ST552 Homework 6

Brian Cervantes Alvarez
March 5, 2024

Problem 1

```
library(faraway)
library(tidyverse)
data(kanga)

missingByCase <- kanga %>%
  rowwise() %>%
  summarise(missingValues = sum(is.na(c_across(where(is.numeric)))))
missingCaseSummary <- missingByCase %>%
  count(missingValues) %>%
  as.data.frame()
missingByVariable <- kanga %>%
  summarise(across(where(is.numeric), ~ sum(is.na(.)))) %>%
  as.data.frame()
print(missingCaseSummary)
```

```
  missingValues   n
1             0 101
2             1  39
3             2   6
4             3   2
```

```
print(missingByVariable)
```

```
  basilar.length occipitonasal.length palate.length palate.width nasal.length
1              1                    2             1           24             1
  nasal.width squamosal.depth lacrymal.width zygomatic.width orbital.width
1           0               1              0               1             0
  .rostral.width occipital.depth crest.width foramina.length mandible.length
1              3              11           0                0              12
  mandible.width mandible.depth ramus.height
1              0              0            0
```

Problem 2

Part A

```r
library(faraway)
library(mice)

data(gala)
data(galamiss)

modelA <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
             data = gala)
summary(modelA)
```

```
Call:
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = gala)

Residuals:
     Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,    Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

Part B

Comparison of using the deletion method for handling missing values in the linear model reveals slight differences in model fit. Retaining missing values yields an adjusted R^2 of 0.7171, while employing deletion reduces it to 0.702. Although coefficients remain relatively consistent, the deletion method increases the model's standard error, indicating slightly decreased precision.

```
galaMissDeleted <- na.omit(galamiss)
modelB <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
             data = galaMissDeleted)
summary(modelB)
```

```
Call:
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = galaMissDeleted)

Residuals:
    Min      1Q  Median      3Q     Max
-115.17  -37.60  -10.08   35.17  172.54

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.32286   27.47417   0.558  0.58391
Area        -0.02765    0.02557  -1.081  0.29388
Elevation    0.32550    0.06476   5.026 8.78e-05 ***
Nearest     -0.11042    1.17784  -0.094  0.92635
Scruz       -0.28427    0.25422  -1.118  0.27818
Adjacent    -0.07880    0.02092  -3.766  0.00141 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 67.36 on 18 degrees of freedom
Multiple R-squared:  0.7668,    Adjusted R-squared:  0.702
F-statistic: 11.83 on 5 and 18 DF,  p-value: 3.54e-05
```

Part C

Looking at both models, the mean imputation model exhibits a lower adjusted R^2 of 0.5774 compared to 0.7171 in Part A, indicating reduced explanatory power. Additionally, while some coefficients remain consistent, others vary notably, suggesting potential bias introduced by mean imputation. Moreover, the mean imputation model shows a higher residual standard error of 74.52, implying increased uncertainty in predictions compared to the original model.

```
galaMissMean <- galamiss
galaMissMean[] <- lapply(galaMissMean,
                         function(x) ifelse(is.na(x),
                                            mean(x, na.rm = TRUE), x))
modelC <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
             data = galaMissMean)
summary(modelC)
```

```
Call:
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = galaMissMean)

Residuals:
    Min      1Q  Median      3Q     Max
-94.710 -42.598  -9.742  26.146 220.893

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.48266   28.62644  -0.436 0.666695
Area         -0.00137    0.02683  -0.051 0.959697
Elevation     0.27388    0.06891   3.975 0.000562 ***
Nearest       0.37776    1.28270   0.295 0.770905
Scruz        -0.08544    0.27140  -0.315 0.755629
Adjacent     -0.06553    0.02215  -2.958 0.006856 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 74.52 on 24 degrees of freedom
Multiple R-squared:  0.6503,    Adjusted R-squared:  0.5774
F-statistic: 8.925 on 5 and 24 DF,  p-value: 6.77e-05
```

Part D

This model reveals minor differences in model fit. The regression imputation model shows a slightly higher adjusted R^2 of 0.7315 compared to 0.7171 in Part A, suggesting slight improvement in explanatory power. While coefficients mostly align, subtle variations suggest some influence from the imputation method, yet the residual standard error remains comparable at 59.4, indicating similar prediction uncertainty. Nevertheless, the regression-based imputation may introduce bias into the dataset if the relationship between predictors and missing values is complex or nonlinear. It worked here, but it is imperative to understand this potential bias that may be introduced.

