Problem Set 5 - Eddie Capone 3/31/23

Code

1.

Hide

```
crickets <- read.table(file.choose(), stringsAsFactors = TRUE)
```

Hide

```
summary(crickets)
```

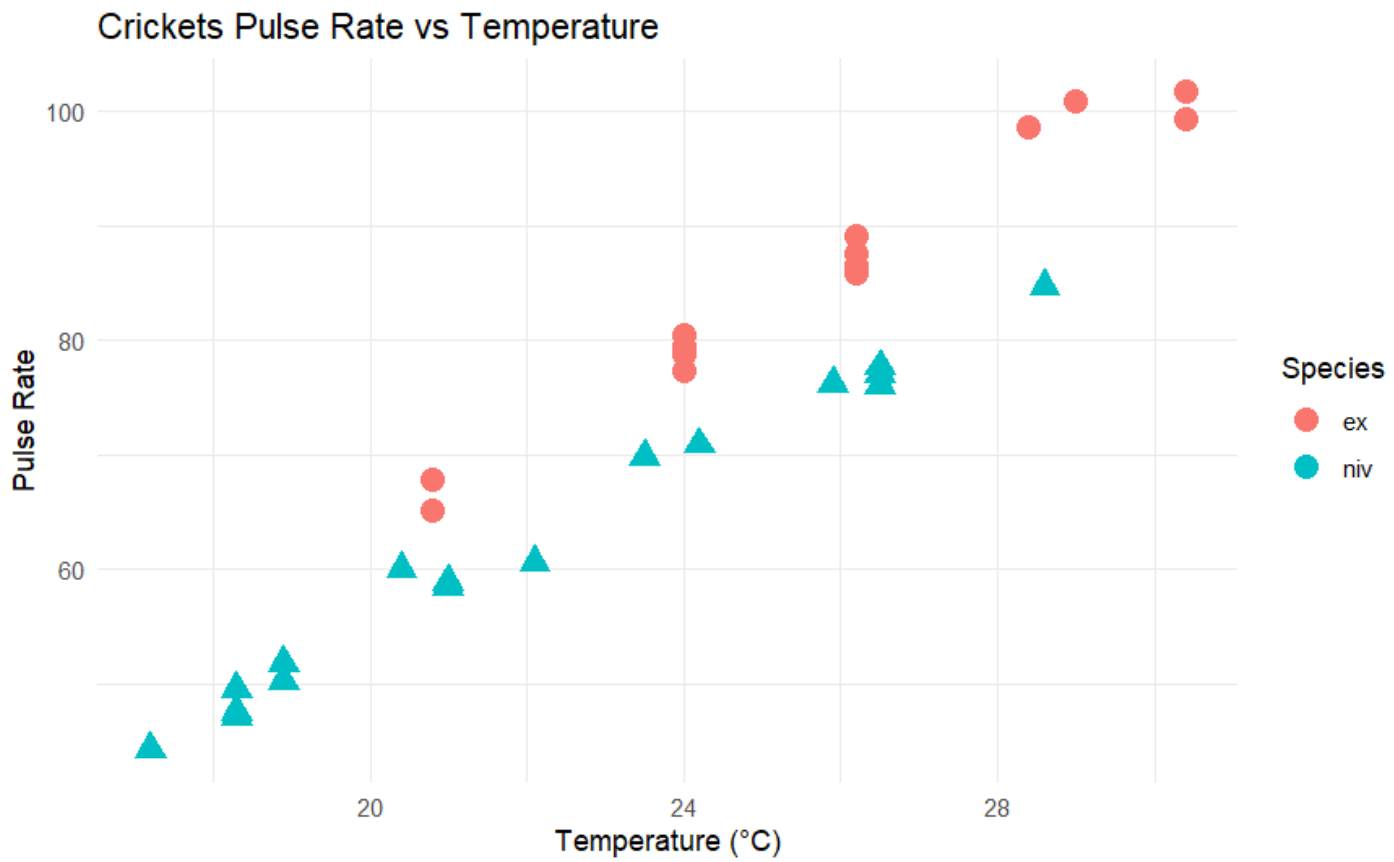
```
 species       temp           pulse       
 ex :14   Min.   :17.20   Min.   : 44.30  
 niv:17   1st Qu.:20.80   1st Qu.: 59.45  
          Median :24.00   Median : 76.20  
          Mean   :23.76   Mean   : 72.89  
          3rd Qu.:26.35   3rd Qu.: 85.25  
          Max.   :30.40   Max.   :101.70  
```

From this data frame we can see the different chirping rates (pulse) of two different species of crickets Oecanthus exclamationis (ex) and Oecanthus niveus (niv) at different temperatures. We notice that the ex species has 14 crickets and the niv species has 17. The minimum temperature used in the dataframe is 17.2, the maximum is 30.4, and the mean is 23.76. The minimum pulse is 44.3, the maximum is 101.7, and the mean is 72.89.

a. Plot the data using different color symbols for the two species of crickets.

Hide

```
library(ggplot2)
ggplot(crickets, aes(x = temp,y = pulse, color = species)) +
  geom_point(size = 4, pch = as.numeric(crickets$species)+15) +
  theme_minimal() +
   labs(title = "Crickets Pulse Rate vs Temperature",
       x = "Temperature (°C)",
       y = "Pulse Rate",
       color = "Species",
       shape = "Species")
```



The graph above shows the pulse on temperature for each species, with ex in red and niv in blue. Both of the species show a relatively strong positive relationship, stating that when temperature increases, so does pulse, and vice versa. Additionally, it looks like niv appears in cooler temperatrues, hence they will have a lower mean pulse rating. Alternatively, ex is found in warmer climates, so they will have a higher overall pulse rating than the niv species.

b. Find summary statistics for the two groups. Clearly the means are very different, but from the graph it is also clear that the temperature is a confounding variable–the pulse rate changes with the temperature.

