Problem set 6

Exercises to hand in: 3.34, 3.36, 3.48 (modified a)

3.34 Fish eggs

a. Simple linear regression

```
data("FishEggs")
m1=lm(PctDM~Age, data=FishEggs)
```

b. Percent of variability

The percentage of variability is about 20%.

```
summary(m1)
```

```
Call:
lm(formula = PctDM ~ Age, data = FishEggs)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9091 -0.8471  0.3822  1.0271  2.1409 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 38.70206    0.86783  44.596   <2e-16 ***
Age         -0.21033    0.07313  -2.876    0.007 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

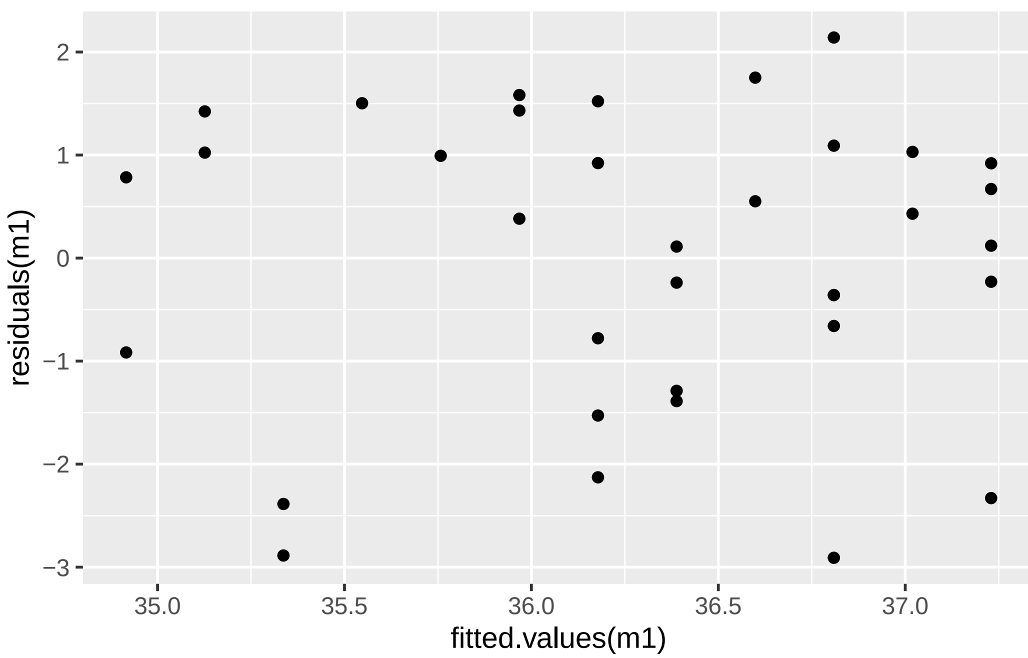
```
Residual standard error: 1.426 on 33 degrees of freedom
Multiple R-squared:  0.2004,    Adjusted R-squared:  0.1762
F-statistic: 8.272 on 1 and 33 DF,  p-value: 0.007001
```

c. Statistically significant?

P-value = 0.07. Greater than 0.05. Therefore can not reject the Ho hypothesis. The results are not statistically significant.

d. Residual v. fitted

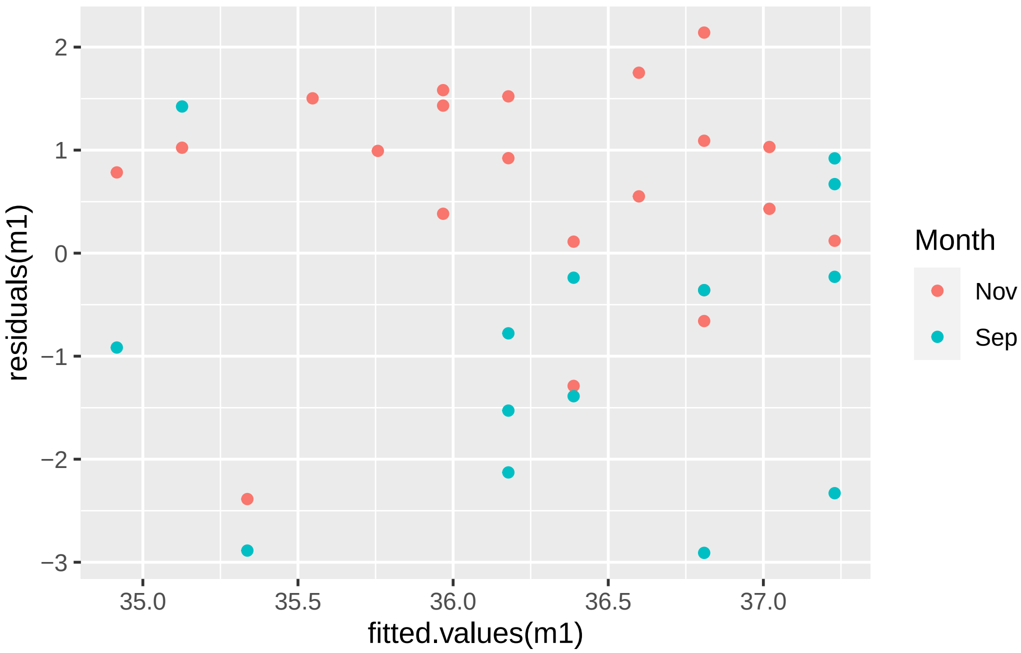
```
ggplot(FishEggs) + geom_point(aes(x = fitted.values(m1), y = residuals(m1)))
```



The graph does not have a regular pattern, because we can see that data plots are not distributed equally through out the dotted lines on the graph.

e. Modified residual v. fitted