```
galaMissRegImpute <- mice(galamiss, method = "norm.predict", m = 1)
```

```
 iter imp variable
  1   1  Elevation
  2   1  Elevation
  3   1  Elevation
  4   1  Elevation
  5   1  Elevation
```

```
galaMissComplete <- complete(galaMissRegImpute, action = "long",
                             include = FALSE)
modelD <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
             data = galaMissComplete)
summary(modelD)
```

```
Call:
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = galaMissComplete)

Residuals:
     Min       1Q   Median       3Q      Max
-119.106  -27.164   -8.814   18.211  176.298

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.05530   18.92647   0.161 0.873106
Area        -0.02918    0.02228  -1.309 0.202771
Elevation    0.33695    0.05420   6.216 2.01e-06 ***
Nearest     -0.10688    1.03035  -0.104 0.918248
Scruz       -0.22151    0.21011  -1.054 0.302276
Adjacent    -0.08047    0.01772  -4.542 0.000133 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 59.4 on 24 degrees of freedom
Multiple R-squared:  0.7778,    Adjusted R-squared:  0.7315
F-statistic:  16.8 on 5 and 24 DF,  p-value: 3.721e-07
```

Part E

By using multiple imputation, this model reveals similar adjusted R^2 values, with Model E showing an adjusted R^2 of 0.7665 compared to 0.7171 in Part A. While most coefficients align closely between the two models, minor variations exist, potentially attributed to the imputation process. Overall, the multiple imputation approach provides comparable model fit to the original model, suggesting its effectiveness in handling missing values without substantial loss of explanatory power. However, caution should be exercised regarding the assumption of missing data mechanism and the potential impact of imputation on parameter estimates, particularly if the missingness is not completely at random or if the imputation model fails to adequately capture the true data generation process.

```
galaMissMultiImpute <- mice(galamiss, m = 5)
```

```
 iter imp variable
  1   1  Elevation
  1   2  Elevation
  1   3  Elevation
  1   4  Elevation
  1   5  Elevation
  2   1  Elevation
  2   2  Elevation
  2   3  Elevation
  2   4  Elevation
  2   5  Elevation
  3   1  Elevation
  3   2  Elevation
  3   3  Elevation
  3   4  Elevation
  3   5  Elevation
  4   1  Elevation
  4   2  Elevation
  4   3  Elevation
  4   4  Elevation
  4   5  Elevation
  5   1  Elevation
  5   2  Elevation
  5   3  Elevation
  5   4  Elevation
  5   5  Elevation
```

```
modelE <- with(galaMissMultiImpute,
               lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent))

summary(pool(modelE))
```

```
         term    estimate   std.error   statistic       df      p.value
1 (Intercept)  6.57202632 20.78211586  0.31623471 17.49480 7.555662e-01
2        Area -0.02562805  0.02272062 -1.12796439 22.11417 2.714246e-01
3   Elevation  0.32547850  0.05519453  5.89693413 21.86211 6.358086e-06
4     Nearest -0.06095599  1.05777774 -0.05762646 22.19954 9.545617e-01
5       Scruz -0.23337475  0.22093861 -1.05628772 20.94200 3.028705e-01
6    Adjacent -0.07800361  0.01832634 -4.25636508 21.53115 3.358043e-04
```

```
pool.r.squared(modelE)
```

```
          est     lo 95     hi 95        fmi
R^2 0.7659623 0.5620281 0.8834709 0.02846054
```

Problem 3

Part A

```r
data(prostate)
backModel <- step(lm(lpsa ~ ., data = prostate), direction = "backward")
```

```
Start:  AIC=-58.32
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45

           Df Sum of Sq    RSS     AIC
- gleason  1    0.0412 44.204 -60.231
- pgg45    1    0.5258 44.689 -59.174
- lcp      1    0.6740 44.837 -58.853
<none>                44.163 -58.322
- age      1    1.5503 45.713 -56.975
- lbph     1    1.6835 45.847 -56.693
- lweight  1    3.5861 47.749 -52.749
- svi      1    4.9355 49.099 -50.046
- lcavol   1   22.3721 66.535 -20.567

