HW 7

SDS348 Spring 2021

Rose Hedderman EID: rrh2298

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`.

```
# load ggplot2 library
library(ggplot2)
# get dataset msleep
data(msleep)

# regression model with interaction
fit <- lm(formula = sleep_rem ~ brainwt * vore, data = msleep)
summary(fit)
```

```
##
## Call:
## lm(formula = sleep_rem ~ 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. When the diet doesn't matter and brainwt is 0, the expected rem sleep is 2.685 hours (t = 5.575, df = 35 , p = 2.8e-06). b. For every 1 kg the brainwt increases, REM sleep decreases by 3.426 hours (t = -0.981, df = 35 , p = 0.333). c. If the subject is an insectivore, REM sleep decreases by 0.041 hours (t = -0.057, df = 35 , p = 0.955). d. The slope for diet on brain weight is 46.017 higher for insectivores compared to the other diets (t = 3.298, df = 35 , p = 0.002).

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.

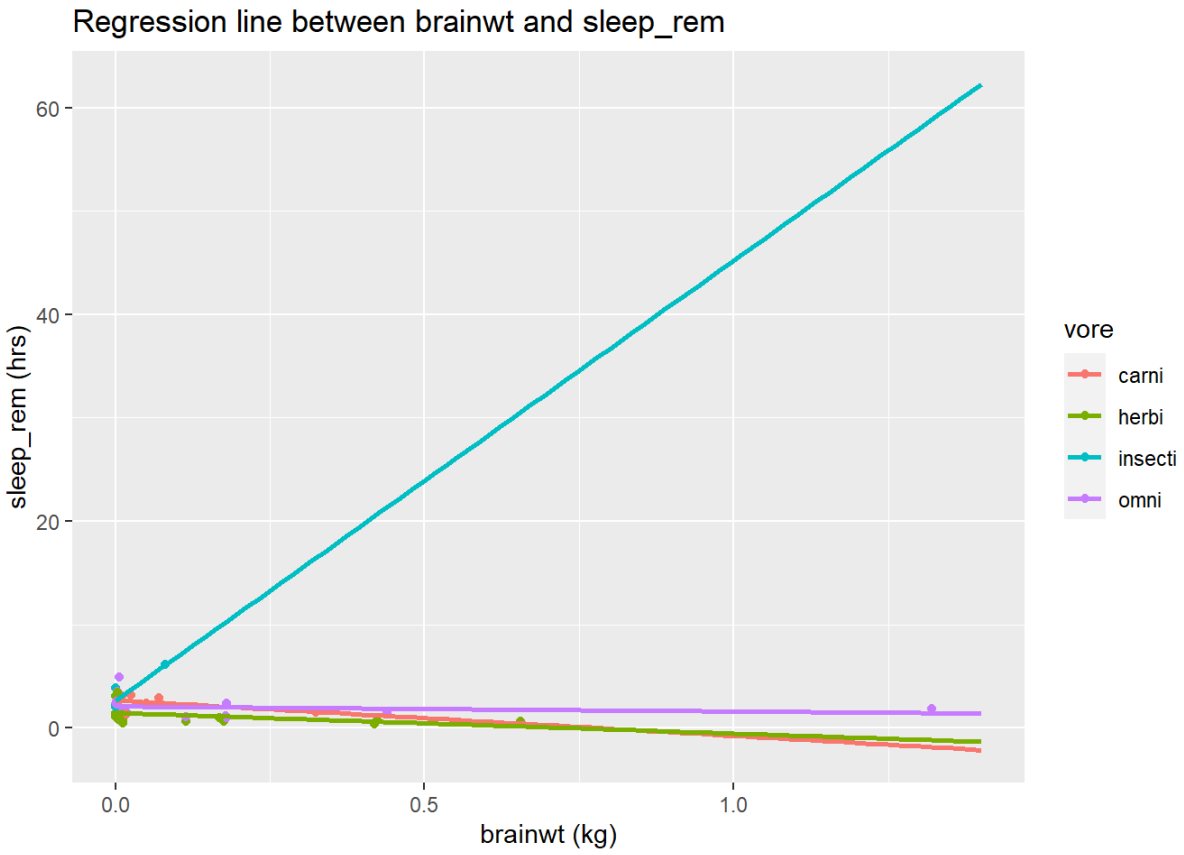
```
# subtract the mean to each observation
msleep$brainwt_c <- msleep$brainwt - mean(msleep$brainwt, na.rm = T)

