STAT302 Assignment 3 Solution

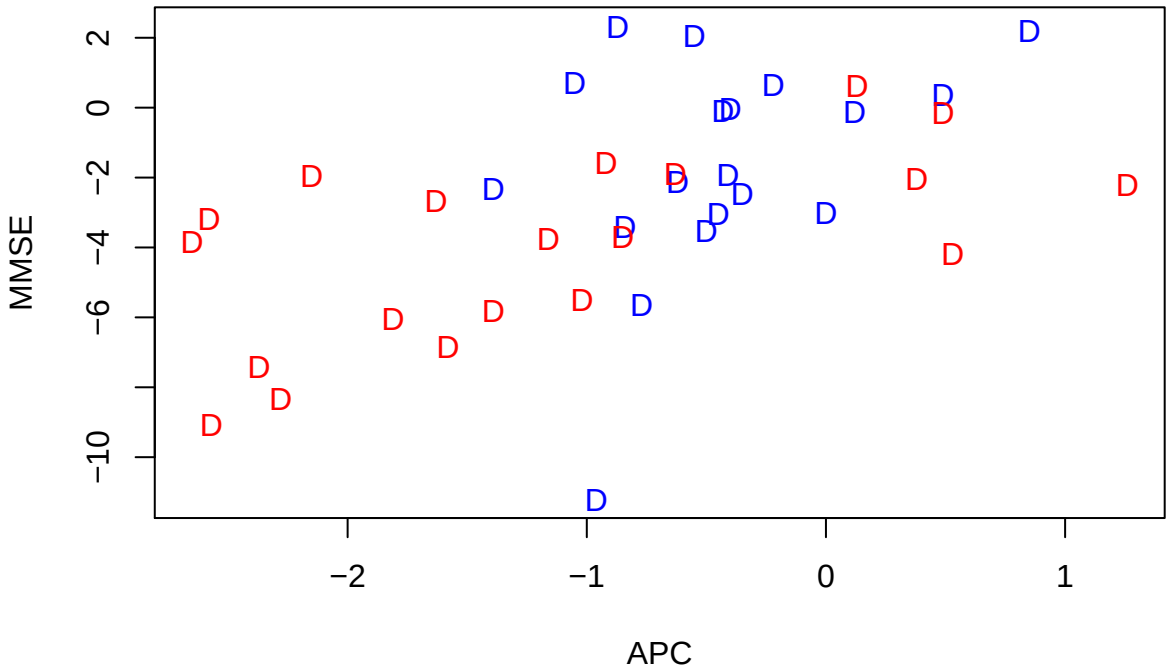
Daisy Yu

7/23/2023

Q3.24 (7 points)

(a). (2 points)

```
library(Stat2Data)
data("LewyBody2Groups")
plot(MMSE ~ APC, type="n", data=LewyBody2Groups, ylab="MMSE", xlab="APC")
data_DLB <- LewyBody2Groups[LewyBody2Groups$Type=='DLB',]
data_DLB_AD <- LewyBody2Groups[LewyBody2Groups$Type=='DLB/AD',]
points(MMSE ~ APC, pch="DLB", col="blue", data=data_DLB)
points(MMSE ~ APC, pch="DLB/AD", col="red", data=data_DLB_AD)
```



We can see from the scatterplot that there is a linear relationship between MMSE and APC.

(b). (3 points)

```
model1 <- lm(MMSE ~ APC, data=LewyBody2Groups)
summary(model1)
```

```
##
```

```
## Call:
## lm(formula = MMSE ~ APC, data = LewyBody2Groups)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1022 -1.7043  0.2174  1.9484  5.2706
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.4214     0.5528  -2.572 0.014277 *
## APC           1.7462     0.4401   3.968 0.000321 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.664 on 37 degrees of freedom
## Multiple R-squared:  0.2985, Adjusted R-squared:  0.2795
## F-statistic: 15.74 on 1 and 37 DF,  p-value: 0.0003208
```

t-statistics $= 3.968$
p-value $= 0.000321 < 0.05$
So we conclude that there is a significant linear relationship between MMSE and APC.