Hide

```
tapply(crickets$temp, crickets$species, summary)
```

```
$ex
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  20.80   24.00   26.20   25.76   27.85   30.40 

$niv
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  17.20   18.90   21.00   22.12   25.90   28.60 
```

Hide

```
tapply(crickets$pulse, crickets$species, summary)
```

```
$ex
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  65.10   78.88   86.20   85.59   96.22  101.70

$niv
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  44.30   50.30   60.00   62.43   76.10   84.70
```

These summary statistics prove the strength of the analysis we came up with regarding the graph in part a. When looking at species ex, we can see that the mean temperature is 25.76, and for niv the mean temperature is 22.12. This data correlates with the findings we had above, which was that the ex species tends to be found in warmer climates. Moving on, the mean for the pulse of the ex species is 85.59 and 62.43 for the niv species. The ex species lives in warmer climates and therefore has a faster chirping rate than the niv species, and vice versa. Therefore, parts a and b demonstrate that there is a correlation between the temperature the crickets are exposed to and the pulse of their chirping.

c. Perform an ANCOVA to test whether or not the slopes of the regression lines modeling pulse on temp for each species of cricket are the same. Superimpose the lines from both models on the same scatterplot of pulse on temp.

Hide

```
model1 <- lm(pulse ~ temp * species, data = crickets)
model1$coef -> c1
summary(model1)
```

```

Call:
lm(formula = pulse ~ temp * species, data = crickets)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7031 -1.3417 -0.1235  0.8100  3.6330

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11.0408     4.1515  -2.659    0.013 *
temp              3.7514     0.1601  23.429   <2e-16 ***
speciesniv       -4.3484     4.9617  -0.876    0.389
temp:speciesniv  -0.2340     0.2009  -1.165    0.254
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.775 on 27 degrees of freedom
Multiple R-squared:  0.9901,    Adjusted R-squared:  0.989
F-statistic: 898.9 on 3 and 27 DF,  p-value: < 2.2e-16
```

This model includes the interaction term between temperature (temp) and species. Based on the output of the linear regression model, the R-squared value is 0.9901. This indicates that the model explains approximately 99.01% of the variance in the pulse variable and it suggests the model is a good fit to the data. For ‘(Intercept)’, the p-value is 0.013, which is greater than 0.01, thus indicating that the intercept is not statistically significant. The

'temp' variable has a p-value of <2e-16, which is much smaller than 0.01, hence indicating that the variable is statistically significant and suggesting a strong relationship between temperature and pulse. The p-value of 'speciesniv' is 0.389, thus indicating that the variable is not statistically significant and concluding that there is not a significant difference between the intercepts of the two cricket species. Lastly, 'temp:speciesniv:' has a p-value of .254, which tell us that the interaction term (temp:speciesniv) is not statistically significant and concludes that the slopes for the two cricket species are different.

Hide

```
model2 <- lm(pulse ~ temp + species, data = crickets)
model2$coef -> c2
summary(model2)
```

```

Call:
lm(formula = pulse ~ temp + species, data = crickets)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0128 -1.1296 -0.3912  0.9650  3.7800 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -7.21091    2.55094  -2.827  0.00858 ** 
temp          3.60275    0.09729  37.032  < 2e-16 ***
speciesniv  -10.06529    0.73526 -13.689 6.27e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.786 on 28 degrees of freedom
Multiple R-squared:  0.9896,    Adjusted R-squared:  0.9888 
F-statistic:  1331 on 2 and 28 DF,  p-value: < 2.2e-16
```

This model does not include the interaction term between temperature (temp) and species. It assumes that the relationship between pulse and temperature is the same for both species of cricket. Based on the output, the R-squared value is 0.9896. This indicates that the model explains approximately 98.96% of the variance in the pulse variable. The high R-squared value suggests a good fit of the model to the data. The p-value of (Intercept) is 0.00858, which indicates that the intercept is statistically significant since the p-value is less than 0.01. The 'temp' variable has a p-value of 2e-16, which indicates that the variable is statistically significant and suggests that there is a strong relationship between temperature and pulse. The 'speciesniv' variable has a p-value of 6.27e-14, which indicates that the variable is statistically significant and concludes that there is a significant difference in the pulse between the two cricket species when accounting for temperature.

Hide

```
anova(model2, model1)
```

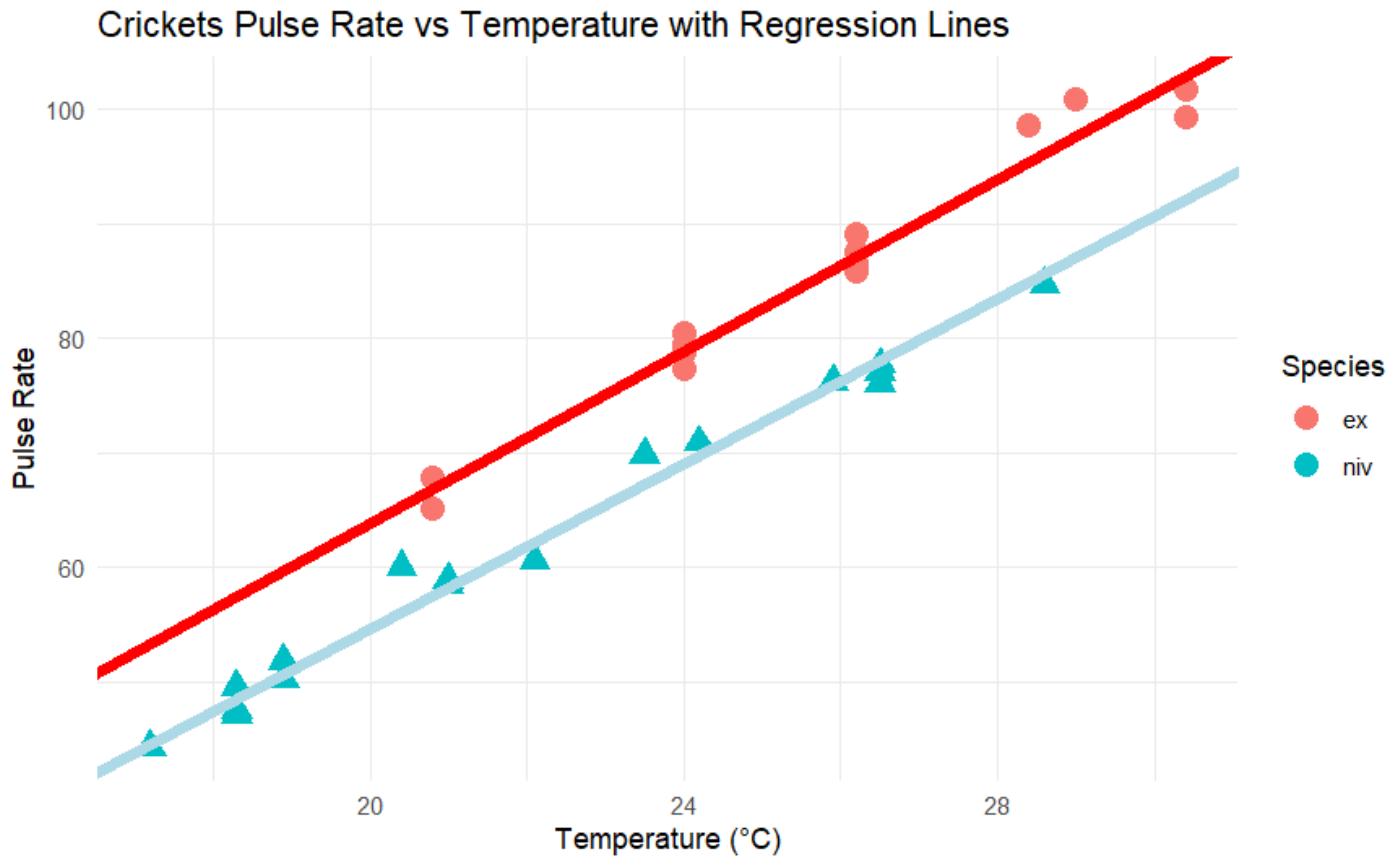
```
Analysis of Variance Table

Model 1: pulse ~ temp + species
Model 2: pulse ~ temp * species
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     28 89.350
2     27 85.074  1    4.2758 1.357 0.2542
```

