Guided Question Set 7 Solutions

1)

```
library(faraway) ##for seatpos dataset and vif function
library(tidyverse)
result<-lm(hipcenter~., data=seatpos) ##full model
summary(result)
```

```
##
## 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
```

The p-value for the ANOVA F test is very small. However, none of the individual t tests suggest any of the predictors is significant, given the other predictors.

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)

The pairwise correlations between all the variables are shown below

```
round(cor(seatpos),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
```

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

4)

The VIFs for this regression are shown below

```
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
```

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.

5)

The pairwise correlations between the 6 variables regarding length are shown below

```
round(cor(seatpos[,3:8]),3)
```

```
##         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
```

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

6)

The correlation matrix and VIFs suggest that just one of these predictors is linearly dependent on the other predictors. We could decide to pick `Ht`, the height of the driver, since it has the highest VIF, meaning it is has the strongest linear dependence on the rest of the predictors. From a practical standpoint, `Ht` can be chosen since it 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)

```
reduced<-lm(hipcenter~Age+Weight+Ht, data=seatpos)
vif(reduced)
```

```
##      Age   Weight       Ht
## 1.093018 3.457681 3.463303
```

For the model with only `Age`, `Weight`, and `Ht` as predictors, we can see that the VIFs have been drastically reduced. The VIFs for this reduced model are all below 4, suggesting we do not have a huge issue with multicollinearity.

8)

The null hypothesis for the partial 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.

The results of the partial F test are shown below

```
anova(reduced,result)
```

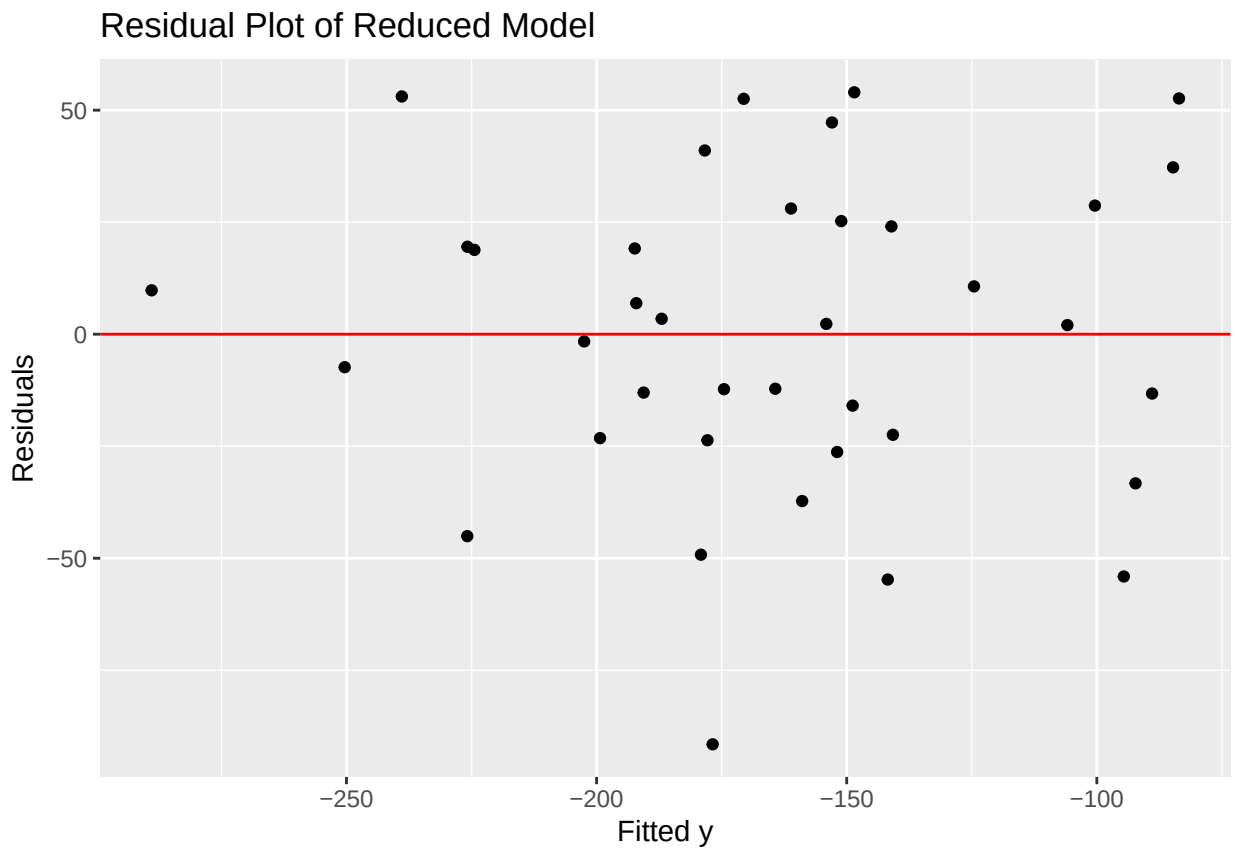
```
## Analysis of Variance Table
##
## Model 1: hipcenter ~ Age + Weight + Ht
## Model 2: hipcenter ~ Age + Weight + HtShoes + Ht + Seated + Arm + Thigh +
##     Leg
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     34 45262
## 2     29 41262  5    4000.3 0.5623 0.7279
```