(c). (2 points)

```
model2 <- lm(MMSE ~ APC + Type, data=LewyBody2Groups)
summary(model2)
```

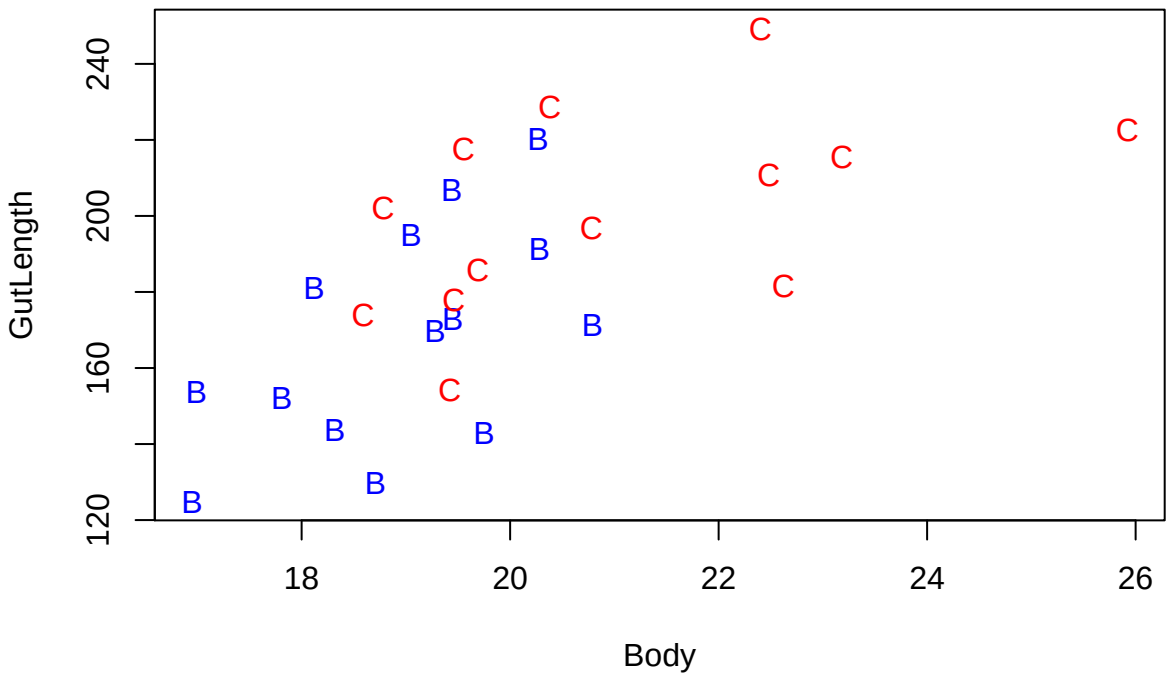
```
##
## Call:
## lm(formula = MMSE ~ APC + Type, data = LewyBody2Groups)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8153 -1.6382 -0.1469  1.9103  4.5796
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9433     0.6358  -1.484  0.14662
## APC           1.5015     0.4650   3.229  0.00265 **
## TypeDLB/AD   -1.3135     0.9017  -1.457  0.15385
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.624 on 36 degrees of freedom
## Multiple R-squared:  0.3375, Adjusted R-squared:  0.3007
## F-statistic:  9.17 on 2 and 36 DF,  p-value: 0.0006042
```

DLB: $\hat{MMSE} = -0.9433 + 1.5015APC$
DLB/AD: $\hat{MMSE} = -0.9433 + 1.5015APC - 1.3135 = -2.2568 + 1.5015APC$

Q3.26 (9 points)

a. (2 points)

```
data(Tadpoles)
plot(GutLength ~ Body, type="n", data=Tadpoles, ylab="GutLength", xlab="Body")
data_bd <- Tadpoles[Tadpoles$Treatment=='Bd',]
data_control <- Tadpoles[Tadpoles$Treatment=='Control',]
points(GutLength ~ Body, pch="Bd", col="blue", data=data_bd)
points(GutLength ~ Body, pch="Control", col="red", data=data_control)
```



We can see from the scatterplot that there is a linear relationship between GutLength and Body.

