# Hw8

*Manuel Alejandro Garcia Acosta*

*11/5/2019*

We require the following packages for using the function vif().

```
library(car)
```

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

```
library(carData)
```

```
setwd('/home/noble_mannu/Documents/PhD/First/STAT_2131_Applied_Statistical_Methods_I/HW8')
Data <- read.table('SeatPos.txt', header = TRUE)
```

# Exercise (e)

# Part (i)

## Making the multilinear regression model

```
linearMod <- lm(hipcenter ~ ., data=Data)
```

## Displaying the summary of our model

Next we display the summary of our model.

```
summary(linearMod)
```

```
##
## Call:
## lm(formula = hipcenter ~ ., data = Data)
##
## 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
```

The P-Value for the null hypothesis that the coefficients for all predictors are 0 is $1.306 * 10^{-5}$. At a confidence level $\alpha = 0.05$ we'll reject the null hypothesis. We have a fairly high coefficient of determination $R^2 = 0.6866$. Here we have mixed signals since the test indicates that some of the predictors actually do a good job but while checking the P-Values individually they are really high (they aren't significant).

Since we have high $R^2$ and high P-Values for all covariates we can think that our predictor might have collinearity. We'll confirm this in the following parts of the homework. In the last part (vi) we'll see that if we drop all but one of the predictors that are highly correlated our model will improve.

## Part (ii)

With the model in part (i) and its output I would not say that there are any individual covariates that appear to be significantly related to hipcenter since all of them have high P-Values (all of them above the confidence level $\alpha = 0.05$).

## Part (iii)

Here I computed the Variance Inflation Factors (VIF) for all 8 predictors. As you'll see some of them were really large.

```r
vif(lm(hipcenter ~ ., data=Data))
```

```
##        Age     Weight    HtShoes         Ht     Seated        Arm
##   1.997931   3.647030 307.429378 333.137832   8.951054   4.496368
##      Thigh        Leg
##   2.762886   6.694291
```

In class we saw that any VIF above 10 is considered too large. This is the case of HtShoes (Height with shoes in cm) and Ht (Height when bare foot in cm). Both of them have VIF larger than 300.

## Part (iv)

```r
test <- read.table('SeatPos.txt', header = TRUE)
beta_v <- NULL

for (i in 1:100){
  copy <- data.frame(test)
  noise <- rnorm( length(copy$hipcenter), mean = 0, sd = 1 )
  copy$hipcenter <- copy$hipcenter+noise
  linear_loop <- lm(hipcenter ~ ., data = copy)
  beta_hs <- coef(linear_loop)[4]
  beta_v <- c(beta_v,beta_hs)
}

beta <- coef(linearMod)[4]
```

```
sqrt( (1/100)*sum( ((beta_v - beta)/beta)^2 ) )
```

```
## [1] 0.09606563
```

We saw in class that $var(\hat{\beta}_k) = \frac{\sigma^2}{n} VIF_k$. This implies $VIF_k = \frac{n}{\sigma^2} var(\hat{\beta}_k)$ Therefore, a large inflation factor implies a high variance for the estimate of the corresponding coefficient.

