Emma Butzler Final

1. Find one or two commands that taken together will calculate:

```
> j<-(5:27)
> sum((6*j-2)/(sqrt(3+2*j)))
[1] 354.5068
```

2.a By simulation, find the estimate of the probability of

Without replacement

```
> bag<-c(rep(1,5),rep(2,7),rep(3,9),rep(4,11))
> selection<-sample(bag,2)
> same<-selection[1]==selection[2]
> same
[1] FALSE
> bag<-c(rep(1,5),rep(2,7),rep(3,9),rep(4,11))
> same<-rep(0,1000000)
> for(i in 1:1000000){
+ selection<-sample(bag,2)
+ same[i]<-selection[1]==selection[2]
+ }
> table(same)/1000000
same
       0        1
0.753625 0.246375
```

The probability the two balls are the same color without replacement is .246.

2.b

```
> bag1<-c(rep(1,5),rep(2,7),rep(3,9),rep(4,11))
> same<-(rep(0,1000000))
> for(i in 1:1000000){
+ selection1<-sample(bag1,2)
+ bag2<-c(1,2,2,3,3,3,4,4,4,4,selection1)
+ selection2<-sample(bag2,2)
+ same[i]<-selection2[1]==selection2[2]
```

```
+ }
>
> table(same)/1000000
same
       0        1
0.759347 0.240653
```

The probability that we pick balls of the same color when we place them in a second bag is .24.

3 Write a function..

```
> my.skewness.EB<-function(x){
+ top<-sum(((x-mean(x)))^3)
+ bottom<-(sum((x-mean(x))^2))^(3/2)
+ out<-top/bottom
+ return(out)
+ }
> x<-c(1,2,3,3,4,4,5,5,5,6,6,6,7,7,7,7)
> my.skewness.EB(x)
[1] -0.1365575
```

4.

4.a. The p-value of p=.23 indicates there is little to no evidence that the mean weight of rabbits and hares is different.

4.b There is suggestive but inconclusive evidence (p=.07) that the mean weight of rabbits and hares are different

4.c There is moderate evidence (p=.035) that the mean weight of rabbits and hares is different.

4.d There is good evidence (p=.012) that the mean weight of rabbits and hares are different

4.e There is very strong evidence (p=.0007) that the mean weight of rabbits and hares are different.

5. a observational, prospective

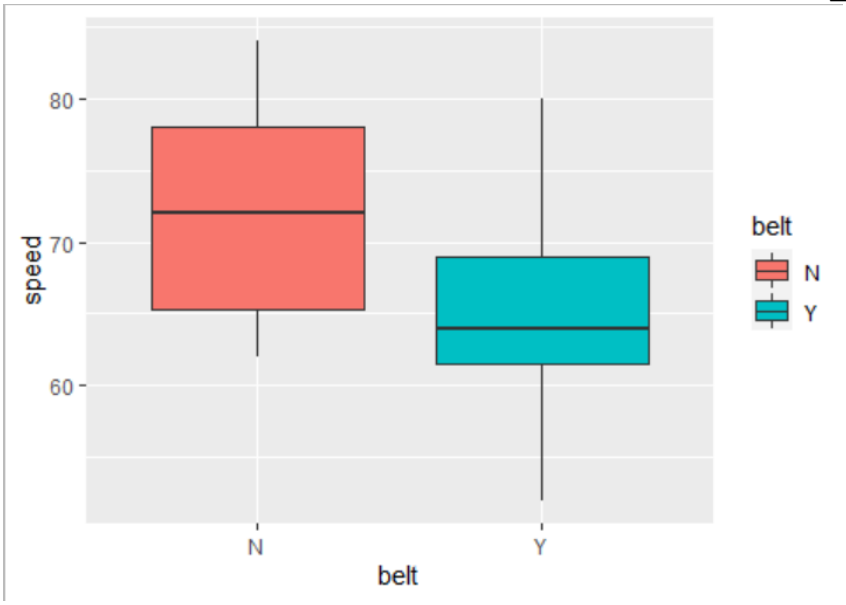
5.b experimental, double blinded

5.c. experimental, blinded

5.d observational, prospective

6.

```
> ggplot(driver,aes(belt,speed,fill=belt))+geom_boxplot()
```



The data is reasonable to assume a t.test because there doesn't appear to be any outliers. There is no apparent skew in either group so there isn't a reason to assume it isn't normal. By taking the t.test we are finding the true difference in means. This test is to determine if there is a statistically significant difference.

