# Inclass

Hyun Suk (Max) Ryoo (hr2ee)

10/5/2021

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.2
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0
## v readr   1.4.0
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
## Warning: package 'readr' was built under R version 4.0.2
```

```
## Warning: package 'stringr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

```
## -- Conflicts ------------------------------------------------------------------------------------ 
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## x MASS::select()  masks dplyr::select()
```

```
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 4.0.2
```

```
setwd("/Users/maxryoo/Documents/MSDS/STAT6021/Module7")
data <- seatpos
head(data)
```

```
##   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
```

## 1) Fit the full model with all the predictors. Using the summary() function, comment on

the results of the t tests and ANOVA F test from the output.

```
result <- lm(hipcenter~., data=data)
summary(result)
```

```
##
## 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
```

This model is good for predicting hte hipcenter however, each predictor has a very high p-value, which means it is insignicant. We need to do more testing to drop the predictors

## 2) Briefly explain why, based on your output from part 1, you suspect the model shows

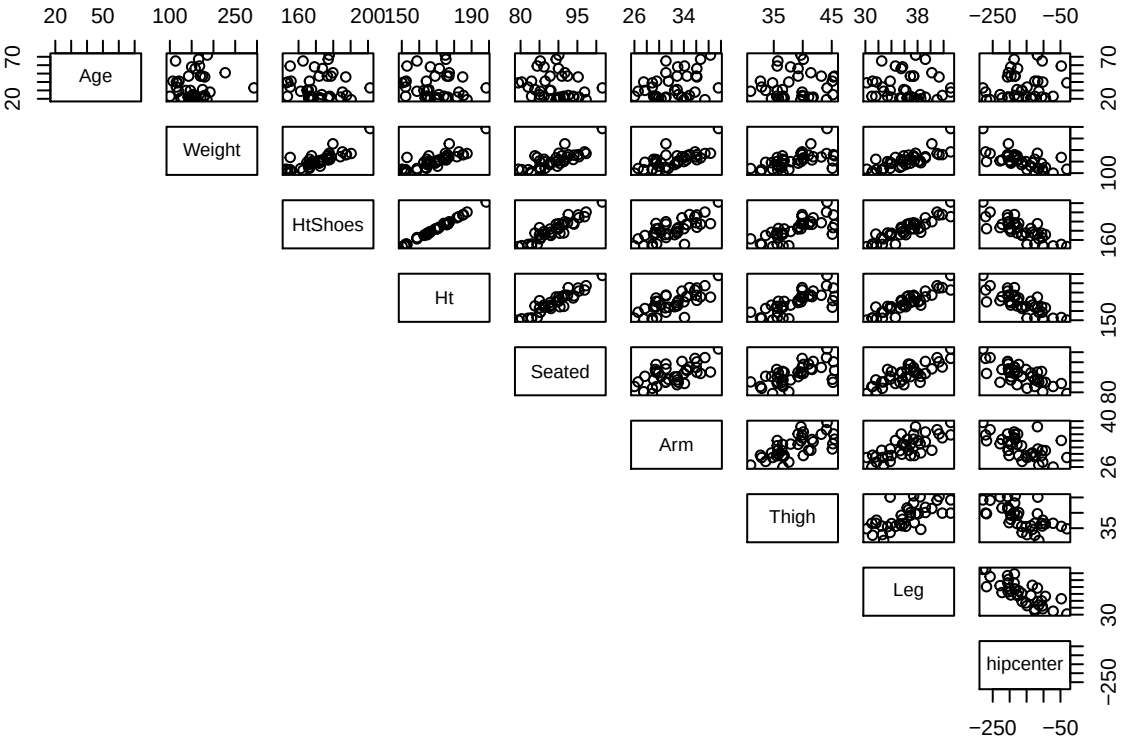
signs of multicollinearity.

This model is good for predicting hte hipcenter however, each predictor has a very high p-value, which means it is insignicant. We need to do more testing to drop the predictors

## 3) . Provide the output for all the pairwise correlations among the predictors. Comment

briefly on the pairwise correlations

```
pairs(data, lower.panel = NULL)
```




```
cor(data)
```

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

```
## Seated     0.60709067  0.81191429 -0.7312537
## Arm        0.67109849  0.75381405 -0.5850950
## Thigh      1.00000000  0.64954120 -0.5912015
## Leg        0.64954120  1.00000000 -0.7871685
## hipcenter -0.59120155 -0.78716850  1.0000000
```

Lots of correlations that high between the predictors. I.E. Weight, HtShoes, Ht, Seated .. All of the variables seem to be correlated with one noather to be frankly honest.

