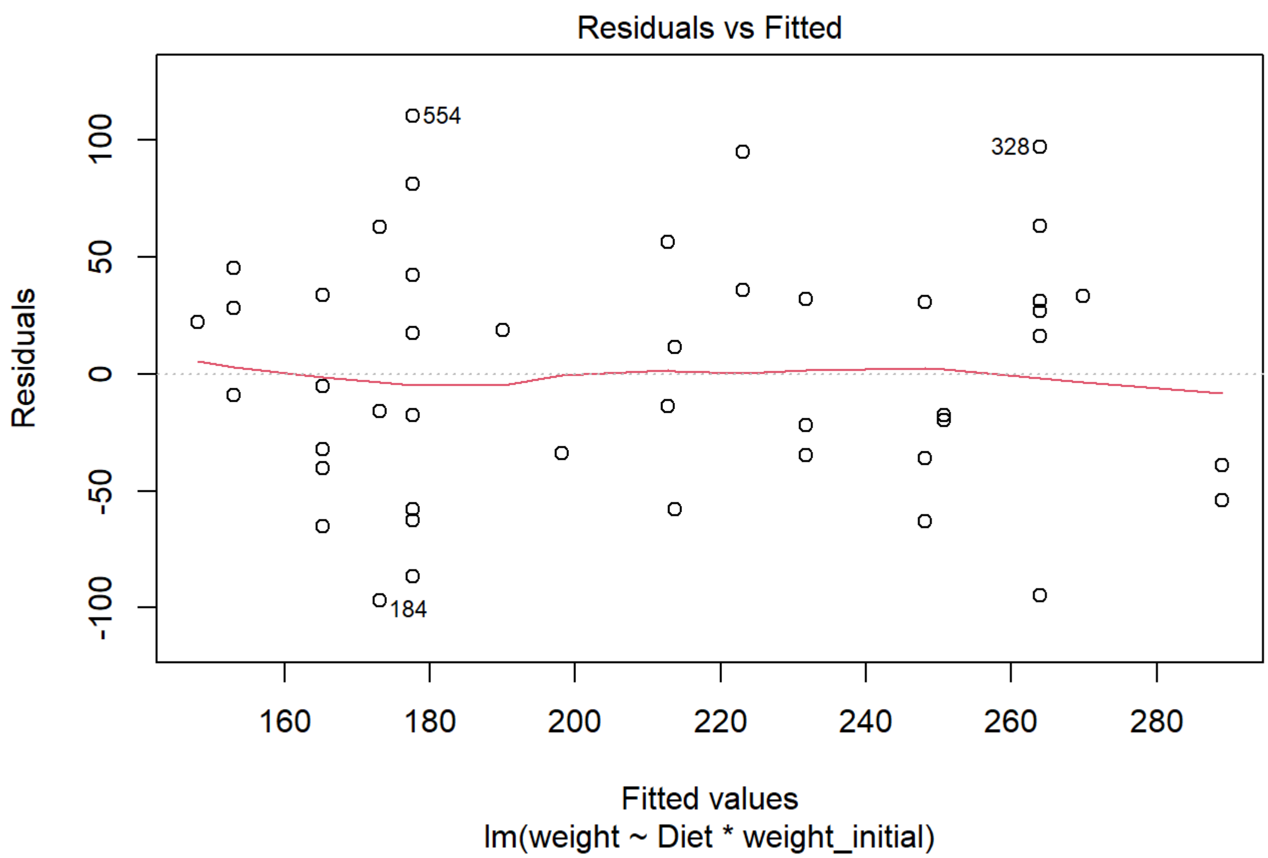
```
shapiro.test(resid(aov(weight ~ Diet + weight_initial, data_21)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(aov(weight ~ Diet + weight_initial, data_21))
## W = 0.99117, p-value = 0.9792
```