# regression model with centered brainwt variable
fit2 <- lm(formula = sleep_rem ~ brainwt_c * vore, data = msleep)
summary(fit2)
```

```
##
## Call:
## lm(formula = sleep_rem ~ brainwt_c * 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)             1.7204     0.7920   2.172  0.03670 *
## brainwt_c              -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_c:voreherbi     1.4288     3.6960   0.387  0.70141
## brainwt_c:voreinsecti  46.0171    13.9522   3.298  0.00224 **
## brainwt_c: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
```

The intercept and voreinsecti changed. Intercept: When the diet doesn't matter and centered brainwt is 0, the expected rem sleep is 1.7204 hours (t = 2.172, df = 35 , p = 0.03670). voreinsecti: If the subject is an insectivore, REM sleep increases by 12.9163 hours (t = 3.546 , df = 35 , p = 0.00113).

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?

```
# remove missing values of vore
msleep <- msleep %>%
  filter(!is.na(vore))

# plot sleep_rem and brainwt for differents diets (vore)
ggplot(msleep, aes(x = brainwt, y = sleep_rem, color = vore)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)+
  # set limits on x axis btween 0,1.4
  xlim(0,1.4) +
  labs(title ="Regression line between brainwt and sleep_rem",
       x = "brainwt (kg)", y = "sleep_rem (hrs)")
```




```
mean(msleep$brainwt, na.rm = T)
```

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

The mean value of brain weight is 0.308 kgs. It does not make sense to interpret the coefficient estimate of insectivores in terms of the mean value of brain weight because it follows a different trend than the other 3 diets.

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.

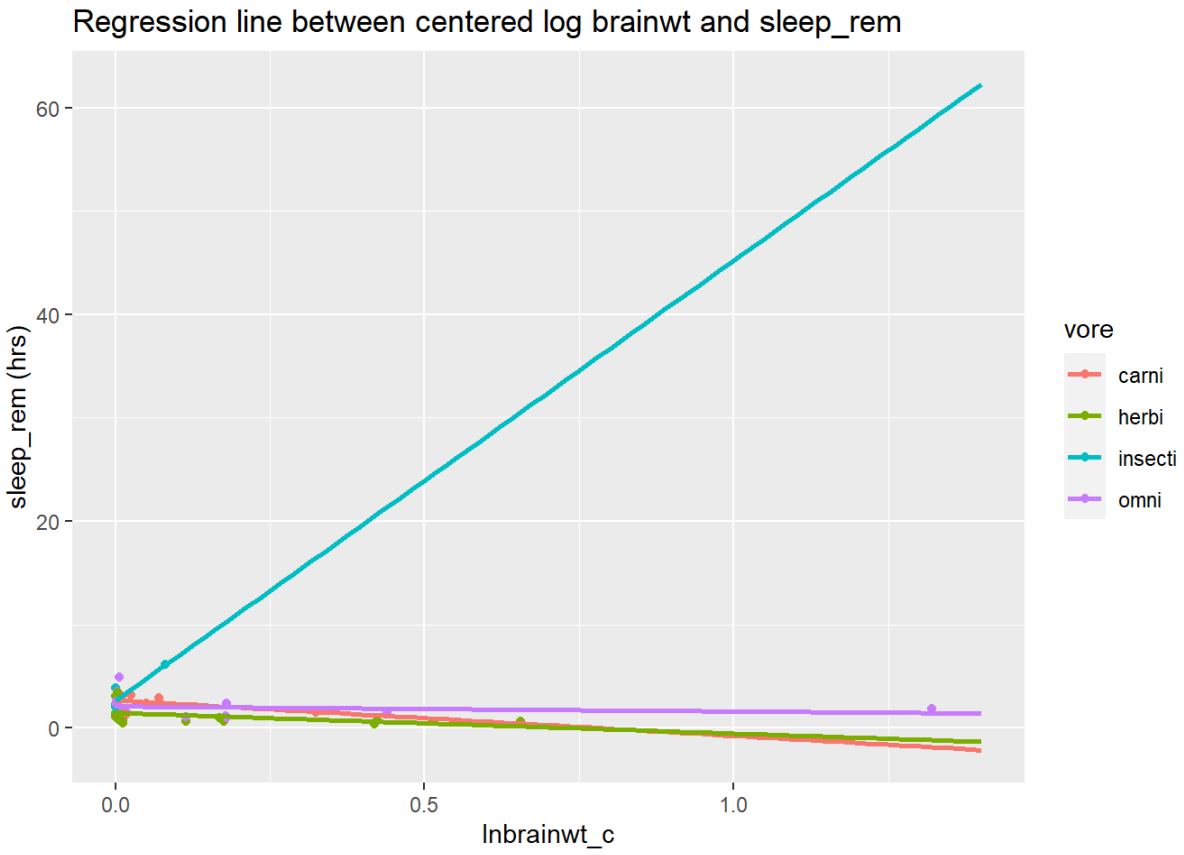
```
# consider the natural log of brainwt
msleep <- msleep %>%
  mutate(lnbrainwt = log(brainwt))