b. (3 points)

```
m1 <- lm(GutLength ~ Body, data=Tadpoles)
summary(m1)
```

```
##
## Call:
## lm(formula = GutLength ~ Body, data = Tadpoles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.575 -22.245   0.027  17.815  39.998
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -20.764     49.002  -0.424 0.675384
## Body          10.280      2.445   4.204 0.000293 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 24.83 on 25 degrees of freedom
## Multiple R-squared:  0.4141, Adjusted R-squared:  0.3907
## F-statistic: 17.67 on 1 and 25 DF,  p-value: 0.0002931
```

t-statistics = 4.204
p-value = 0.000293 < 0.05
So we conclude that there is a significant linear relationship between GutLength and Body.

c. (2 points)

```
m2 <- lm(GutLength ~ Body + Treatment, data=Tadpoles)
summary(m2)
```

```
##
## Call:
## lm(formula = GutLength ~ Body + Treatment, data = Tadpoles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.462 -15.006  -2.545  19.117  41.298
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         15.47      53.81   0.288  0.77618
## Body                 8.07       2.82   2.862  0.00859 **
## TreatmentControl    16.28      11.03   1.476  0.15286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.26 on 24 degrees of freedom
## Multiple R-squared:  0.4629, Adjusted R-squared:  0.4181
## F-statistic: 10.34 on 2 and 24 DF,  p-value: 0.0005762
```

Bd: $\hat{GutLength} = 15.47 + 8.07Body$
Control: $\hat{GutLength} = 15.47 + 8.07Body + 16.28 = 31.75 + 8.07Body$

d. (2 points)

```
m3 <- lm(GutLength ~ Body + Treatment + MouthpartDamage, data=Tadpoles)
summary(m3)
```

```
##
## Call:
## lm(formula = GutLength ~ Body + Treatment + MouthpartDamage,
##     data = Tadpoles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.422 -17.701  -6.771  16.338  40.877
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
```

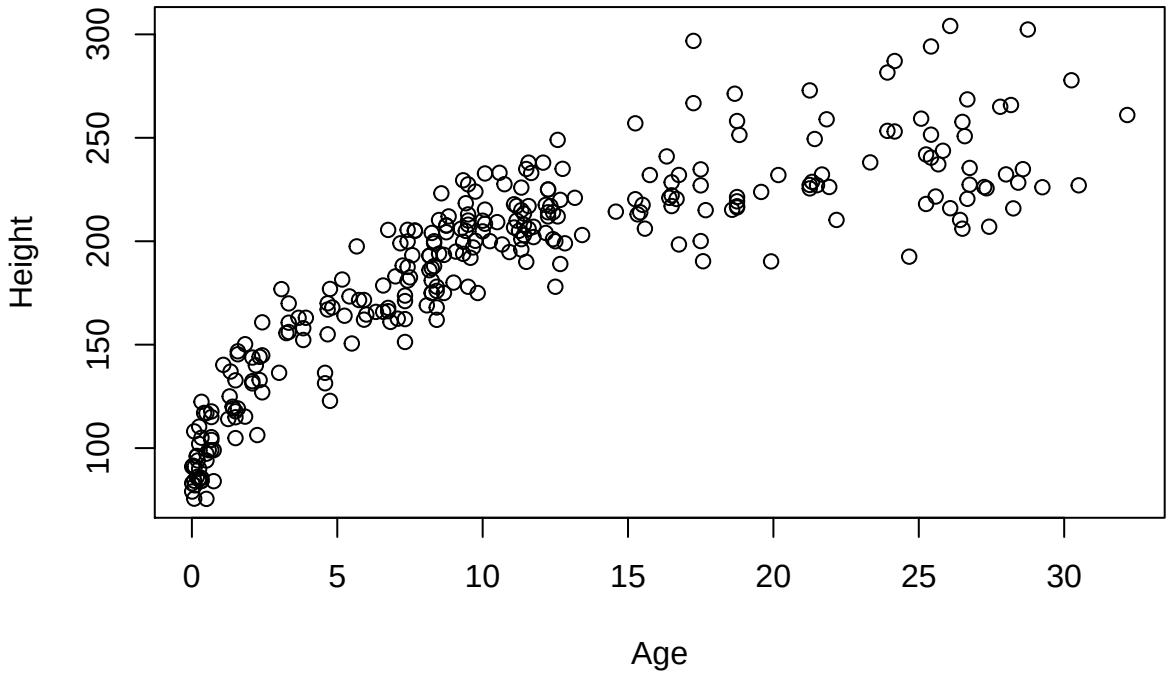
```
## (Intercept)       -20.258     53.070  -0.382   0.7062
## Body                6.442      2.746   2.346   0.0280 *
## TreatmentControl   25.412     11.177   2.274   0.0326 *
## MouthpartDamage    96.839     45.839   2.113   0.0457 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.68 on 23 degrees of freedom
## Multiple R-squared:  0.5502, Adjusted R-squared:  0.4915
## F-statistic: 9.378 on 3 and 23 DF,  p-value: 0.0003092
```

The new fitted model which includes MouthpartDamage as a predictor supports biologists' hypothesis that GutLength has a positive relationship with both Treatment and MouthpartDamage.