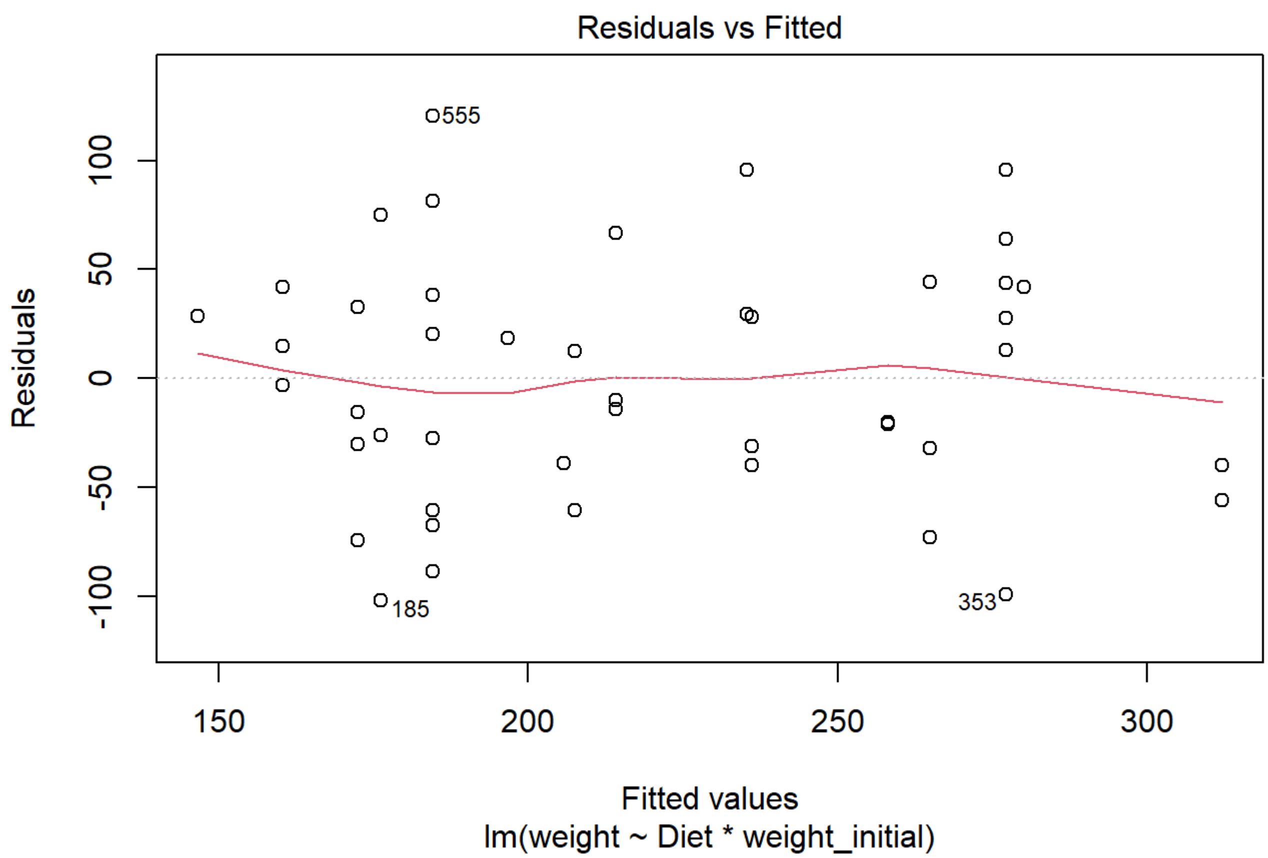
```
# homoscedasticity, outliers
plot(lm(weight ~ Diet * weight_initial, data = data_16), which = 1)
```




```
plot(lm(weight ~ Diet * weight_initial, data = data_20), which = 1)
```




```
plot(lm(weight ~ Diet * weight_initial, data = data_21), which = 1)
```




```
bartlett.test(weight ~ Diet,data_16)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  weight by Diet
## Bartlett's K-squared = 4.4411, df = 3, p-value = 0.2176
```

```
bartlett.test(weight ~ Diet,data_20)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  weight by Diet
## Bartlett's K-squared = 3.2498, df = 3, p-value = 0.3547
```

```
bartlett.test(weight ~ Diet,data_21)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  weight by Diet
## Bartlett's K-squared = 3.0524, df = 3, p-value = 0.3836
```

```r
# Parallelism
library(car)
```

```
## Warning: package 'car' was built under R version 4.2.3
```

```r
Anova(aov(weight~Diet*weight_initial, data_16))
```

```
## Anova Table (Type II tests)
##
## Response: weight
##                     Sum Sq Df F value  Pr(>F)
## Diet                 14578  3  2.5532 0.06937 .
## weight_initial        4201  1  2.2073 0.14540
## Diet:weight_initial   2787  3  0.4882 0.69250
## Residuals            74228 39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(aov(weight~Diet*weight_initial, data_20))
```

```
## Anova Table (Type II tests)
##
## Response: weight
##                     Sum Sq Df F value   Pr(>F)
## Diet                 42138  3  4.4674 0.008784 **
## weight_initial        6672  1  2.1219 0.153417
## Diet:weight_initial  17043  3  1.8069 0.162349
## Residuals           119476 38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(aov(weight~Diet*weight_initial, data_21))
```

```
## Anova Table (Type II tests)
##
## Response: weight
##                     Sum Sq Df F value  Pr(>F)
## Diet                 43763  3  4.0730 0.01349 *
## weight_initial        7137  1  1.9926 0.16642
## Diet:weight_initial  28185  3  2.6232 0.06495 .
## Residuals           132517 37
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q3_Answer_ANCOVA:

Linearity: there is obvious curvature for day 16, 20, and 21 in three scatter plots. Linearity assumptions are not satisfied.

Normality: by Shapiro test, normality assumptions are satisfied.

Homoscedasticity: from residual versus fitted value plots and bartlett test, homoscedasticity are satisfied without significant outlines.

Paralleism: because the interaction term is not significant, the regression slopes are roughly equal. There are no parallelism.

In sum, only linearity is not satisfied for ANCOVA test.

