M07Guided

Alanna Hazlett

2024-03-18

x1: Age. Age in years
x2: Weight. Weight in pounds
x3: HtShoes. Height with shoes in cm
x4: Ht. Height without shoes in cm
x5: Seated. Seated height in cm
x6: Arm. Arm length in cm
x7: Thigh. Thigh length in cm
x8: Leg. Lower leg length in cm
y: hipcenter

Problem 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.

```
#Capitalized Hipcenter, because all of the other variables were capitalized
Data<- Data %>%
  rename(Hipcenter = hipcenter)
```

```
result.full<-lm(Hipcenter~.,data=Data)
summary(result.full)
```

```
##
## 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
```

```
qt(1-(0.05/2),(38-10))
```

```
## [1] 2.048407
```

Our t values and p-values for all of our predictors indicate that they are statistically insignificant, meaning we would fail to reject the null hypothesis for each predictor equaling zero. This indicates that we should drop each predictor where we fail to reject the null hypothesis. Since all of the predictors indicate this, we would drop all of our predictors.
Our F statistic on 8 and 29 df is 7.94, we compare this to a F 8, 29 distribution.

```
qf(1-(0.05/2),8,29)
```

```
## [1] 2.668562
```

Our F statistic is larger than our critical value and our p-value is smaller than our significance level , so we reject the null hypothesis of all coefficients equaling zero. The data support the claim that our model with the eight predictors is useful in predicting Hipcenter.

Problem 2

Briefly explain why, based on your output from part 1, you suspect the model shows signs of multicollinearity. The two tests are contradictory, one is stating that none of the predictors are good for the model and the other states that they are good for the model. It also appears that Ht (Height bare foot in cm) has a large standard error for it's coefficient.

Problem 3

Provide the output for all the pairwise correlations among the predictors. Comment briefly on the pairwise correlations.

```
round(cor(Data),3)
```

```
##              Age Weight HtShoes     Ht Seated    Arm  Thigh    Leg Hipcenter
## Age        1.000  0.081  -0.079 -0.090 -0.170  0.360  0.091 -0.042     0.205
## Weight     0.081  1.000   0.828  0.829  0.776  0.698  0.573  0.784    -0.640
## HtShoes   -0.079  0.828   1.000  0.998  0.930  0.752  0.725  0.908    -0.797
## Ht        -0.090  0.829   0.998  1.000  0.928  0.752  0.735  0.910    -0.799
## Seated    -0.170  0.776   0.930  0.928  1.000  0.625  0.607  0.812    -0.731
## Arm        0.360  0.698   0.752  0.752  0.625  1.000  0.671  0.754    -0.585
## Thigh      0.091  0.573   0.725  0.735  0.607  0.671  1.000  0.650    -0.591
## Leg       -0.042  0.784   0.908  0.910  0.812  0.754  0.650  1.000    -0.787
## Hipcenter  0.205 -0.640  -0.797 -0.799 -0.731 -0.585 -0.591 -0.787     1.000
```

Almost all of the correlation pairs are high, except: All predictors with Age

Problem 4

Check the variance inflation factors (VIFs). What do these values indicate about multicollinearity?

```
round(faraway::vif(result.full),2)
```

```
##     Age  Weight HtShoes      Ht  Seated     Arm   Thigh     Leg
##    2.00    3.65  307.43  333.14    8.95    4.50    2.76    6.69
```

Some level of multicollinearity is noted by Seated and Leg. There is extremely high level of multicollineartiy noted by HtShoes and Ht.

Problem 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. These are all highly positively correlated with one another.

Problem 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.
I would choose Thigh, which is the length of the thigh in cm. I would choose this because Hipcenter is the horizontal distance of the hips to a specified point in the car. The thigh length I believe would play the largest role in this, because it is also a horizontal distance.

Problem 7

Based on your choice in part 6, fit a multiple regression with your choice of predictor to keep, along with the predictors x1 = Age and x2 =Weight. Check the VIFs for this model. Comment on whether we still have an issue with multicollinearity.

```
result.reduced<-lm(Hipcenter~ Age + Weight + Thigh, data=Data)
round(faraway::vif(result.reduced),4)
```

```
##    Age Weight  Thigh
## 1.0096 1.4897 1.4924
```

No, it appears that we no longer have an issue with multicolinearity based on the VIFs.