Q3.36 (4 points)

a. (2 point)

```
data("ElephantsMF")
plot(Height ~ Age, data=ElephantsMF, ylab="Height", xlab="Age")
```



We can see from the scatterplot that the relationship between Height and Age is quadratic instead of linear.

b. (1 point)

```
ElephantsMF$Age2 <- (ElephantsMF$Age)^2
fit <- lm(Height ~ Age+Age2, data=ElephantsMF)
summary(fit)
```

```
##
## Call:
```

```
## lm(formula = Height ~ Age + Age2, 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) 102.48332    2.54514   40.27   <2e-16 ***
## Age          12.56560    0.45204   27.80   <2e-16 ***
## Age2         -0.27628    0.01582  -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
```

$\hat{Height} = 102.483 + 12.566Age - 0.276Age^2$

c. (1 point)

```
newdata = data.frame(Age=10,Age2=10^2)
predict(fit, newdata)
```

```
##        1
## 200.5113
```

Q3.52 (8 points)

a. (2 points)

```
full_model <- lm(MMSE ~ APC * Type, data=LewyBody2Groups)
summary(full_model)
```

```
##
## Call:
## lm(formula = MMSE ~ APC * Type, data = LewyBody2Groups)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3905 -1.5841 -0.1014  1.6959  4.9309
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.5846     0.7927  -0.738   0.4657
## APC              2.3176     1.1640   1.991   0.0543 .
## TypeDLB/AD      -1.8513     1.1471  -1.614   0.1155
## APC:TypeDLB/AD  -0.9732     1.2712  -0.766   0.4490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.64 on 35 degrees of freedom
## Multiple R-squared:  0.3484, Adjusted R-squared:  0.2926
## F-statistic: 6.239 on 3 and 35 DF,  p-value: 0.001656
```

DLB: $\hat{MMSE} = -0.5846 + 2.3176APC$
DLB/AD: $\hat{MMSE} = -0.5846 + 2.3176APC - 1.8513 - 0.9732APC = -2.4359 + 1.3444APC$

b. (3 points)

t-statistic = -0.766
p-value = $0.4490 > 0.05$
So we reject null hypothesis and conclude that the interaction term is not needed.

c. (3 points)

```
reduced_model <- lm(MMSE ~ APC, data=LewyBody2Groups)
anova(full_model,reduced_model)
```

