Multiple-linear-regression-analysis.R

Kgomotso Xaba

```r
#Multiple Linear Regression Analysis
data(seatpos,package = "faraway") #loading the data
library(faraway)
#attach(seatpos)
head(seatpos)
```

```
##   Age Weight HtShoes    Ht Seated  Arm Thigh  Leg hipcenter
## 1  46    180   187.2 184.9   95.2 36.1  45.3 41.3  -206.300
## 2  31    175   167.5 165.5   83.8 32.9  36.5 35.9  -178.210
## 3  23    100   153.6 152.2   82.9 26.0  36.6 31.0   -71.673
## 4  19    185   190.3 187.4   97.3 37.4  44.1 41.0  -257.720
## 5  23    159   178.0 174.1   93.9 29.5  40.1 36.9  -173.230
## 6  47    170   178.7 177.0   92.4 36.0  43.2 37.4  -185.150
```

```r
summary(seatpos)
```

```
##       Age            Weight         HtShoes            Ht
##  Min.   :19.00   Min.   :100.0   Min.   :152.8   Min.   :150.2
##  1st Qu.:22.25   1st Qu.:131.8   1st Qu.:165.7   1st Qu.:163.6
##  Median :30.00   Median :153.5   Median :171.9   Median :169.5
##  Mean   :35.26   Mean   :155.6   Mean   :171.4   Mean   :169.1
##  3rd Qu.:46.75   3rd Qu.:174.0   3rd Qu.:177.6   3rd Qu.:175.7
##  Max.   :72.00   Max.   :293.0   Max.   :201.2   Max.   :198.4
##      Seated            Arm            Thigh            Leg
##  Min.   : 79.40   Min.   :26.00   Min.   :31.00   Min.   :30.20
##  1st Qu.: 85.20   1st Qu.:29.50   1st Qu.:35.73   1st Qu.:33.80
##  Median : 89.40   Median :32.00   Median :38.55   Median :36.30
##  Mean   : 88.95   Mean   :32.22   Mean   :38.66   Mean   :36.26
##  3rd Qu.: 91.62   3rd Qu.:34.48   3rd Qu.:41.30   3rd Qu.:38.33
##  Max.   :101.60   Max.   :39.60   Max.   :45.50   Max.   :43.10
##    hipcenter
##  Min.   :-279.15
##  1st Qu.:-203.09
##  Median :-174.84
##  Mean   :-164.88
##  3rd Qu.:-119.92
##  Max.   : -30.95
```

```r
#dim(seatpos)
library(car)
```

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

```
##
## Attaching package: 'car'
```

```
## The following objects are masked from 'package:faraway':
##
##     logit, vif

#install.packages("faraway")

#fitting the model

model<-lm(hipcenter~., data = seatpos)

summary(model)

##
## Call:
## lm(formula = hipcenter ~ ., 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

#1. testing for multicollinearity

#i) condition numbers

c <- model.matrix(model)[,-1]  #eigendecomposition of the predictor space exc
luding the intercept
e <- eigen(t(c) %*% c)
e$val

## [1] 3.653671e+06 2.147948e+04 9.043225e+03 2.989526e+02 1.483948e+02
## [6] 8.117397e+01 5.336194e+01 7.298209e+00
```

```r
sqrt(e$val[1]/e$val)
```

```
## [1]   1.00000  13.04226  20.10032 110.55123 156.91171 212.15650 261.66698
## [8] 707.54911
```

```r
#ii) correlation matrix to determine the strength of the relationships
#between predictors
round(cor(seatpos),2)
```

```
