HW 7

SDS348 Spring 2021

Name: Santhosh Saravanan

EID: sks3648

This homework is due on April 5, 2021 at 8am. Submit a pdf file on Gradescope.

For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.

Question 1: (13 pts)

Recall the built-in dataset `msleep` (in `ggplot2`) about the sleeping habits and other characteristics of 83 mammals.

1.1 (3 pts) Fit a regression model with the REM sleep (`sleep_rem`, in hours) as the response, and use the variables for brain weight (`brainwt` in kg) and the type of diet (`vore`) as predictors, as well as their interaction. Interpret **in context** (regardless of the significance of the coefficient estimates):

a. the intercept,

b. the coefficient of `brainwt`,

c. the coefficient for `voreinsecti`,

d. the coefficient for `brainwt:voreinsecti`.

```
library(ggplot2)
# Fit a multiple linear regression model with both predictors
fit <- lm(sleep_rem ~ brainwt + vore + brainwt*vore, data = msleep)
mySleep <- msleep
summary(fit)
```

```
##
## Call:
## lm(formula = sleep_rem ~ brainwt + vore + brainwt * vore, data = msleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32515 -0.65057 -0.09685  0.43587  2.80300
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.68511    0.48162   5.575  2.8e-06 ***
## brainwt             -3.42607    3.49182  -0.981  0.33324
## voreherbi           -1.20881    0.55723  -2.169  0.03693 *
## voreinsecti         -0.04121    0.72892  -0.057  0.95524
## voreomni            -0.58505    0.54330  -1.077  0.28891
## brainwt:voreherbi    1.42876    3.69597   0.387  0.70141
## brainwt:voreinsecti 46.01707   13.95225   3.298  0.00224 **
## brainwt:voreomni     2.94134    3.56621   0.825  0.41508
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9408 on 35 degrees of freedom
##   (40 observations deleted due to missingness)
## Multiple R-squared:  0.5087, Adjusted R-squared:  0.4104
## F-statistic: 5.176 on 7 and 35 DF,  p-value: 0.0004084
```

a.The mean REM sleep is 2.68511 hours if an animal has 0kg of brainweight and is a carnivore. b. The mean REM sleep decreases by 3.42607 hours with every unit increase in brainweight and if an animal is a carnivore. c. The mean REM sleep decreases by 0.04121 hours if an animal has 0kg of brain weight and is an insectivore.

```
yInsectivore <- 2.68511 - 3.42607 - 0.04121 + 46.01707 # if animal is an insectivore
yHerbivore <- 2.68511 - 3.42607 - 1.20881 + 1.42876 # if animal isa herbivore
yCarnivore <- 2.68511 - 3.42607  # if animal is a carnivore
yOmnivore <- 2.68511 - 3.42607 - 0.58505 + 2.94134
diffInsectiOmni <- yInsectivore - yOmnivore
diffInsectiCarni <- yInsectivore - yCarnivore
diffInsectiHerbi <- yInsectivore - yHerbivore
```

d. The difference in mean REM sleep is 43.39962 hours if an animal is an insectivore compared to when an animal is an Omnivore. The difference in mean REM sleep is 45.97586 hours if an animal is an insectivore compared to when an animal is a carnivore. The difference in mean REM sleep is 45.75591 hours if an animal is an insectivore compared to when an animal is a herbivore.

1.2 (3 pts) Fit the same regression model as previously, but center the `brainwt` variable first by subtracting the mean to each observation (using `na.rm = TRUE`). Which coefficients that you interpreted in the previous question (1.1) have changed? Why? Reinterpret any coefficient from question 1.1 that has changed.