6.b The null hypothesis is that there is no true difference the mean speed of drivers with seatbelts and mean of drivers without seatbelts.

The alternative hypothesis is that there is a true difference in the mean speed of driver with seatbelts and the mean drivers without seatbelts.

6.c `> t.test(speed~belt,data=driver)`

```
        Welch Two Sample t-test

data:  speed by belt
t = 1.885, df = 13.623,
p-value = 0.08094
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9914159 15.0747492
sample estimates:
mean in group N mean in group Y
```

```
        72.37500         65.33333
```

6.d

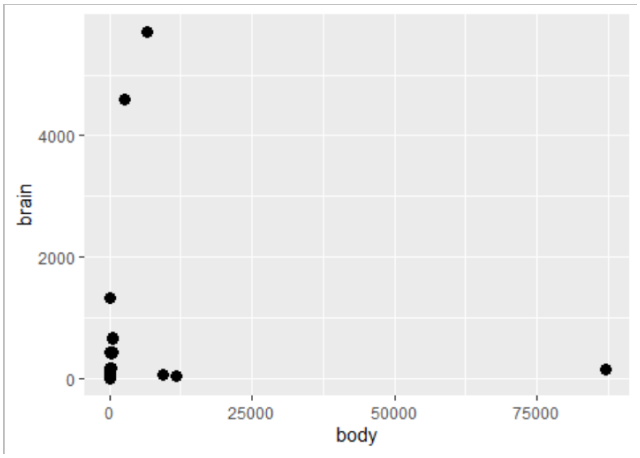
Rstudio subtracted the difference N-Y because it is alphabetical order. This means that the t.test subtracted the drivers wearing seatbelts from the drivers without the seatbelts. The 95% confidence interval (-0.99 to 15.07) includes zero. That means we can't reasonably conclude that the two means of drivers with seatbelts and drivers without seatbelts are different.

7.

7.a.

Plot of brain size on body without a transformation:

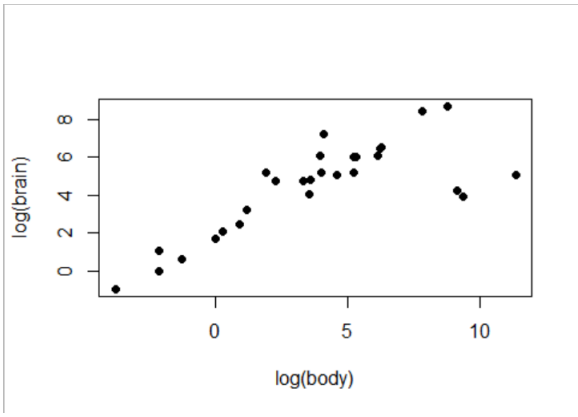
```
> ggplot(Animals,aes(body,brain))+geom_point(size=3)
```



Plot of Brain on body with transformations.