**4) Check the variance inflation factors (VIFs). What do these values indicate about**

multicollinearity?

```
vif(result)
```

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

## 5) Looking at the data, we may want to look at the correlations for the variables that

describe length of body parts: HtShoes, Ht, Seated, Arm, Thigh, and Leg. Comment on the correlations of these six predictors

Highly Correlated

## 6) Since all the six predictors from the previous part are highly correlated, you may decide

to just use one of the predictors and remove the other five from the model. Decide which predictor out of the six you want to keep, and briefly explain your choice. Thigh because it has a lower vif value after arm SHOULD BE HEIGHT SINCE HIGHER VIF

```
result2 <- lm(hipcenter~Thigh, data=data)
summary(result2)
```

```
##
## Call:
## lm(formula = hipcenter ~ Thigh, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.708 -30.030  -9.213  30.879 106.534
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  186.891     80.373   2.325   0.0258 *
## Thigh         -9.100      2.069  -4.398 9.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.77 on 36 degrees of freedom
## Multiple R-squared:  0.3495, Adjusted R-squared:  0.3315
## F-statistic: 19.34 on 1 and 36 DF,  p-value: 9.29e-05
```

## 7) Since all the six predictors from the previous part are highly correlated, you may decide

to just use one of the predictors and remove the other five from the model. Decide which predictor out of the six you want to keep, and briefly explain your choice.

```
result3 <- lm(hipcenter~Ht + Arm + Weight, data=data)
summary(result3)
```

```
##
## Call:
## lm(formula = hipcenter ~ Ht + Arm + Weight, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.865 -27.968   4.019  22.776  68.378
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 589.7231   129.2991   4.561 6.32e-05 ***
## Ht           -4.6457     1.0889  -4.266  0.00015 ***
## Arm           0.4538     2.8202   0.161  0.87311
## Weight        0.1047     0.3127   0.335  0.73987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.33 on 34 degrees of freedom
## Multiple R-squared:   0.64,  Adjusted R-squared:  0.6083
## F-statistic: 20.15 on 3 and 34 DF,  p-value: 1.104e-07
```

```
vif(result3)
```

```
##       Ht      Arm   Weight
## 3.930209 2.400173 3.324415
```

Should be okay

## 8) Conduct a partial F test to investigate if the predictors you dropped from the full

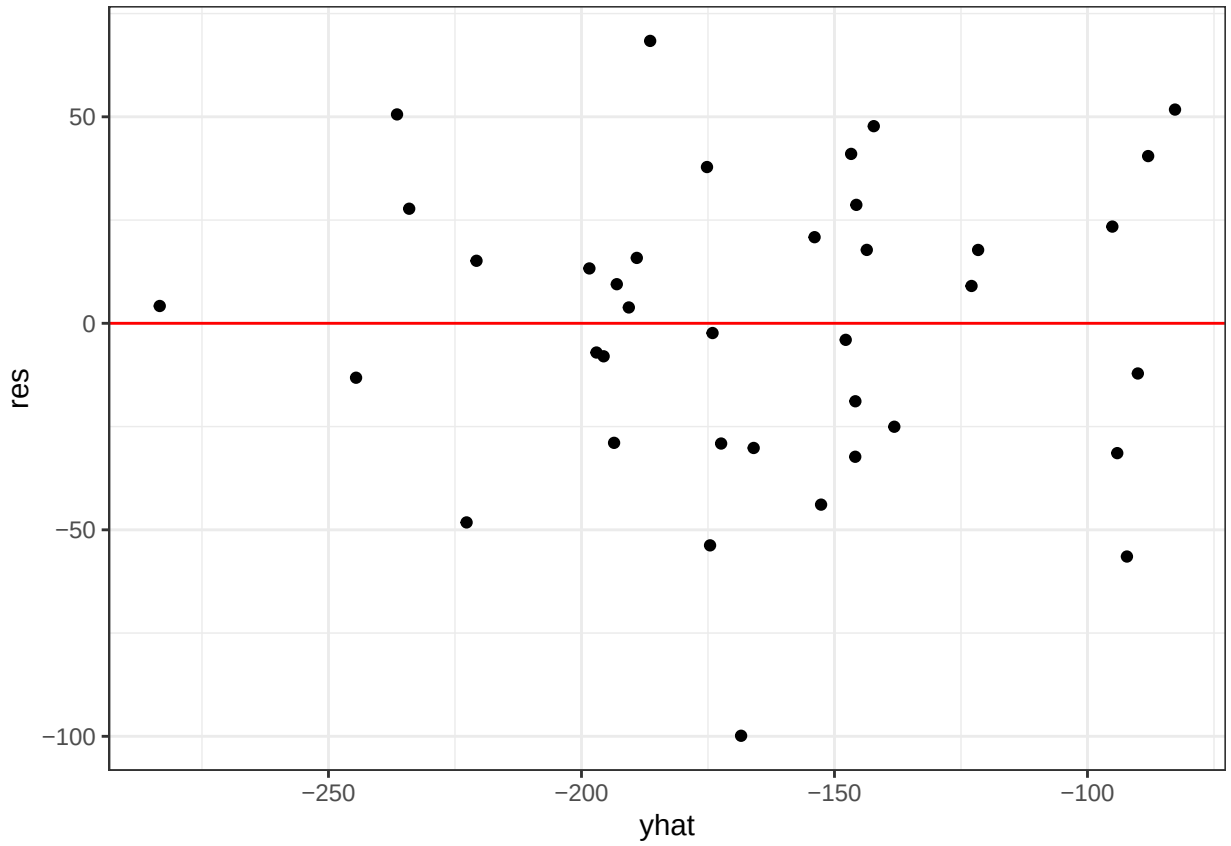
model were jointly insignificant. Be sure to state a relevant conclusion.

```
anova(result3,result)
```