Problem 8

Conduct a general linear F test to investigate if the predictors you dropped from the full model were jointly insignificant. Be sure to state a relevant conclusion.

## **Two Methods to Solve General Linear F test***

$$H_0 : \hat{\beta}_0 = \hat{\beta}_{Age} = \hat{\beta}_{Weight} = \hat{\beta}_{Thigh} = \hat{\beta}_{HtShoes} = \hat{\beta}_{Ht} = \hat{\beta}_{Seated} = \hat{\beta}_{Arm} = 0$$
$$H_a : at\ least\ one\ \neq 0$$

### 1. Compare F0 statistic to F (r, n-p) distribution:

```
anova(result.reduced,result.full)
```

```
## Analysis of Variance Table
##
## Model 1: Hipcenter ~ Age + Weight + Thigh
## Model 2: Hipcenter ~ Age + Weight + HtShoes + Ht + Seated + Arm + Thigh +
##     Leg
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     34 57963
## 2     29 41262  5     16702 2.3477 0.06611 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F_0 = \frac{(SS_R(F) - SS_R(R)) \ / \ r}{(SS_{res}(F)) \ / \ (n-p)} = \frac{(SS_{res}(R) - SS_{res}(F)) \ / \ r}{(SS_{res}(F)) \ / \ (n-p)}$$

$$F_0 = \frac{(57963 - 41262) \ / \ 5}{41262 \ / \ (38-9)} = 2.3477$$

```
qf(1-0.05,5,38-9)
```

```
## [1] 2.545386
```

Our F0 statistic is less than our critical value, and our p-value of 0.06611 is greater than our significance level. We fail to reject the null hypothesis, so we go with the reduced model.

2. Sequential Sum of Squares/Extra Sums of Squares

```
result.seq<-lm(Hipcenter ~ Age + Weight + Thigh + HtShoes + Ht + Seated + Arm + Leg, data = Data)
anova(result.seq)
```

```
## Analysis of Variance Table
##
## Response: Hipcenter
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Age        1   5541    5541  3.8947 0.058036 .
## Weight     1  57175   57175 40.1840 6.31e-07 ***
## Thigh      1  10960   10960  7.7028 0.009551 **
## HtShoes    1  12900   12900  9.0663 0.005350 **
## Ht         1     54      54  0.0380 0.846722
## Seated     1    419     419  0.2942 0.591687
## Arm        1    674     674  0.4738 0.496694
## Leg        1   2655    2655  1.8659 0.182445
## Residuals 29  41262    1423
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