# Part (v)

Next I'll compute the pearson correlation matrix for the dataset. I also extract the correlation coefficient pair-wise for the predictors (HtShoes,Ht,Seaterd,Arm,Thigh and Leg). As you'll see all of these are highly correlated to one another.

```
# This is the correlation matrix between all covariates
cor(Data, method = 'pearson')
```

```
##                   Age      Weight     HtShoes          Ht     Seated
## Age        1.00000000  0.08068523 -0.07929694 -0.09012812 -0.1702040
## Weight     0.08068523  1.00000000  0.82817733  0.82852568  0.7756271
## HtShoes   -0.07929694  0.82817733  1.00000000  0.99814750  0.9296751
## Ht        -0.09012812  0.82852568  0.99814750  1.00000000  0.9282281
## Seated    -0.17020403  0.77562705  0.92967507  0.92822805  1.0000000
## Arm        0.35951115  0.69755240  0.75195305  0.75214156  0.6251964
## Thigh      0.09128584  0.57261442  0.72486225  0.73496041  0.6070907
## Leg       -0.04233121  0.78425706  0.90843341  0.90975238  0.8119143
## hipcenter  0.20517217 -0.64033298 -0.79659640 -0.79892742 -0.7312537
##                  Arm       Thigh         Leg hipcenter
## Age        0.3595111  0.09128584 -0.04233121  0.2051722
## Weight     0.6975524  0.57261442  0.78425706 -0.6403330
## HtShoes    0.7519530  0.72486225  0.90843341 -0.7965964
## Ht         0.7521416  0.73496041  0.90975238 -0.7989274
## Seated     0.6251964  0.60709067  0.81191429 -0.7312537
## Arm        1.0000000  0.67109849  0.75381405 -0.5850950
## Thigh      0.6710985  1.00000000  0.64954120 -0.5912015
## Leg        0.7538140  0.64954120  1.00000000 -0.7871685
## hipcenter -0.5850950 -0.59120155 -0.78716850  1.0000000
```

```
# Here we extract the correlation between the predictors mentioned above
corr <- cor(Data, method = 'pearson')

corr[3,4] # HtShoes,Ht
```

```
## [1] 0.9981475
```

```
corr[3,5] # HtShoes,Seaterd
```

```
## [1] 0.9296751
```

```
corr[3,6] # HtShoes,Arm
```

```
## [1] 0.751953
```

```
corr[3,7] # HtShoes,Thigh
```

```
## [1] 0.7248622
```

```
corr[3,8] # HtShoes,Leg
```

```
## [1] 0.9084334
```

```
corr[4,5] # Ht,Seaterd
```

```
## [1] 0.9282281
```

```
corr[4,6] # Ht,Arm
```

```
## [1] 0.7521416
```

```
corr[4,7] # Ht,Thigh
```

```
## [1] 0.7349604
```

```
corr[4,8] # Ht,Leg
```

```
## [1] 0.9097524
```

```
corr[5,6] # Seaterd,Arm
```

```
## [1] 0.6251964
```

```
corr[5,7] # Seaterd,Thigh
```

```
## [1] 0.6070907
```

```
corr[5,8] # Seaterd,Leg
```

```
## [1] 0.8119143
```

```
corr[6,7] # Arm,Thigh
```

```
## [1] 0.6710985
```

```
corr[6,8] # Arm,Leg
```

```
## [1] 0.753814
```

```
corr[7,8] # Thigh,Leg
```

```
## [1] 0.6495412
```

# Part (vi)

Here I run another multilinear regression model dropping the predictors HtShoes, Seated, Arm, Thigh and Leg.

### Making the new multilinear regression model by dropping several predictors

```
linearMod1 <- lm(hipcenter ~ Age+Weight+Ht, data=Data)
```

### Displaying the summary of our model

Next we display the summary of our model.

```r
summary(linearMod1)
```

```
##
## Call:
## lm(formula = hipcenter ~ Age + Weight + Ht, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.526 -23.005   2.164  24.950  53.982
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 528.297729 135.312947   3.904 0.000426 ***
## Age           0.519504   0.408039   1.273 0.211593
## Weight        0.004271   0.311720   0.014 0.989149
## Ht           -4.211905   0.999056  -4.216 0.000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.49 on 34 degrees of freedom
## Multiple R-squared:  0.6562, Adjusted R-squared:  0.6258
## F-statistic: 21.63 on 3 and 34 DF,  p-value: 5.125e-08
```

As we can see, the coefficient of determination $R_A^2 = 0.6562$ remained almost the same as in the model with all 8 predictors $R_S^2 = 0.6866$. The difference is $R_S^2 - R_A^2 = 0.0304$. The small difference can be explained by the fact that the predictors we dropped are highly correlated to Ht.

The same argument holds for the change of the P-Value for Ht. Since in the other model we considered 5 predictors -HtShoes, Seated, Arm, Thigh and Leg- that were highly correlated with Ht (we can talk about multicollinearity here) it had a large P-Value as a result. After we dropped those predictors the P-Value became small and made Ht significant in the new model.