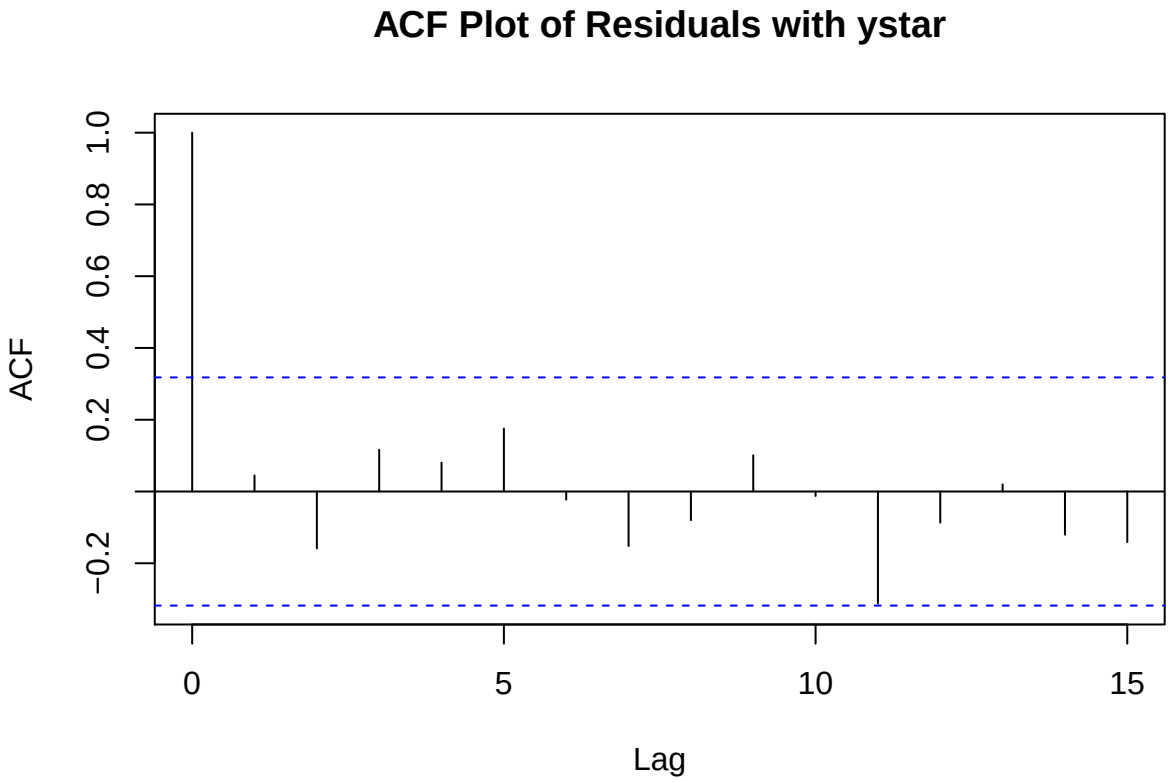
```
## Analysis of Variance Table
##
## Model 1: hipcenter ~ Ht + Arm + Weight
## Model 2: hipcenter ~ Age + Weight + HtShoes + Ht + Seated + Arm + Thigh +
##     Leg
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     34 47384
## 2     29 41262  5    6122.1 0.8606 0.5192
```

Produce a plot of residuals against fitted values for your model from part 7. Based on the residual plot, comment on the assumptions for the multiple regression model. Also produce an ACF plot and QQ plot of the residuals, and comment on the plots.