The following transformation appeared to make the data reasonably linear.

```
> plot(log(body),log(brain),pch=19,cex=1)
```



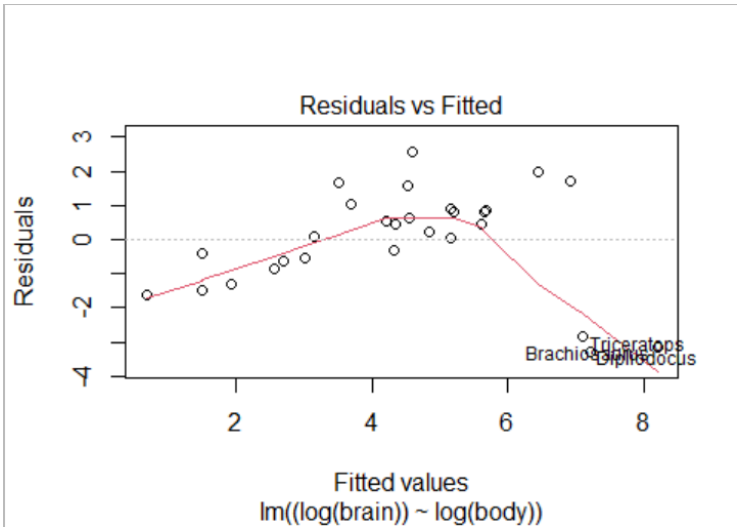
7.b Regress your transformed brain size on transformed body size. Look at the summary .

```
> b.lm<-lm((log(brain))~log(body),data=Animals)
```

```
> summary(b.lm)

Call:
lm(formula = (log(brain)) ~ log(body), data = Animals)

Residuals:
    Min      1Q  Median      3Q
-3.2890 -0.6763  0.3316  0.8646
    Max
 2.5835

Coefficients:
            Estimate Std. Error
(Intercept)  2.55490    0.41314
log(body)    0.49599    0.07817
            t value Pr(>|t|)
(Intercept)   6.184 1.53e-06 ***
log(body)     6.345 1.02e-06 ***
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
  0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.532 on 26 degrees of freedom
Multiple R-squared:  0.6076,     Adjusted R-squared:  0.5925
F-statistic: 40.26 on 1 and 26 DF,  p-value: 1.017e-06
```

The y intercept is 2.55 and the slope is .49. The p-value (p=1.017e-06) indicates there is very strong evidence there is a difference in means between body size and brain size and there is a strong correlation between brain size and body size. The r squared value is .59 meaning that 59% of the variation in Brain size can be explained by knowing what the associated body size is. The p-values for the coefficients (p=1.53e-06) and (p=1.02e-06) are both significant.

```
> plot(b.lm,1)
```



There appears to be a pattern of a reverse quadratic and heteroskdacaity. There also appears to be outliers on the bottom right. Without the outliers there appears to also be a positive linear pattern.

7.c

The three animals that form the outlying cluster is the Dipliodocus, Brachiosaurus, Triceratops.The difference in these outliers is they have massive bodies compared to their brain (head).

7.d

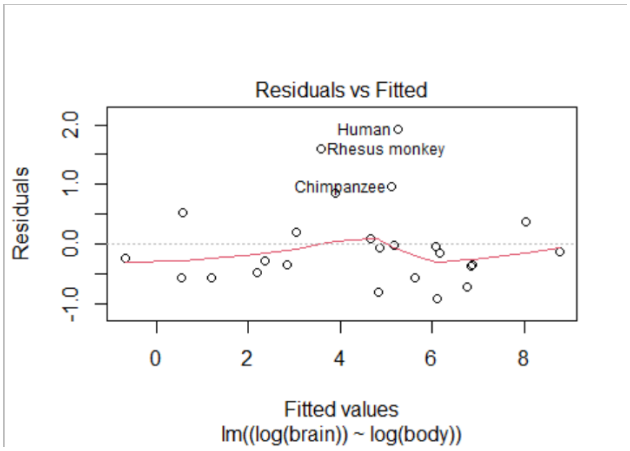
To remove the outliers

```
> outliers.removed<-Animals[-c(6,16,26),]

> outliers.removed

> r.lm<-lm((log(brain))~log(body),data=outliers.removed)

> summary(r.lm)


Call:

lm(formula = (log(brain)) ~ log(body), data = outliers.removed)


Residuals:

    Min      1Q  Median      3Q     Max

-0.9125 -0.4752 -0.1557  0.1940  1.9303


Coefficients:

            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  2.15041    0.20060   10.72 2.03e-10 ***
log(body)    0.75226    0.04572   16.45 3.24e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.7258 on 23 degrees of freedom
Multiple R-squared:  0.9217,     Adjusted R-squared:  0.9183
F-statistic: 270.7 on 1 and 23 DF,  p-value: 3.243e-14
```