```
ggplot(FishEggs)+ geom_point(aes(x=fitted.values(m1),y = residuals(m1), color = Month))
```



f. Need both terms?

```
m2=lm(PctDM~Age + Sept, data=FishEggs)
summary(m2)
```

```
Call:
lm(formula = PctDM ~ Age + Sept, data = FishEggs)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9100 -0.5869  0.2974  0.7599  2.4380 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 39.51922    0.77827  50.778  < 2e-16 ***
```

```
Age         -0.22870    0.06292  -3.635 0.000965 ***
Sept        -1.51929    0.42342  -3.588 0.001096 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.223 on 32 degrees of freedom
Multiple R-squared:  0.4298,    Adjusted R-squared:  0.3942
F-statistic: 12.06 on 2 and 32 DF,  p-value: 0.0001248
```

We would not use both the graphs because they are not significant, because of their p value which is about 0.000 for age and the p value for Sept would be 0.001 which is less than 0.005.

g. Percent of variability

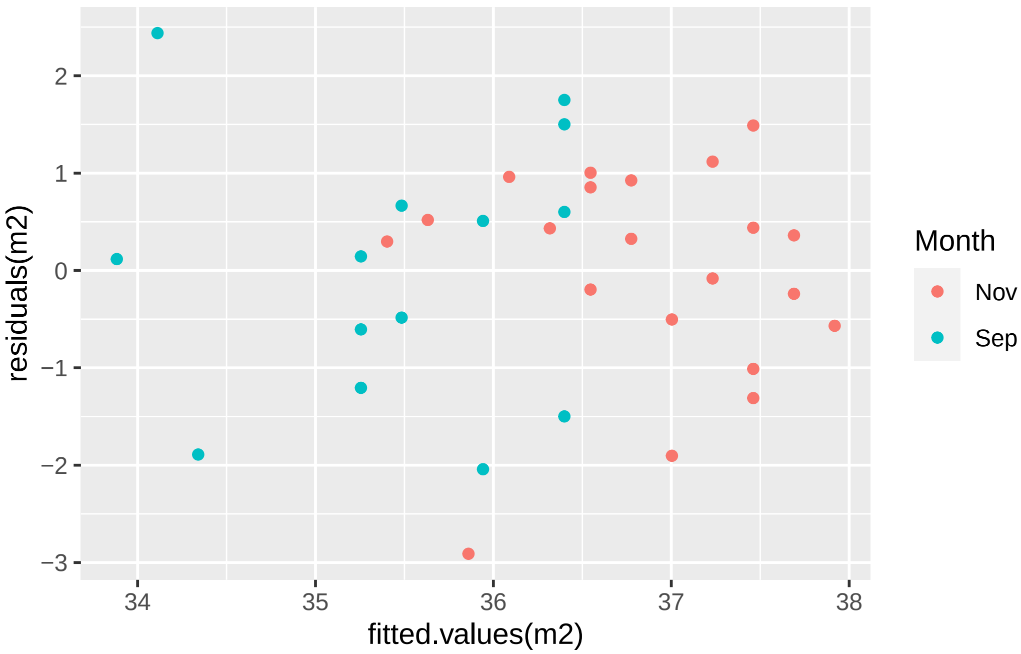
In this new model, there is a multiple R-squared value of 0.4298, which indicates that 42.98% of the data fit the regression model.