# center the log variable
msleep$lnbrainwt_c <- msleep$lnbrainwt - mean(msleep$lnbrainwt, na.rm = T)

# fit a model with centered log variable, vore, and interaction
fit3 <- lm(formula = sleep_rem ~ lnbrainwt_c * vore, data = msleep)
summary(fit3)
```

```
##
## Call:
## lm(formula = sleep_rem ~ lnbrainwt_c * vore, data = msleep)
##
## 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)              2.59866    0.44973   5.778 1.51e-06 ***
## lnbrainwt_c             -0.26848    0.33800  -0.794  0.43236
## voreherbi               -1.40347    0.50478  -2.780  0.00868 **
## voreinsecti              2.39131    0.80837   2.958  0.00552 **
## voreomni                -0.61094    0.50264  -1.215  0.23233
## lnbrainwt_c:voreherbi    0.03404    0.35198   0.097  0.92350
## lnbrainwt_c:voreinsecti  0.85380    0.39093   2.184  0.03575 *
## lnbrainwt_c: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
##   (33 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 most significant effect is that of the intercept for when all other variables go to 0 or are disregarded, the expected REM sleep is 2.598 hours (t = 5.778, df = 35, p = 1.51e-06). The p-value is 1.51e-06 which is practically 0 and much smaller than 0.05 which allows us to conclude that there is no significant interaction between the intercept and REM sleep.

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.

```
# update plot from question 1.3 using brainwt_log instead of brainwt
ggplot(msleep, aes(x = brainwt, y = sleep_rem, color = vore)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)+
  # set limits on x axis btween 0,1.4
  xlim(0,1.4) +
  labs(title ="Regression line between centered log brainwt and sleep_rem",
       x = "lnbrainwt_c", y = "sleep_rem (hrs)")
```