Dicussion: The y intercept is 2.51 and the slope is 0.77. Both coefficients p-value are significantly different from zero (p= 2.03e-10) and (p=3.24e-14). The r-square value is .932 meaning that 93% of the variation in Brain size can be explained by knowing what the associated body size is after the three outliers were removed. The r squared suggests a very strong correlation. There is strong evidence (p=3.243e-14) that there is a difference in means of body size and brain size.

```
>plot(r.lm,1)
```



The residual plot indicate no real pattern and there appears to be no heteroskedasticity. There are three positive outliers (Humans, Rhesus monkey, and Chimpanzee). The three animals being largely positive indicates that our brain size is large in proportion to our body size. This is understandable because humans are the most intelligent animals on earth and monkeys and chimpanzees are closely related to us. We deem monkeys and chimpanzees as intelligent as well therefore have larger brains.

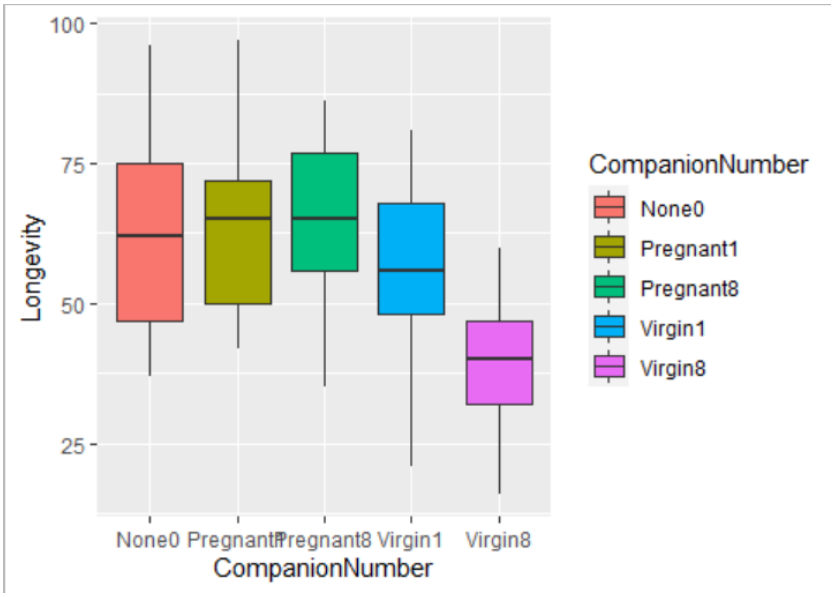
8.

8.a

```
> fruitfly<-read.csv(file.choose())
> fruitfly
```

8.b

```
>
ggplot(fruitfly,aes(CompanionNumber,Longevity,fill=CompanionNumber))+g
eom_boxplot()
```



The variance between the groups look reasonably similar. There are no huge outliers. The range of spread is similar. There is no apparent skew in either group therefore there isn't a reason to assume it isn't normal.

8.c

```
> anova(lm(Longevity~CompanionNumber,data=fruitfly))

Analysis of Variance Table


Response: Longevity

                 Df Sum Sq Mean Sq F value    Pr(>F)

CompanionNumber   4  11939 2984.82  13.612 3.516e-09 ***

Residuals       120  26314  219.28

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value of (p= 3.516e-09) suggests that there is a difference in means between at least one of the fruit fly groups. There is significant evidence for at least one mean being different from others.

8.d.