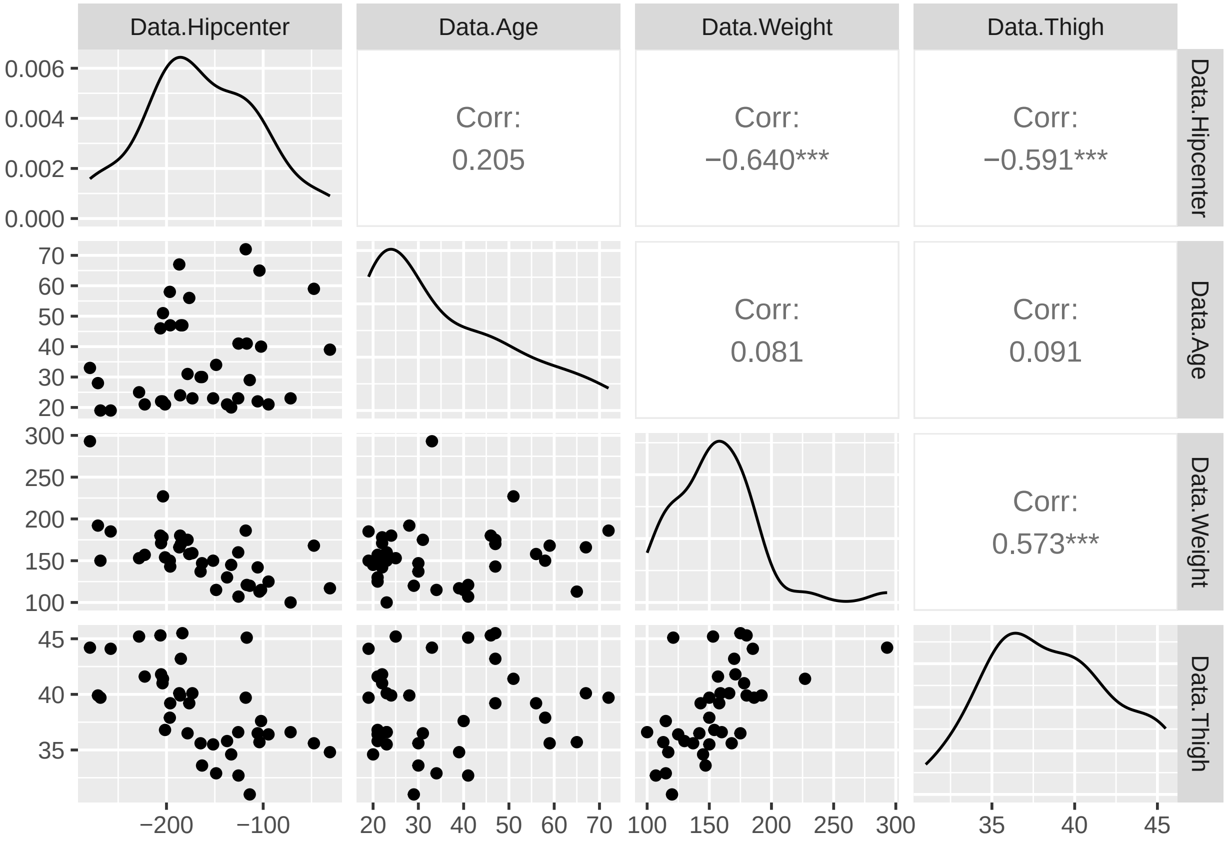
$$F_0 = \frac{(131640 - 114938) \ / \ 5}{(41262) \ / \ (38-9)} = 2.3477$$

Our F0 statistic is less than our critical value, and our p-value of 0.06611 is greater than our significance level. We fail to reject the null hypothesis, so we go with the reduced model.

