STAT5120: Homework 5

Yunlu Li

Problem 1

(a)

```
## Analysis of Variance Table
##
## Response: InfctRsk
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Stay        1  57.305  57.305 58.1676 1.044e-11 ***
## Cultures    1  33.397  33.397 33.8995 6.154e-08 ***
## Age         1   0.136   0.136  0.1376   0.71144
## Census      1   5.101   5.101  5.1781   0.02487 *
## Beds        1   0.028   0.028  0.0279   0.86759
## Residuals 107 105.413   0.985
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA table above, we have SSR($\beta_5|\beta_1$,β_2,β_3,β_4)=0.028.

(b)

The increment in the variability of InfctRsk that is explained by the predictors, by adding Beds to an existing set of four predictors (Stay, Cultures, Age, Census), is 0.028.

(c)

```
##
## Call:
## lm(formula = InfctRsk ~ Stay + Cultures + Age + Census + Beds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1658 -0.8085  0.1343  0.5928  2.4293
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.2051282  1.2075929   0.170   0.8654
## Stay        0.2055252  0.0660885   3.110   0.0024 **
## Cultures    0.0590369  0.0103096   5.726  9.5e-08 ***
## Age         0.0173637  0.0229966   0.755   0.4519
## Census      0.0010306  0.0034942   0.295   0.7686
## Beds        0.0004476  0.0026781   0.167   0.8676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9926 on 107 degrees of freedom
## Multiple R-squared:  0.4765, Adjusted R-squared:  0.4521
## F-statistic: 19.48 on 5 and 107 DF,  p-value: 9.424e-14
```

Age, Census, Beds appear to be not significant based on t-statistics above.

(d)

```
## Analysis of Variance Table
##
## Model 1: InfctRsk ~ Stay + Cultures
## Model 2: InfctRsk ~ Stay + Cultures + Age + Census + Beds
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    110 110.68
## 2    107 105.41  3    5.2644 1.7812 0.1551
```

$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$. H_a : at least one of $\beta_3, \beta_4, \beta_5$ is non zero. The F statistic is 1.7812 and the p-value is 0.1551, so we cannot reject the null hypothesis. This means Age, Census, Beds can be dropped from the model.

(e)

```
##
## Call:
## lm(formula = InfctRsk ~ Stay + Cultures)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1822 -0.7275  0.1040  0.6847  2.7143
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.805491   0.487756   1.651    0.102
## Stay        0.275472   0.052465   5.251 7.46e-07 ***
## Cultures    0.056451   0.009798   5.761 7.70e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.003 on 110 degrees of freedom
## Multiple R-squared:  0.4504, Adjusted R-squared:  0.4404
## F-statistic: 45.07 on 2 and 110 DF,  p-value: 5.04e-15
```

The estimated regression equation is $\hat{y} = 0.805491 + 0.275472x_1 + 0.056451x_2$.

Problem 2

(a)

```
## The following object is masked from data:
##
##     Age
```

```
##
## 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 associated with F statistic is very small. However, individual t-statistic indicates that none of predicators is significant given the presence of other predicators. Rˆ2 is 0.6866.

(b)

The samll p-value associated with F statistic suggests that the response is significantly linearly related to at least one of the predictors, but individual t-statistic indicates that none of predicators is significant given the presence of other predicators. This shows the sign of multicollinearity.

(c)

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

Some pairs of predicators show strong pairwise correlation.

(d)

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

HtShoes and Ht have very high VIF, indicating that there is serious multicollinearity.

(e)

```
##          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
```

The six preddicators are highly correlated to each other.

(f)

I would like to keep HtShoes, since it is most highly correlated to other predicators.

(g)

```
##      Age   Weight  HtShoes
## 1.080473 3.418028 3.417264
```

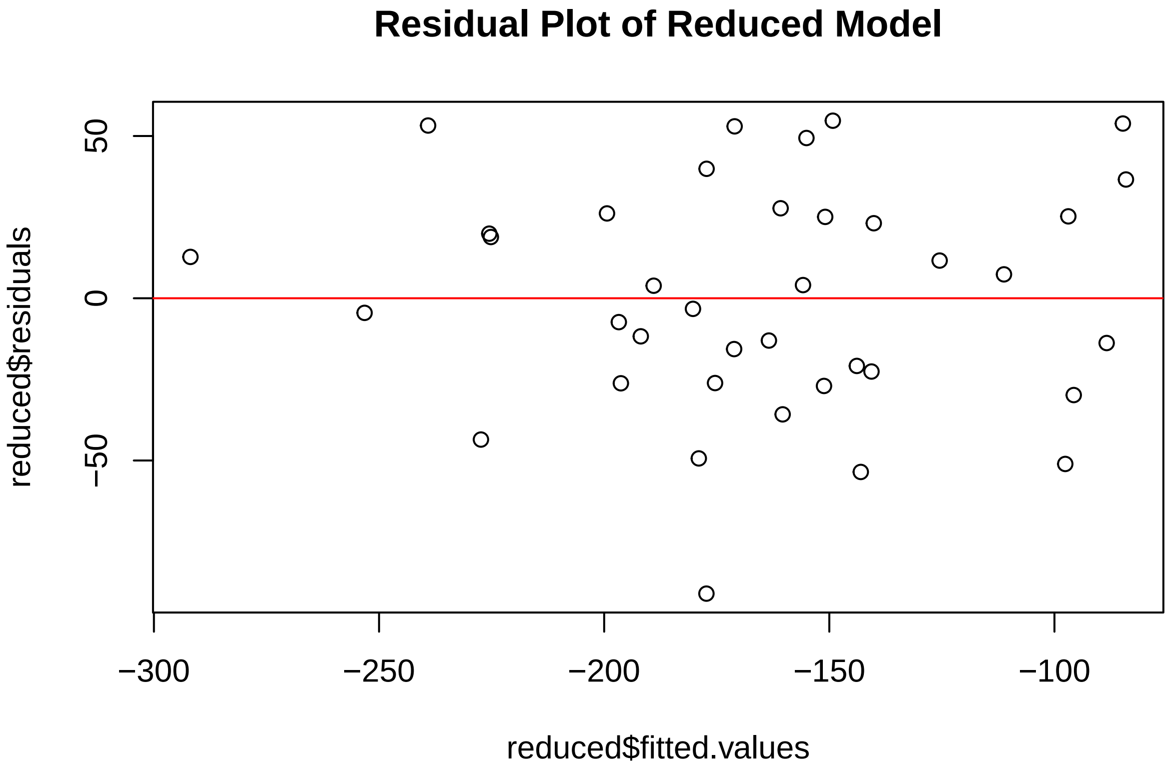
Since the VIFs are less than 5, so multicollinearity is not a concern.