The F statistic is 0.5623, with p-value 0.7279. We do not reject the null hypothesis. Our data suggest we can drop the predictors $x_3 =$ `HtShoes`, $x_5 =$ `Seated`, $x_6 =$ `Arm`, $x_7 =$ `Thigh`, $x_8 =$ `Leg` and go with the reduced model.

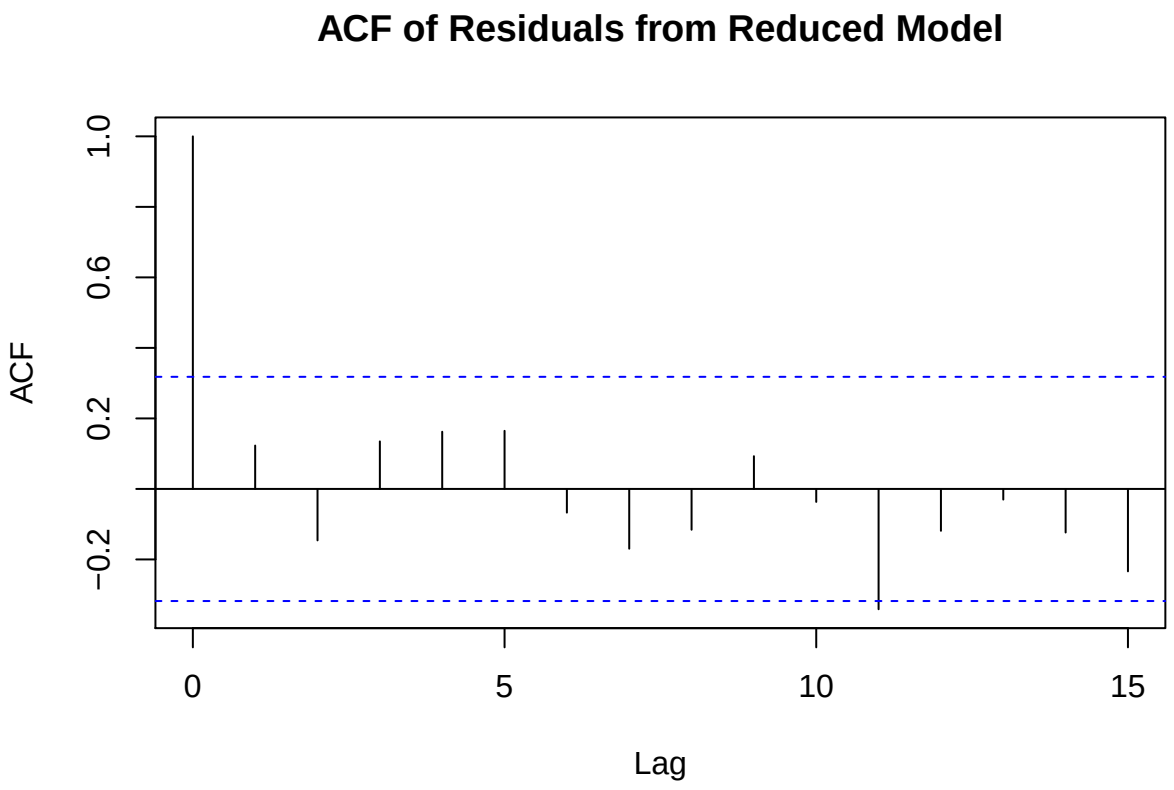
9)