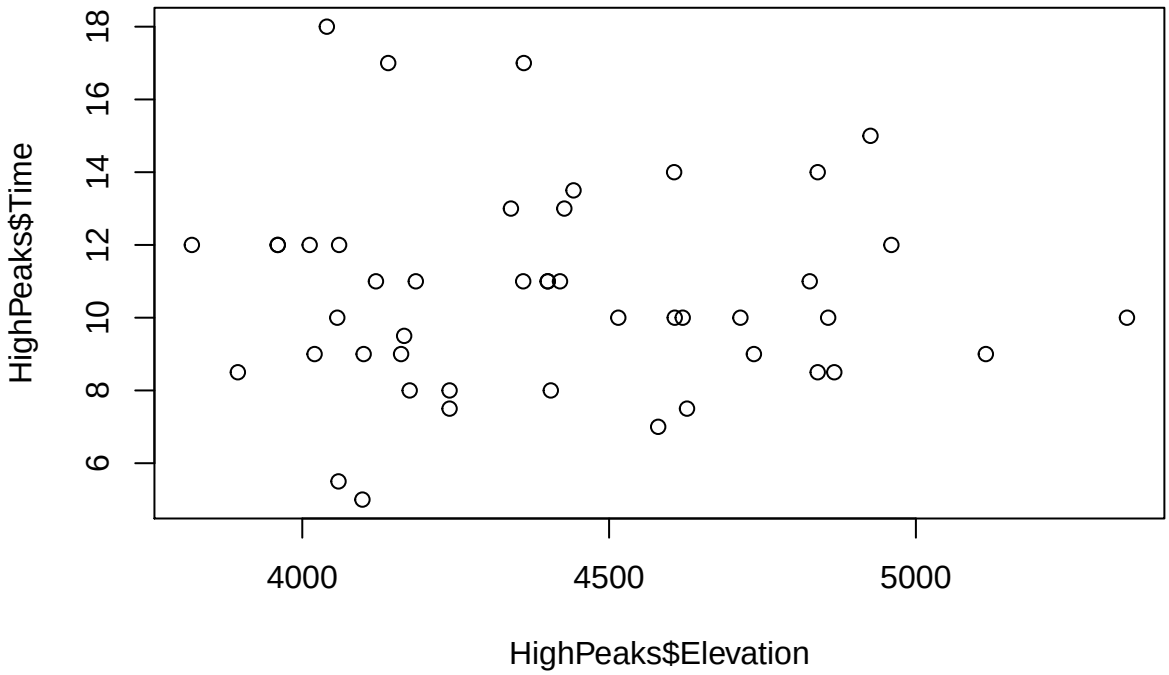
```
## Analysis of Variance Table
##
## Model 1: MMSE ~ APC * Type
## Model 2: MMSE ~ APC
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     35 243.88
## 2     37 262.58 -2   -18.701 1.342 0.2744
```

F-statistics = 1.342
P-value = 0.2744
So we reject null hypothesis and conclude that neither of the terms involving Type is needed and and a common regression line for both levels of Type is adequate for modeling how MMSE depends on APC.

Q4.2 (5 points)

a. (3 points)

```
data("HighPeaks")
plot(HighPeaks$Elevation,HighPeaks$Time)
```




```
cor(HighPeaks$Elevation,HighPeaks$Time)
```

```
## [1] -0.0162768
```

The correlation between Elevation and Time is -0.0162768. The scatterplot and the correlation show that Elevation is not helpful in predicting Time.

b. (2 points)

```
f1 <- lm(Time~Elevation, data=HighPeaks)
f2 <- lm(Time~Length, data=HighPeaks)
f3 <- lm(Time~Elevation+Length, data=HighPeaks)
summary(f1)
```

```
##
## Call:
## lm(formula = Time ~ Elevation, data = HighPeaks)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6912 -1.6985 -0.5639  1.2963  7.3015
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.2113764  5.1953800   2.158   0.0364 *
## Elevation   -0.0001269  0.0011756  -0.108   0.9145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.826 on 44 degrees of freedom
## Multiple R-squared:  0.0002649,  Adjusted R-squared:  -0.02246
## F-statistic: 0.01166 on 1 and 44 DF,  p-value: 0.9145
```

```
summary(f2)
```

```
##
## Call:
## lm(formula = Time ~ Length, data = HighPeaks)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4491 -0.6687 -0.0122  0.5590  4.0034
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.04817    0.80371   2.548   0.0144 *
## Length       0.68427    0.06162  11.105 2.39e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.449 on 44 degrees of freedom
## Multiple R-squared:  0.737,  Adjusted R-squared:  0.7311
## F-statistic: 123.3 on 1 and 44 DF,  p-value: 2.39e-14
```

```
summary(f3)
```

```
##
## Call:
## lm(formula = Time ~ Elevation + Length, data = HighPeaks)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5924 -0.8050 -0.1959  0.6380  3.8432
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.0753787  2.5327132   3.188  0.00267 **
## Elevation   -0.0014483  0.0005805  -2.495  0.01653 *
## Length       0.7123344  0.0593330  12.006 2.54e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.37 on 43 degrees of freedom
## Multiple R-squared:  0.7703, Adjusted R-squared:  0.7596
## F-statistic: 72.09 on 2 and 43 DF,  p-value: 1.844e-14
```

Elevation is important in this multiple regression model. This two-predictor model is substantially better at explaining Time than either Elevation or Length alone.