```
> fruitfly.lm<-lm(Longevity~CompanionNumber,data=fruitfly)

> fit.contrast(fruitfly.lm,"CompanionNumber",c(-
1,1/2,1/2,0,0),conf.int=.95)
```

```
                                        Estimate Std. Error   t value
Pr(>|t|)  lower CI upper CI
CompanionNumber c=( -1 0.5 0.5 0 0 )     0.52   3.627225 0.1433603
0.8862461
lower CI upper CI
 -6.661653 7.701653
attr(,"class")
[1] "fit_contrast"
```

The p-value is (which is highlighted) is .88 indicating there is no evidence that the male flies group that were individually isolated from other flies is different than the mean of male flies placed with 1 pregnant fruit fly and male flies placed with 8 pregnant fruit flies. The confidence interval is from -6.66 to 7.70 which includes the value of zero meaning there isn't a difference in true means. I conclude there is no difference in the three control groups.

The fit contrast function subtracted the None0 group from the two pregnant groups. I chose this order (seen in my order of my coefficients) because the means from the pregnant groups appeared to be slightly higher than the mean for the none0 group in the boxplot.

6.e

```
> fit.contrast(fruitfly.lm,"CompanionNumber",c(1/3,1/3,1/3,-
1,0),conf.int=.95)


CompanionNumber c=( 0.333333333333333 0.333333333333333
0.333333333333333 -1 0 )
Estimate Std. Error  t value   Pr(>|t|)  lower CI upper CI
7.146667    3.41978 2.089803 0.03874761 0.3757395 13.91759
attr(,"class")
[1] "fit_contrast"
```

The p-value (p=0.038) indicates there is good or moderate evidence that the true mean of longevity of male flies placed with one virgin female fly is different than the mean longevity of all three control groups. The confidence interval (0.375 to 13.91) indicates that 95% of the true mean values will be above zero meaning will be different than the control groups. This relates to the greater question because the mean longevity of the control groups is larger than the mean longevity of the male flies placed with one virgin fly which is supported by the confidence interval and p-value. The male flies that were surrounded by the virgin flies had a shorter lifespan than the control groups.

6.f

```
> fit.contrast(fruitfly.lm,"CompanionNumber",c(1/3,1/3,1/3,0,-
1),conf.int=.95)


CompanionNumber c=( 0.333333333333333 0.333333333333333
0.333333333333333 0 -1 )

Estimate Std. Error  t value     Pr(>|t|) lower CI upper CI

25.18667    3.41978 7.364995 2.448098e-11 18.41574 31.95759

attr(,"class")

[1] "fit_contrast"
```

The p-value (p=2.44e-1) indicates there is very strong evidence that the true mean of longevity of male flies placed with eight virgin female flies is different than the mean longevity of all three control groups. The confidence interval (18.41 to 31.95) indicates that 95% of the true mean values will be above zero by a significant amount . This relates to the greater question because the mean longevity of the control groups is larger than the mean longevity of the male flies placed with eight virgin fly which is supported by the confidence interval and p-value. The p-value is stronger than the comparison from above. The male flies that were surrounded by the virgin flies had a shorter lifespan than the control groups which does not support the hypothesis that impregnating females will also shorten the lifespan of male fruit flies.

6.g

```
> fit.contrast(fruitfly.lm,"CompanionNumber",c(0,0,0,1,-
1),conf.int=.95)


CompanionNumber c=( 0 0 0 1 -1 )

 Estimate Std. Error  t value Pr(>|t|) lower CI upper CI


  18.04   4.188358 4.307177 3.401023e-05 9.747342 26.33266

CompanionNumber c=( 0 0 0 1 -1 ) 26.33266

attr(,"class")