The ANOVA table shows that the p-value for the F-test of the interaction term is much larger than 0.01. This suggests that there is not significant evidence that the slopes of the regression lines are significantly different between the two species of crickets. In other words, the slopes are the same for Oecanthus exclamationis and Oecanthus niveus.

Hide

```
library(ggplot2)
ggplot(crickets, aes(x = temp,y = pulse, color = species)) +
  geom_point(size = 4, pch = as.numeric(crickets$species)+15) +
  theme_minimal() +
   labs(title = "Crickets Pulse Rate vs Temperature with Regression Lines",
       x = "Temperature (°C)",
       y = "Pulse Rate",
       color = "Species",
       shape = "Species") +
  geom_abline(intercept = c1[1], slope = c1[2], color = "red", size = 2) +
  geom_abline(intercept = c2[1] + c2[3], slope = c2[2], color = "lightblue", size = 2)
```



d. Ok, the slopes are the same. Are the intercepts the same? Model pulse on temp without species, and compare this reduced model to the parallel lines model from the previous part.

To test if the intercepts are the same for both species, we can compare model2 without species (assuming the same intercept for both species) to the parallel lines model (assuming different intercepts for both species but the same slope) from the previous part using an ANOVA test.

Hide

```
model3 <- lm(pulse ~ temp, data = crickets)
model3$coef -> c3
summary(model3)
```

```

Call:
lm(formula = pulse ~ temp, data = crickets)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.7044 -2.5789 -0.4466  4.5353  7.5916 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -27.9476     5.5937  -4.996 2.56e-05 ***
temp          4.2431     0.2325  18.251  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.868 on 29 degrees of freedom
Multiple R-squared:  0.9199,    Adjusted R-squared:  0.9171 
F-statistic: 333.1 on 1 and 29 DF,  p-value: < 2.2e-16
```

Based on the output of (model3), the R-squared value is 0.9199. This indicates that the model explains approximately 91.99% of the variance in the pulse variable. The high R-squared value suggests a good fit of the model to the data. However, it is lower than the R-squared values from the models with the species variable (model1 and model2), which indicates that this model might not capture all the relevant information about the relationship between pulse and temperature for different cricket species. ‘(Intercept)’ has a p-value of 2.56e-05, which indicates that the intercept is statistically significant. The ‘temp’ variable has a p-value of 2e-16, which indicates that the temp variable is statistically significant and suggests that there is a strong relationship between temperature and pulse.