```
#check for ANOVA
#sphericity
library("rstatix")
```

```
## Warning: package 'rstatix' was built under R version 4.2.3
```

```
##
## Attaching package: 'rstatix'
```

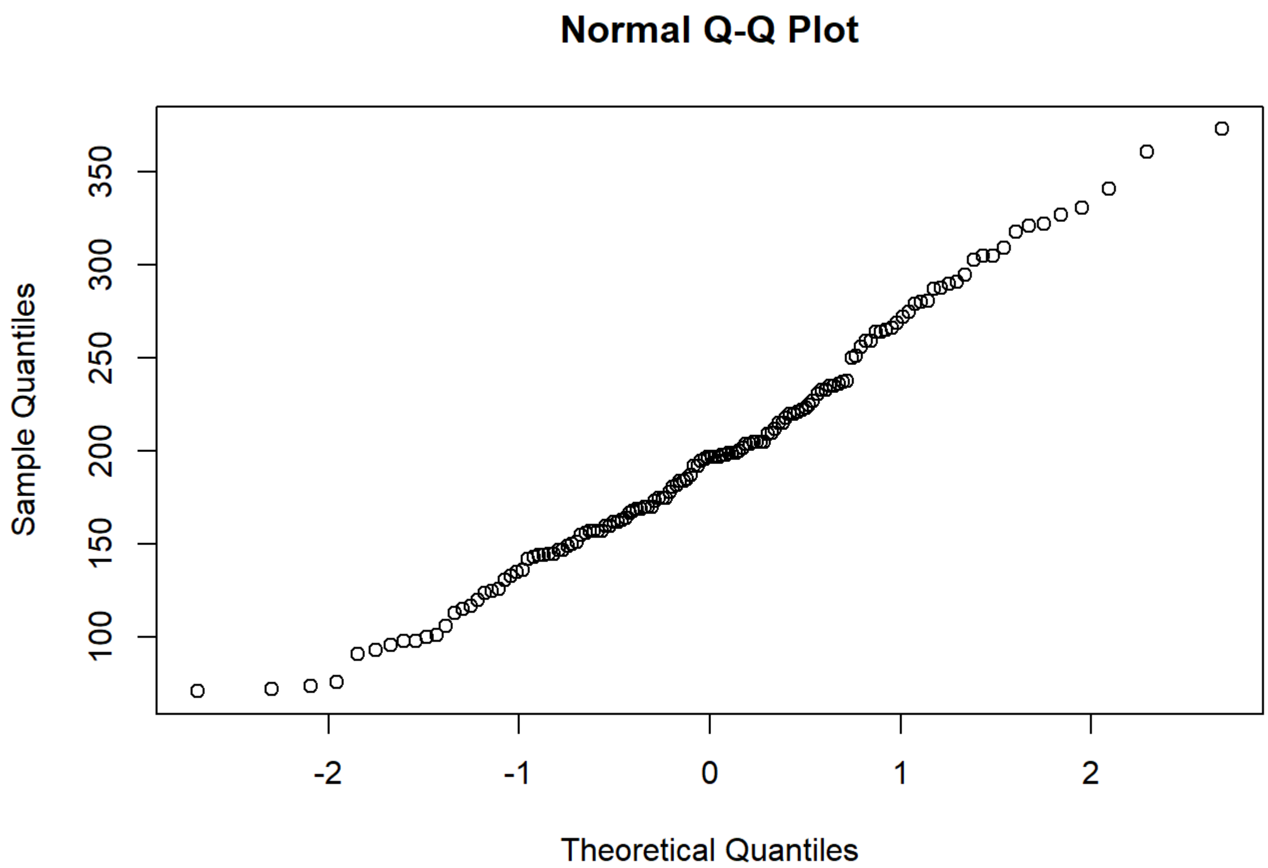
```
## The following object is masked from 'package:stats':
##
##     filter
```

