Stat 6021: Guided Question Set 5 Solutions

1. The p-value for the F test is very small, and the R^2 is fairly high, around 69%. However, none of the individual t tests suggest any of the predictors is significant, given the other predictors.

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

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
```

2. The p-value for the F test suggests our model is useful in predicting the response. However, the individual t tests suggests none of the predictors are significant (given the presence of the other predictors). Also, the standard errors for some of the estimated coefficients are large. These observations suggest the presence of multicollinearity.

3. There are several large pairwise correlations between some of the predictors, as well as between predictors and the response.

```
             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
```

4. We have some high VIFs, for *HtShoes* and *Ht*. For example, the VIF for *HtShoes* is 307.429378, which tells us that the variance for *HtShoes* is 307 times larger than it would have been without collinearity. Note: you cannot apply this as a correction, the VIF just gives a sense of the effect.

```
> 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. These six predictors that relate to length are highly correlated with each other, as expected.

```
          HtShoes    Ht Seated   Arm Thigh   Leg
HtShoes     1.000 0.998  0.930 0.752 0.725 0.908
Ht          0.998 1.000  0.928 0.752 0.735 0.910
Seated      0.930 0.928  1.000 0.625 0.607 0.812
Arm         0.752 0.752  0.625 1.000 0.671 0.754
Thigh       0.725 0.735  0.607 0.671 1.000 0.650
Leg         0.908 0.910  0.812 0.754 0.650 1.000
```

6. The correlation matrix suggests that perhaps just one of these predictors will do a good job of representing the other predictors. We could decide to pick *Ht*, the height of the driver, since that is the easiest predictor to measure, when compared to the others. Your choice might be different, and depending on the context, you may have a compelling reason to choose another predictor.

7. I chose to fit *hipcenter* with the predictors $x_1 = Age$, $x_2 = Weight$, and $x_4 = Ht$. The VIFs for this reduced model are all below 4, suggesting we do not have a huge issue with multicollinearity.

```
> vif(reduced)
     Age   Weight       Ht
1.093018 3.457681 3.463303
```

8. The null hypothesis for the general linear F test to drop the other predictors is H_0 : $\beta_3 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$.
   H_a : not all $\beta_3, \beta_5, \beta_6, \beta_7, \beta_8$ are zero.

   There are two equivalent approaches in carrying out the partial F test.

   Approach 1: Fit and compare the full and reduced models.

```
> reduced<-lm(hipcenter~Age+Weight+Ht)
> anova(reduced,result)
Analysis of Variance Table

Model 1: hipcenter ~ Age + Weight + Ht
Model 2: hipcenter ~ Age + Weight + HtShoes + Ht + Seated + Arm + Thigh +
    Leg
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     34 45262
2     29 41262  5    4000.3 0.5623 0.7279
```

   The F statistic is 0.5623, with p-value 0.7279. We do not reject the null hypothesis. Our data suggests we can drop the predictors $x_3 = \mathit{HtShoes}$, $x_5 = \mathit{Seated}$, $x_6 = \mathit{Arm}$, $x_7 = \mathit{Thigh}$, $x_8 = \mathit{Leg}$.

   Approach 2: Fit the full model, and list the predictors you want to drop last in the lm() function.

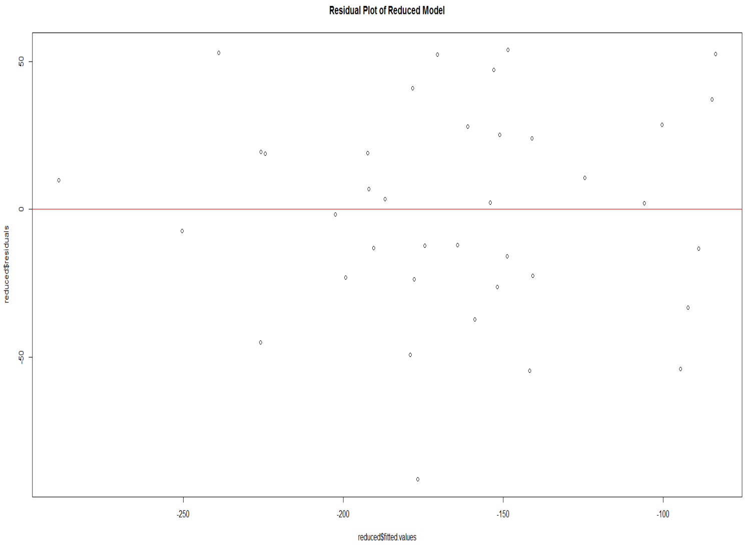
```
> result2<-lm(hipcenter~Age+Weight+Ht+HtShoes+Seated+Arm+Thigh+Leg)
> ##note the order of the predictors
> anova(result2)
Analysis of Variance Table

Response: hipcenter
          Df Sum Sq Mean Sq F value    Pr(>F)
Age        1   5541    5541  3.8947 0.0580359 .
Weight     1  57175   57175 40.1840  6.31e-07 ***
Ht         1  23661   23661 16.6296 0.0003236 ***
HtShoes    1     12      12  0.0087 0.9264796
Seated     1    538     538  0.3779 0.5435008
Arm        1    726     726  0.5105 0.4806345
Thigh      1     69      69  0.0485 0.8272673
Leg        1   2655    2655  1.8659 0.1824453
Residuals 29  41262    1423
```