Hide

```
anova(model3, model2)
```

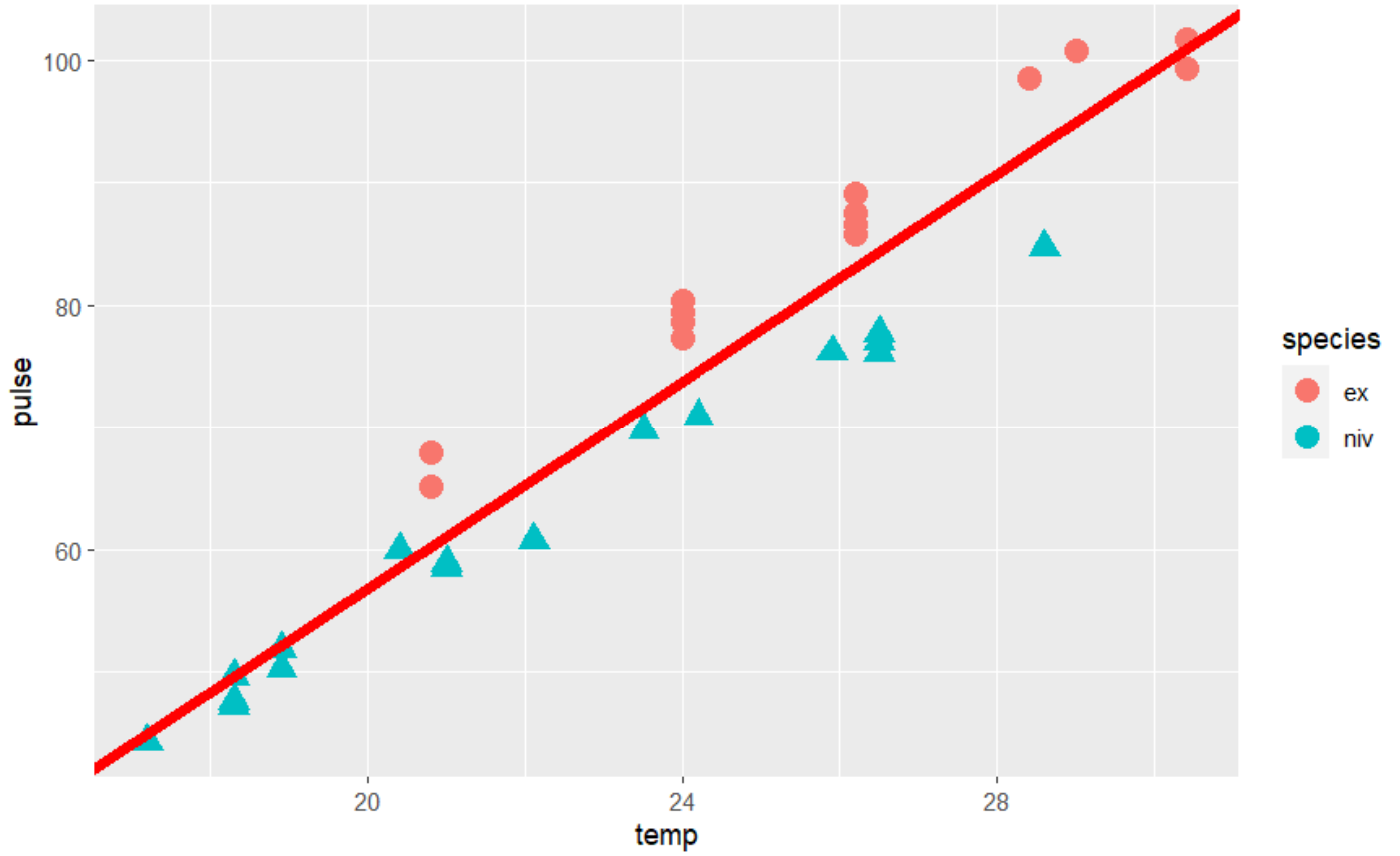
```
Analysis of Variance Table

Model 1: pulse ~ temp
Model 2: pulse ~ temp + species
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1     29 687.35
2     28  89.35  1       598 187.4 6.272e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA table shows that the p-value for the F-test of the interaction term is much smaller than 0.01. This suggests that there is sufficient evidence that the intercepts of the regression lines are significantly different between the two species of crickets. In other words, the intercepts are not the same for Oecanthus exclamationis (ex) and Oecanthus niveus (niv).

Hide

```
ggplot(crickets, aes(temp, pulse, color = species)) +
  geom_point(size = 4, pch = as.numeric(crickets$species) + 15) +
  geom_abline(intercept = c3[1], slope = c3[2], color = "red", size = 2)
```



The slope of this regression line is very similar to those of the dual regression lines for each species seen above. If we put all of these lines on the same graph, the one in part d would fit right in between the lines in part c.

e. What is a good estimate for the mean difference in chirping rates after controlling for temperature?

Hide

```
c2
```

```
(Intercept)        temp  speciesniv
  -7.210906    3.602753  -10.065291
```

A good estimate for the mean difference in chirping rates between the two species of crickets after controlling for temperature is approximately 10.07 units, which is model2, with Oecanthus niveus (niv) having a lower mean pulse rate than Oecanthus exclamationis (ex).

2.

Hide

```
ff <- read.csv(file.choose(), stringsAsFactors = T)
```

Hide

```
summary(ff)
```

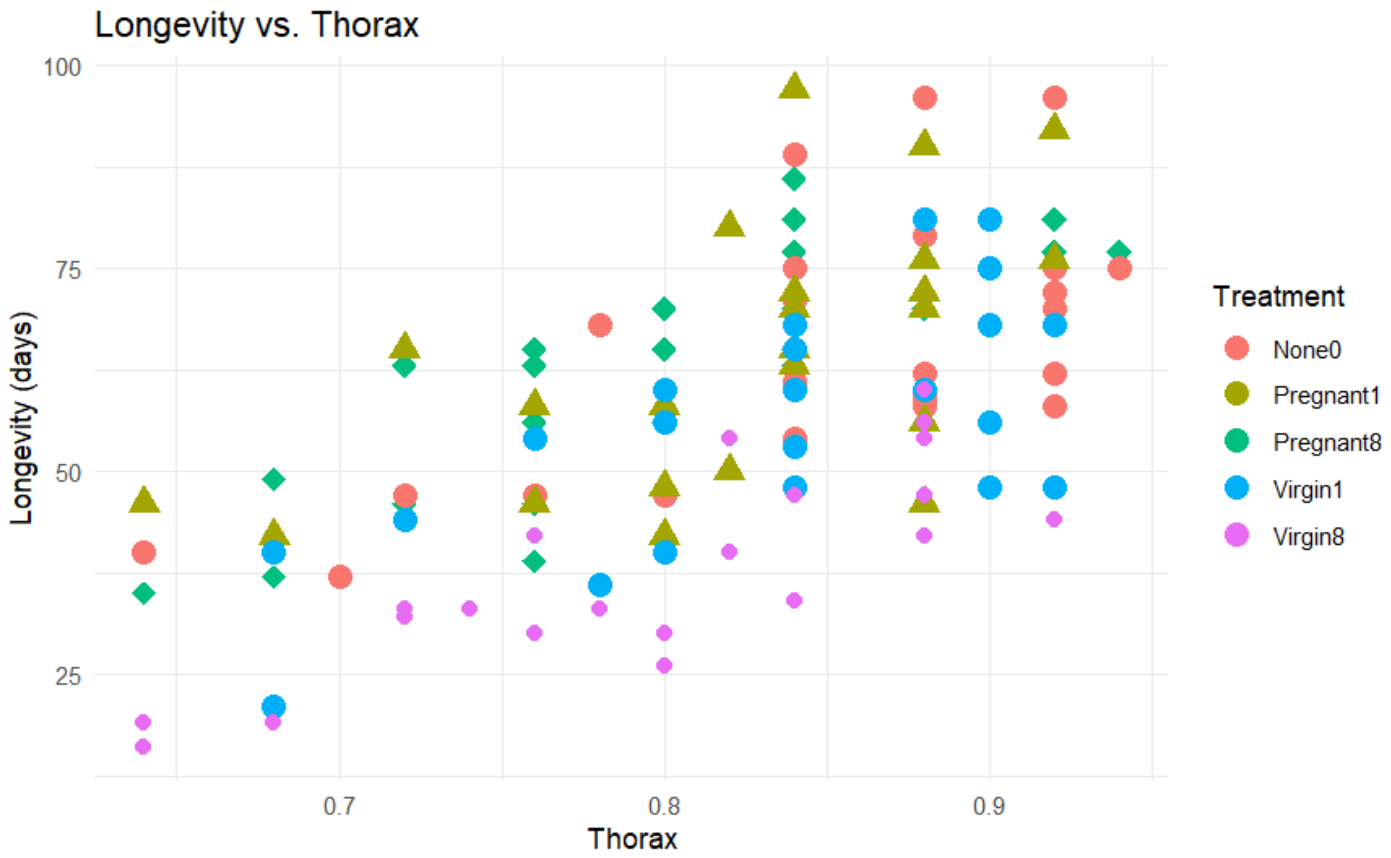
```
   Longevity      CompanionNumber     Thorax     
 Min.   :16.00   None0    :25     Min.   :0.640  
 1st Qu.:46.00   Pregnant1:25     1st Qu.:0.760  
 Median :58.00   Pregnant8:25     Median :0.840  
 Mean   :57.44   Virgin1  :25     Mean   :0.821  
 3rd Qu.:70.00   Virgin8  :25     3rd Qu.:0.880  
 Max.   :97.00                    Max.   :0.940  
```

In this dataset we are measuring the longevity of the life span of a fruitfly in days depending on its thorax length and companion status. Whether they are kept alone, with one pregnant fly, eight pregnant flies, one virgin fly, or eight virgin flies. In this data, the minimum lifespan is 16, the maximum is 97, and the mean is 57.44. The minimum thorax size is 0.64, the maximum is 0.94, and the mean is 0.821.

a. Plot Longevity against Thorax coloring by treatment: CompanionNumber

Hide