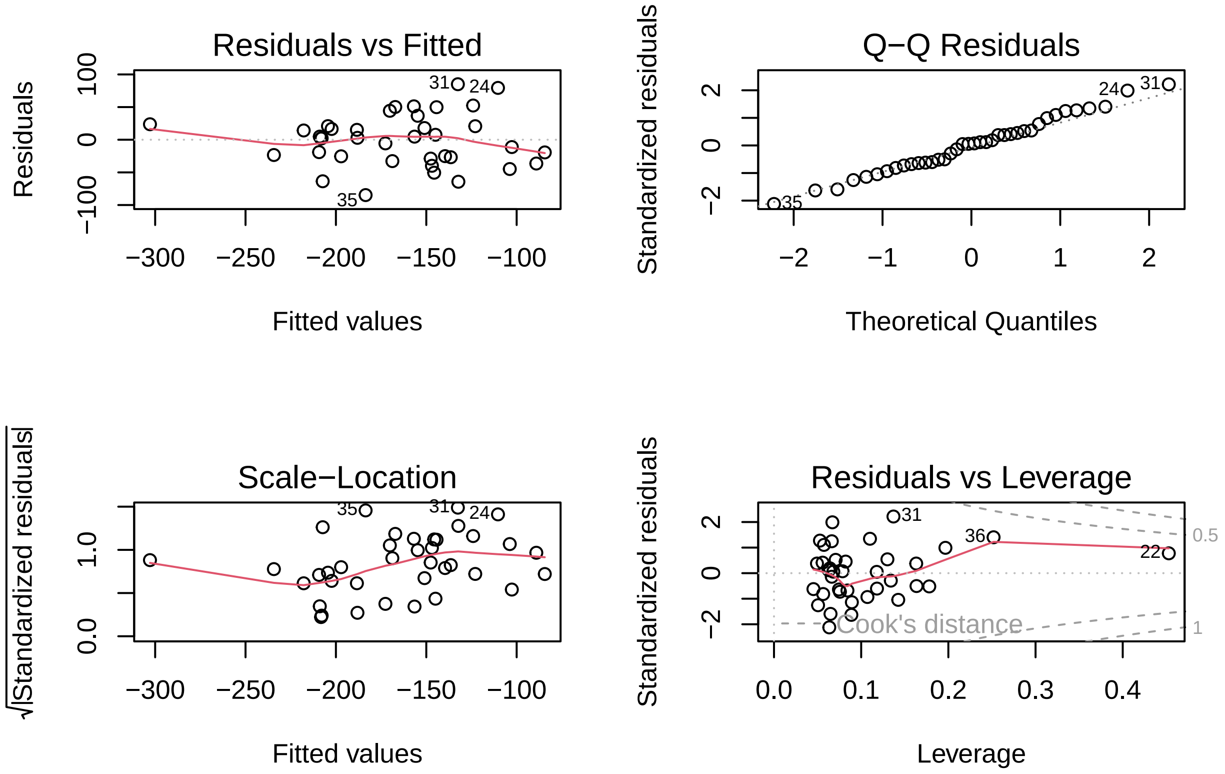
Problem 9

Produce the diagnostic plots for your model from part 7. Comment on whether the regression assumptions are met.

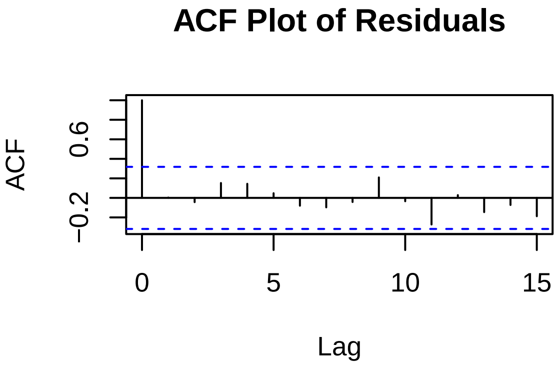
```
Data.Reduced<-data.frame(Data$Hipcenter, Data$Age, Data$Weight, Data$Thigh)
ggpairs(Data.Reduced)
```



```
par(mfrow=c(2,2))
plot(result.reduced)
```



```r
acf(result.reduced$residuals, main="ACF Plot of Residuals")
```



In Residuals vs Fitted the errors appear to have a mean near 0, asssumption 1 is met.
In Residuals vs Fitted we can see that the variance or the errors is relatvely constant lef to right, assumption 2 is met.
In ACF Plot confirms the belief that the observations are uncorrelated and the correlations between the vector of observations and lagged versions of these observations are very near zero, assumption 3 is met.
In Q-Q we can see that the observations are normally distribute, assumption 4 is met.

Problem 10

Based on your results, write your estimated regression equation from part 7. Also report the R2 of this model, and compare with the R2 you reported in part 1, for the model with all predictors. Also comment on the adjusted R2 for both models.

```r
summary(result.reduced)
```

```
##
```

```
## Call:
## lm(formula = Hipcenter ~ Age + Weight + Thigh, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.764 -26.436   2.596  20.809  84.995
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 126.7917    69.6700   1.820  0.07759 .
## Age           1.0654     0.4438   2.401  0.02198 *
## Weight       -0.7679     0.2315  -3.316  0.00218 **
## Thigh        -5.4259     2.1400  -2.535  0.01599 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.29 on 34 degrees of freedom
## Multiple R-squared:  0.5597, Adjusted R-squared:  0.5208
## F-statistic: 14.41 on 3 and 34 DF,  p-value: 3.194e-06
```

The estimated regression equation for this model is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_{Age}x + \hat{\beta}_{Weight}x + \hat{\beta}_{Thigh}x$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_{Age}x + \hat{\beta}_{Weight}x + \hat{\beta}_{Thigh}x$$

$$\hat{y} = 126.79 + 1.07x_{Age} + -0.77x_{Weight} + -5.43x_{Thigh}$$

### Rˆ2

Rˆ2 for the reduced model is 0.5597.
Rˆ2 for the full model is 0.6866.
Rˆ2 is the proportion of variance in the resonse variable that is explained by the predictors. We notice a slight decline in Rˆ2 from the full model to the reduced model, this is due to the removal of predictors.

Rˆ2 Adjusted

Rˆ2 adjusted for the reduced model is 0.5208.
Rˆ2 adjusted for the full model is 0.6001 .
Rˆ2 adjusted is not affected by the increase of predictors and only increases if the model is more useful.