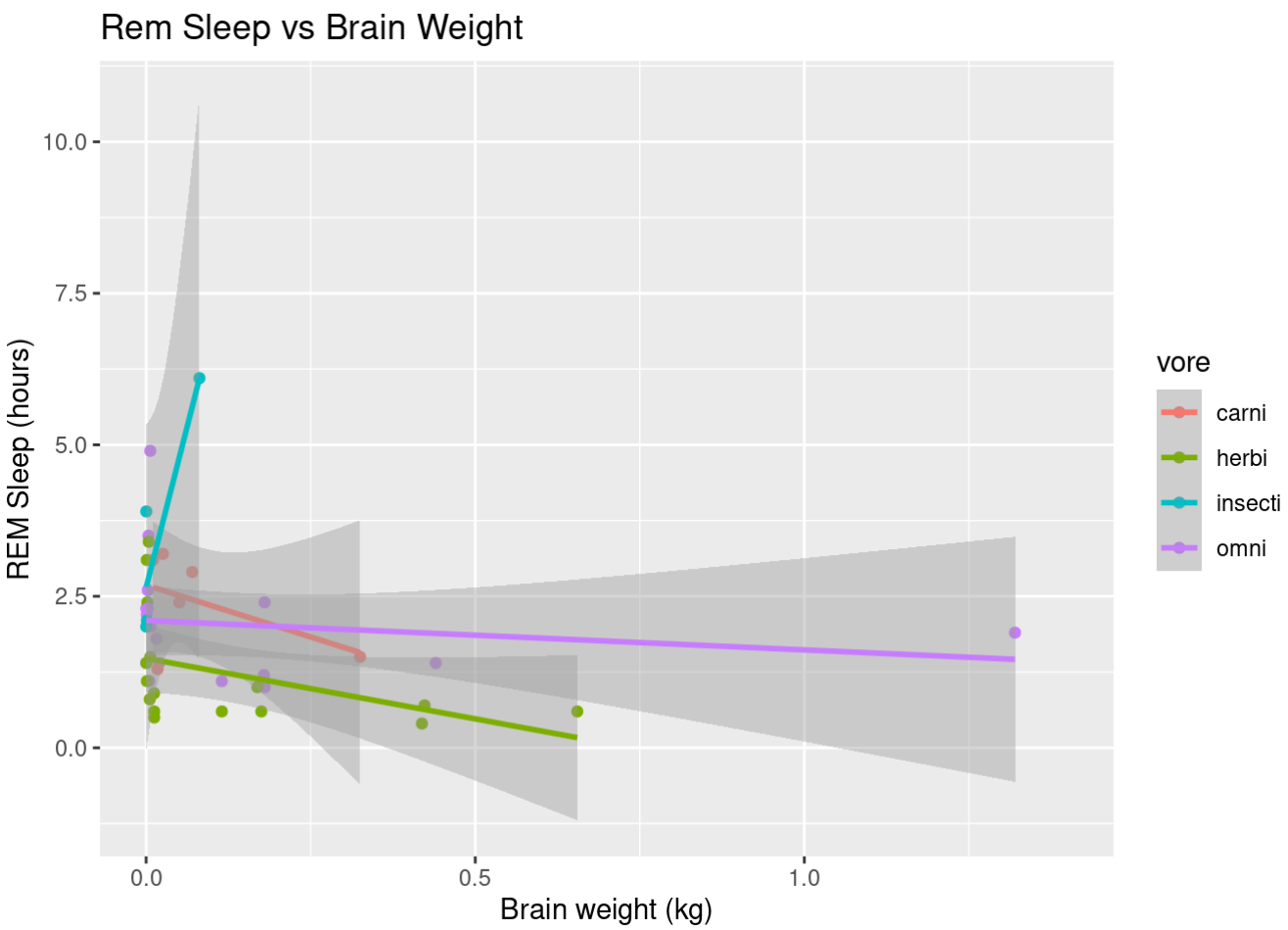
```
mySleep$brainwt <- mySleep$brainwt - mean(mySleep$brainwt,na.rm = TRUE)
fit <- lm(sleep_rem ~ brainwt + vore + brainwt*vore, data = mySleep)
summary(fit)
```

```
##
## Call:
## lm(formula = sleep_rem ~ brainwt + vore + brainwt * vore, data = mySleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32515 -0.65057 -0.09685  0.43587  2.80300
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.7204     0.7920   2.172  0.03670 *
## brainwt              -3.4261     3.4918  -0.981  0.33324
## voreherbi            -0.8065     0.8475  -0.952  0.34780
## voreinsecti          12.9163     3.6426   3.546  0.00113 **
## voreomni              0.2432     0.8301   0.293  0.77129
## brainwt:voreherbi     1.4288     3.6960   0.387  0.70141
## brainwt:voreinsecti  46.0171    13.9522   3.298  0.00224 **
## brainwt:voreomni      2.9413     3.5662   0.825  0.41508
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9408 on 35 degrees of freedom
##   (40 observations deleted due to missingness)
## Multiple R-squared:  0.5087, Adjusted R-squared:  0.4104
## F-statistic: 5.176 on 7 and 35 DF,  p-value: 0.0004084
```

a.The mean REM sleep is 1.7204 hours if an animal has 0kg of brainweight and is a carnivore. b. The mean REM sleep decreases by 3.42607 hours with every unit increase in brainweight and if an animal is a carnivore. c. The mean REM sleep increases by 12.9163 hours if an animal has 0kg of brain weight and is an insectivore.