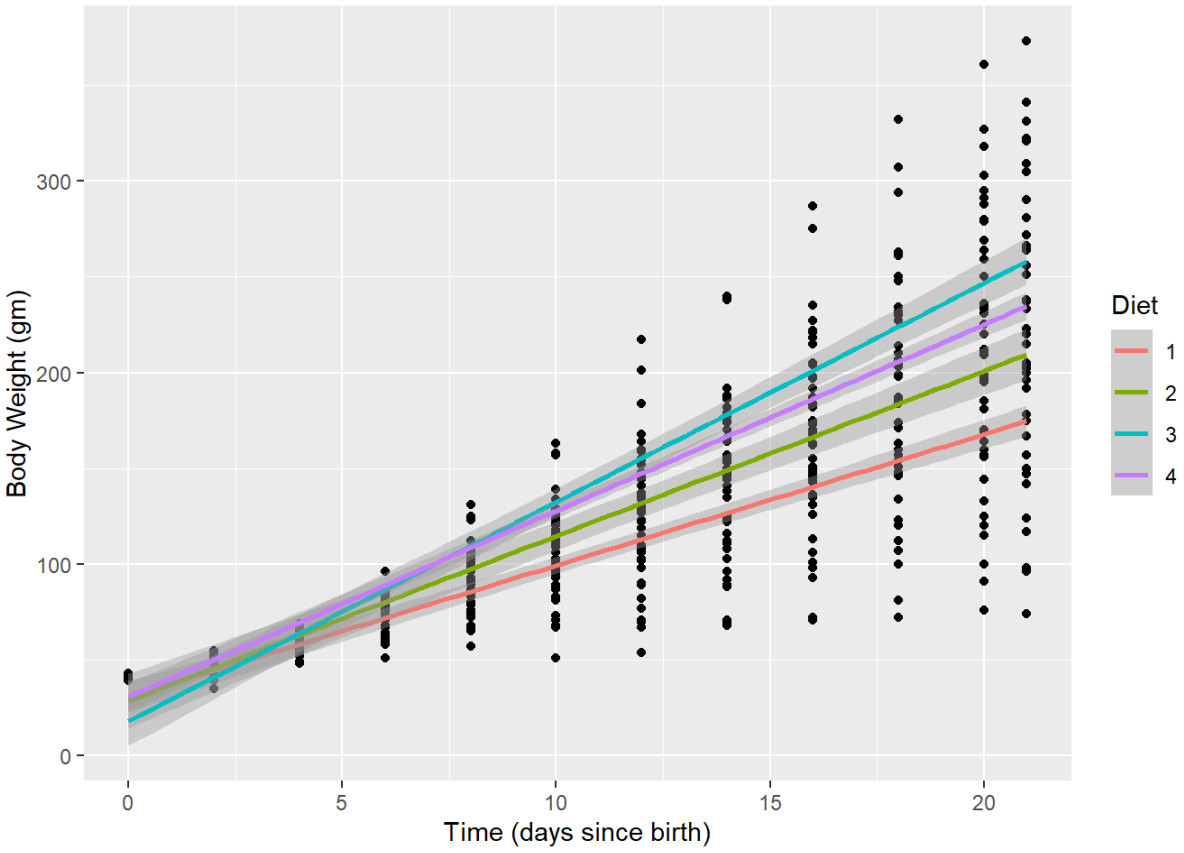
The interaction between the centered log brain weight and the insectivore diet seems to be the most important because of the dramatic line on the graph as well as the interaction's p-value of 0.03575 which is lower than 0.05. This indicates tha there is no significant interaction between the two.

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?

```
# recall ChickWeight
data(ChickWeight)

# copy graph from 2.7 on HW2
ggplot(ChickWeight, aes(Time, weight)) + geom_point() +
  xlab("Time (days since birth)") +
  ylab("Body Weight (gm)") +
  geom_smooth(aes(color = Diet), method = "lm", formula = y ~ x)
```



I think there will be a significant interaction between time and type of diet due to the similar regression lines across all diets.

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?