```
anova_test(data = data_repeated, dv = weight, wid = Chick, between=Diet, within = Time)
```

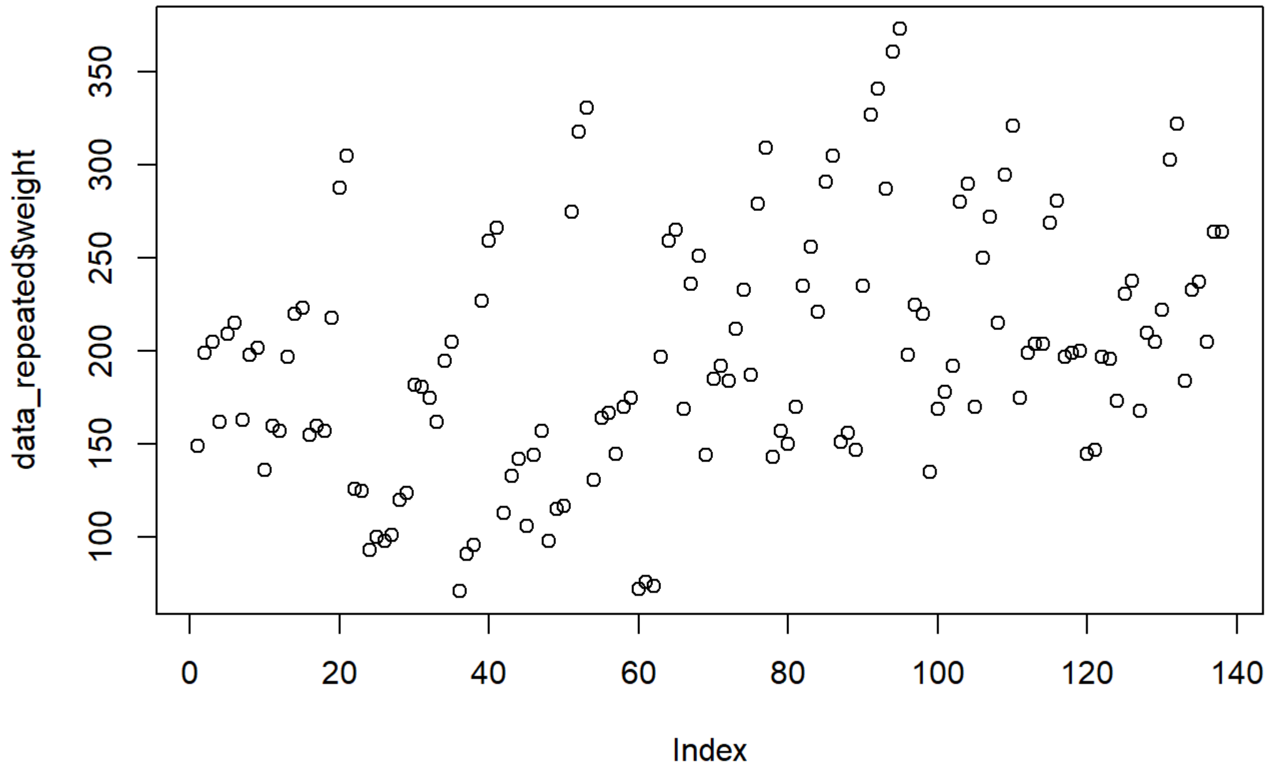
```
## ANOVA Table (type III tests)
##
## $ANOVA
##      Effect DFn DFd       F       p p<.05   ges
## 1      Diet   3  41   4.525 8.0e-03     * 0.238
## 2      Time   2  82 125.532 1.1e-25     * 0.148
## 3 Diet:Time   6  82   3.811 2.0e-03     * 0.016
##
## $`Mauchly's Test for Sphericity`
##      Effect     W        p p<.05
## 1      Time 0.091 1.45e-21     *
## 2 Diet:Time 0.091 1.45e-21     *
##
## $`Sphericity Corrections`
##      Effect   GGe      DF[GG]    p[GG] p[GG]<.05   HFe     DF[HF]    p[HF]
## 1      Time 0.524 1.05, 42.95 1.31e-14         * 0.526 1.05, 43.1 1.19e-14
## 2 Diet:Time 0.524 3.14, 42.95 1.50e-02         * 0.526 3.15, 43.1 1.50e-02
##   p[HF]<.05
## 1         *
## 2         *
```