##             Age Weight HtShoes    Ht Seated   Arm Thigh   Leg hipcenter
## Age        1.00   0.08   -0.08 -0.09  -0.17  0.36  0.09 -0.04      0.21
## Weight     0.08   1.00    0.83  0.83   0.78  0.70  0.57  0.78     -0.64
## HtShoes   -0.08   0.83    1.00  1.00   0.93  0.75  0.72  0.91     -0.80
## Ht        -0.09   0.83    1.00  1.00   0.93  0.75  0.73  0.91     -0.80
## Seated    -0.17   0.78    0.93  0.93   1.00  0.63  0.61  0.81     -0.73
## Arm        0.36   0.70    0.75  0.75   0.63  1.00  0.67  0.75     -0.59
## Thigh      0.09   0.57    0.72  0.73   0.61  0.67  1.00  0.65     -0.59
## Leg       -0.04   0.78    0.91  0.91   0.81  0.75  0.65  1.00     -0.79
## hipcenter  0.21  -0.64   -0.80 -0.80  -0.73 -0.59 -0.59 -0.79      1.00
```

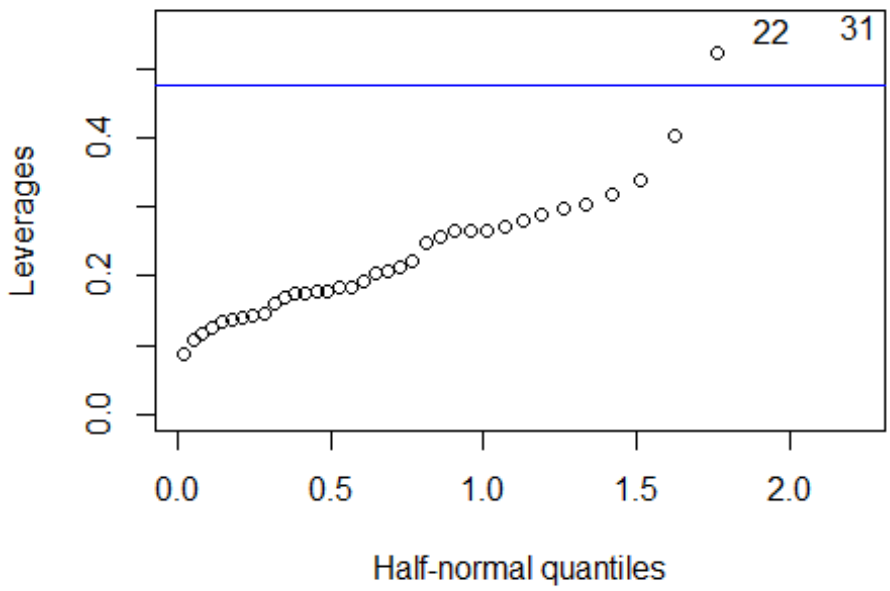
```r
#iv) variance inflation factor
vif(model)
```

```
##        Age     Weight    HtShoes         Ht     Seated        Arm       Thi
gh
##   1.997931   3.647030 307.429378 333.137832   8.951054   4.496368   2.7628
86
##        Leg
##   6.694291
```

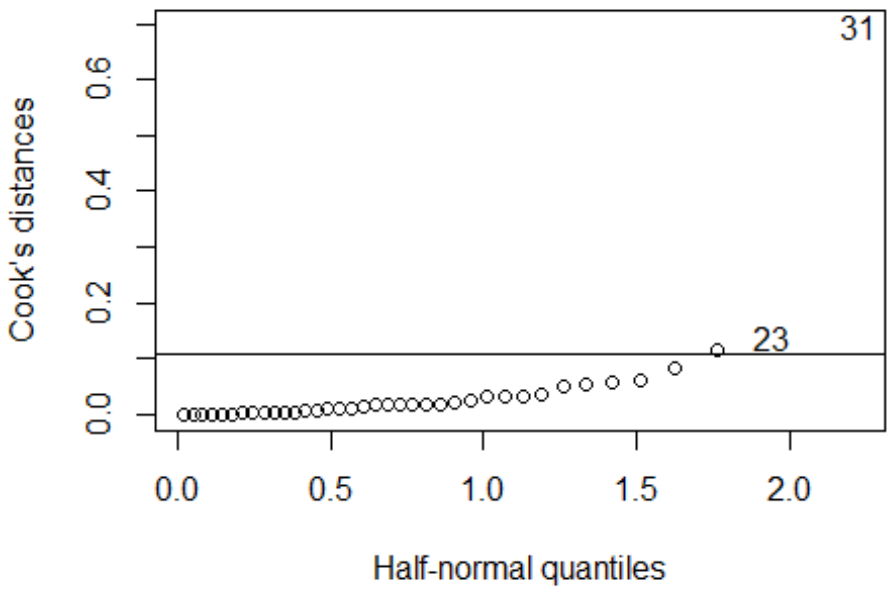
```r
#based on the results HtShoes and Ht are correlated with all other variables
#And Leg and seated variables have a strong correlation to other variables

set.seed(133)
#adding random noise to test the effect of collinearity
noise = rnorm(n = nrow(seatpos), mean = 0, sd = 10)
model_noise = lm(hipcenter + noise ~ ., data = seatpos)
summary(model_noise)
```

```
##
## Call:
## lm(formula = hipcenter + noise ~ ., data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.980 -20.542  -6.721  25.573  71.947
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 399.02324  169.42664   2.355   0.0255 *
## Age           0.82637    0.58010   1.425   0.1650
```

```
## Weight       -0.00111    0.33664  -0.003   0.9974
## HtShoes      -1.92471    9.92020  -0.194   0.8475
## Ht           -0.59384   10.30350  -0.058   0.9544
## Seated        1.31649    3.82637   0.344   0.7333
## Arm           1.41767    3.96705   0.357   0.7234
## Thigh        -2.02133    2.70562  -0.747   0.4610
## Leg          -6.77551    4.79465  -1.413   0.1683
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.37 on 29 degrees of freedom
## Multiple R-squared:  0.6754, Adjusted R-squared:  0.5858
## F-statistic: 7.542 on 8 and 29 DF,  p-value: 2.076e-05

coef(model)

##  (Intercept)          Age       Weight      HtShoes           Ht       Sea
ted
## 436.43212823   0.77571620   0.02631308  -2.69240774   0.60134458   0.53375
170
##          Arm        Thigh          Leg
##  -1.32806864  -1.14311888  -6.43904627

coef(model_noise)

##   (Intercept)           Age        Weight       HtShoes            Ht
## 399.023242920   0.826369569  -0.001110362  -1.924708007  -0.593842884
##        Seated           Arm         Thigh           Leg
##   1.316490384   1.417672091  -2.021331743  -6.775506344

# no significant change based on the coefficient comparison

#2. Checking unusual observations
par(mfrow = c(1, 1))