1.3 (3 pts) Remove missing values for the `vore` variable only. Make a plot of `sleep_rem` by `brainwt` and explore the relationship between these two variables for different types of diets (using color and `geom_smooth(method = "lm")`). To make it more readable, set the limits of x-axis between 0 and 1.4 (using `xlim(,)`). What is the mean value of brain weight? Does it make sense to interpret the coefficient estimate of insectivores in terms of the mean value of brain weight? Why/Why not?

```
mySleep <- msleep
mySleep <- mySleep[!is.na(mySleep$vore),]
ggplot(mySleep,aes(y=sleep_rem,x=brainwt,color=vore)) +
  geom_point() + xlim(0,1.4) + geom_smooth(method = "lm") + xlab("Brain weight (kg)") + ylab("R
EM Sleep (hours)") + ggtitle("Rem Sleep vs Brain Weight")
```




```
print(mean(mySleep$brainwt,na.rm = TRUE))
```

```
## [1] 0.3084398
```

The mean value of brain weight is 0.3084398 kg. It doesn't make sense to interpret the coefficient estimates of insectivores in terms of mean value of brain weight simply because the trend observed in the above graph doesn't necessarily follow the same trend as the other diets for 1.1 and 1.2.

1.4 (2 pts) Consider the natural log of the variable `brainwt` , then center the log variable (*Note: you can't just take the log of the centered variable*) and then fit a model with that centered log variable, the `vore` variable, and the interaction. Interpret the most significant effect and discuss your decision with respect to the null hypothesis.

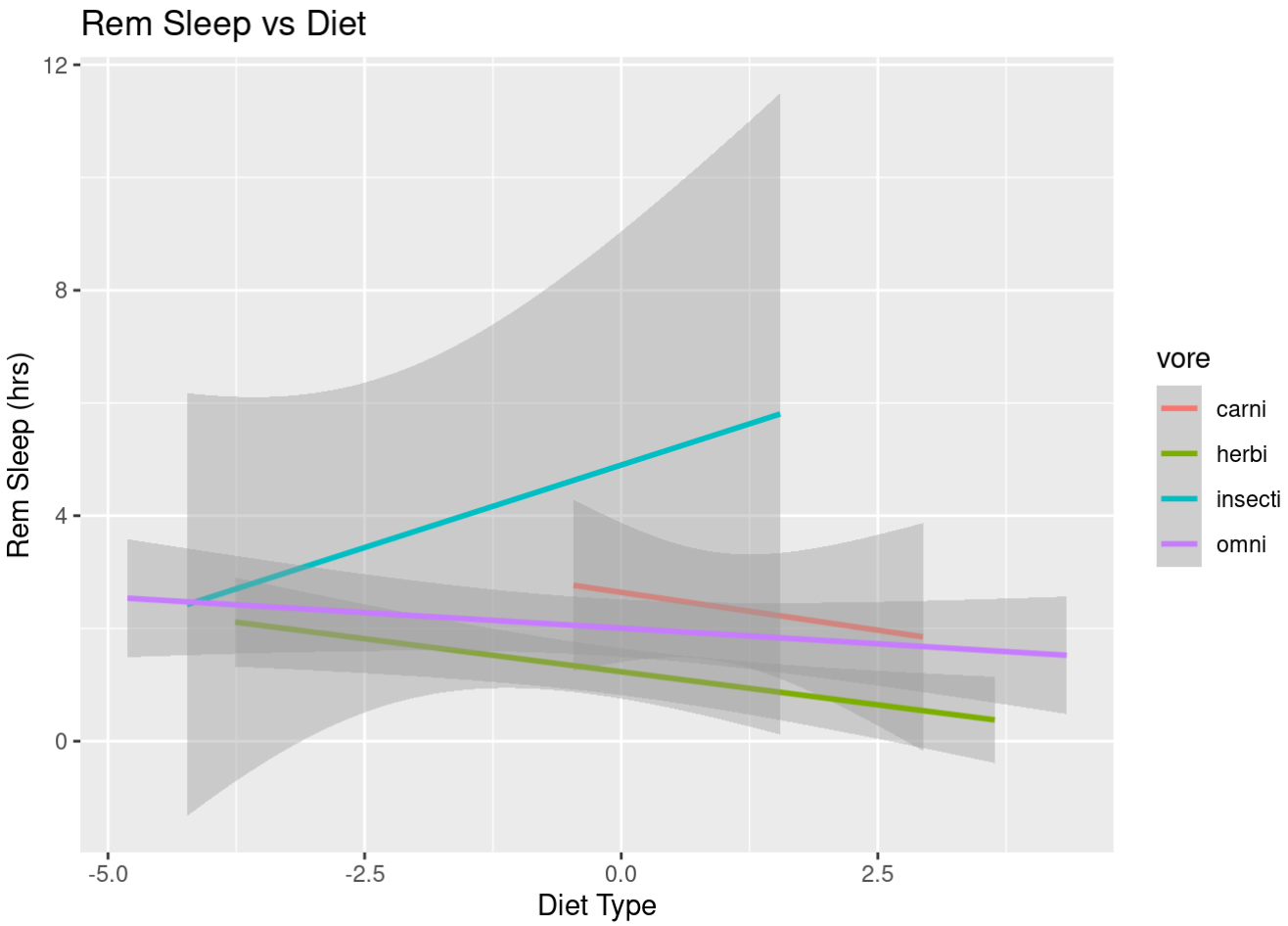
```
mySleep <- msleep
mySleep$brainwt <- log(mySleep$brainwt)
mySleep$brainwt <- mySleep$brainwt - mean(mySleep$brainwt,na.rm = TRUE)
fit <- lm(sleep_rem ~ brainwt + vore + brainwt*vore, data = mySleep)
summary(fit)
```

```
##
## Call:
## lm(formula = sleep_rem ~ brainwt + vore + brainwt * vore, data = mySleep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3355 -0.6414 -0.1108  0.3551  2.7841
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.64015    0.48074   5.492 3.61e-06 ***
## brainwt             -0.26848    0.33800  -0.794  0.43236
## voreherbi           -1.40873    0.53246  -2.646  0.01213 *
## voreinsecti          2.25937    0.80833   2.795  0.00837 **
## voreomni            -0.63533    0.52990  -1.199  0.23860
## brainwt:voreherbi    0.03404    0.35198   0.097  0.92350
## brainwt:voreinsecti  0.85380    0.39093   2.184  0.03575 *
## brainwt:voreomni     0.15788    0.34911   0.452  0.65390
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9155 on 35 degrees of freedom
##   (40 observations deleted due to missingness)
## Multiple R-squared:  0.5347, Adjusted R-squared:  0.4416
## F-statistic: 5.746 on 7 and 35 DF,  p-value: 0.0001762
```