[1] "fit_contrast"
```

The p-value (p=3.40e-05) indicates there is very strong evidence that the true mean of longevity of males placed with one female virgin fly is different than the true mean of longevity of males placed with eight female flies. The confidence interval of (9.74 to 26.33) indicates that 95% of the true mean values will be above zero by a significant amount. This relates to the larger question because the longevity decreases as the male fly is surrounded by more virgin flies. The mean longeveity in male flies surrounded by only one virgin fly was significantly greater than the mean longevity in male flies sour rounded by eight virgin female flies.

8.h

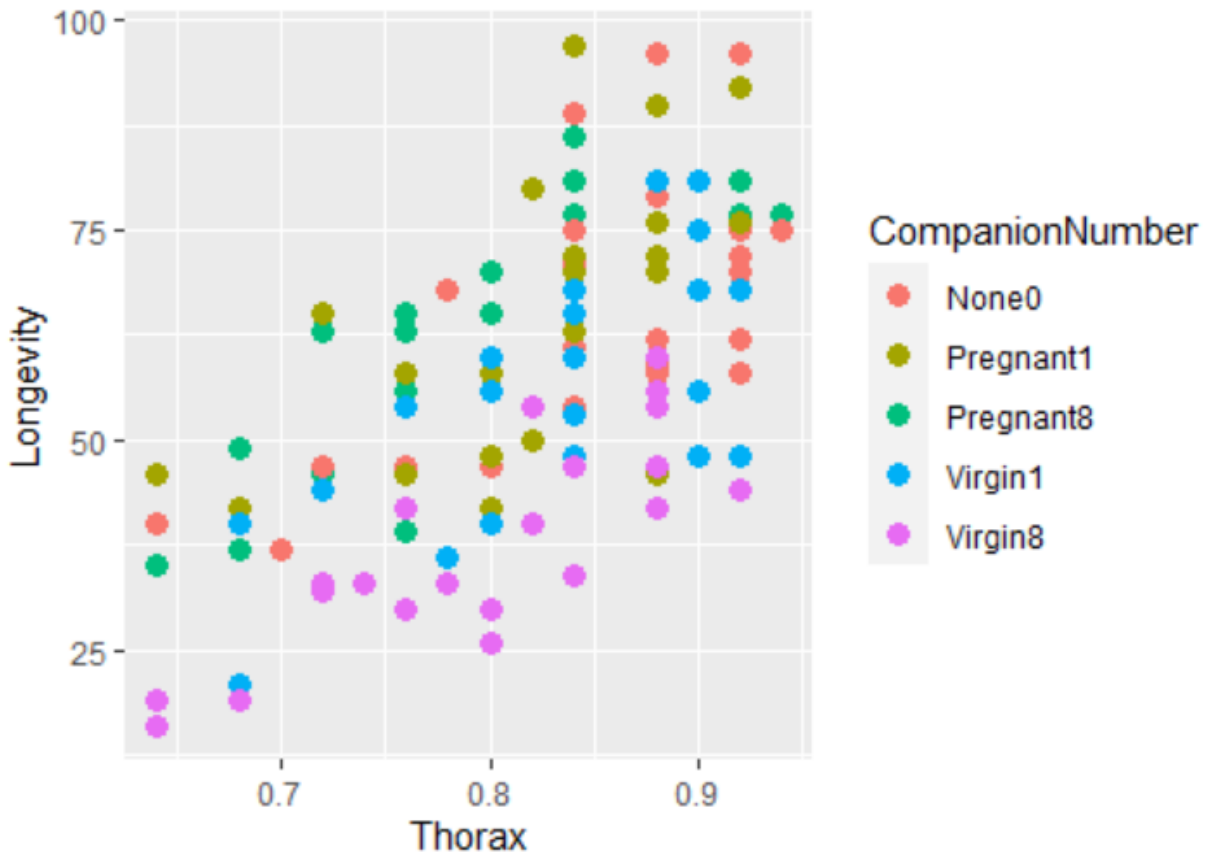
```
> thorax.lm<-lm(Longevity~CompanionNumber+Thorax,data=fruitfly)

> anova(fruitfly.lm,thorax.lm)

Analysis of Variance Table


Model 1: Longevity ~ CompanionNumber

Model 2: Longevity ~ CompanionNumber + Thorax

  Res.Df   RSS Df Sum of Sq      F    Pr(>F)

1    120 26314

2    119 13145  1     13169 119.22 < 2.2e-16 ***

---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The p value (p=2.2e-16) is essentially zero. We have strong evidence that the model that incorporates thorax length explains significantly more variation than the model that doesn't not incorporate thorax length. I think it is important to look at the coefficients of the Thorax.lm model as well to get a better idea of the data.