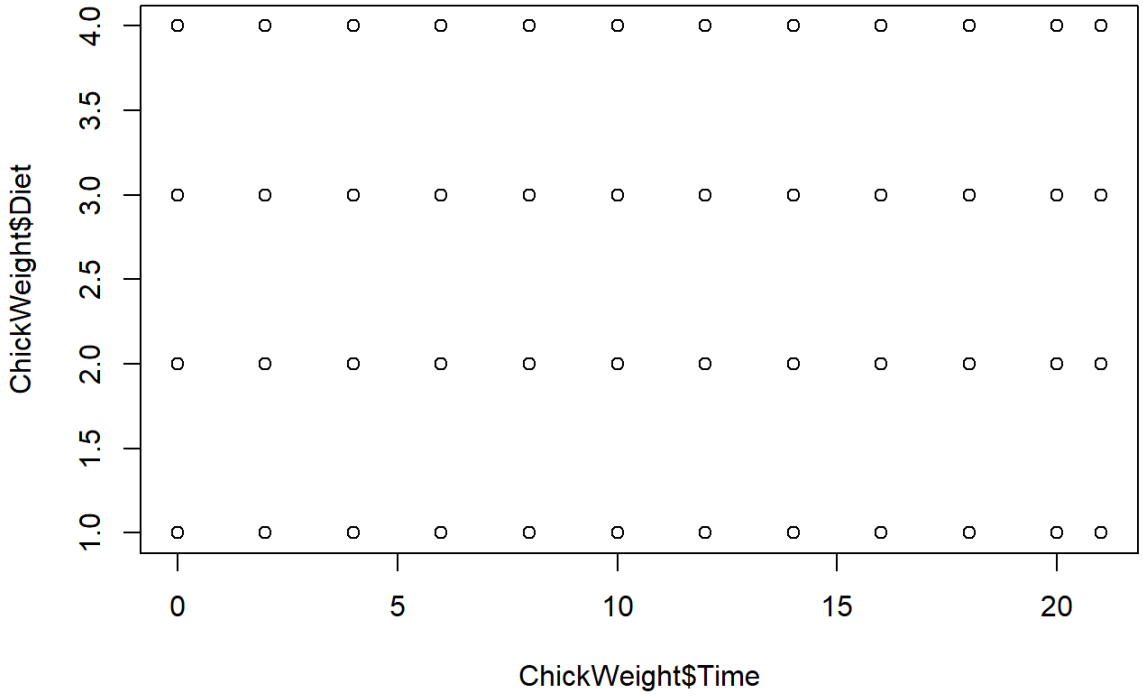
```
# Fit a regression model to predict weights of chicks based on num days since birth and type of diet,
# including the interaction
fit4 <- lm(formula = weight ~ Time * Diet, data = ChickWeight)
summary(fit4)
```

```
##
## Call:
## lm(formula = weight ~ 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 individual effects of the different diets were