The null hypothesis is that there is no significant correlation between REM sleep and any of the predictor variables. The fit of the observed REM sleep values compared to what is predicted by the multiple regression equation is no better than what one would expect by chance. Given that the p-value is less than 0.05, we reject the null hypothesis and claim that with sufficient evidence, that there is indeed a statistically significant correlation between REM sleep and whether an animal is an insectivore.

1.5 (2 pts) Update your plot from question 1.3 by representing the centered log brain weight on the x-axis. The interaction between the centered log brain weight and which type of diet seem to be the most important? Refer to the previous question to check for significance.

```
mySleep <- msleep
mySleep$brainwt <- log(mySleep$brainwt)
mySleep$brainwt <- mySleep$brainwt - mean(mySleep$brainwt,na.rm = TRUE)
mySleep %>% drop_na(vore) %>% ggplot(aes(brainwt,sleep_rem))+ geom_smooth(aes(col=vore),method=
"lm") + ggtitle("Rem Sleep vs Diet") + xlab("Diet Type") + ylab("Rem Sleep (hrs)")
```



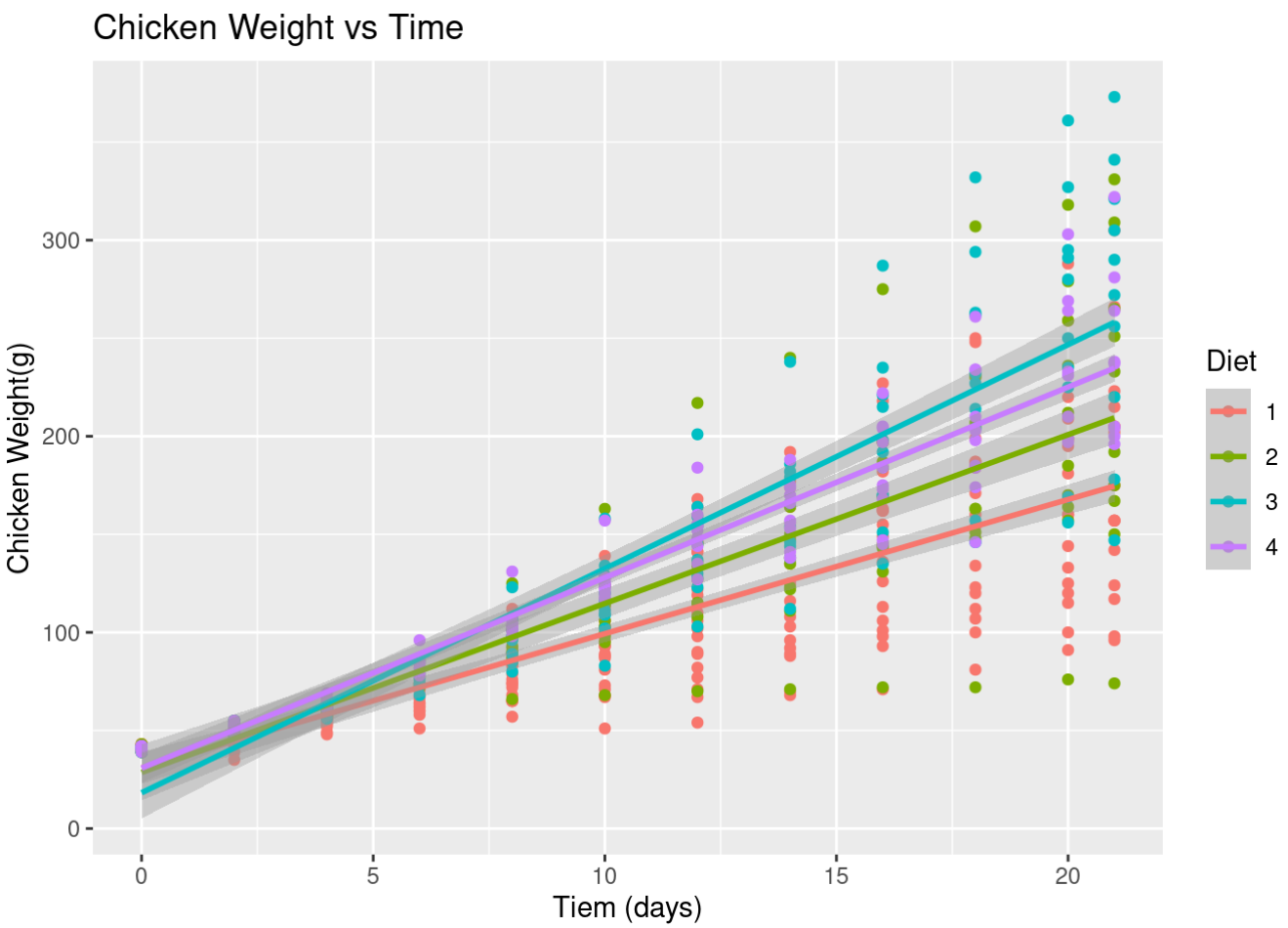
The interaction between the centered log brain weight and the insectivore diet seem to be the most important