```
# normality of y
hist(data_repeated$weight)
```

```
qqnorm(data_repeated$weight)
```




```
#outlier
plot(data_repeated$weight)
```




```r
library("outliers")
grubbs.test(data_repeated$weight)
```

```
##
##  Grubbs test for one outlier
##
## data:  data_repeated$weight
## G = 2.65315, U = 0.94824, p-value = 0.5027
## alternative hypothesis: highest value 373 is an outlier
```

```r
library("rstatix")
identify_outliers(weight,data=data_repeated)
```

```
##   weight Time Chick Diet is.outlier is.extreme
## 1    361   20    35    3       TRUE      FALSE
## 2    373   21    35    3       TRUE      FALSE
```

Q3_Answer_repeated measure_sphericity+normality+outlier:

The assumptions of a repeated measures ANOVA are that the continuous dependent variable is approximately **normally** distributed is not satisfied (histogram). The categorical independent variable “Diet” has equal or more than three levels is satisfied. No **outlier** is not satisfied (Grubbs test). **Sphericity** is not satisfied, variance of group difference are not equal (Mauchly’s test for sphericity).

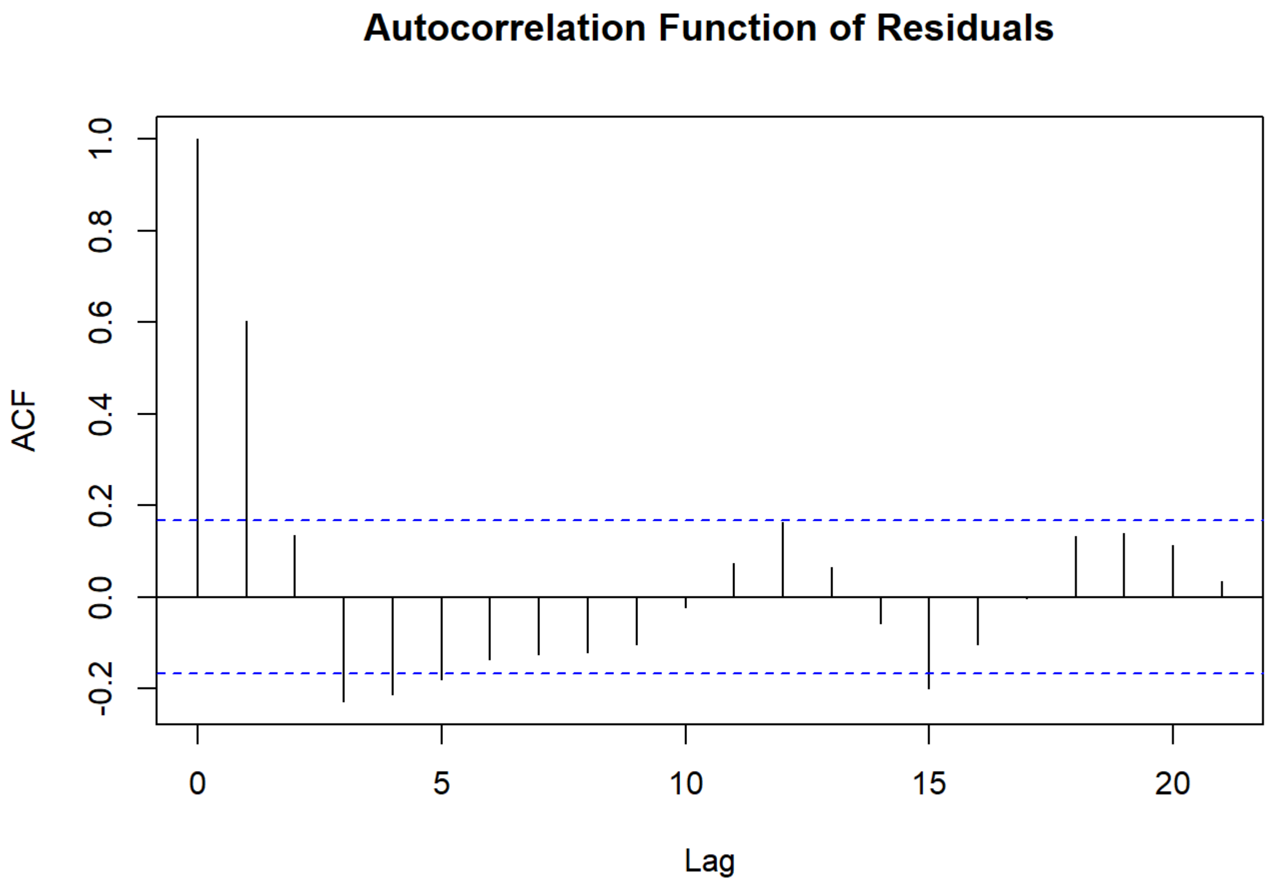
```
# AR(1)
library(nlme)
model_ar1 <- lme(weight ~ Diet * Time, random = ~1 | Chick, correlation = corAR1(form = ~Time |
Chick), data = data_repeated)

summary(model_ar1)
```

```