```
# scatterplot of Longevity against Thorax, colored by the treatment (CompanionNumber)
ggplot(ff, aes(x = Thorax, y = Longevity, color = CompanionNumber)) +
  geom_point(size = 4, pch = as.numeric(ff$CompanionNumber)+15) +
  theme_minimal() +
  labs(title = "Longevity vs. Thorax",
       x = "Thorax",
       y = "Longevity (days)",
       color = "Treatment")
```



This graph shows there is not much correlation between thorax size and companion number on the longevity of a fruitfly. The data points are scattered and do not have any significant relation.

b. Regress Longevity against Thorax and CompanionNumber twice, once with no interaction and once without. Use the anova function to decide which is the better model.

Hide

```
# model without interaction
nointeractionmodel <- lm(Longevity ~ Thorax + CompanionNumber, data = ff)
summary(nointeractionmodel)
```

```
Call:
lm(formula = Longevity ~ Thorax + CompanionNumber, data = ff)

Residuals:
    Min      1Q  Median      3Q     Max
-26.189  -6.599  -0.989   6.408  30.244

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              -49.984     10.609  -4.711 6.73e-06 ***
Thorax                   135.819     12.439  10.919  < 2e-16 ***
CompanionNumberPregnant1   2.653      2.975   0.891   0.3745
CompanionNumberPregnant8   3.929      2.997   1.311   0.1923
CompanionNumberVirgin1    -7.017      2.973  -2.361   0.0199 *
CompanionNumberVirgin8    -19.951      3.006  -6.636 1.00e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.51 on 119 degrees of freedom
Multiple R-squared:  0.6564,    Adjusted R-squared:  0.6419
F-statistic: 45.46 on 5 and 119 DF,  p-value: < 2.2e-16
```

The R-squared value is 0.6564, which means that around 65.64% of the variance in the Longevity can be explained by the model that includes Thorax and CcompanionNumber as predictors. For the ‘Thorax’ variable, The p-value is less than 2e-16, indicating that Thorax is a highly significant predictor of Longevity. For ‘CompanionNumberPregnant1’ The p-value is 0.3745, which is not statistically significant. This suggests that there is no significant difference in Longevity between the reference group (None0) and the Pregnant1 group. For ‘CompanionNumberPregnant8’ The p-value is 0.1923, which is not statistically significant, hence indicating that there is no significant difference in Longevity between the reference group (None0) and the Pregnant8 group. For ‘CompanionNumberVirgin1’ the p-value is 0.0199, which is not statistically significant. This suggests that there is not a significant difference in Longevity between the reference group (None0) and the Virgin1 group. For ‘CompanionNumberVirgin8’ the p-value is 1.00e-09, which is statistically significant, therefore indicating that there is a significant difference in Longevity between the reference group (None0) and the Virgin8 group.

Hide

```
# model with interaction
interactionmodel <- lm(Longevity ~ Thorax * CompanionNumber, data = ff)
summary(interactionmodel)
```

```
Call:
lm(formula = Longevity ~ Thorax * CompanionNumber, data = ff)

Residuals:
     Min       1Q   Median       3Q      Max
-25.9509  -6.5324  -0.7693   6.3792  30.3071