Q4.4 (8 points)

a. (3 points)

```
data("Fertility")
round(cor(Fertility[,-2]), digit=3)
```

```
##               Age MeanAFC    FSH     E2  MaxE2 MaxDailyGn TotalGn Oocytes
```

```
## Age         1.000  -0.230  0.274 -0.023 -0.102      0.569   0.521  -0.113
## MeanAFC    -0.230   1.000 -0.296 -0.127  0.246     -0.397  -0.384   0.417
## FSH         0.274  -0.296  1.000 -0.071 -0.224      0.443   0.473  -0.285
## E2         -0.023  -0.127 -0.071  1.000 -0.030     -0.024  -0.007  -0.117
## MaxE2      -0.102   0.246 -0.224 -0.030  1.000     -0.291  -0.274   0.504
## MaxDailyGn  0.569  -0.397  0.443 -0.024 -0.291      1.000   0.908  -0.278
## TotalGn     0.521  -0.384  0.473 -0.007 -0.274      0.908   1.000  -0.265
## Oocytes    -0.113   0.417 -0.285 -0.117  0.504     -0.278  -0.265   1.000
## Embryos    -0.128   0.346 -0.223 -0.087  0.434     -0.218  -0.208   0.758
##            Embryos
## Age         -0.128
## MeanAFC      0.346
## FSH         -0.223
## E2          -0.087
## MaxE2        0.434
## MaxDailyGn  -0.218
## TotalGn     -0.208
## Oocytes      0.758
## Embryos      1.000
```

Oocytes has the strongest correlation with MeanAFC and E2 has the weakest correlation with MeanAFC.

b. (1 point)

```
summary(lm(MeanAFC~E2,data=Fertility[,-2]))
```

```
##
## Call:
## lm(formula = MeanAFC ~ E2, data = Fertility[, -2])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.917  -4.917  -1.290   3.213  38.204
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.09163    1.16918  13.763   <2e-16 ***
## E2          -0.06212    0.02660  -2.336   0.0201 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.378 on 331 degrees of freedom
## Multiple R-squared:  0.01621,    Adjusted R-squared:  0.01324
## F-statistic: 5.455 on 1 and 331 DF,  p-value: 0.02011
```

Even though E2 has the weakest correlation with MeanAFC, it is still effective for predicting MeanAFC.

c. (2 points)

```
library(leaps)
all <- regsubsets(MeanAFC~.,data=Fertility[,-2])
summary(all)
```

```
## Subset selection object
## Call: regsubsets.formula(MeanAFC ~ ., data = Fertility[, -2])
## 8 Variables  (and intercept)
##            Forced in Forced out
## Age            FALSE      FALSE
## FSH            FALSE      FALSE
## E2             FALSE      FALSE
## MaxE2          FALSE      FALSE
## MaxDailyGn     FALSE      FALSE
## TotalGn        FALSE      FALSE
## Oocytes        FALSE      FALSE
## Embryos        FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          Age FSH E2  MaxE2 MaxDailyGn TotalGn Oocytes Embryos
## 1  ( 1 ) " " " " " " " " "   " "        " "     "*"     " "
## 2  ( 1 ) " " " " " " " " "   "*"        " "     "*"     " "
## 3  ( 1 ) " " " " " "*" " "   "*"        " "     "*"     " "
## 4  ( 1 ) " " "*" "*" " "   "*"        " "     "*"     " "
## 5  ( 1 ) " " "*" "*" " "   "*"        " "     "*"     "*"
## 6  ( 1 ) " " "*" "*" " "   "*"        "*"     "*"     "*"
## 7  ( 1 ) "*" "*" "*" " "   "*"        "*"     "*"     "*"
## 8  ( 1 ) "*" "*" "*" "*"   "*"        "*"     "*"     "*"
```