#i) leverage
hatv <- hatvalues(model)
head(hatv)

##         1         2         3         4         5         6
## 0.1763370 0.2805628 0.2042970 0.2552597 0.3367341 0.1326991

sum(hatv)    #sum of all leverages equal number of parameters in the model

## [1] 9

drivers<- row.names(seatpos)
halfnorm(hatv,labs=drivers,ylab="Leverages")
abline(h=0.474, col="blue") # cutoff point = 2*p/n, any point > h, is a lever
age
```



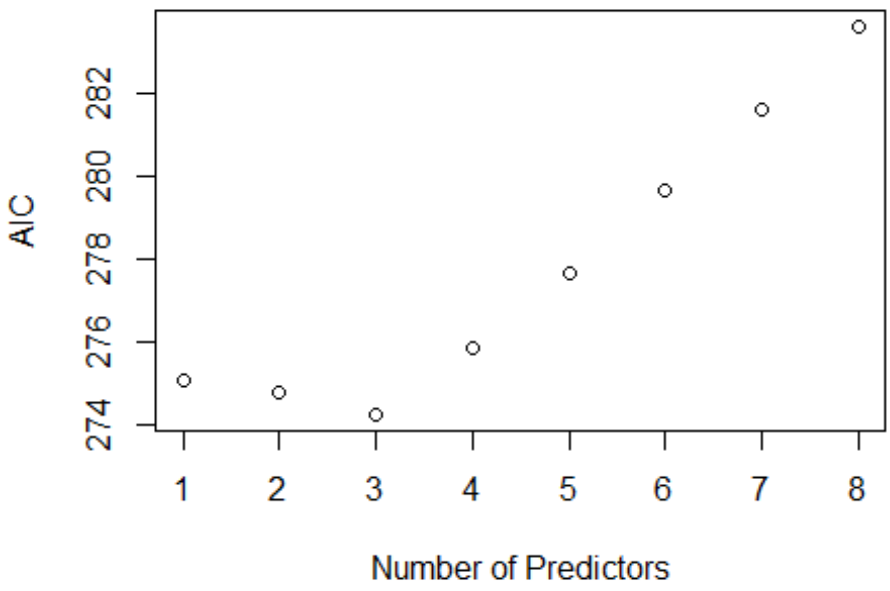

```
#22 and 31 stick out most as influential points


#ii) outliers

stud <- rstudent(model)
stud[which.max(abs(stud))]

##       31
## 2.389611

qt(.05/(38*2),28)              #Bonferroni critical value

## [1] -3.56932

#since 2.39 < |-3.57|, observation 31 is not an outlier

#iii) influential points - use Cook's distance
cook <- cooks.distance(model)
halfnorm(cook,2,labs=drivers,ylab="Cook's distances")
abline(h=0.11)
```




```
#the plot suggests that age 31 may significantly affect regression results an
d
#potentially distort conclusions drawn

#removing the observation with the largest's cooks distance to check how it i
nfluences the fit
modcook <- lm(hipcenter~.,seatpos,subset=(cook < max(cook)))
sumary(modcook)

##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 250.08839  173.06130  1.4451  0.15953
## Age           0.88623    0.53100  1.6690  0.10627
## Weight       -0.34659    0.34437 -1.0065  0.32281
## HtShoes      -0.18932    9.10647 -0.0208  0.98356
## Ht            3.51633    9.47441  0.3711  0.71333
## Seated       -3.74070    3.92098 -0.9540  0.34824
## Arm          -5.08089    3.94366 -1.2884  0.20816
## Thigh        -2.13845    2.50211 -0.8547  0.39999
## Leg         -10.44082    4.68189 -2.2300  0.03394
##
## n = 37, p = 9, Residual SE = 34.98590, R-Squared = 0.71

coef(model)# coeffiecients related to each predictor before removing largest

##  (Intercept)          Age       Weight      HtShoes           Ht       Sea
ted
## 436.43212823   0.77571620   0.02631308  -2.69240774   0.60134458   0.53375
```

```
170
##          Arm        Thigh          Leg
##  -1.32806864  -1.14311888  -6.43904627

#cooks distance
coef(modcook)# coeffiecients related to each predictor after removing largest

## (Intercept)         Age      Weight     HtShoes          Ht      Seated
## 250.0883943   0.8862323  -0.3465932  -0.1893213   3.5163265  -3.7406957
##         Arm       Thigh         Leg
##  -5.0808904  -2.1384459 -10.4408228

#cooks distance
#coefficients significantly change which suggests the 31 does in fact
# distort results

require(lmtest)

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

shapiro.test(residuals(model))  #test for normality

##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.97152, p-value = 0.4341

#p-value=0.4341>0.05 therefore it suggests the residuals
#likey Normally distributed