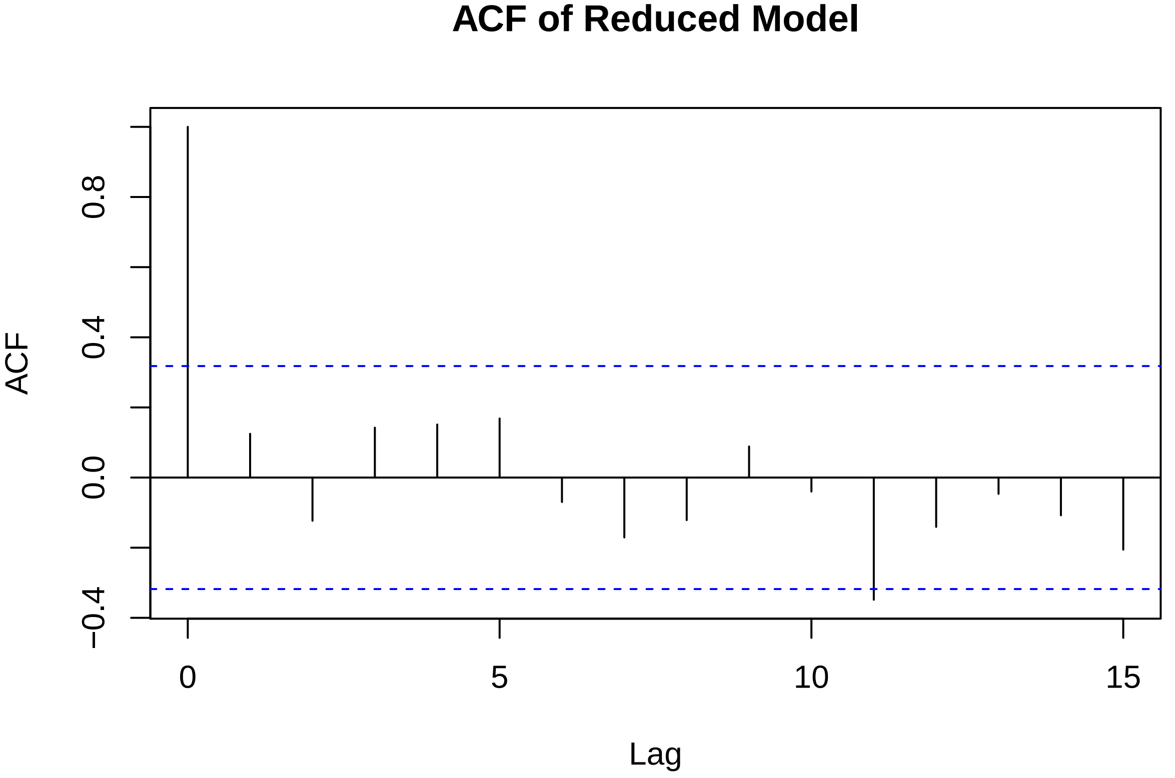
(h)

```
## Analysis of Variance Table
##
## Model 1: hipcenter ~ Age + Weight + HtShoes
## Model 2: hipcenter ~ Age + Weight + HtShoes + Ht + Seated + Arm + Thigh +
##     Leg
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     34 45433
## 2     29 41262  5    4171.2 0.5863 0.7103
```

$H_0 : \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$. H_a : at least one of $\beta_4, \beta_5, \beta_6, \beta_7, \beta_8$ is non zero. The F statistic is 0.5863 and the p-value is 0.7103, so we cannot reject the null hypothesis. This means Ht, Seated, Arm, Thigh, and Leg can be dropped from the model.

(i)





From the residual plot, the assumptions for the multiple regression model are satisfied. The residuals fall in a horizontal band around 0 with constant variance, and have no apparent pattern. The ACF plot indicates the residuals are uncorrelated.

(j)

```
##
## Call:
## lm(formula = hipcenter ~ Age + Weight + HtShoes)
##
## Coefficients:
## (Intercept)          Age       Weight      HtShoes
##  532.877125     0.557597    -0.008688    -4.178042
```

$\hat{y} = 532.877125 + 0.557597x_1 - 0.008688x_2 - 4.178042x_3$