Step:  AIC=-60.23
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45

           Df Sum of Sq    RSS     AIC
- lcp      1    0.6623 44.867 -60.789
<none>                44.204 -60.231
- pgg45    1    1.1920 45.396 -59.650
- age      1    1.5166 45.721 -58.959
- lbph     1    1.7053 45.910 -58.560
- lweight  1    3.5462 47.750 -54.746
- svi      1    4.8984 49.103 -52.037
- lcavol   1   23.5039 67.708 -20.872

Step:  AIC=-60.79
lpsa ~ lcavol + lweight + age + lbph + svi + pgg45

           Df Sum of Sq    RSS     AIC
- pgg45    1    0.6590 45.526 -61.374
<none>                44.867 -60.789
- age      1    1.2649 46.131 -60.092
- lbph     1    1.6465 46.513 -59.293
```

```
- lweight  1    3.5647 48.431 -55.373
- svi      1    4.2503 49.117 -54.009
- lcavol   1   25.4189 70.285 -19.248

Step:  AIC=-61.37
lpsa ~ lcavol + lweight + age + lbph + svi

          Df Sum of Sq    RSS     AIC
<none>                 45.526 -61.374
- age      1    0.9592 46.485 -61.352
- lbph     1    1.8568 47.382 -59.497
- lweight  1    3.2251 48.751 -56.735
- svi      1    5.9517 51.477 -51.456
- lcavol   1   28.7665 74.292 -15.871
```

```
summary(backModel)
```

```
Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.83505 -0.39396  0.00414  0.46336  1.57888

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.95100    0.83175   1.143 0.255882
lcavol       0.56561    0.07459   7.583 2.77e-11 ***
lweight      0.42369    0.16687   2.539 0.012814 *
age         -0.01489    0.01075  -1.385 0.169528
lbph         0.11184    0.05805   1.927 0.057160 .
svi          0.72095    0.20902   3.449 0.000854 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7073 on 91 degrees of freedom
Multiple R-squared:  0.6441,    Adjusted R-squared:  0.6245
F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

Part B

```r
aicMod <- step(lm(lpsa ~ ., data = prostate), direction = "both", k = 2)
```

```
Start:  AIC=-58.32
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45

           Df Sum of Sq    RSS     AIC
- gleason   1    0.0412 44.204 -60.231
- pgg45     1    0.5258 44.689 -59.174
- lcp       1    0.6740 44.837 -58.853
<none>                  44.163 -58.322
- age       1    1.5503 45.713 -56.975
- lbph      1    1.6835 45.847 -56.693
- lweight   1    3.5861 47.749 -52.749
- svi       1    4.9355 49.099 -50.046
- lcavol    1   22.3721 66.535 -20.567

Step:  AIC=-60.23
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45

           Df Sum of Sq    RSS     AIC
- lcp       1    0.6623 44.867 -60.789
<none>                  44.204 -60.231
- pgg45     1    1.1920 45.396 -59.650
- age       1    1.5166 45.721 -58.959
- lbph      1    1.7053 45.910 -58.560
+ gleason   1    0.0412 44.163 -58.322
- lweight   1    3.5462 47.750 -54.746
- svi       1    4.8984 49.103 -52.037
- lcavol    1   23.5039 67.708 -20.872

Step:  AIC=-60.79
lpsa ~ lcavol + lweight + age + lbph + svi + pgg45

           Df Sum of Sq    RSS     AIC
- pgg45     1    0.6590 45.526 -61.374
<none>                  44.867 -60.789
+ lcp       1    0.6623 44.204 -60.231
- age       1    1.2649 46.131 -60.092
- lbph      1    1.6465 46.513 -59.293
+ gleason   1    0.0296 44.837 -58.853
- lweight   1    3.5647 48.431 -55.373
```

```
- svi      1    4.2503 49.117 -54.009
- lcavol   1   25.4189 70.285 -19.248

Step:  AIC=-61.37
lpsa ~ lcavol + lweight + age + lbph + svi

          Df Sum of Sq    RSS     AIC
<none>                 45.526 -61.374
- age      1    0.9592 46.485 -61.352
+ pgg45    1    0.6590 44.867 -60.789
+ gleason  1    0.4560 45.070 -60.351
+ lcp      1    0.1293 45.396 -59.650
- lbph     1    1.8568 47.382 -59.497
- lweight  1    3.2251 48.751 -56.735
- svi      1    5.9517 51.477 -51.456
- lcavol   1   28.7665 74.292 -15.871
```

```
summary(aicMod)
```

```
Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.83505 -0.39396  0.00414  0.46336  1.57888

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.95100    0.83175   1.143 0.255882
lcavol       0.56561    0.07459   7.583 2.77e-11 ***
lweight      0.42369    0.16687   2.539 0.012814 *
age         -0.01489    0.01075  -1.385 0.169528
lbph         0.11184    0.05805   1.927 0.057160 .
svi          0.72095    0.20902   3.449 0.000854 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7073 on 91 degrees of freedom
Multiple R-squared:  0.6441,    Adjusted R-squared:  0.6245
F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

Part C