Question 2: (12 pts)

Recall the built-in dataset `ChickWeight` about the weights (in grams) of chicks on 4 different diets over time (at 2-day intervals).

2.1 (2 pts) In HW2, question 2.7, you created a scatterplot, representing the weight of chicks over time, and fitting a regression line for each diet. Recreate the graph below. What do you expect in terms of interaction between time and type of diet?

```
ggplot(data=ChickWeight,aes(x=`Time`,y=`weight`,color = Diet)) + geom_point(size=1.5)+ggtitle(
"Chicken Weight vs Time") + labs(y="Chicken Weight(g)",x="Tiem (days)") + geom_smooth(method =
"lm",aes(color = Diet))
```



I expect a somewhat reasonable correlation between time and type of diet.

2.2 (2 pts) Fit a regression model to predict weights of chicks based on the number of days since birth and on the type of diet, including the interaction. Notice that the individual effects of the different diets were not significant while the interactions between types of diets and time are. Why is that so?

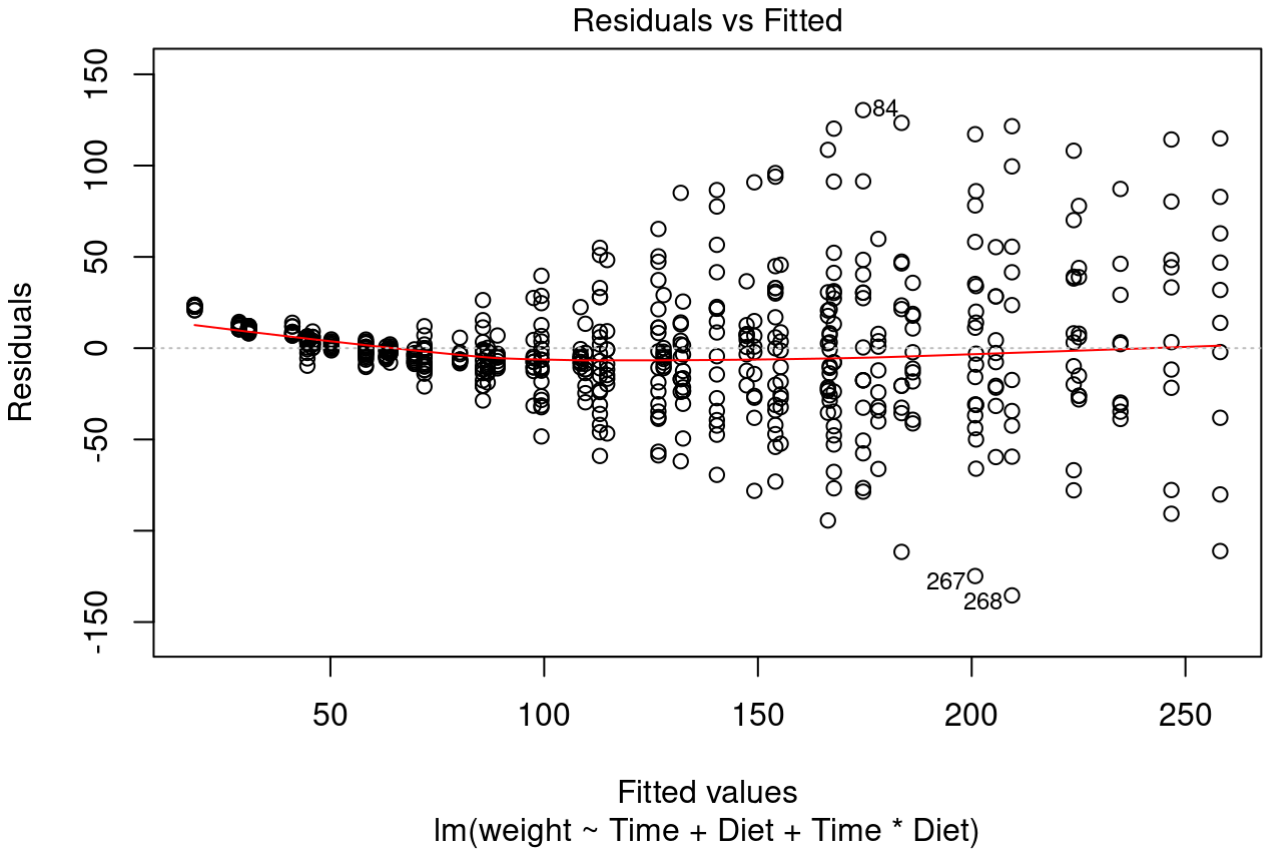
```
fit <- lm(weight ~ Time + Diet + Time*Diet, data = ChickWeight)
summary(fit)
```

```
##
## Call:
## lm(formula = weight ~ Time + Diet + Time * Diet, data = ChickWeight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -135.425  -13.757   -1.311   11.069  130.391
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.9310     4.2468   7.283 1.09e-12 ***
## Time          6.8418     0.3408  20.076  < 2e-16 ***
## Diet2        -2.2974     7.2672  -0.316  0.75202
## Diet3       -12.6807     7.2672  -1.745  0.08154 .
## Diet4        -0.1389     7.2865  -0.019  0.98480
## Time:Diet2    1.7673     0.5717   3.092  0.00209 **
## Time:Diet3    4.5811     0.5717   8.014 6.33e-15 ***
## Time:Diet4    2.8726     0.5781   4.969 8.92e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.07 on 570 degrees of freedom
## Multiple R-squared:  0.773,  Adjusted R-squared:  0.7702
## F-statistic: 277.3 on 7 and 570 DF,  p-value: < 2.2e-16
```