## Linear mixed-effects model fit by REML
##   Data: data_repeated
##        AIC      BIC   logLik
##   1232.606 1264.149 -605.303
##
## Random effects:
##  Formula: ~1 | Chick
##         (Intercept) Residual
## StdDev:  0.01589593 54.13759
##
## Correlation Structure: ARMA(1,0)
##  Formula: ~Time | Chick
##  Parameter estimate(s):
##      Phi1
## 0.9805524
## Fixed effects:  weight ~ Diet * Time
##                  Value Std.Error DF   t-value p-value
## (Intercept)  47.37236  24.69681 87  1.918157  0.0584
## Diet2       -42.65283  40.49514 43 -1.053283  0.2981
## Diet3       -83.18338  40.49514 43 -2.054157  0.0461
## Diet4       -34.31618  41.22278 43 -0.832457  0.4098
## Time          6.08175   1.14272 87  5.322191  0.0000
## Diet2:Time    3.91833   1.87050 87  2.094803  0.0391
## Diet3:Time    8.49854   1.87050 87  4.543457  0.0000
## Diet4:Time    4.48483   1.93159 87  2.321831  0.0226
##  Correlation:
##            (Intr) Diet2  Diet3  Diet4  Time   Dt2:Tm Dt3:Tm
## Diet2      -0.610
## Diet3      -0.610  0.372
## Diet4      -0.599  0.365  0.365
## Time       -0.855  0.521  0.521  0.512
## Diet2:Time  0.522 -0.854 -0.318 -0.313 -0.611
## Diet3:Time  0.522 -0.318 -0.854 -0.313 -0.611  0.373
## Diet4:Time  0.506 -0.308 -0.308 -0.859 -0.592  0.361  0.361
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.59932595 -0.63673113  0.01986824  0.64761106  2.39964250
##
## Number of Observations: 138
## Number of Groups: 47
```