```r
# (c) Adjusted R^2
adjR2Mod <- step(lm(lpsa ~ ., data = prostate), direction = "both", trace = 0)
summary(adjR2Mod)
```

```
Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.83505 -0.39396  0.00414  0.46336  1.57888

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.95100    0.83175   1.143 0.255882
lcavol       0.56561    0.07459   7.583 2.77e-11 ***
lweight      0.42369    0.16687   2.539 0.012814 *
age         -0.01489    0.01075  -1.385 0.169528
lbph         0.11184    0.05805   1.927 0.057160 .
svi          0.72095    0.20902   3.449 0.000854 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7073 on 91 degrees of freedom
Multiple R-squared:  0.6441,    Adjusted R-squared:  0.6245
F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

Part D

```
library(leaps)
mallowsMod <- regsubsets(lpsa ~ ., data = prostate, nvmax = ncol(prostate))
summaryMallows <- summary(mallowsMod)
bestCp <- which.min(summaryMallows$cp)
summary(mallowsMod)$which[bestCp, ]
```

```
(Intercept)      lcavol     lweight         age        lbph         svi
       TRUE        TRUE        TRUE       FALSE        TRUE        TRUE
        lcp     gleason       pgg45
      FALSE       FALSE       FALSE
```

Part E

```r
library(leaps)
forwardModel <- regsubsets(lpsa ~ ., data = prostate, method = "forward")
bestModel <- which.max(summary(forwardModel)$adjr2)
selectedPreds <- names(coef(forwardModel, id = bestModel))
summary(forwardModel)
```

```
Subset selection object
Call: regsubsets.formula(lpsa ~ ., data = prostate, method = "forward")
8 Variables  (and intercept)
        Forced in Forced out
lcavol      FALSE      FALSE
lweight     FALSE      FALSE
age         FALSE      FALSE
lbph        FALSE      FALSE
svi         FALSE      FALSE
lcp         FALSE      FALSE
gleason     FALSE      FALSE
pgg45       FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: forward
         lcavol lweight age lbph svi lcp gleason pgg45
1  ( 1 ) "*"    " "     " " " "  " " " " " " "     " "
2  ( 1 ) "*"    "*"     " " " "  " " " " " " "     " "
3  ( 1 ) "*"    "*"     " " " "  "*" " " " " "     " "
4  ( 1 ) "*"    "*"     " " "*"  "*" " " " " "     " "
5  ( 1 ) "*"    "*"     "*" "*"  "*" " " " " "     " "
6  ( 1 ) "*"    "*"     "*" "*"  "*" " " " " "     "*"
7  ( 1 ) "*"    "*"     "*" "*"  "*" "*" " " "     "*"
8  ( 1 ) "*"    "*"     "*" "*"  "*" "*" "*"     "*"
```

Problem 4

Part A

The model indicates that leg length does not have a statistically significant effect on hipcenter in the regression model. The coefficient for leg length is not significant (p = 0.1824), suggesting that after accounting for other variables, leg length may not strongly influence hipcenter. However, we it may still have some influence through model selection.

```
library(MASS)
data(seatpos)

modelAll <- lm(hipcenter ~ Age + Weight + HtShoes + Ht +
                 Seated + Arm + Thigh + Leg, data = seatpos)
summary(modelAll)
```

```
Call:
lm(formula = hipcenter ~ Age + Weight + HtShoes + Ht + Seated +
    Arm + Thigh + Leg, data = seatpos)

Residuals:
    Min      1Q  Median      3Q     Max
-73.827 -22.833  -3.678  25.017  62.337

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 436.43213  166.57162   2.620   0.0138 *
Age           0.77572    0.57033   1.360   0.1843
Weight        0.02631    0.33097   0.080   0.9372
HtShoes      -2.69241    9.75304  -0.276   0.7845
Ht            0.60134   10.12987   0.059   0.9531
Seated        0.53375    3.76189   0.142   0.8882
Arm          -1.32807    3.90020  -0.341   0.7359
Thigh        -1.14312    2.66002  -0.430   0.6706
Leg          -6.43905    4.71386  -1.366   0.1824
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.72 on 29 degrees of freedom
Multiple R-squared:  0.6866,    Adjusted R-squared:  0.6001
F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

Part B