not significant while the interactions between types of diets and time are because the individual diet type does not make a large difference in weight without comparing it overtime.

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?

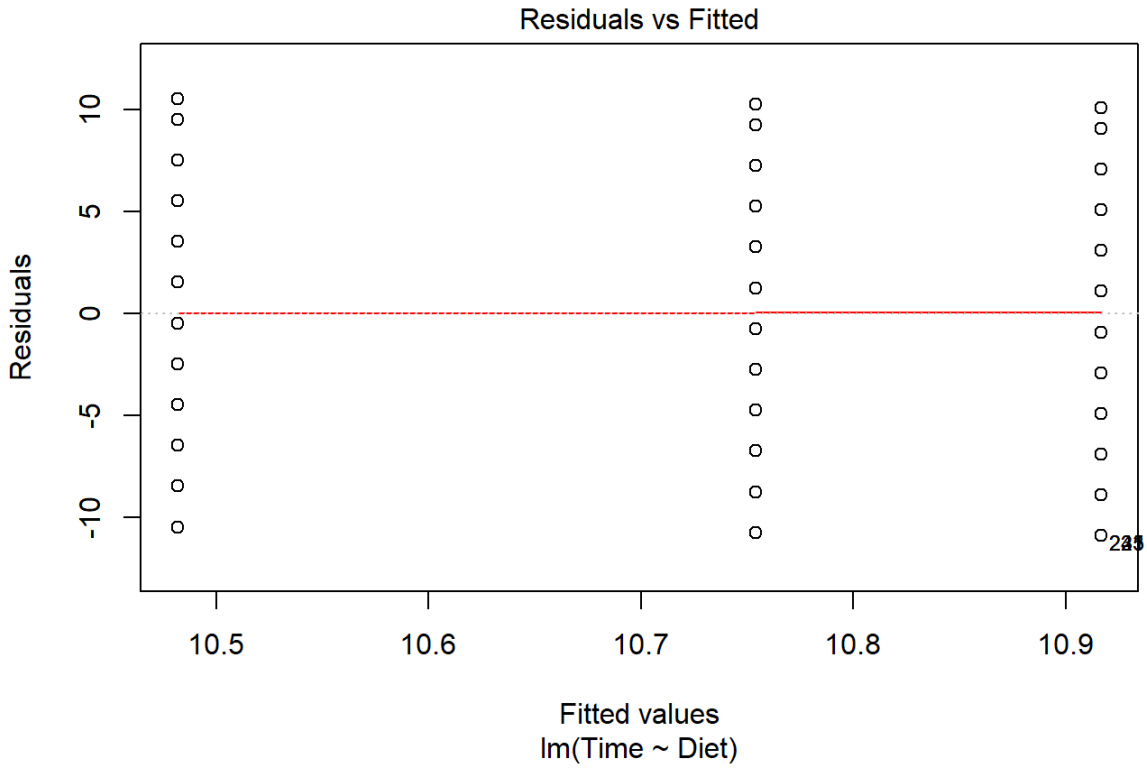
```
# check assumptions visually
plot(ChickWeight$Time, ChickWeight$Diet)
```



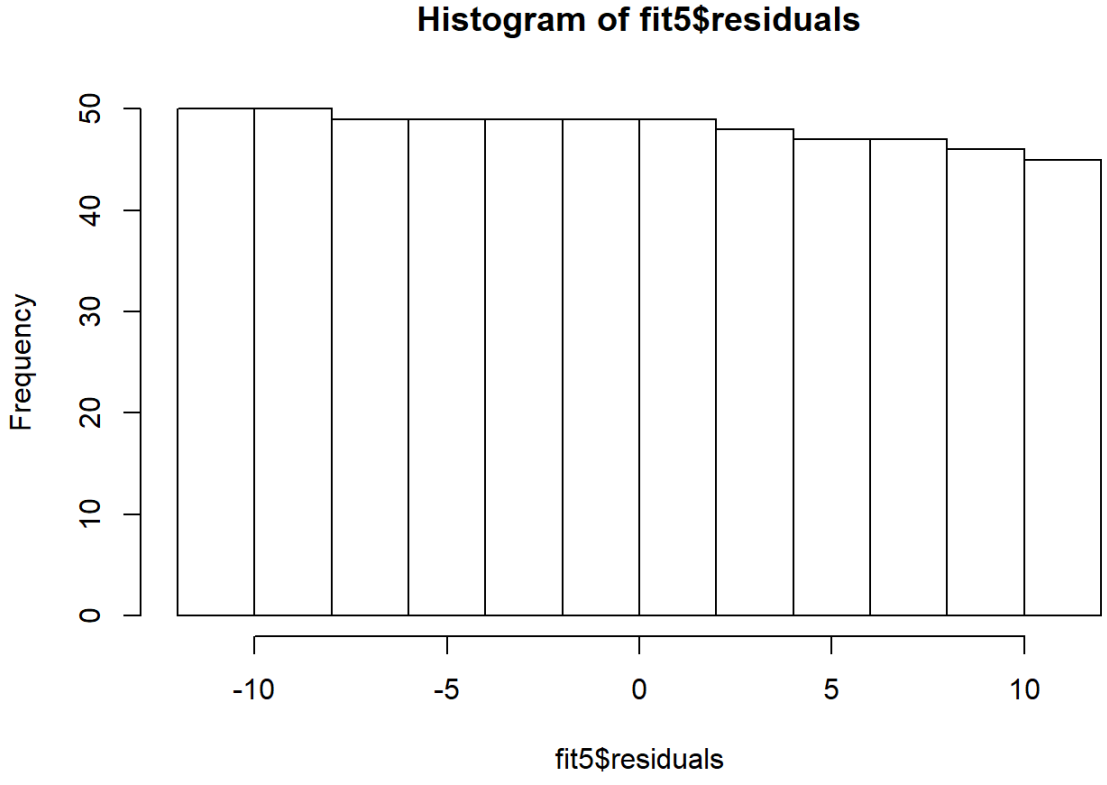

