Lab Assignment 1

Charlie Lock

2023-03-28

Setup

```
library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(car)
```

```
## Loading required package: carData
```

Question 1

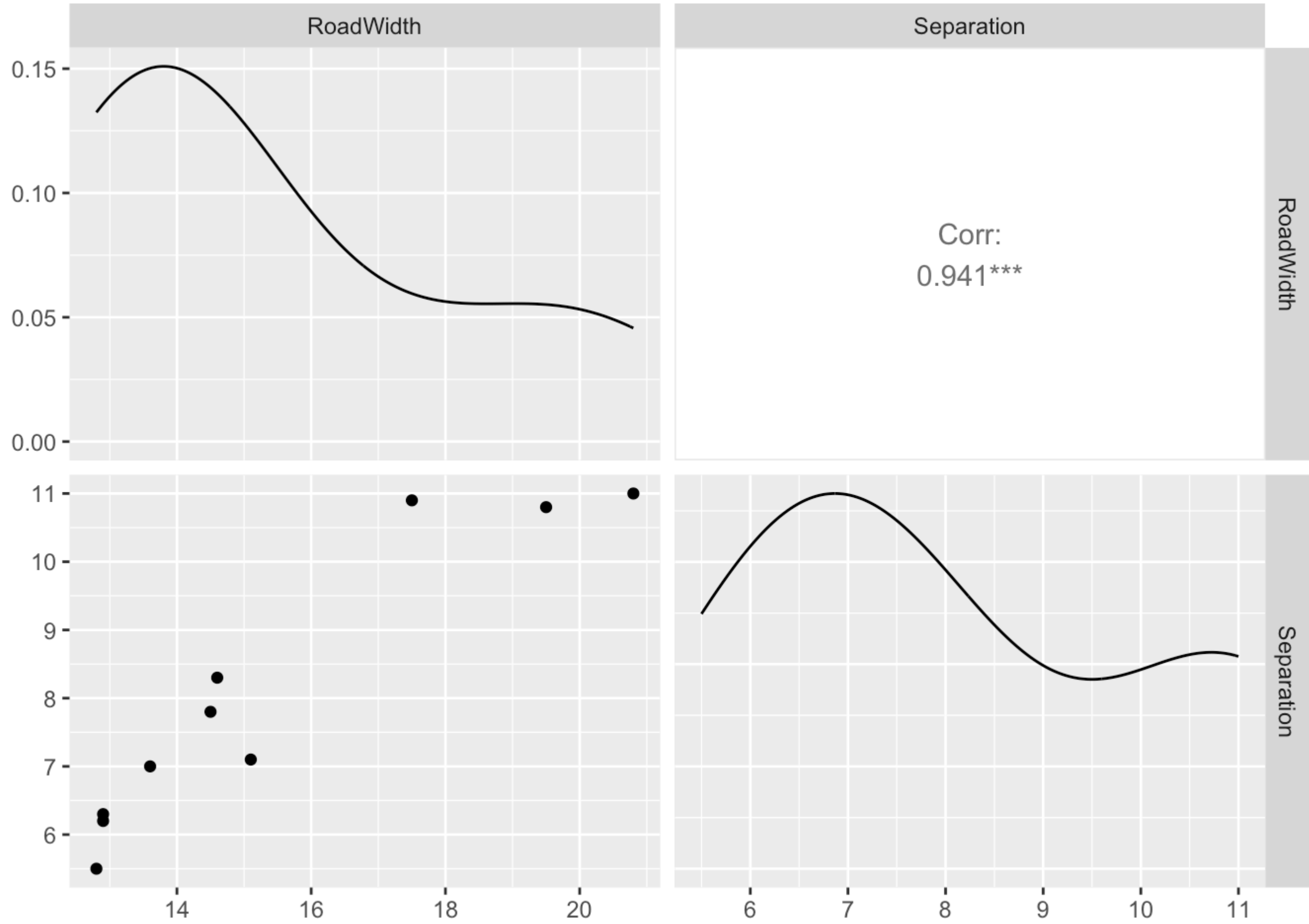
```
BikeLanes <- data.frame(RoadWidth = c(12.8, 12.9, 12.9, 13.6, 14.5, 14.6, 15.1, 17.5,
19.5, 20.8),
                        Separation = c(5.5, 6.2, 6.3, 7.0, 7.8, 8.3, 7.1, 10.9, 10.8,
11.0))
```