```
predictorsMean <- colMeans(seatpos[, c("Age", "Weight", "HtShoes",
                                       "Ht", "Seated", "Arm",
                                       "Thigh", "Leg")], na.rm = TRUE)
predMean <- as.data.frame(t(predictorsMean))
predictionInterval <- predict(modelAll, newdata = predMean,
                              interval = "prediction", level = 0.95)
print(predictionInterval)
```

```
       fit     lwr       upr
1 -164.8849 -243.04 -86.72972
```

Part C

The model selection process using AIC, identified a final model for predicting hipcenter, by using the predictors Age, HtShoes, and Leg as predictors. With the lowest AIC value of 274.26, this selection indicates optimal balance between model simplicity and predictive accuracy. Additionally, the final selected model exhibits a prediction interval of $(-237.21, -92.56)$, indicating improved precision compared to the original model's interval of $(-243.04, -86.73)$ with point estimates of -164.88.. Hence, the chosen model provides the most effective framework for predicting hipcenter based on the given variables.

```
modelAIC <- step(modelAll, direction = "both")
```

```
Start:  AIC=283.62
hipcenter ~ Age + Weight + HtShoes + Ht + Seated + Arm + Thigh +
    Leg

           Df Sum of Sq   RSS    AIC
- Ht        1      5.01 41267 281.63
- Weight    1      8.99 41271 281.63
- Seated    1     28.64 41290 281.65
- HtShoes   1    108.43 41370 281.72
- Arm       1    164.97 41427 281.78
- Thigh     1    262.76 41525 281.87
<none>                  41262 283.62
- Age       1   2632.12 43894 283.97
- Leg       1   2654.85 43917 283.99

Step:  AIC=281.63
hipcenter ~ Age + Weight + HtShoes + Seated + Arm + Thigh + Leg

           Df Sum of Sq   RSS    AIC
- Weight    1     11.10 41278 279.64
- Seated    1     30.52 41297 279.66
- Arm       1    160.50 41427 279.78
- Thigh     1    269.08 41536 279.88
- HtShoes   1    971.84 42239 280.51
<none>                  41267 281.63
- Leg       1   2664.65 43931 282.01
- Age       1   2808.52 44075 282.13
+ Ht        1      5.01 41262 283.62

Step:  AIC=279.64
hipcenter ~ Age + HtShoes + Seated + Arm + Thigh + Leg

           Df Sum of Sq   RSS    AIC
```

```
- Seated   1     35.10 41313 277.67
- Arm      1    156.47 41434 277.78
- Thigh    1    285.16 41563 277.90
- HtShoes  1    975.48 42253 278.53
<none>                 41278 279.64
- Leg      1   2661.39 43939 280.01
- Age      1   3011.86 44290 280.31
+ Weight   1     11.10 41267 281.63
+ Ht       1      7.12 41271 281.63

Step:  AIC=277.67
hipcenter ~ Age + HtShoes + Arm + Thigh + Leg

          Df Sum of Sq   RSS    AIC
- Arm      1    172.02 41485 275.83
- Thigh    1    344.61 41658 275.99
- HtShoes  1   1853.43 43166 277.34
<none>                 41313 277.67
- Leg      1   2871.07 44184 278.22
- Age      1   2976.77 44290 278.31
+ Seated   1     35.10 41278 279.64
+ Weight   1     15.68 41297 279.66
+ Ht       1     10.02 41303 279.66

Step:  AIC=275.83
hipcenter ~ Age + HtShoes + Thigh + Leg

          Df Sum of Sq   RSS    AIC
- Thigh    1     472.8 41958 274.26
<none>                 41485 275.83
- HtShoes  1    2340.7 43826 275.92
- Age      1    3501.0 44986 276.91
- Leg      1    3591.7 45077 276.98
+ Arm      1     172.0 41313 277.67
+ Seated   1      50.6 41434 277.78
+ Weight   1      11.5 41474 277.82
+ Ht       1       2.6 41482 277.83

Step:  AIC=274.26
hipcenter ~ Age + HtShoes + Leg

          Df Sum of Sq   RSS    AIC
<none>                 41958 274.26
- Age      1    3108.8 45067 274.98
- Leg      1    3476.3 45434 275.28
```