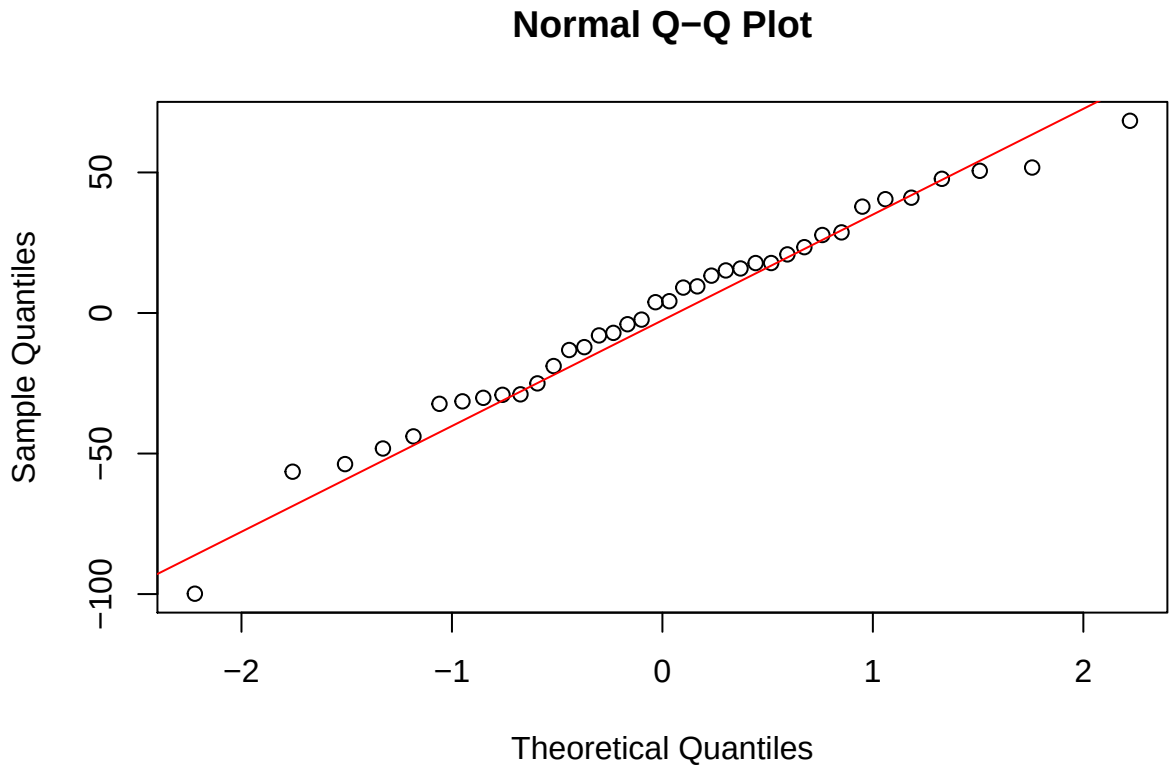
```
yhat = result3$fitted.values
res = result3$residuals
data %>%
  ggplot(aes(yhat, res)) + geom_point() + theme_bw() + geom_hline(yintercept=0, color="red")
```




```
acf(result3$residuals, main="ACF Plot of Residuals with ystar")
```




```
{
qqnorm(result3$residuals)
qqline(result3$residuals, col="red")
}
```



Produce a plot of residuals against fitted values for your model from part 7. Based on the residual plot, comment on the assumptions for the multiple regression model. Also produce an ACF plot and QQ plot of the residuals, and comment on the plots.

```
summary(result3)
```

```
##
## Call:
## lm(formula = hipcenter ~ Ht + Arm + Weight, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.865 -27.968   4.019  22.776  68.378
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 589.7231   129.2991   4.561 6.32e-05 ***
## Ht           -4.6457     1.0889  -4.266  0.00015 ***
## Arm           0.4538     2.8202   0.161  0.87311
## Weight        0.1047     0.3127   0.335  0.73987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.33 on 34 degrees of freedom
## Multiple R-squared:   0.64,  Adjusted R-squared:  0.6083
## F-statistic: 20.15 on 3 and 34 DF,  p-value: 1.104e-07
```