1.1.

```
plot(BikeLanes$RoadWidth, BikeLanes$Separation, main = "Road Width against separation
between bike and car", xlab = "Road Width", ylab = "Separation between Bike and passi
ng Car")
```


```
ggpairs(data = BikeLanes)
```

```
## Warning in geom_point(): All aesthetics have length 1, but the data has 4 rows.
## ℹ Please consider using `annotate()` or provide this layer with data containing
##   a single row.
```



Plot suggests a strong linear relationship between y and x and a large correlation coefficient of 0.941 supports that so it is definitely reasonable to use simple linear regression as a model when relating y to x.

1.2.

```
LOBestFit <- lm(BikeLanes$Separation~BikeLanes$RoadWidth)
LOBestFit
```

```
##
## Call:
## lm(formula = BikeLanes$Separation ~ BikeLanes$RoadWidth)
##
## Coefficients:
##         (Intercept)  BikeLanes$RoadWidth
##             -2.4804               0.6855
```

The output is -2.4804 for the constant and 0.6855 for the slope which means the least squared line of best fit is ‘y = 0.6855 * x - 2.4804’ where x is the road width and y is the separation between bike and car.

1.3.

```
anova(LOBestFit)
```

```
## Analysis of Variance Table
##
## Response: BikeLanes$Separation
##                     Df Sum Sq Mean Sq F value    Pr(>F)
## BikeLanes$RoadWidth  1 34.969  34.969  61.886 4.927e-05 ***
## Residuals            8  4.520   0.565
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is less than 0.05 which suggests that a relationship between road width and gap between car and bike does in fact exist.

1.4.

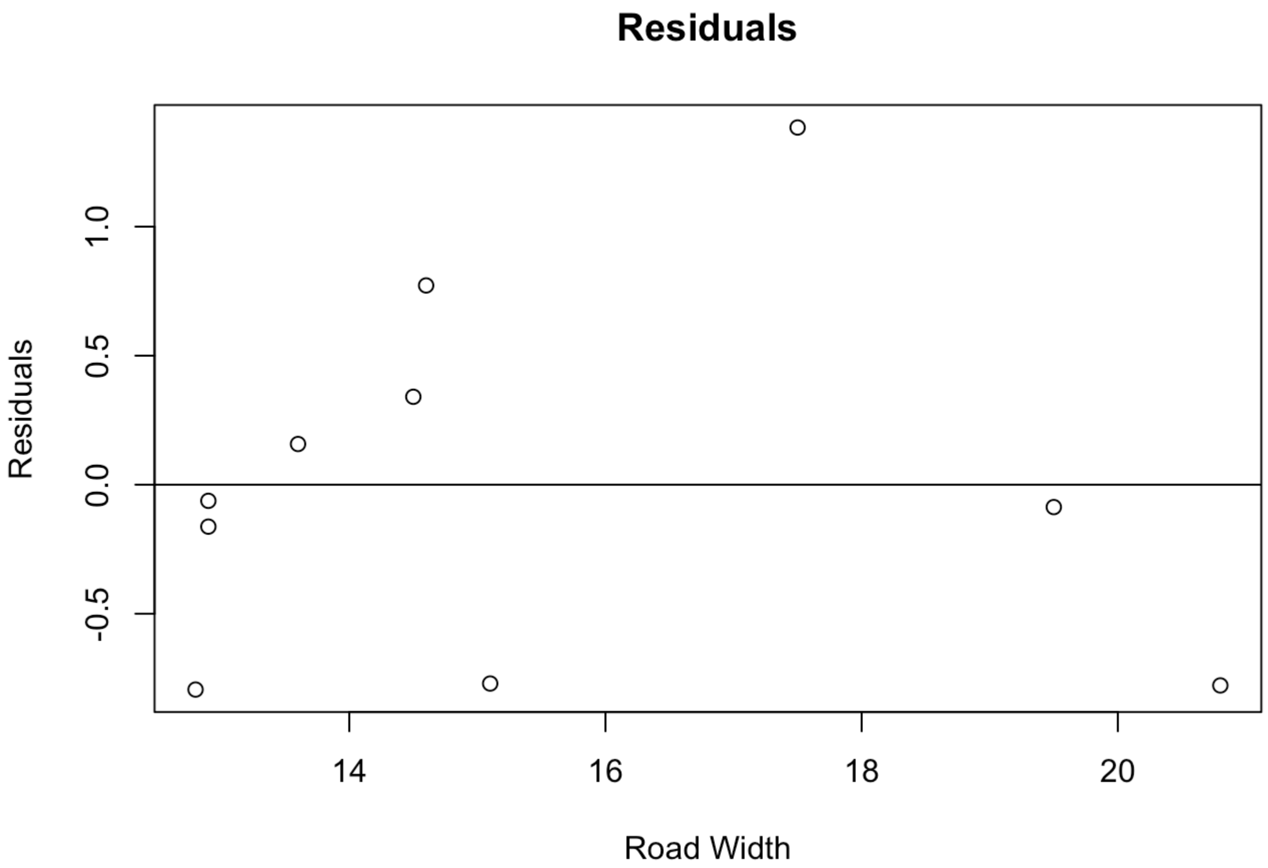
```
stats::confint(LOBestFit)
```