```
m2=lm(PctDM~Age + Sept, data=FishEggs)
summary(m2)
```

```
Call:
lm(formula = PctDM ~ Age + Sept, data = FishEggs)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9100 -0.5869  0.2974  0.7599  2.4380

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.51922    0.77827  50.778  < 2e-16 ***
Age         -0.22870    0.06292  -3.635 0.000965 ***
Sept        -1.51929    0.42342  -3.588 0.001096 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.223 on 32 degrees of freedom
Multiple R-squared:  0.4298,    Adjusted R-squared:  0.3942
F-statistic: 12.06 on 2 and 32 DF,  p-value: 0.0001248
```

h. Redo plot

```
ggplot(FishEggs)+ geom_point(aes(x=fitted.values(m2),y = residuals(m2), color = Month))
```



This new model is a better way to predict the PctDM points since the residual points in this plot are more centered or closer to 0 which shows increased accuracy.

3.36 Elephants

```
data("ElephantsMF")
```

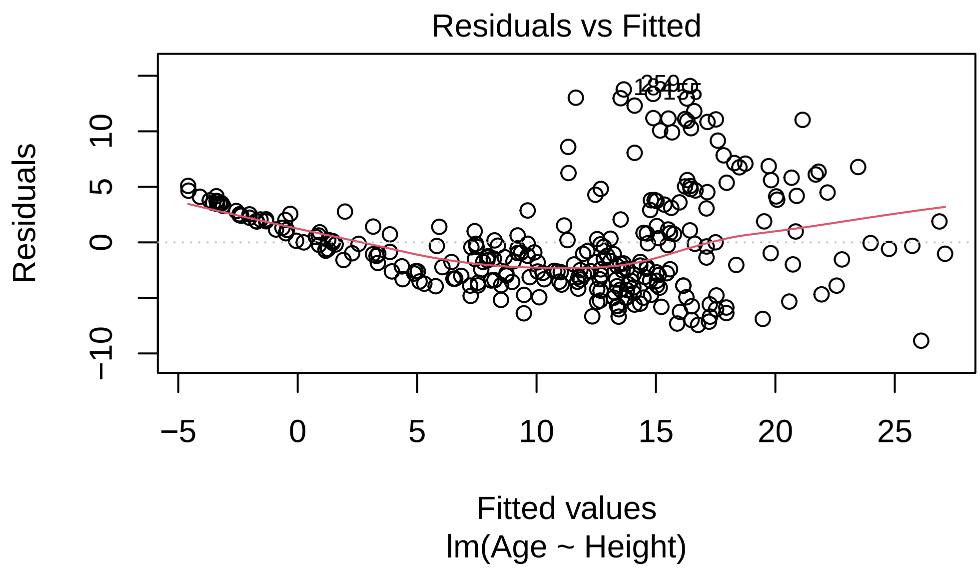
a. Plot

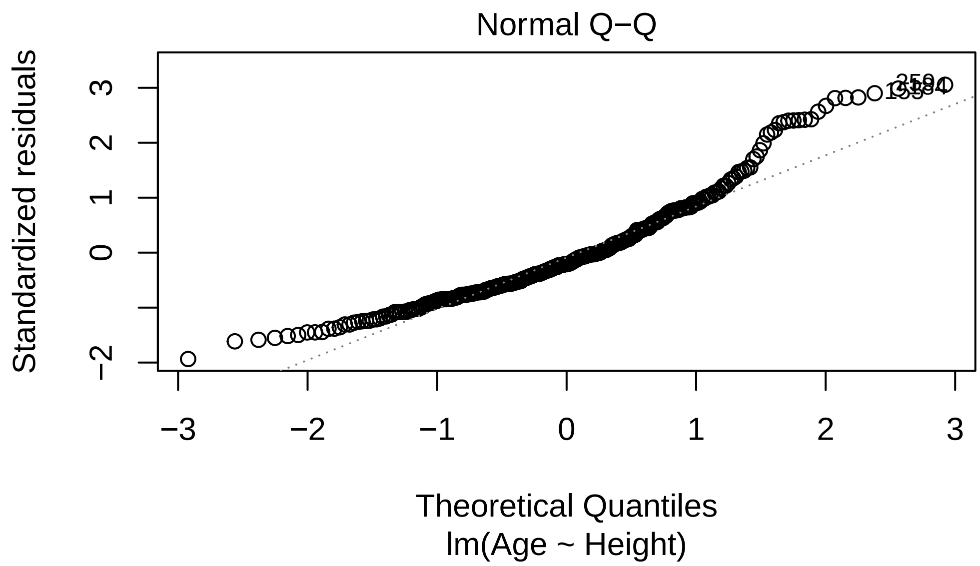
The pattern does not look linear. The distribution of the data plots are not equally distributed. Normal Q-Q plots at the start and end of the graph the data tends to move away from the dotted line. The graph in general looks curved.

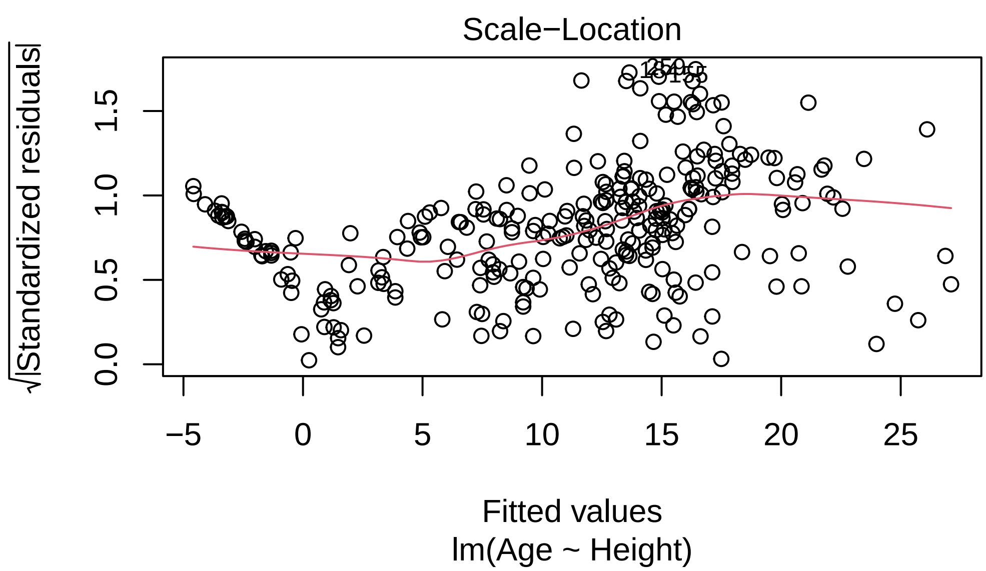
```
m1 <- lm(Age~Height, data = ElephantsMF)
summary(m1)
```

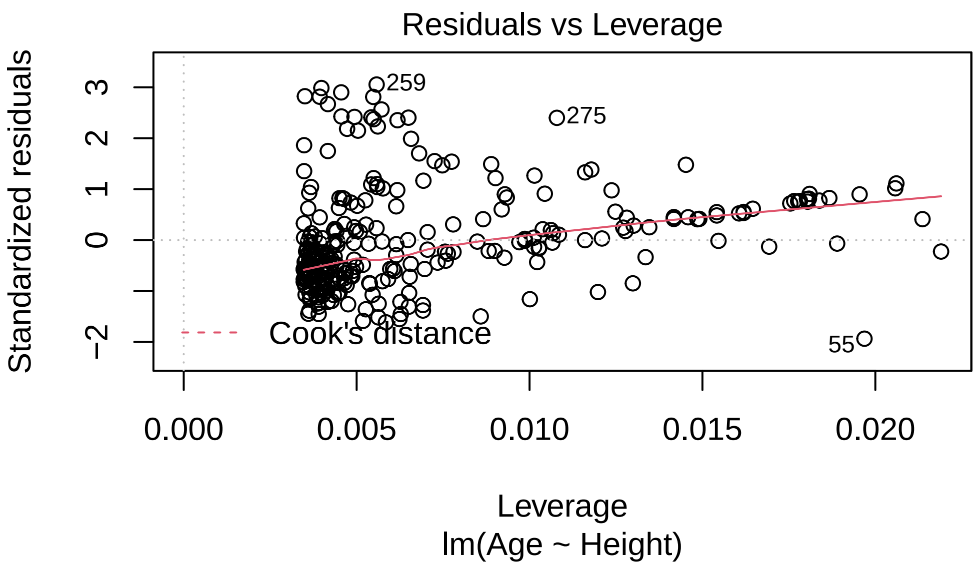
```
Call:
lm(formula = Age ~ Height, data = ElephantsMF)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8549 -3.3159 -0.9629  2.4614 14.0736 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -15.049133   1.047306  -14.37   <2e-16 ***
Height        0.138640   0.005388   25.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.619 on 286 degrees of freedom
Multiple R-squared:  0.6983,	Adjusted R-squared:  0.6973 
F-statistic:   662 on 1 and 286 DF,  p-value: < 2.2e-16
```