```
+ Thigh    1     472.8 41485 275.83
- HtShoes  1    4218.6 46176 275.90
+ Arm      1     300.2 41658 275.99
+ Seated   1     144.0 41814 276.13
+ Weight   1      38.7 41919 276.23
+ Ht       1      33.1 41925 276.23
```

```r
summary(modelAIC)
```

```
Call:
lm(formula = hipcenter ~ Age + HtShoes + Leg, data = seatpos)

Residuals:
    Min      1Q  Median      3Q     Max
-79.269 -22.770  -4.342  21.853  60.907

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 456.2137   102.8078   4.438 9.09e-05 ***
Age           0.5998     0.3779   1.587   0.1217
HtShoes      -2.3023     1.2452  -1.849   0.0732 .
Leg          -6.8297     4.0693  -1.678   0.1024
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.13 on 34 degrees of freedom
Multiple R-squared:  0.6813,    Adjusted R-squared:  0.6531
F-statistic: 24.22 on 3 and 34 DF,  p-value: 1.437e-08
```

```r
predictorsMeanAIC <- colMeans(seatpos[, c("Age", "HtShoes", "Leg")],
                              na.rm = TRUE)
predMeanAIC <- as.data.frame(t(predictorsMeanAIC))
predictionIntervalAIC <- predict(modelAIC, newdata = predMeanAIC,
                                 interval = "prediction", level = 0.95)
print(predictionIntervalAIC)
```

```
       fit      lwr       upr
1 -164.8849 -237.209 -92.56072
```

Problem 5

Train-Test Split Set up

```
library(faraway)
library(glmnet)
library(MASS)

set.seed(123)
data(fat)
fat <- fat[, setdiff(names(fat), c("brozek", "density"))]

# Split the data into training and test sets
n <- nrow(fat)
test_index <- seq(1, n, by = 10)
train_index <- setdiff(1:n, test_index)

# Data Prep
trainDs <- fat[train_index, ]
testDs <- fat[test_index, ]

predictors <- setdiff(names(fat), "siri")
X_train <- trainDs[, predictors]
y_train <- trainDs$siri
X_test <- testDs[, predictors]
y_test <- testDs$siri
```

Part A

```
# Linear regression using all predictors
fullRegModel <- lm(siri ~ ., data = trainDs)
```

Part B

```
# Linear regression with variables selected using BIC
BICModel <- stepAIC(fullRegModel, direction = "both",
                    k = log(nrow(trainDs)), trace = FALSE)
```

Part C

```
# Ridge regression with lambda = 0.005
lambda <- 0.005
ridgeModel <- glmnet(X_train, y_train, alpha = 0, lambda = lambda)
```

Part D

In evaluating these three regression models, where $siri$ represents the response and the remaining variables (excluding $brozek$ and $density$) are predictors, we found different performance among them. The basic (insert my beloved linear regression) lm model yielded an RMSE of 1.946023, bringing solid predictive accuracy. However, the model using BIC, which tends towards simplicity due to its penalty on complexity, demonstrated poorer performance with an RMSE of 2.049515 (Sus); this suggests it might not have sufficiently captured the data's variance. In contrast, the Ridge regression model, set with $\lambda = 0.005$, achieved the best result, presenting the lowest RMSE of 1.926002. These results indicate that Ridge regression, which uses the approach to balance bias and variance, may offer enhanced predictive accuracy in scenarios involving complex data or when dealing with multiple predictor variables. Henceforth, I will vote out the BIC model and consider the normal regression and ridge regression models (since they are not sus).

```
predFullModel <- predict(fullRegModel, newdata = X_test)
predBICModel <- predict(BICModel, newdata = X_test)
predRidgeModel <- predict(ridgeModel, s = lambda, newx = as.matrix(X_test))

# Calculate RMSE
rmseLm <- sqrt(mean((predFullModel - y_test)^2))
rmseBic <- sqrt(mean((predBICModel - y_test)^2))
rmseRidge <- sqrt(mean((predRidgeModel - y_test)^2))

rmseDataFrame <- data.frame(
  Model = c("LM", "BIC", "Ridge"),
  RMSE = c(rmseLm, rmseBic, rmseRidge)
)
print(rmseDataFrame)
```

```
  Model     RMSE
1    LM 1.946023
2   BIC 2.049515
3 Ridge 1.926002
```