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -50.2420    21.7221  -2.313   0.0225 *
Thorax                         136.1268    25.8576   5.264 6.61e-07 ***
CompanionNumberPregnant1         6.5172    33.7479   0.193   0.8472
CompanionNumberPregnant8        -5.4574    30.6537  -0.178   0.8590
CompanionNumberVirgin1          -7.7501    33.8457  -0.229   0.8193
CompanionNumberVirgin8         -11.0380    31.1731  -0.354   0.7239
Thorax:CompanionNumberPregnant1 -4.6771    40.5042  -0.115   0.9083
Thorax:CompanionNumberPregnant8 11.6629    37.1806   0.314   0.7543
Thorax:CompanionNumberVirgin1    0.8743    40.2786   0.022   0.9827
Thorax:CompanionNumberVirgin8  -11.1268    37.9816  -0.293   0.7701
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.67 on 115 degrees of freedom
Multiple R-squared:  0.6575,    Adjusted R-squared:  0.6307
F-statistic: 24.53 on 9 and 115 DF,  p-value: < 2.2e-16
```

The R-squared value is 0.6575, which means that around 65.75% of the variance in the Longevity can be explained by the model that includes Thorax and CcompanionNumber as predictors. The p-value for ‘Thorax’ is 6.61e-07, indicating that Thorax is a highly significant predictor of Longevity. ‘CompanionNumberPregnant1’ has a p-value is 0.8472, which is not statistically significant. This suggests that there is no significant difference in Longevity between the reference group (None0) and the Pregnant1 group. ‘CompanionNumberPregnant8’ has a p-value is 0.8590, which is not statistically significant. This indicates that there is no significant difference in Longevity between the reference group (None0) and the Pregnant8 group. ‘CompanionNumberVirgin1’ has a p-value of 0.8193, which is not statistically significant. This suggests that there is no significant difference in Longevity between the reference group (None0) and the Virgin1 group. ‘CompanionNumberVirgin8’ has a p-value of 0.7239, which is not statistically significant, thus indicating that there is no significant difference in Longevity between the reference group (None0) and the Virgin8 group. ‘Thorax:CompanionNumberPregnant1’ has a p-value of 0.9083, which is not statistically significant, suggesting that there is no significant interaction effect between Thorax and CompanionNumberPregnant1 on Longevity. ‘Thorax:CompanionNumberPregnant8’ has a p-value of 0.7543, which is not statistically significant. This indicates that there is no significant interaction effect between Thorax and CompanionNumberPregnant8 on Longevity. ‘Thorax:CompanionNumberVirgin1’ has a p-value of 0.9827, which is not statistically significant and suggests that there is no significant interaction effect between Thorax and CompanionNumberVirgin1 on Longevity. ‘Thorax:CompanionNumberVirgin8’ has a p-value of 0.7701, which is not statistically significant and indicates that there is no significant interaction effect between Thorax and CompanionNumberVirgin8 on Longevity.

Hide

```
anova(nointeractionmodel, interactionmodel)
```

```
Analysis of Variance Table

Model 1: Longevity ~ Thorax + CompanionNumber
Model 2: Longevity ~ Thorax * CompanionNumber
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    119 13145
2    115 13102  4    42.523 0.0933 0.9844
```

The p-value of 0.9844 tells us that there is not a significant amount of evidence to suggest that the interaction model is better at explaining the variation in Longevity, meaning there is no significant difference between the models, and the simpler model without interaction is preferred.

c. Look at a one-way anova plot. That is, use granovagg.1w from the package granovaGG to plot the five groups. It certainly looks like the three treatments none, pregnant1, and pregnant8, are quite similar. Use the multiple comparison command (HSD.test) we used earlier this semester to compare the means of the five groups. You should conclude that indeed, the three groups mentioned are all similar and could be combined into one group

Hide

```
library(granovaGG)
granovagg.1w(ff$Longevity, group = ff$CompanionNumber, ylab = "Longevity")
```

```
By-group summary statistics for your input data (ordered by group means)
```

	group <fctr>	group.mean <dbl>	trimmed.mean <dbl>	contrast <dbl>	variance <dbl>	standard.deviation <dbl>	group.size <dbl>
5	Virgin8	38.72	38.67	-18.72	146.46	12.10	25
4	Virgin1	56.76	56.80	-0.68	222.86	14.93	25
3	Pregnant8	63.36	65.27	5.92	211.41	14.54	25
1	None0	63.56	62.60	6.12	270.67	16.45	25
2	Pregnant1	64.80	64.20	7.36	245.00	15.65	25

5 rows

```
The following groups are likely to be overplotted
```

	group <fctr>	group.mean <dbl>	contrast <dbl>
3	Pregnant8	63.36	5.92
1	None0	63.56	6.12

2 rows

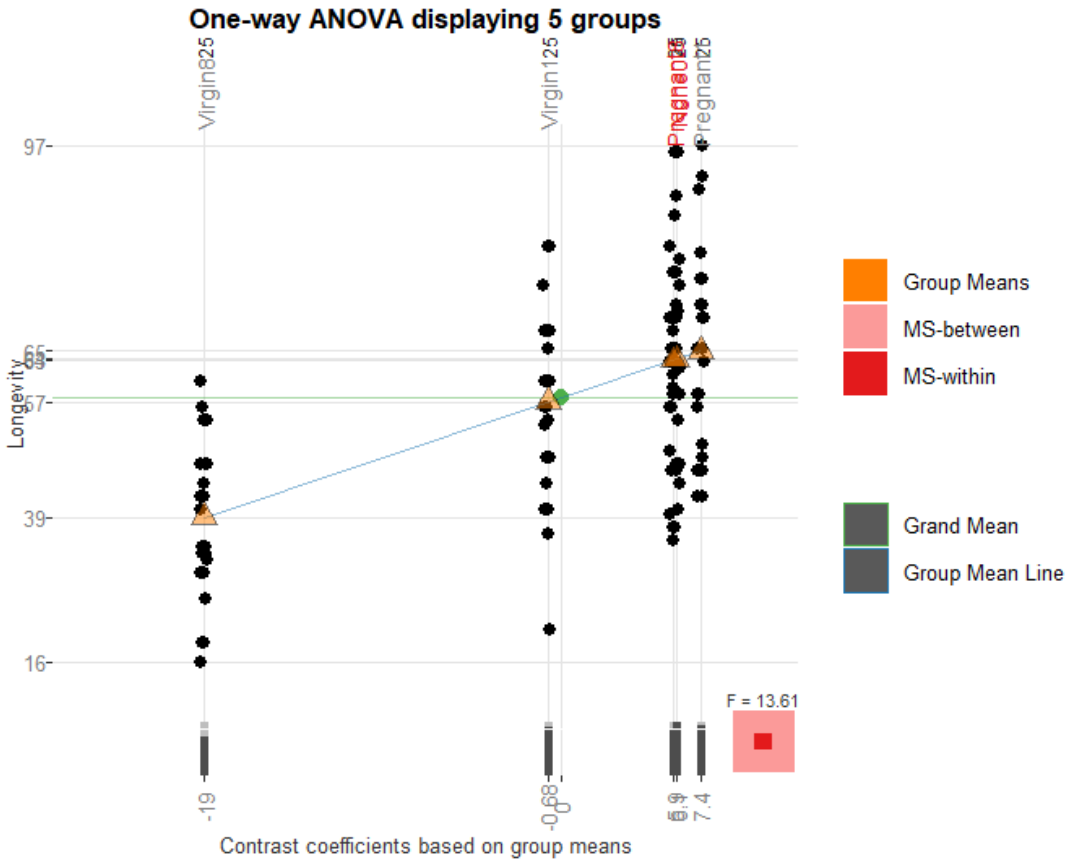
Below is a linear model summary of your input data