```
##                          2.5 %    97.5 %
## (Intercept)         -5.6270076 0.6662572
## BikeLanes$RoadWidth  0.4845561 0.8864393
```

95% confidence interval for β0 is (-5.6270, 0.6663) and the 95% confidence interval for β1 is (0.4846, 0.8864)

1.5.

```
BikeLanes_residuals = resid(LOBestFit)
plot(BikeLanes$RoadWidth, BikeLanes_residuals, ylab = "Residuals", xlab = "Road Widt
h", main = "Residuals")
abline(0,0)
```



Question 2

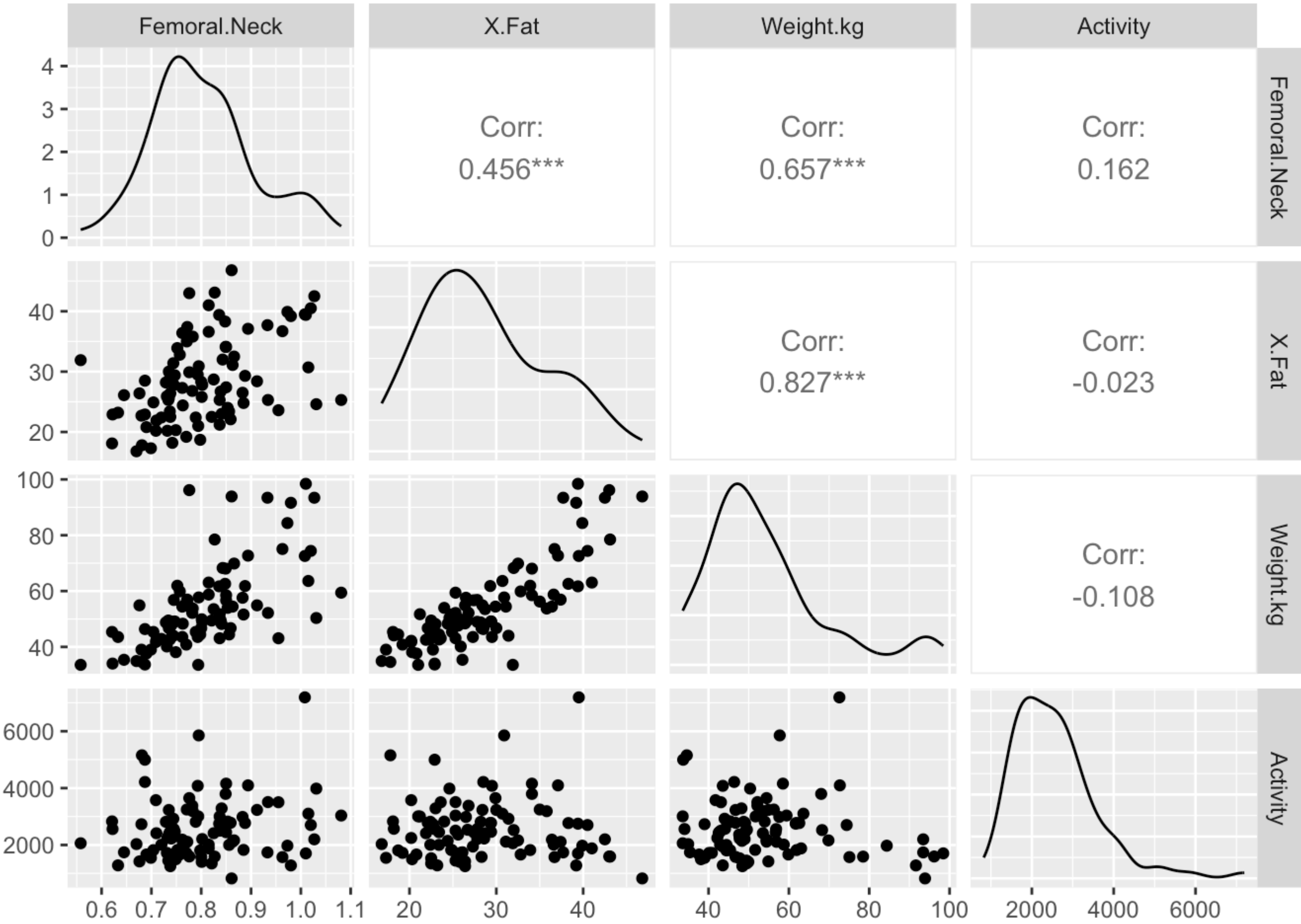
```
lab.data <- read.csv("/Users/charlielock/Documents/R/Datasets/DataLab.csv")
head(lab.data)
```

```
##   Femoral.Neck X.Fat Weight.kg Activity
## 1        0.934  25.3  52.16313  3508.44
## 2        0.888  29.3  61.80196  2773.54
## 3        0.933  37.7  93.44003  1738.97
## 4        0.757  32.8  59.87420  1665.29
## 5        1.031  24.6  50.34876  3982.95
## 6        0.883  26.5  57.60623  2985.74
```

2.1. Multicollinearity

```
ggpairs(data = lab.data, columns = c(1,2,3,4))
```

```
## Warning in geom_point(): All aesthetics have length 1, but the data has 16 rows.
## ℹ Please consider using `annotate()` or provide this layer with data containing
##   a single row.
```



Using the output from the ggpairs() function we can see that the variables Weight.kg (weight) and X.Fat (body fat) are very strongly correlated by 0.827. This indicates multicollinearity.

2.2 ANOVA

```
lm.femoral <- lm(Femoral.Neck ~ X.Fat + Weight.kg + Activity + Weight.kg * X.Fat, la
b.data)
anova(lm.femoral)
```

```
## Analysis of Variance Table
##
## Response: Femoral.Neck
##                  Df  Sum Sq  Mean Sq F value    Pr(>F)
## X.Fat             1 0.20514 0.205137 41.2591 6.835e-09 ***
## Weight.kg         1 0.24506 0.245059 49.2886 4.610e-10 ***
## Activity          1 0.06384 0.063843 12.8408 0.0005585 ***
## X.Fat:Weight.kg   1 0.04175 0.041745  8.3962 0.0047565 **
## Residuals        87 0.43256 0.004972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value is >0.05 for all 3 predictor variables as well as for the interaction between body fat and weight. This suggests that they all have a relationship with the bone density of the femoral neck.