```
fit5 <- lm(Time ~ Diet, data = ChickWeight)
summary(fit5)
```

```
##
## Call:
## lm(formula = Time ~ Diet, data = ChickWeight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9167  -6.4818  -0.4818   5.5182  10.5182
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.4818     0.4567  22.954   <2e-16 ***
## Diet2         0.4348     0.7687   0.566    0.572
## Diet3         0.4348     0.7687   0.566    0.572
## Diet4         0.2724     0.7729   0.352    0.725
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.773 on 574 degrees of freedom
## Multiple R-squared:  0.0008309,  Adjusted R-squared:  -0.004391
## F-statistic: 0.1591 on 3 and 574 DF,  p-value: 0.9238
```

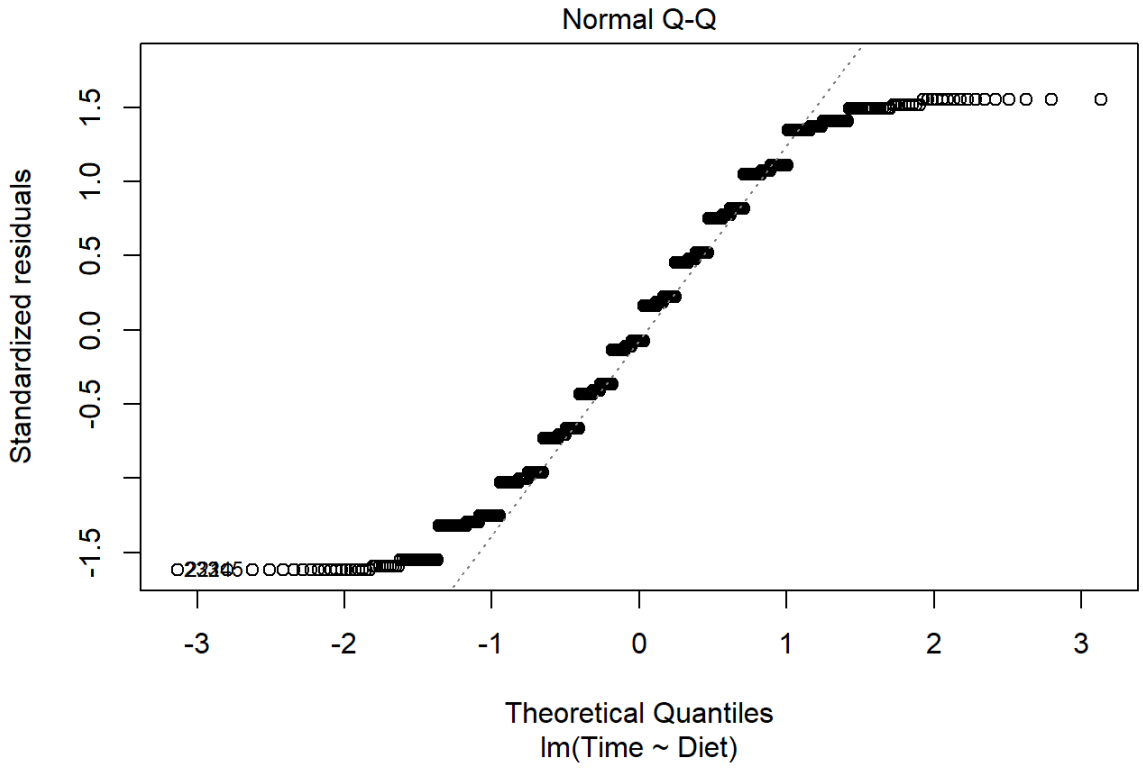
```
# Residuals vs Fitted values plot
plot(fit5, which = 1)
```




```
# Histogram of residuals
hist(fit5$residuals)
```



```
# Q-Q plot for the residuals
plot(fit5, which = 2)
```



No assumptions are met. The original plot is completely unrelated. The only significant regression value was that of the intercept. The residual plot is not linear, the histogram is not normal, and the QQ plot of residuals appears to be in a log function, not linear.

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.

```
# create a 'newdata' with specified variables
newdata <- data.frame(Time = 15, Diet = '1')

# using fit4 construct confidence interval for mean weight of chick after 15 days after diet 1
predict(fit4, newdata = newdata, interval = "confidence")
```

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

For the dataset, 95% of the mean weight of a chick after 15 days following Diet 1 falls between the interval of 128.127 gm and 138.989 gm.

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?

```
# When assumptions are violated (homoscedasticity, normality, small sample size)
# use bootstrap samples to estimate coefficients, SEs, fitted values, ...

# Use the function replicate to repeat the process (similar to a for loop)
samp_chick <- replicate(5000, {
  # Bootstrap your data (resample observations)
  chick_data <- sample_frac(ChickWeight, replace = TRUE)
  # For each bootstrap sample, fit a regression model
  fitchick <- lm(weight ~ Time * Diet, data = ChickWeight)
  # calculate the mean weight of a chick after 15 days following Diet 1
  newd <- data.frame(Time = 15, Diet = '1')
  predict(fitchick, newdata = newd, interval = "confidence")
})
# find the bootstrap 95% confidence interval for the mean weight of a chick after 15 days following Diet 1
quantile(samp_chick, 0.025)
```

```
##     2.5%
## 128.1267
```

```
quantile(samp_chick, 0.975)
```

```
##    97.5%
## 138.9891
```

This confidence interval is not different from that in question 2.4.

```
##        sysname        release        version       nodename        machine
##      "Windows"       "10 x64"  "build 19042"     "ROSE-XPS"       "x86-64"
##          login           user effective_user
##        "roseh"        "roseh"        "roseh"
```