```
Call:
lm(formula = score ~ group, data = owp$data)

Residuals:
   Min     1Q Median     3Q    Max
-35.76  -8.76   0.20  11.20  32.44

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      63.560      2.962  21.461  < 2e-16 ***
groupPregnant1    1.240      4.188   0.296    0.768
groupPregnant8   -0.200      4.188  -0.048    0.962
groupVirgin1     -6.800      4.188  -1.624    0.107
groupVirgin8    -24.840      4.188  -5.931 2.98e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.81 on 120 degrees of freedom
Multiple R-squared:  0.3121,    Adjusted R-squared:  0.2892
F-statistic: 13.61 on 4 and 120 DF,  p-value: 3.516e-09
```



Hide

```
library(agricolae)
model <- lm(Longevity ~ CompanionNumber, ff)
HSD.test(model, "CompanionNumber", console = T)
```

```
Study: model ~ "CompanionNumber"

HSD Test for Longevity

Mean Square Error:  219.2793

CompanionNumber,  means
```

	Longevity <dbl>	std <dbl>	r <int>	Min <dbl>	Max <dbl>
None0	63.56	16.45215	25	37	96
Pregnant1	64.80	15.65248	25	42	97
Pregnant8	63.36	14.53983	25	35	86
Virgin1	56.76	14.92838	25	21	81
Virgin8	38.72	12.10207	25	16	60

5 rows

```
Alpha: 0.05 ; DF Error: 120
Critical Value of Studentized Range: 3.916938

Minimun Significant Difference: 11.60047

Treatments with the same letter are not significantly different.
```

	Longevity <dbl>	groups <chr>
Pregnant1	64.80	a
None0	63.56	a
Pregnant8	63.36	a
Virgin1	56.76	a
Virgin8	38.72	b

5 rows

From the above one way anova analysis we can see that the three treatments none0, pregnant1, and pregnant8, are very similar, meaning the three groups mentioned could all be combined into one group. The longevity for None0 is 63.56, Pregnant1 is 64.8 and Pregnant8 is 63.36. These are not evidently not statistically different.

d. Add a new factor to fly that renames the pregnant groups to none. Run this code below:

Hide

```
ff$CN <- factor(ff$CompanionNumber,
                      labels = c("None0", "None0", "None0", "Virgin1", "Virgin8"))
```

e. Regress Longevity on the new factor CN both with and without an interaction. Which regression is to be preferred? (Use anova.)

Hide

```
nointeractionmodel2 <- (lm(Longevity ~ Thorax + CN, data = ff))
summary(nointeractionmodel2)
```

```

Call:
lm(formula = Longevity ~ Thorax + CN, data = ff)

Residuals:
     Min       1Q   Median       3Q      Max 
-25.6157  -7.4270  -0.9692   6.5053  30.7378 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -46.161     10.209  -4.522 1.44e-05 ***
Thorax       133.837     12.326  10.858  < 2e-16 ***
CNVirgin1     -9.181      2.432  -3.775  0.00025 ***
CNVirgin8    -22.189      2.441  -9.091 2.36e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.5 on 121 degrees of freedom
Multiple R-squared:  0.6512,    Adjusted R-squared:  0.6425 
F-statistic:  75.3 on 3 and 121 DF,  p-value: < 2.2e-16
```

The R-squared value is 0.6512, indicating that 65.12% of the variation in Longevity can be explained by the model, which includes Thorax and CN as predictors. The intercept term is statistically significant with a p-value of 1.44e-05, which is less than 0.01. The Thorax variable is statistically significant with a p-value of less than 2e-16. The CNVirgin1 variable is statistically significant with a p-value of 0.00025. The CNVirgin8 variable is statistically significant with a p-value of 2.36e-15. All the predictor variables in this model are statistically significant at the 0.01 level.

Hide

```
interactionmodel2 <- (lm(Longevity ~ Thorax * CN, data = ff))
summary(interactionmodel2)
```

```
Call:
lm(formula = Longevity ~ Thorax * CN, data = ff)

Residuals:
    Min      1Q  Median      3Q     Max
-25.732  -7.165  -1.165   6.872  30.702

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       -47.814     12.902  -3.706 0.000321 ***
Thorax            135.847     15.618   8.698 2.22e-14 ***
CNVirgin1         -10.178     28.788  -0.354 0.724304
CNVirgin8         -13.466     25.650  -0.525 0.600570
Thorax:CNVirgin1    1.154     34.373   0.034 0.973283
Thorax:CNVirgin8  -10.847     31.699  -0.342 0.732799
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.58 on 119 degrees of freedom
Multiple R-squared:  0.6516,    Adjusted R-squared:  0.6369
F-statistic: 44.51 on 5 and 119 DF,  p-value: < 2.2e-16
```

The R-squared value is 0.6516, indicating that 65.16% of the variation in Longevity can be explained by the model, which includes Thorax, CN, and their interaction as predictors. The intercept term is statistically significant with a p-value of 0.000321. The Thorax variable is statistically significant with a p-value of 2.22e-14. The CNVirgin1 variable is not statistically significant with a p-value of 0.724304. The CNVirgin8 variable is not statistically significant with a p-value of 0.600570. The interaction term between Thorax and CNVirgin1 (Thorax:CNVirgin1) is not statistically significant with a p-value of 0.973283. The interaction term between Thorax and CNVirgin8 (Thorax:CNVirgin8) is not statistically significant with a p-value of 0.732799.

Hide