2.3. Significant Variables

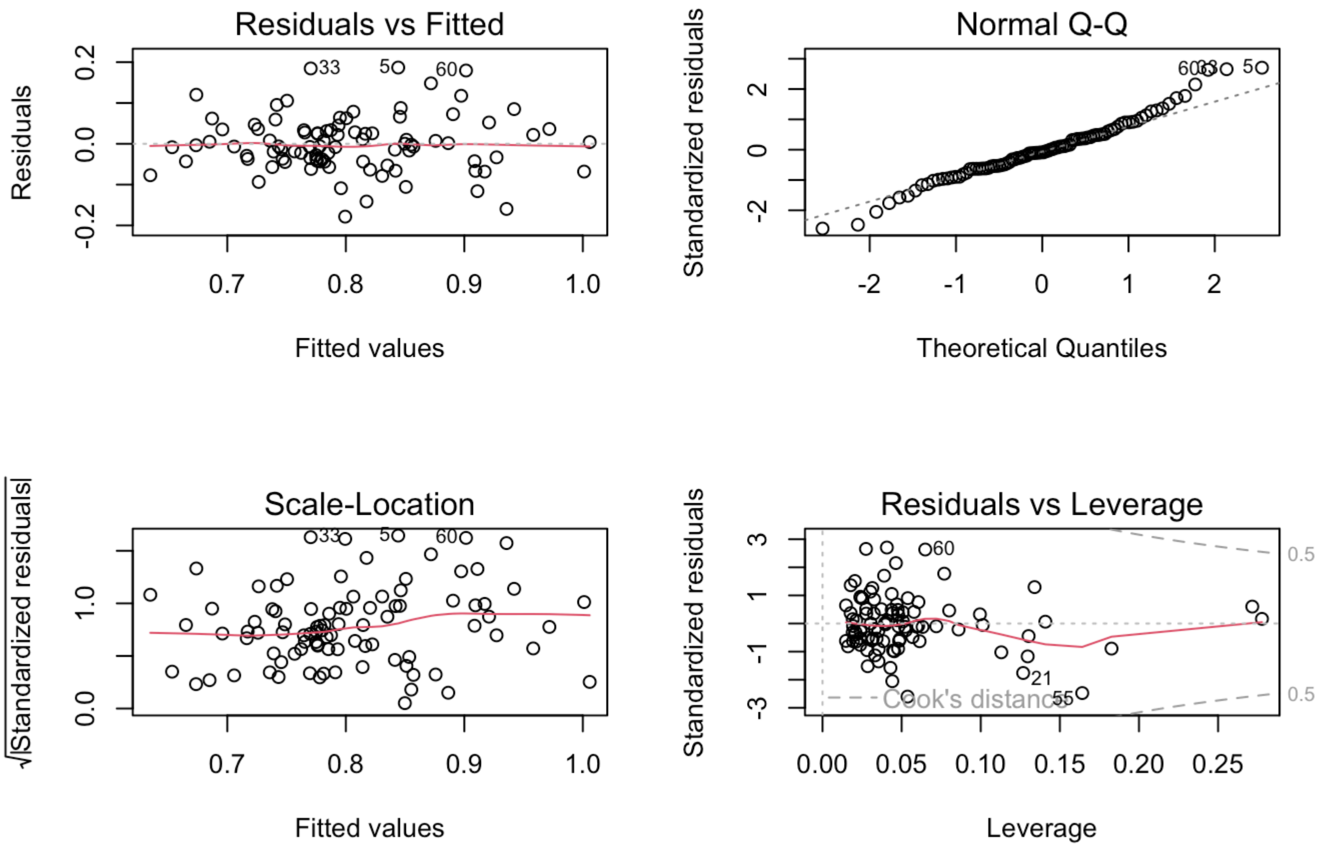
```
summary(lm.femoral)
```

```
##
## Call:
## lm(formula = Femoral.Neck ~ X.Fat + Weight.kg + Activity + Weight.kg *
##     X.Fat, data = lab.data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.178453 -0.042754 -0.006129  0.033937  0.186795
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.549e-01  1.317e-01   1.176  0.24274
## X.Fat            5.571e-03  4.087e-03   1.363  0.17632
## Weight.kg        1.447e-02  2.852e-03   5.073 2.19e-06 ***
## Activity         2.238e-05  7.276e-06   3.075  0.00281 **
## X.Fat:Weight.kg -2.142e-04  7.394e-05  -2.898  0.00476 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07051 on 87 degrees of freedom
## Multiple R-squared:  0.5623, Adjusted R-squared:  0.5422
## F-statistic: 27.95 on 4 and 87 DF,  p-value: 6.242e-15
```