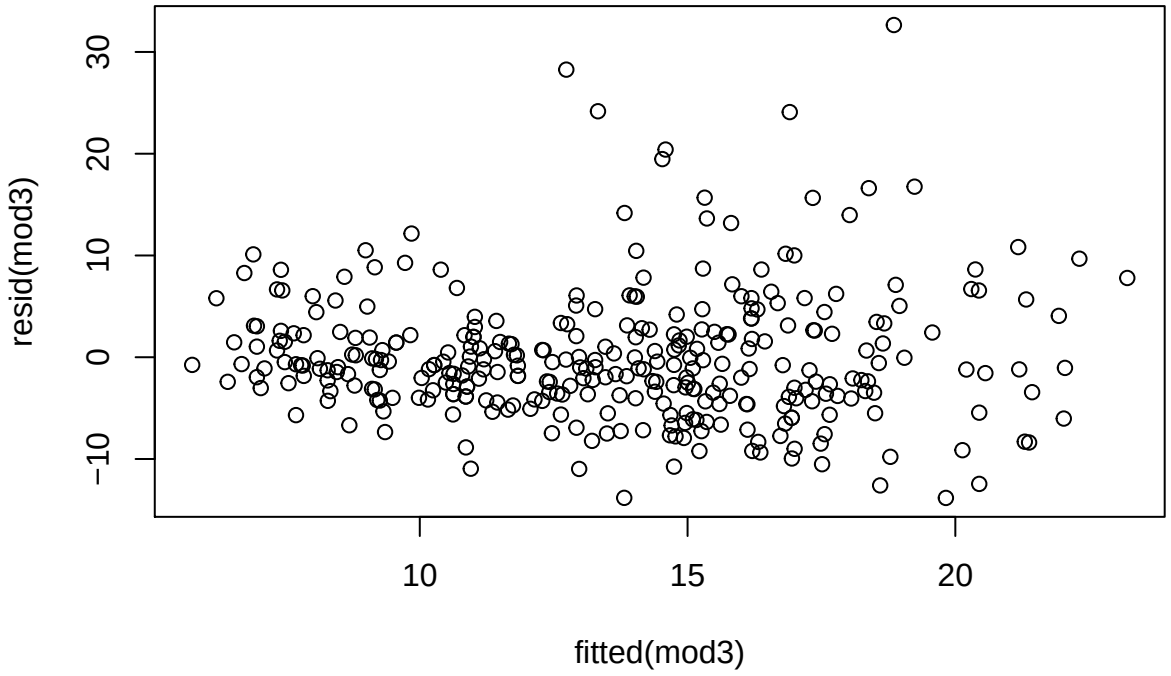
```
mod3 <- lm(MeanAFC~E2+MaxDailyGn+Oocytes,data=Fertility[,-2])
summary(mod3)
```

```
##
## Call:
## lm(formula = MeanAFC ~ E2 + MaxDailyGn + Oocytes, data = Fertility[,
##     -2])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.824  -3.791  -0.905   2.593  32.647
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.915874   1.799695   9.399  < 2e-16 ***
## E2          -0.047433   0.023199  -2.045   0.0417 *
## MaxDailyGn  -0.019902   0.003154  -6.311 8.96e-10 ***
## Oocytes      0.401693   0.062163   6.462 3.72e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.38 on 329 degrees of freedom
## Multiple R-squared:  0.2688, Adjusted R-squared:  0.2622
## F-statistic: 40.32 on 3 and 329 DF,  p-value: < 2.2e-16
```