The residual plot, ACF plot of residuals, and QQ plot of residuals are shown below

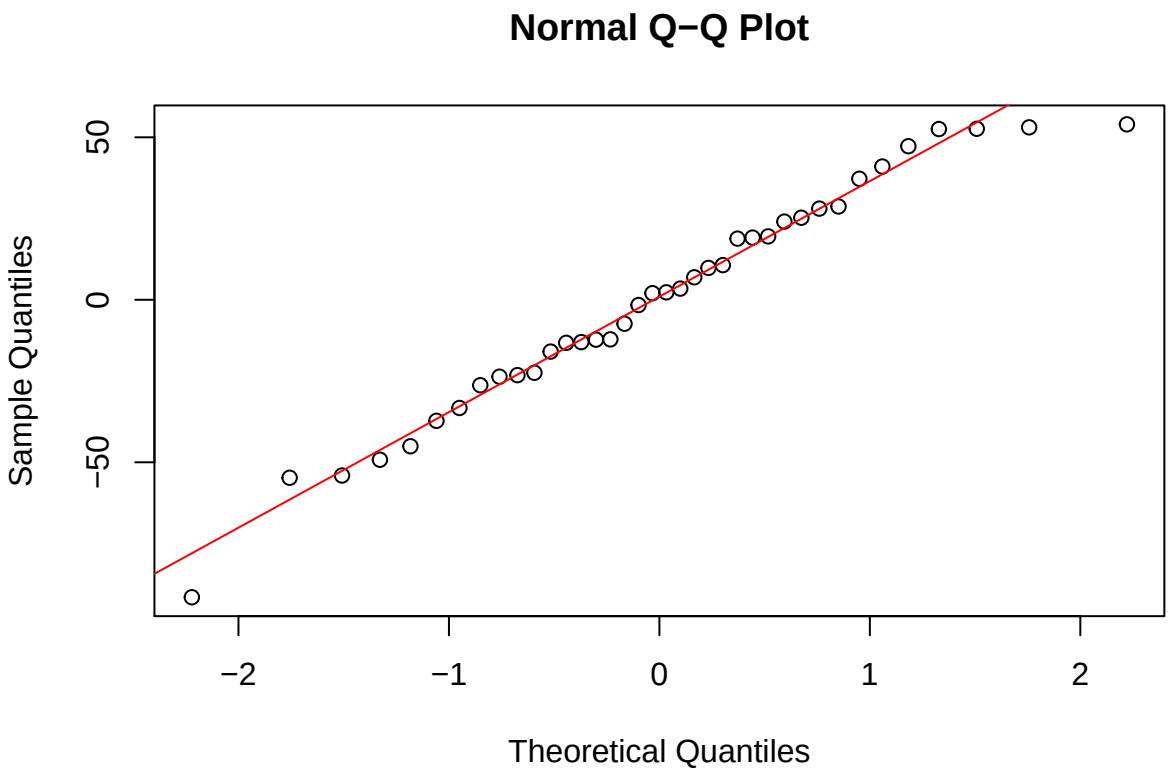
```
yhat<-reduced$fitted.values
res<-reduced$residuals

Data<-data.frame(seatpos, yhat, res)

##residual plot
ggplot(Data, aes(x=yhat,y=res))+
  geom_point()+
  geom_hline(yintercept=0, color="red")+
  labs(x="Fitted y", y="Residuals", title="Residual Plot of Reduced Model")
```




```
acf(res, main="ACF of Residuals from Reduced Model")
```




```
qqnorm(res)
qqline(res, col="red")
```



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. Since the plots on the QQ plot fall close to the diagonal line, the normality assumption of the errors are met.

10)

```
summary(reduced)
```

```
##
## Call:
## lm(formula = hipcenter ~ Age + Weight + Ht, data = seatpos)
##
## 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
```

The estimated regression equation for the reduced model is $\hat{y} = 528.300 + 0.520\text{Age} + 0.004\text{Weight} - 4.212\text{Ht}$.

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 reduced 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.