Using the p-value again, since it is more than 0.05 for body fat it is implied that body fat is not a significant variable. Weight, physical activity and the interaction between body fat and weight are considered significant variables as they all output a p-value < 0.05.

2.4. Residual Plots and Tests

```
par(mfrow = c(2,2))
plot(lm.femoral)
```




```
ncvTest(lm.femoral)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 3.927862, Df = 1, p = 0.047492
```

H0: Errors have constant variance H1: Errors do not have constant variance

The p-value is just < 0.05 so H0 is rejected which implies that the assumption of constant error variance is violated.

```
shapiro.test(lm.femoral$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm.femoral$residuals
## W = 0.97758, p-value = 0.1138
```

H0: Errors are normally distributed H1: Errors are not normally distributed

The p-value is >0.05 so H0 cannot be rejected and the assumption of normality error is not violated.

```
durbinWatsonTest(lm.femoral)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.02873784      2.035368   0.916
##  Alternative hypothesis: rho != 0
```

H0: Errors are not correlated H1: Errors are correlated

2.5. Comparing results using different methods

```
full=lm(Femoral.Neck~., data = lab.data)
null=lm(Femoral.Neck~1, data = lab.data)
```

Stepwise Regression Method

```
step(null, scope = list(upper = full), direction = "both")
```

```
## Start:  AIC=-415.08
## Femoral.Neck ~ 1
## 
##              Df Sum of Sq     RSS     AIC
## + Weight.kg  1   0.42635 0.56199 -465.02
## + X.Fat      1   0.20514 0.78320 -434.49
## + Activity   1   0.02588 0.96246 -415.52
## <none>                   0.98834 -415.08
## 
## Step:  AIC=-465.02
## Femoral.Neck ~ Weight.kg
## 
##              Df Sum of Sq     RSS     AIC
## + Activity   1   0.05407 0.50793 -472.33
## + X.Fat      1   0.02385 0.53815 -467.01
## <none>                   0.56199 -465.02
## - Weight.kg  1   0.42635 0.98834 -415.08
## 
## Step:  AIC=-472.33
## Femoral.Neck ~ Weight.kg + Activity
## 
##              Df Sum of Sq     RSS     AIC
## + X.Fat      1   0.03362 0.47430 -476.63
## <none>                   0.50793 -472.33
## - Activity   1   0.05407 0.56199 -465.02
## - Weight.kg  1   0.45454 0.96246 -415.52
## 
## Step:  AIC=-476.63
## Femoral.Neck ~ Weight.kg + Activity + X.Fat
## 
##              Df Sum of Sq     RSS     AIC
## <none>                    0.47430 -476.63
## - X.Fat      1  0.033623 0.50793 -472.33
## - Activity   1  0.063843 0.53815 -467.01
## - Weight.kg  1  0.279621 0.75392 -435.99
```

```
##
## Call:
## lm(formula = Femoral.Neck ~ Weight.kg + Activity + X.Fat, data = lab.data)
##
## Coefficients:
## (Intercept)    Weight.kg     Activity        X.Fat
##   5.214e-01    6.608e-03    2.574e-05   -4.923e-03
```

The Stepwise Regression method suggests that the model with the best fit includes all 3 predictor variables (weight, body fat and physical activity) as this has the minimum AIC value of -476.63.