```
plot(m1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Age ~ Height)


Scale−Location
√|Standardized residuals|
Fitted values
lm(Age ~ Height)



b. Quadratic regression

Height = 187.683 + (716.385 * Age) - (338.586 * Age^2)

```
m7 <- lm(Height~poly(x=Age,degree = 2), data = ElephantsMF)
summary(m7)
```

```
Call:
lm(formula = Height ~ poly(x = Age, degree = 2), data = ElephantsMF)

Residuals:
    Min      1Q  Median      3Q     Max
-52.910 -13.337  -1.226  11.900  66.968

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 187.683      1.142  164.32   <2e-16 ***
poly(x = Age, degree = 2)1  716.385     19.383   36.96   <2e-16 ***
poly(x = Age, degree = 2)2 -338.586     19.383  -17.47   <2e-16 ***
---
```

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

Residual standard error: 19.38 on 285 degrees of freedom
Multiple R-squared:  0.8543,    Adjusted R-squared:  0.8533
F-statistic: 835.5 on 2 and 285 DF,  p-value: < 2.2e-16
```

c. Prediction

```
predict(m7, newdata = data.frame(Age = 10))
```

```
       1
200.5113
```

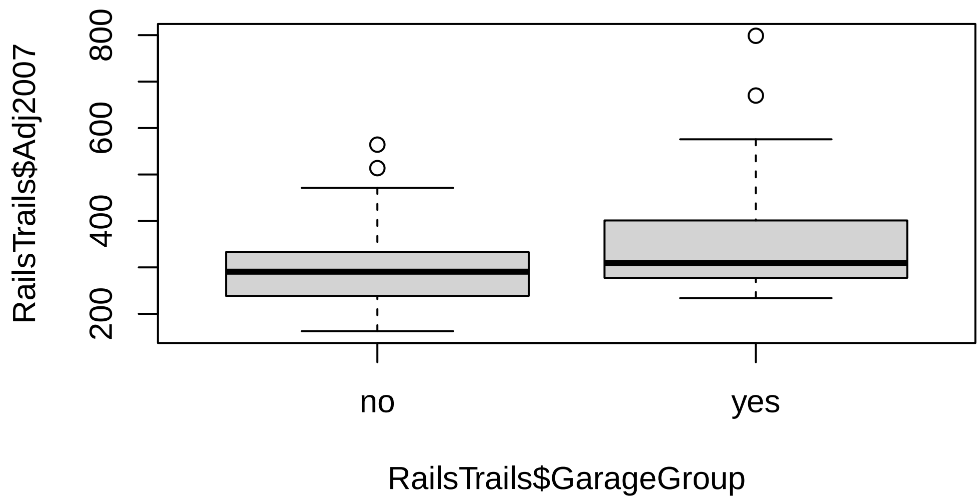
Based on the model, a 10-year old elephant would have the height of approximately 200.51 cm.

3.48 Real estate near Rails to Trails: nested F-test.