bptest(model)                   #test for homoscedasticity Breusch-Pagan

##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 14.037, df = 8, p-value = 0.0808

#p-value=0.081>0.05 therefore it suggests the variance of the errors are
#likey constant
dwtest (model)               #test for correlated errors

##
##  Durbin-Watson test
```

```
##
## data:  model
## DW = 1.7688, p-value = 0.2441
## alternative hypothesis: true autocorrelation is greater than 0

#p-value=0.2441>0.05 therefore it suggests there is
#likey no significant autocorrelation



#5. Selection of the "best" model
#i) AIC                          #AIC = nlog(RSS/n) + 2p
require(leaps)

## Loading required package: leaps

par(mfrow = c(1, 1)) #
b <- regsubsets(hipcenter~., data = seatpos)
rs <- summary(b)
rs$which

##   (Intercept)   Age Weight HtShoes    Ht Seated   Arm Thigh   Leg
## 1        TRUE FALSE  FALSE   FALSE  TRUE  FALSE FALSE FALSE FALSE
## 2        TRUE FALSE  FALSE   FALSE  TRUE  FALSE FALSE FALSE  TRUE
## 3        TRUE  TRUE  FALSE   FALSE  TRUE  FALSE FALSE FALSE  TRUE
## 4        TRUE  TRUE  FALSE    TRUE FALSE  FALSE FALSE  TRUE  TRUE
## 5        TRUE  TRUE  FALSE    TRUE FALSE  FALSE  TRUE  TRUE  TRUE
## 6        TRUE  TRUE  FALSE    TRUE FALSE   TRUE  TRUE  TRUE  TRUE
## 7        TRUE  TRUE   TRUE    TRUE FALSE   TRUE  TRUE  TRUE  TRUE
## 8        TRUE  TRUE   TRUE    TRUE  TRUE   TRUE  TRUE  TRUE  TRUE

AIC <- 38*log(rs$rss/38) + (2:9)*2
AIC

## [1] 275.0667 274.7798 274.2418 275.8291 277.6712 279.6389 281.6286 283.624
0

plot(AIC ~ I(1:8), ylab="AIC", xlab="Number of Predictors")
```




```
# since the model 3 has the lowest AIC value we can assume its the best model
# i.e. HtShoes is the best predictor

model2<-lm(hipcenter~Age+Ht+Leg, data = seatpos)
summary(model2)

##
## Call:
## lm(formula = hipcenter ~ Age + Ht + Leg, data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.715 -22.758  -4.102  21.394  60.576
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 452.1976   100.9482   4.480 8.04e-05 ***
## Age           0.5807     0.3790   1.532   0.1347
## Ht           -2.3254     1.2545  -1.854   0.0725 .
## Leg          -6.7390     4.1050  -1.642   0.1099
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.12 on 34 degrees of freedom
## Multiple R-squared:  0.6814, Adjusted R-squared:  0.6533
## F-statistic: 24.24 on 3 and 34 DF,  p-value: 1.426e-08
```

```r
#significance of regression coefficients using CI #if B_j = 0 falls within a
CI
confint(model2)
```

```
##                    2.5 %      97.5 %
## (Intercept) 247.0461589 657.3490391
## Age          -0.1894153   1.3508993
## Ht           -4.8747940   0.2240374
## Leg         -15.0813810   1.6034314
```

```r
#Zero falls within the confidence region for all predictors
# i.e.there is insufficient evidence to conclude the predictors
#significantly affect the outcome
```