8. I

```
>
ggplot(fruitfly,aes(Thorax,Longevity,color=CompanionNumber))+geom_poin
t(size=3)
```



There appears to be a positive linear association when you look at each group separately. It appears that longevity increases for each group depending on the length of the thorax. As a whole the data appears to have a positive linear association between longeveity and thorax length.

8j.

```
> anova(f.lm,r.lm)
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
```

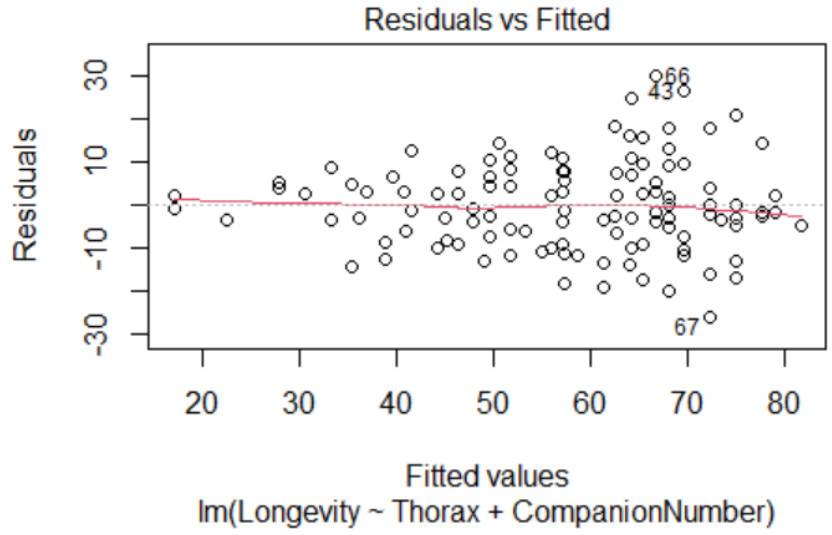
The p value of .9844 indicates there is no evidence that model 2 (the five arbitrary line model) explains more variation than model 1 (the five line parallel model).

8.k

```
> summary(f.lm)

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

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
```

Thorax coefficient (p=2e-16) is significant so the association appears to be real. The p- value of the coefficients for pregnant 1, pregnant 8 and virgin 1 are not significant. The y intercept is -49.98 and the slope is 135. The p-value for the coefficient Virgin8 (p=1.0e-09) appear to be significant. The p- value (p=2.2e-16) is very small meaning there is strong evidence that this is different from zero. The r squared is 0.64 meaning the model explains 64% of the variation. The r squared is relevant and moderately high.

```
> plot(f.lm,1)
```



The residual plot indicate no real pattern and there appears to be some heteroskedasticity in a funnel pattern.

8.i

```
>  fruitfly$Three<-
factor(fruitfly$CompanionNumber,labels=c("None0","None0","None0","Virg
in1","Virgin8"))
```

```
> three.lm<-lm(Longevity~Thorax+Three,data=fruitfly)
```