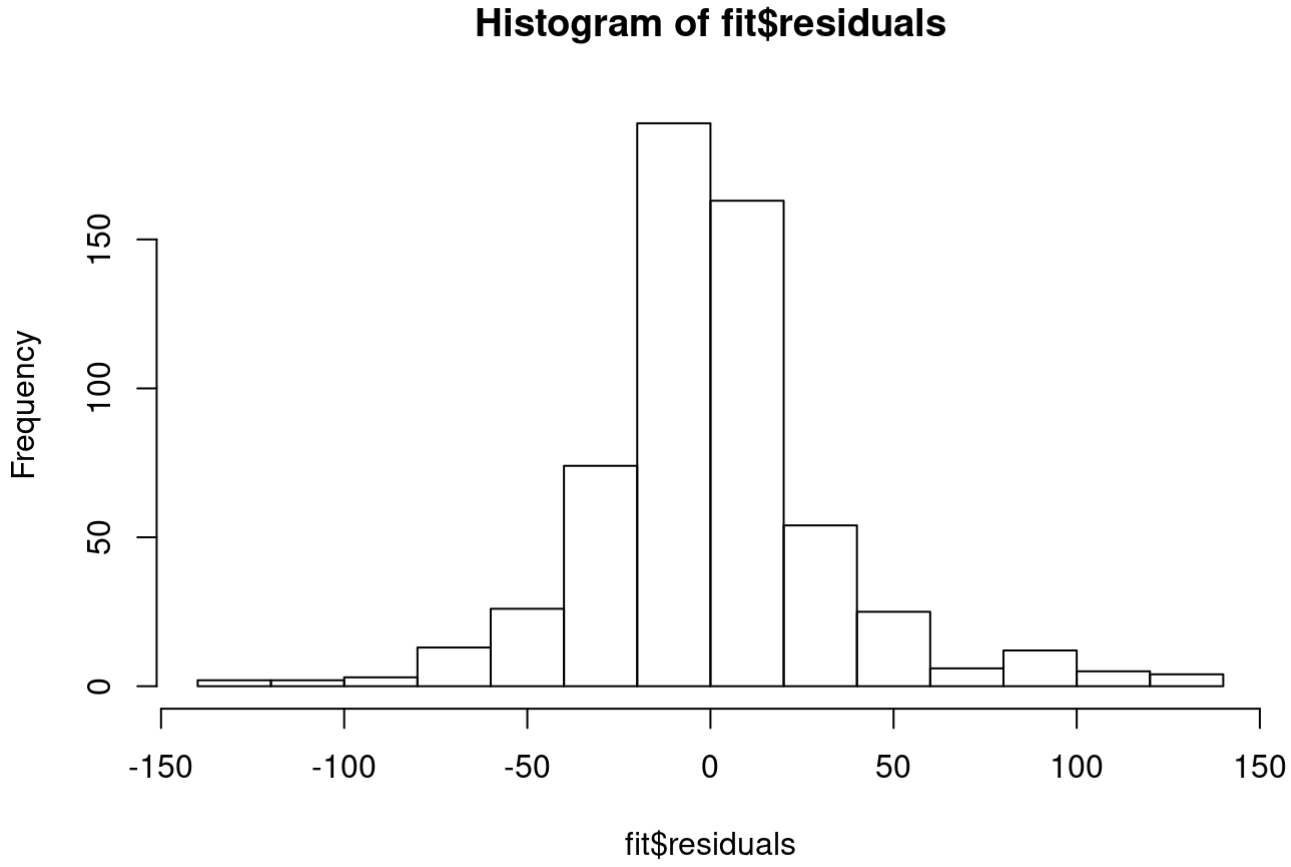
The distinctive means from the different diets cross over each other in different situations. This is known as a cross-over interaction. The different factors of diet have one kind of effect on weight for one condition, but have an opposite effect on weight for another condition

2.3 (3 pts) Check the assumptions visually for the regression model created in question 2.2. Which assumption(s) may or may not be met?

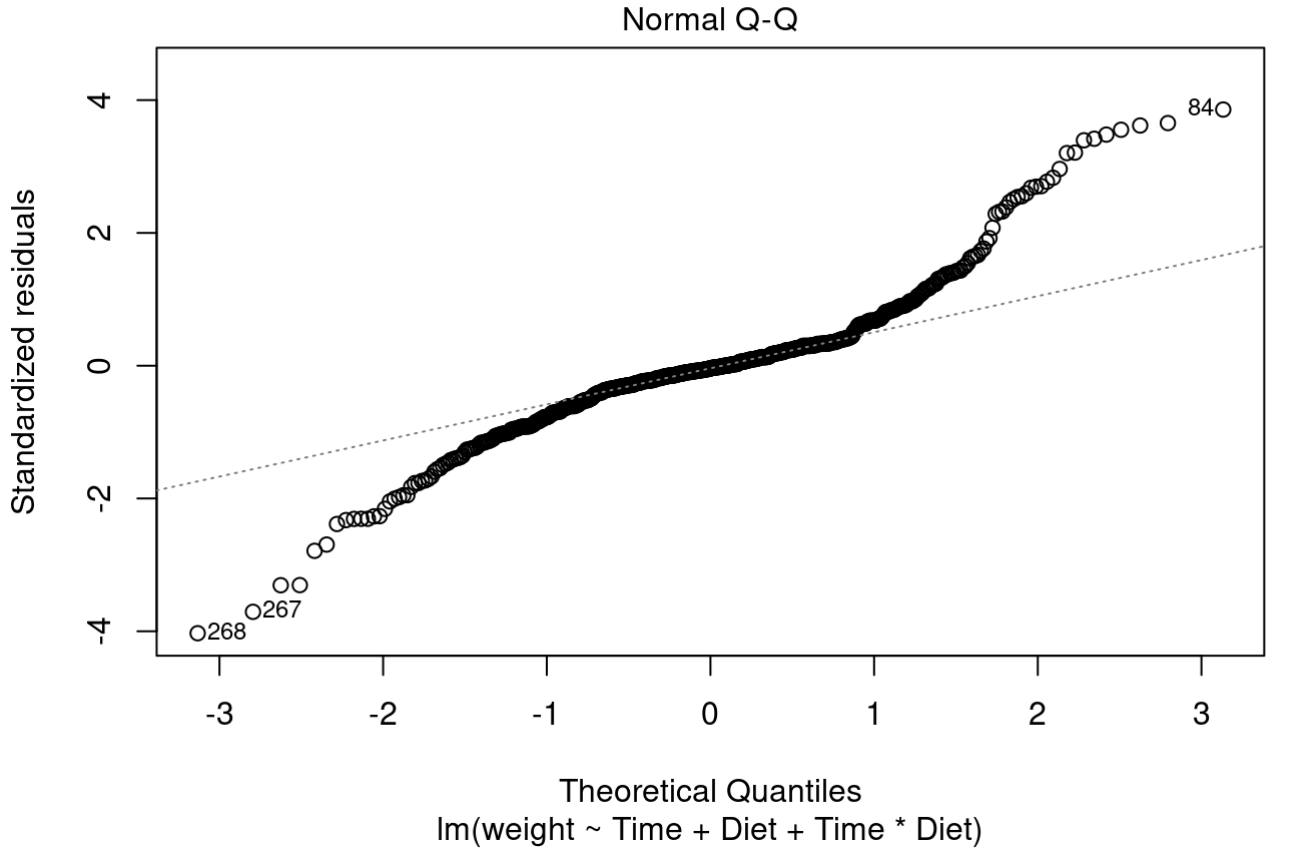
```
# Residuals vs Fitted values plot
plot(fit, which = 1)
```




```
# Histogram of residuals
hist(fit$residuals)
```




```
# Q-Q plot for the residuals
plot(fit, which = 2)
```




```
shapiro.test(fit$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit$residuals
## W = 0.92507, p-value = 2.252e-16
```

```
# Kolmogorov-Smirnov test
# H0: normality
ks.test(fit$residuals, "pnorm", mean=0, sd(fit$residuals))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  fit$residuals
## D = 0.14063, p-value = 2.359e-10
## alternative hypothesis: two-sided
```