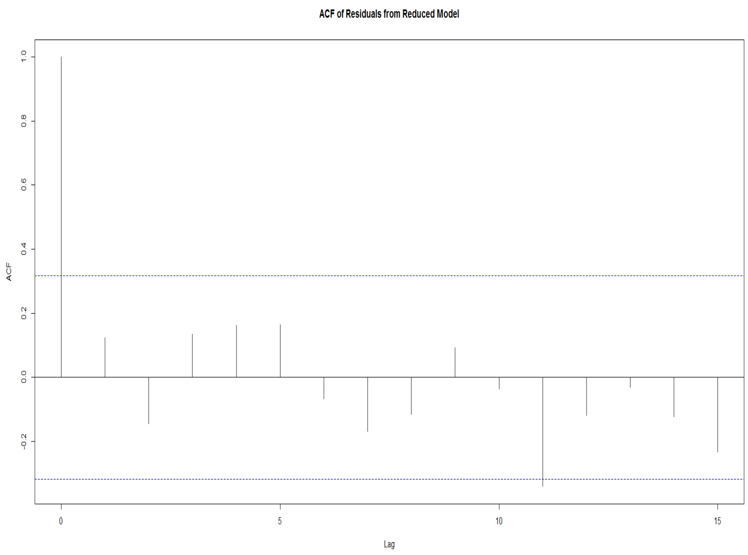
$$\begin{aligned} F &= \frac{\text{SSR}(\beta_3, \beta_5, \beta_6, \beta_7, \beta_8 | \beta_1, \beta_2, \beta_4)}{(8-3) \times \text{MSE}(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8)} \\ &= \frac{12 + 538 + 726 + 69 + 2655}{5 \times 1423} \\ &= 0.5621926 \end{aligned}$$

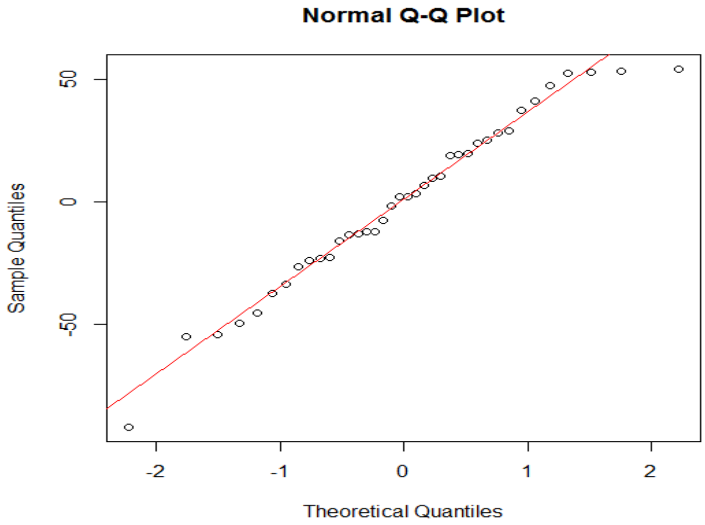
The corresponding p-value is $1 - pf(0.5621926, 5, 29) = 0.7280221$. Using the F table, the critical region is 2.545. So we fail to reject the null. Our data suggests we can drop the predictors $x_3 = \textit{HtShoes}$, $x_5 = \textit{Seated}$, $x_6 = \textit{Arm}$, $x_7 = \textit{Thigh}$, $x_8 = \textit{Leg}$.

Again, note that the F statistic in both approaches are the same (slight difference due to rounding off, theoretically they are the same).

9. Based on the residual plot, the assumptions for the multiple regression model appear to be satisfied. The residuals generally fall in a horizontal band around 0, have constant variance, and have no apparent curvature or pattern. There may be one residual that is fairly large in magnitude, but by and large, the assumptions are met. The ACF is slightly significant at lag 11, but this could be due to sampling variation (false positive), given that the data were likely not collected in a sequence and are likely to be uncorrelated.







10. The estimated regression equation is

$$\hat{hipcenter} = 528.297729 + 0.519504 Age + 0.004271\, Weight - 4.211905 Ht.$$

The R^2 for this model is 0.6562, which is only slightly less than the R^2 for the model with all predictors. The adjusted R^2 for this simplified model is 0.6258, which is higher than the adjusted R^2 for the full model, which is 0.6001. One thing to note is that adding predictors to a model never decreases the R^2, so the adjusted R^2 is a better way to compare models with different number of predictors.

```
> summary(reduced)

Call:
lm(formula = hipcenter ~ Age + Weight + Ht)
```

```
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 ***

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
```

11. Although the R^2 and standard error are very similar to the model with no measurement error, a number of the estimated coefficients are quite different, indicating their sensitivity to the accuracy in the measurement of the response variables. This sensitivity is another indication of multicollinearity.

```
> result.error<-lm(hipcenter+10*rnorm(38)˜.,seatpos)
> summary(result.error)

Call:
lm(formula = hipcenter + 10 * rnorm(38) ˜ ., data = seatpos)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 387.91411  163.82395   2.368   0.0248 *
Age           1.11390    0.56092   1.986   0.0566 .
Weight       -0.09135    0.32551  -0.281   0.7810
HtShoes      -2.89308    9.59215  -0.302   0.7651
Ht            0.63992    9.96278   0.064   0.9492
Seated        1.38912    3.69984   0.375   0.7101
Arm          -2.32291    3.83586  -0.606   0.5495
Thigh        -0.48836    2.61615  -0.187   0.8532
Leg          -6.10860    4.63610  -1.318   0.1979

Residual standard error: 37.1 on 29 degrees of freedom
Multiple R-squared:  0.7121,    Adjusted R-squared:  0.6327
F-statistic: 8.967 on 8 and 29 DF,  p-value: 4.199e-06
```