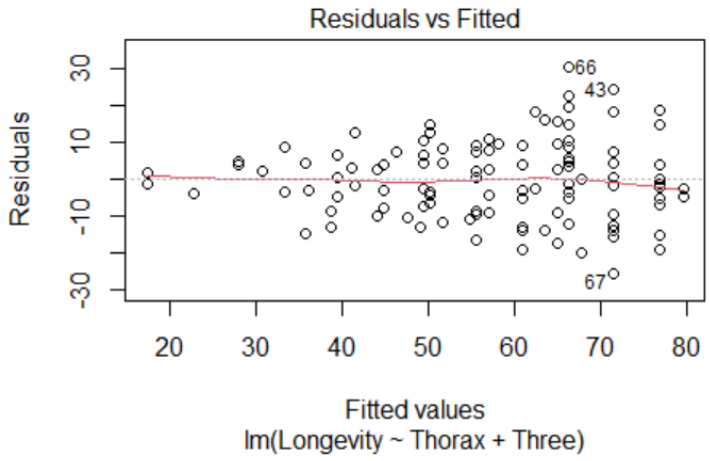
8.m

```
> summary(three.lm)

Call:

lm(formula = Longevity ~ Thorax + Three, data = fruitfly)

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 ***

ThreeVirgin1   -9.181      2.432  -3.775  0.00025 ***

ThreeVirgin8  -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
```

It appears that all the coefficients are significant using the three line parallel model and the indicator variables are significant as well. The adjusted r value is 0.6425 which is higher than the 0.6419 from the 5 parallel model above. The p value (2.2e-16) is very strong indicating there is a significance. The p value and the r squared of the three line model and the five line model are the same (p=2.2e-16) while the r squared is essentially the same value as well. The coeffencents are more significant than the five line model.

```
>plot(three.lm,1)
```



There doesn’t appear to be any pattern but there is a funnel shape and heteroskactidty.

8.n

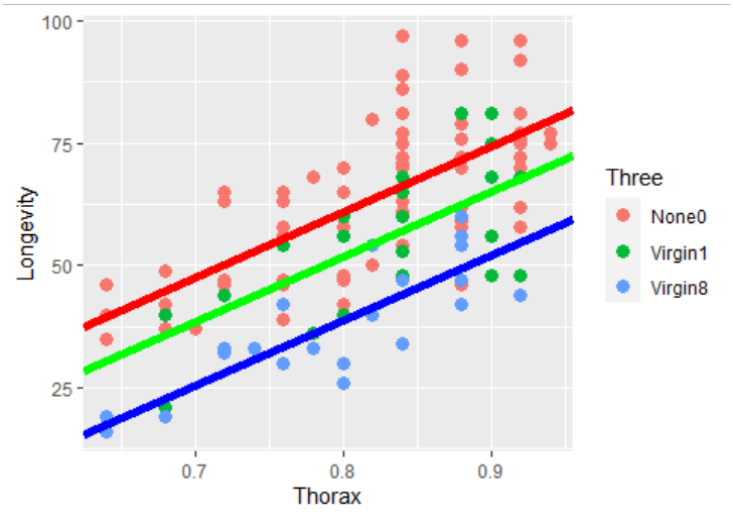
```
> anova(three.lm,f.lm)
Analysis of Variance Table


Model 1: Longevity ~ Thorax + Three
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 second model (the five parallel line model) does not explain a significant difference (p=.4108) amount of variation so we should go with the less complicated model which is the three lined model. We should go with three.lm model.

8.o.

```
>
ggplot(fruitfly,aes(Thorax,Longevity,color=Three))+geom_point(size=3)+
geom_abline(intercept =
c1[1],slope=c1[2],size=2,color='red')+geom_abline(intercept =
c1[1]+c1[3],slope=c1[2],size=2,color="green")+geom_abline(intercept=c1
[1]+c1[4],slope=c1[2],size=2,color="blue")
```



8.p

```
> three.lm$coefficients
 (Intercept)        Thorax ThreeVirgin1 ThreeVirgin8
  -46.161214   133.837404    -9.180995   -22.188709
> co<-three.lm$coefficients
> co[1]-(co[1]+co[3])
(Intercept)
   9.180995
```

It is estimated that the three control groups ‘None0” can live 9 longevity’s (days?) longer than the group of male flies surrounded by one virgin fly.

8.q

```
> co[1]-(co[1]+co[4])
(Intercept)
```

```
22.18871
```

It is estimates that the three control groups “None0” can live 22 longevity’s longer than the group of male flies surrounded by 8 virgin females