```
data("RailsTrails")
```

a. Use comparative boxplots and a simple linear regression model to determine if having a garage is related to the price of a home. In other words, fit a simple linear regression using GarageGroup as a predictor. Use the t-value and p-value associated with the coefficient to perform a hypothesis test. (This is exactly the same as doing a t-test, but we are not focusing on t-tests in this class.)

```
boxplot(RailsTrails$Adj2007~RailsTrails$GarageGroup)
```



b. Simple linear regression

We would expect the selling price of a house to go down. Adj2007 = 388.204 - (54.427 * Distance)

```
fit_model1 = lm(Adj2007~Distance,data = RailsTrails)
summary(fit_model1)
```

```
Call:
lm(formula = Adj2007 ~ Distance, data = RailsTrails)

Residuals:
    Min      1Q  Median      3Q     Max
-190.55  -58.19  -17.48   25.22  444.41

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  388.204     14.052  27.626  < 2e-16 ***
Distance     -54.427      9.659  -5.635 1.56e-07 ***
---
```

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

Residual standard error: 92.13 on 102 degrees of freedom
Multiple R-squared:  0.2374,    Adjusted R-squared:  0.2299
F-statistic: 31.75 on 1 and 102 DF,  p-value: 1.562e-07
```

c. Multiple regression

We would expect interpretations for each of Garage Group and Distance is held constant.

```r
fit_model2 = lm (Adj2007~Distance+GarageGroup,data = RailsTrails)
summary(fit_model2)
```

```
Call:
lm(formula = Adj2007 ~ Distance + GarageGroup, data = RailsTrails)

Residuals:
    Min      1Q  Median      3Q     Max
-167.88  -51.55  -21.88   36.79  427.49

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     365.103     17.661  20.673   <2e-16 ***
Distance        -51.025      9.638  -5.294    7e-07 ***
GarageGroupyes   37.892     18.032   2.101   0.0381 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 90.62 on 101 degrees of freedom
Multiple R-squared:  0.2693,    Adjusted R-squared:  0.2549
F-statistic: 18.62 on 2 and 101 DF,  p-value: 1.311e-07
```

d. Interaction

Adj^2007noGarage = 359.083 -46.302*Distance

Adj^Garage = 407.945 -56.180 *Distance

```r
fit_model3 = lm (Adj2007~Distance*GarageGroup,data = RailsTrails)
summary(fit_model3)
```

```
Call:
lm(formula = Adj2007 ~ Distance * GarageGroup, data = RailsTrails)

Residuals:
    Min      1Q  Median      3Q     Max
-162.46  -51.65  -17.22   30.04  425.76

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              359.083     21.295  16.862  < 2e-16 ***
Distance                 -46.302     13.391  -3.458 0.000802 ***
GarageGroupyes            48.862     28.108   1.738 0.085222 .
Distance:GarageGroupyes   -9.878     19.366  -0.510 0.611125
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 90.96 on 100 degrees of freedom
Multiple R-squared:  0.2712,	Adjusted R-squared:  0.2494
F-statistic: 12.41 on 3 and 100 DF,  p-value: 5.785e-07
```

e. Nested F-test

```
anova(fit_model1,fit_model2,fit_model3)
```

```
Analysis of Variance Table

Model 1: Adj2007 ~ Distance
Model 2: Adj2007 ~ Distance + GarageGroup
Model 3: Adj2007 ~ Distance * GarageGroup
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    102 865718
2    101 829453  1     36265 4.3835 0.03882 *
3    100 827301  1      2152 0.2602 0.61113
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can conclude that the second model with only the Distance predictor variable and GarageGroup indicator is the best model used out of the 3 because of it's significance. The p-value is 0.03809 which is less than 0.05, this means that GarageGroup is significant when it is used as an indicator that adds significantly to the model of price on distance.