Normality assumption has not been met because even though there isn't a pattern in the residuals, not all of the points lie on the straight line in the QQ plot. Also the Shapiro-Wilk Test fails for the residuals so the residuals originated from a non-normal distribution. The Kolmogorov-Smirnov test rejects the null hypothesis that the distribution of the residuals follow the normal distribution. All in all, I believe the normality assumption has not been met.

```
library(sandwich);
# Install a new package
# install.packages("lmtest")
library(lmtest)

# Breusch-Pagan test
# H0: homoscedasticity
bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 142.91, df = 7, p-value < 2.2e-16
```

The equal variance assumption has also not been met with the results of the Breusch-Pagan test.

2.4 (2 pts) Using the regression model from question 2.2, construct a confidence interval for the mean weight of a chick after 15 days following Diet 1 (Note: create a `newdata` with the values to plug in the model and use `predict`). Interpret the confidence interval.

```
timeDietMatrix <-  data.frame(Time=15,Diet="1")
predict(fit,newdata = timeDietMatrix,interval = "confidence")
```

```
##        fit      lwr      upr
## 1 133.5579 128.1267 138.9891
```

We are 95% confident that the true chick weight for the mean weight of a chick after 15 days following Diet 1 lies between 128.1267g and 138.9891g.

2.5 (3 pts) Bootstrap methods can be used in pretty much any situation and are particularly of interest for calculating a confidence interval when some of the assumptions are not met. Create 5000 bootstrap samples (you can use a `for` loop or the function `replicate`) where the weight is sampled with replacement. For each bootstrap sample, fit a regression model (predicting weight based on time and diet, with the interaction effect) and calculate the mean weight of a chick after 15 days following Diet 1 (using `predict`). Using the empirical distribution of these mean weights, find the bootstrap 95% confidence

interval for the mean weight of a chick after 15 days following Diet 1. How is this confidence interval different from the confidence interval in question 2.4?

```
samp_weights <- replicate(5000, {
  # Bootstrap your data (resample observations)
  boot_data <- sample_frac(ChickWeight, replace = TRUE)
  # Fit regression model
  fitboot <- lm(weight ~ Time + Diet + Time*Diet , data = boot_data)
  # Save the coefficients
  coef(fitboot)
  timeDietMatrix <-  data.frame(Time=15,Diet="1")
  predict(fitboot,newdata = timeDietMatrix)[1]
})
```

```
  quantile(samp_weights,0.025)
```

```
##     2.5%
## 126.0794
```

```
  quantile(samp_weights,0.975)
```

```
##    97.5%
## 141.1758
```

The lower bound of this confidence interval is smaller than the lower bound for the interval in 2.4. The upper bound of this confidence interval is bigger than the upper bound for the interval in 2.4.

```
##                                      sysname
##                                      "Linux"
##                                      release
##                              "5.4.0-70-generic"
##                                      version
## "#78-Ubuntu SMP Fri Mar 19 13:29:52 UTC 2021"
##                                     nodename
##                          "MechaChungus-linux64"
##                                      machine
##                                     "x86_64"
##                                        login
##                                "OmniLordSanta"
##                                         user
##                                "OmniLordSanta"
##                                 effective_user
##                                "OmniLordSanta"
```