```
anova(nointeractionmodel2, interactionmodel2)
```

```
Analysis of Variance Table

Model 1: Longevity ~ Thorax + CN
Model 2: Longevity ~ Thorax * CN
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    121 13343
2    119 13328  2    14.547 0.0649 0.9372
```

The p-value of 0.9372 tells us that there is not a significant amount of evidence to suggest that the interaction model is better at explaining the variation in Longevity, and the R squared values are also very similar and almost even for both models. Therefore, this means there is no significant difference between the models and the simpler model without interaction is preferred.

f. You now have a best regression of Longevity on CompanionNumber and a best regression of Longevity on CN. Which is the smaller model? Which is to be preferred? Again, use anova.

Hide

```
anova(nointeractionmodel2, nointeractionmodel)
```

```
Analysis of Variance Table

Model 1: Longevity ~ Thorax + CN
Model 2: Longevity ~ Thorax + CompanionNumber
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    121 13343
2    119 13145  2    197.99 0.8962 0.4108
```

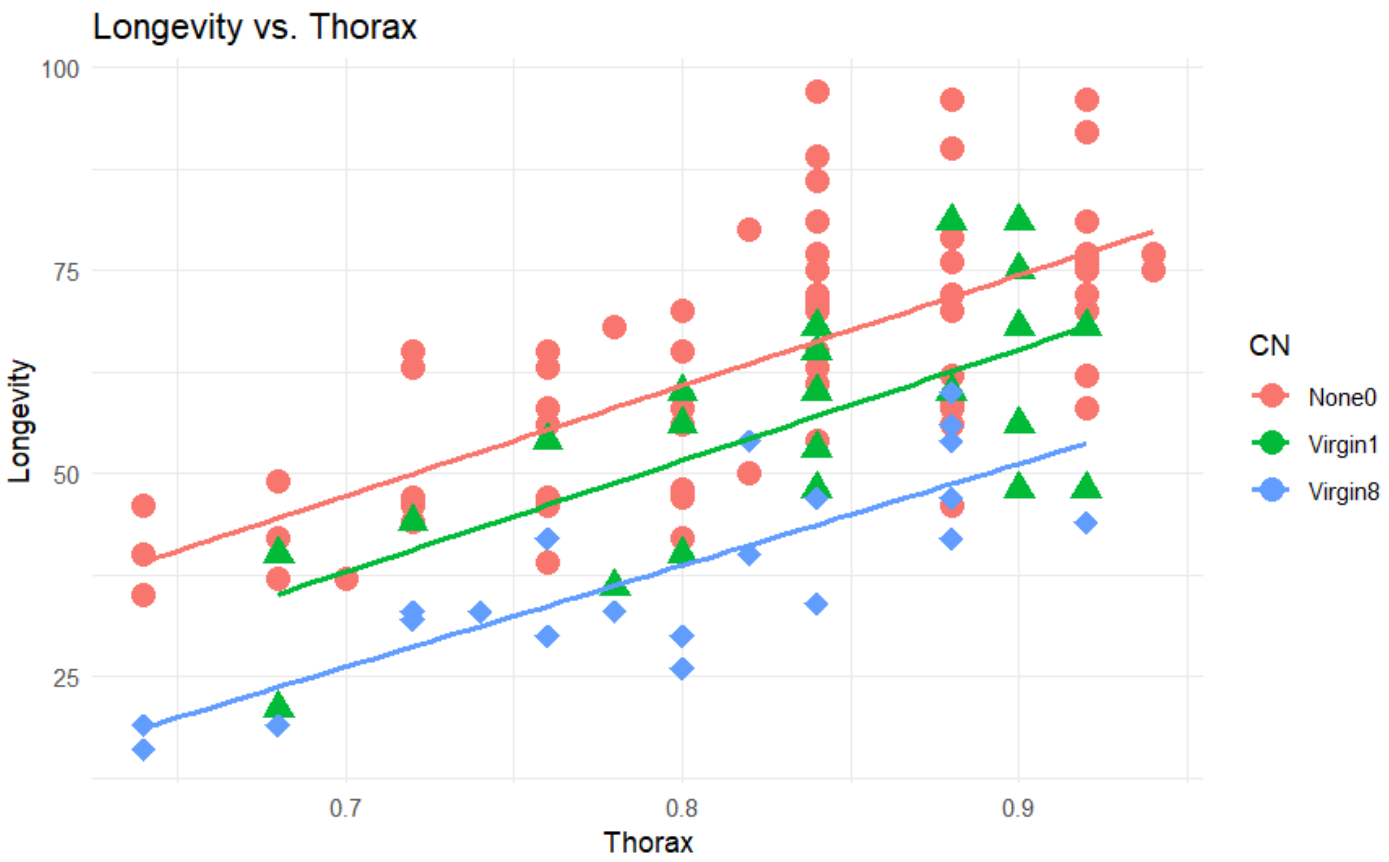
The p-value of 0.4108 tells us that there is not a significant amount of evidence to suggest that the regression on Longevity on CompanionNumber (nointeractionmodel) is better at explaining the variation in Longevity, meaning there is no significant difference between the models, and the simpler model without interaction is preferred, which is nointeractionmodel2.

g. If you decided on the first, superimpose 5 lines on the original plot, making the colors of the lines agree with the colors of the groups. If you decided on the larger, replot the data using only three groups (i.e. with CN), and superimpose the three lines you found, again paying attention to the colors of the lines.

Hide

```
# need to implement c3, c4, c5 like you did with c1 and c2 on the above scatterplot

# Create a scatterplot of Longevity against Thorax, colored by CN
ggplot(data = ff, aes(x = Thorax, y = Longevity, color = CN)) +
  geom_point(size = 4, pch = as.numeric(ff$CN)+15) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Longevity vs. Thorax",
       x = "Thorax",
       y = "Longevity") +
  theme_minimal()
```


Longevity vs. Thorax
100
75
50
25
Longevity
0.7
0.8
0.9
Thorax
CN
None0
Virgin1
Virgin8