```
residuals_ar1 <- residuals(model_ar1)

# Plot the autocorrelation function (ACF) of residuals
acf(residuals_ar1, main = "Autocorrelation Function of Residuals")
```



Q3_Answer_repeated measure_AR(1):

The estimated autoregressive coefficient (phi1), presented as 0.9805524, is significantly different from zero and close to 1, suggests AR(1) is satisfied.

```
# unstructured
library(nlme)
model_unstructured <- lme(weight ~ Diet * Time, random = ~1 | Chick, correlation = corSymm(form = ~1 | Chick), data = data_repeated)
summary(model_unstructured)
```

```
## Linear mixed-effects model fit by REML
##   Data: data_repeated
##        AIC      BIC    logLik
##   1203.984 1241.262 -588.9921
##
## Random effects:
##  Formula: ~1 | Chick
##         (Intercept) Residual
## StdDev:    49.62113  23.7682
##
## Correlation Structure: General
##  Formula: ~1 | Chick
##  Parameter estimate(s):
##  Correlation:
##   1     2
## 2 0.533
## 3 0.113 0.899
## Fixed effects:  weight ~ Diet * Time
##                  Value Std.Error DF   t-value p-value
## (Intercept)   27.01113  25.88408 87  1.043542  0.2996
## Diet2        -33.05406  42.40380 43 -0.779507  0.4400
## Diet3       -100.84947  42.40380 43 -2.378312  0.0219
## Diet4        -80.16674  43.47281 43 -1.844066  0.0721
## Time           6.92284   1.34153 87  5.160409  0.0000
## Diet2:Time     3.53818   2.20166 87  1.607048  0.1117
## Diet3:Time     9.28610   2.20166 87  4.217766  0.0001
## Diet4:Time     6.56802   2.27652 87  2.885109  0.0049
##  Correlation:
##            (Intr) Diet2  Diet3  Diet4  Time   Dt2:Tm Dt3:Tm
## Diet2      -0.610
## Diet3      -0.610  0.373
## Diet4      -0.595  0.363  0.363
## Time       -0.883  0.539  0.539  0.526
## Diet2:Time  0.538 -0.883 -0.328 -0.320 -0.609
## Diet3:Time  0.538 -0.328 -0.883 -0.320 -0.609  0.371
## Diet4:Time  0.520 -0.318 -0.318 -0.888 -0.589  0.359  0.359
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.2212695 -0.2939110  0.3350699  0.9093235  2.1316186
##
## Number of Observations: 138
## Number of Groups: 47
```

Q3_Answer_repeated measure_(unstructure):

In correlation table, off-diagonal elements represent the covariances and correlations between different measurements within the same grouping variable. For each row, we can see correlations take on wide range of value. Unstructured covariance assumption is satisfied.

Q3_Comment on ANCOVA and ANOVA in Q1 and Q2:

Q2 repeated Anova accounts for within-subject variations over multiple timepoints and can detect time-related changes and interactions. Q1 ANCOVA focuses on each single timepoint seperately, accessing the impact of diet and birth weight to weight. They give out different results. We use them depending on the situation and hypothesis that we want to test.