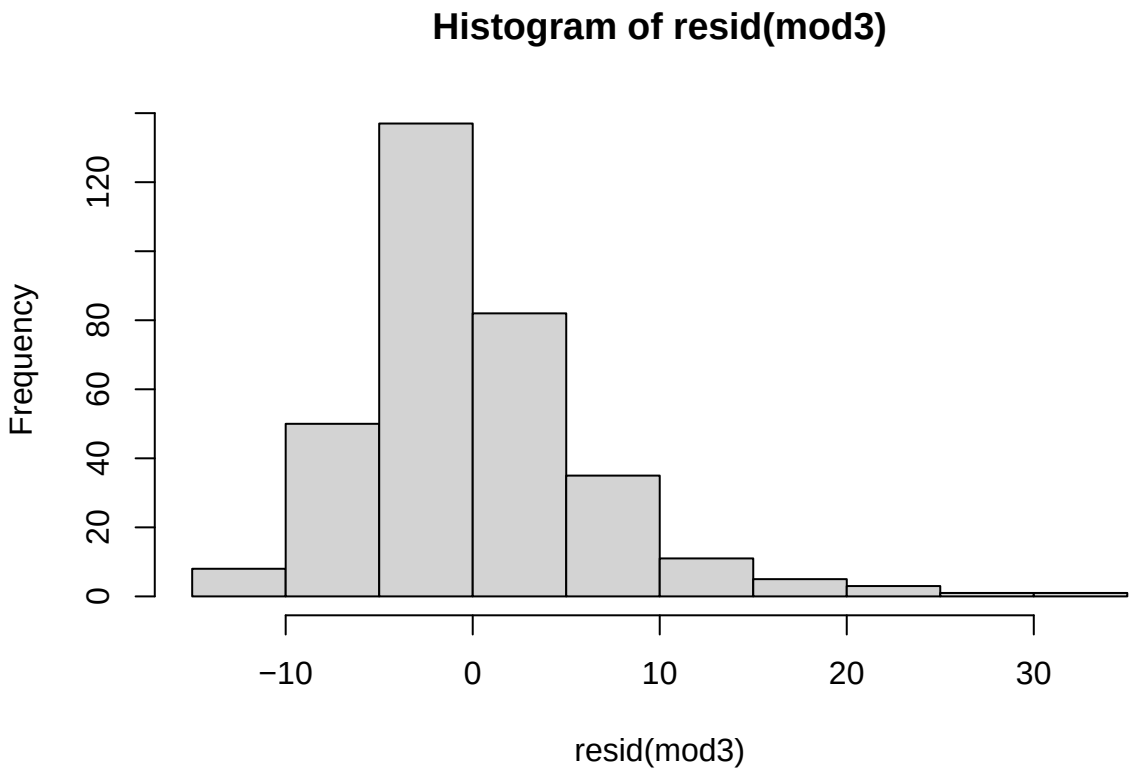
Model: $\hat{MeanAFC} = 16.916 - 0.047E2 - 0.02MaxDailyGn + 0.402Oocytes$
$R^2 = 0.2688$

d. (2 points)

```
plot(fitted(mod3),resid(mod3))
```




```
hist(resid(mod3))
```



The residual plots show the deviation from common variance and normality assumptions. The fit of the three-variable model identified in part (c) is not an appropriate model to predict MeanAFC.

Q4.8 (7 points)

a. (3 points)

```
data("CountyHealth")
train <- CountyHealth[1:35,]
test <- CountyHealth[36:53,]
train_m <- lm(sqrt(MDs) ~ Hospitals, data=train)
summary(train_m)
```

```
##
## Call:
## lm(formula = sqrt(MDs) ~ Hospitals, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.582  -6.362  -2.918   8.277  23.170
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.1695     2.6915  -1.178    0.247
## Hospitals     6.7853     0.5284  12.841 2.19e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.627 on 33 degrees of freedom
## Multiple R-squared:  0.8332, Adjusted R-squared:  0.8282
## F-statistic: 164.9 on 1 and 33 DF,  p-value: 2.194e-14
```

Model: $\hat{\sqrt{MDs}} = -3.1695 + 6.7853 Hospitals$

b. (2 points)

```
new_data <- data.frame(Hospitals=test$Hospitals)
pred = predict(train_m, newdata=new_data)
pred_MDs = pred^2
cor(pred, sqrt(test$MDs))
```

```
## [1] 0.9531439
```

c. (2 points)

```
Shrinkage = 0.8332 - 0.9531^2
Shrinkage
```

```
## [1] -0.07519961
```

The squared cross-validation correlation is close to the R2 value for the training sample, so we can conclude that the model to predict $\sqrt{MDs}$ based on Hospitals works as well for the holdout sample as it did for the training sample.

Q4.14 (13 points)

a. (9 points)

$H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$
t-statistics = -5.15, p-value = $0 < 0.05$
We reject H_0 and conclude that the mother's race as black has significantly effect on the birth weight of a baby.

$H_0 : \beta_2 = 0$ vs $H_a : \beta_2 \neq 0$
t-statistics = 0.34, p-value = 0.731
We reject H_0 and conclude that the mother's race as Hispanic does not have significantly effect on the birth weight of a baby.

$H_0 : \beta_3 = 0$ vs $H_a : \beta_3 \neq 0$
t-statistics = -0.22, p-value = 0.825
We reject H_0 and conclude that the mother's race as other does not have significantly effect on the birth weight of a baby.

b. (1 point)

$R^2 = 1.9\%$

c. (3 points)

F-statistics = 9.53, p-value = $0 < 0.05$
We reject H_0 and conclude that race of the mother significantly effect the birth weight of a baby.