The model is E(y) = 0.5214 + (-0.0049 * X.Fat) + (0.0066 * Weight.kg) + (0.00002 * Activity)

Backwards elimination

```
step(full, data = lab.data, direction = "backward")
```

```
## Start:  AIC=-476.63
## Femoral.Neck ~ X.Fat + Weight.kg + Activity
##
##             Df Sum of Sq     RSS     AIC
## <none>                   0.47430 -476.63
## - X.Fat      1  0.033623 0.50793 -472.33
## - Activity   1  0.063843 0.53815 -467.01
## - Weight.kg  1  0.279621 0.75392 -435.99
```

```
##
## Call:
## lm(formula = Femoral.Neck ~ X.Fat + Weight.kg + Activity, data = lab.data)
##
## Coefficients:
## (Intercept)        X.Fat    Weight.kg     Activity
##   5.214e-01   -4.923e-03    6.608e-03    2.574e-05
```

The Backwards elimination method suggests that based off the minimum AIC value of -476.63 all 3 predictor variables should be used in the best fitted model.

The model is E(y) = 0.5214 + (-0.0049 * X.Fat) + (0.0066 * Weight.kg) + (0.00002 * Activity)

Forward selection

```
step(null, scope = list(lower = null, upper = full, direction = "forward"))
```

```
## Start:  AIC=-415.08
## Femoral.Neck ~ 1
##
##              Df Sum of Sq     RSS     AIC
## + Weight.kg  1   0.42635 0.56199 -465.02
## + X.Fat      1   0.20514 0.78320 -434.49
## + Activity   1   0.02588 0.96246 -415.52
## <none>                   0.98834 -415.08
##
## Step:  AIC=-465.02
## Femoral.Neck ~ Weight.kg
##
##              Df Sum of Sq     RSS     AIC
## + Activity   1   0.05407 0.50793 -472.33
## + X.Fat      1   0.02385 0.53815 -467.01
## <none>                   0.56199 -465.02
## - Weight.kg  1   0.42635 0.98834 -415.08
##
## Step:  AIC=-472.33
## Femoral.Neck ~ Weight.kg + Activity
##
##              Df Sum of Sq     RSS     AIC
## + X.Fat      1   0.03362 0.47430 -476.63
## <none>                   0.50793 -472.33
## - Activity   1   0.05407 0.56199 -465.02
## - Weight.kg  1   0.45454 0.96246 -415.52
##
## Step:  AIC=-476.63
## Femoral.Neck ~ Weight.kg + Activity + X.Fat
##
##              Df Sum of Sq     RSS     AIC
## <none>                    0.47430 -476.63
## - X.Fat      1  0.033623 0.50793 -472.33
## - Activity   1  0.063843 0.53815 -467.01
## - Weight.kg  1  0.279621 0.75392 -435.99
```

```
##
## Call:
## lm(formula = Femoral.Neck ~ Weight.kg + Activity + X.Fat, data = lab.data)
##
## Coefficients:
## (Intercept)    Weight.kg     Activity        X.Fat
##   5.214e-01    6.608e-03    2.574e-05   -4.923e-03
```

The results of the forward selection method suggests that all 3 predictor variables should be used in the best fitted model as the minimum AIC is -476.63 when all 3 variables are included.

The model is E(y) = 0.5214 + (-0.0049 * X.Fat) + (0.0066 * Weight.kg) + (0.00002 * Activity)