Stat 6021: Guided Question Set 7

Tom Lever

10/16/22

Car drivers like to adjust the seat position for their own comfort. Car designers find it helpful to know where different drivers will position the seat. Researchers at HuMoSim laboratory at the University of Michigan collected data on 38 drivers. The response variable is *hipcenter*, the horizontal distance of the midpoint of the hips from a fixed location in the car in *mm*. They measured the following eight predictors:

- x_1: Age: Age in years
- x_2: Weight: Weight in pounds
- x_3: HtShoes: Height with shoes in *cm*.
- x_4: Ht: Height without shoes in *cm*.
- x_5: Seated: Seated height in *cm*.
- x_6: Arm: Arm length in *cm*.
- x_7: Thigh: Thigh length in *cm*.
- x_8: Leg: Lower leg length in *cm*.

The data are from the `faraway` package in R. After installing the faraway package, load the `seatpos` data set.

```
library(faraway)
head(seatpos, n = 3)
```

```
##   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
```

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.

   ```
   library(TomLeversRPackage)
   linear_model <- lm(hipcenter ~ ., data = seatpos)
   summarize_linear_model(linear_model)
   ```

   ```
   ##
   ## 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
##
## E(y | x) =
##     B_0 +
##     B_Age * Age +
##     B_Weight * Weight +
##     B_HtShoes * HtShoes +
##     B_Ht * Ht +
##     B_Seated * Seated +
##     B_Arm * Arm +
##     B_Thigh * Thigh +
##     B_Leg * Leg
## E(y | x) =
##     436.43212822533 +
##     0.775716195411176 * Age +
##     0.0263130815934825 * Weight +
##     -2.69240773927674 * HtShoes +
##     0.601344580352112 * Ht +
##     0.533751697568726 * Seated +
##     -1.32806863757197 * Arm +
##     -1.14311887823954 * Thigh +
##     -6.43904626562725 * Leg
## Number of observations: 38
## Estimated variance of errors: 1422.82012070282
## Multiple R:  0.828585225565444   Adjusted R:  0.774651837544867
## Critical value t(alpha/2 = 0.05/2, DFRes = 29): 2.0452296421327
## Critical value F(alpha = 0.05, DFR = 8, DFRes = 29): 2.27825084905155
```

Since the above F statistic is greater than the above critical value F / the above p-value for an ANOVA F test is less than a significance level $\alpha = 0.05$, we reject a null hypothesis that all regression coefficients in our linear model are 0 / insignificant. We have sufficient evidence to support an alternate hypothesis that at least one regression coefficient in our linear model is not 0 / significant.

Since the magnitudes of the above t statistics for all predictors are less than the above critical value t, we have insufficient evidence to reject a null hypothesis that each regression coefficient is 0 / insignificant in the context of the multiple linear model and the other predictors. We have insufficient evidence to support an alternate hypothesis that each regression coefficient is not 0 / significant in the context of the multiple linear model and the other predictors.

2. Briefly explain why, based on your output from part 1, you suspect the model shows signs of multicollinearity.

   Per section 9.4.4: Multicollinearity: Multicollinearity Diagnostics: Other Diagnostics in *Introduction to Linear Regression Analysis* (Sixth Edition) by Douglas C. Montgomery et al., "if the overall F statistic is significant but the individual t statistics are all nonsignificant, multicollinearity is present".

Multicollinearity is present.

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

```
library(dplyr)
data_frame_of_predictor_values <- seatpos %>% select(-hipcenter)
correlation_matrix <- round(cor(data_frame_of_predictor_values), 3)
correlation_matrix
```

```
##             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
```

```
analyze_correlation_matrix(correlation_matrix)
```

```
## Age
##     V:  Age
##     H:
##     M:
##     L:  Arm
##     N:  Weight, HtShoes, Ht, Seated, Thigh, Leg
## Weight
##     V:  Weight
##     H:  HtShoes, Ht, Seated, Leg
##     M:  Arm, Thigh
##     L:
##     N:  Age
## HtShoes
##     V:  HtShoes, Ht, Seated, Leg
##     H:  Weight, Arm, Thigh
##     M:
##     L:
##     N:  Age
## Ht
##     V:  HtShoes, Ht, Seated, Leg
##     H:  Weight, Arm, Thigh
##     M:
##     L:
##     N:  Age
## Seated
##     V:  HtShoes, Ht, Seated
##     H:  Weight, Leg
##     M:  Arm, Thigh
##     L:
##     N:  Age
## Arm
##     V:  Arm
##     H:  HtShoes, Ht, Leg
##     M:  Weight, Seated, Thigh
```

```
##     L:  Age
##     N:
## Thigh
##     V:  Thigh
##     H:  HtShoes, Ht
##     M:  Weight, Seated, Arm, Leg
##     L:
##     N:  Age
## Leg
##     V:  HtShoes, Ht, Leg
##     H:  Weight, Seated, Arm
##     M:  Thigh
##     L:
##     N:  Age
```

4. Check the Variance Inflation Factors (VIF's). What do these values indicate about multicollinearity?

```
vif(linear_model)
```

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

Let $\boldsymbol{X}$ be the $n \times p$ model matrix. Let C_{jj} be the element of $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ in row j and column j. Let ${R_j}^2$ be the coefficient of multiple determination obtained from regressing x_j on the other predictors / from the model

$$\hat{x}_j = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_{j-1} x_{j-1} + \hat{\beta}_{j+1} x_{j+1} + ... + \hat{\beta}_k x_k$$

Let $Var\left(\hat{\beta}_j\right)$ be the variance of the estimated regression coefficient $\hat{\beta}_j$ of predictor x_j. Let σ^2 be the variance of the linear model for the population of observations. The Variance Inflation Factor for a predictor x_j

$$VIF_j = C_{jj} = \frac{1}{1 - {R_j}^2} = \frac{Var\left(\hat{\beta}_j\right)}{\sigma^2}$$

is the factor by which the variance of the estimated regression coefficient of that predictor increases due to near-linear dependencies among predictors. The Variance Inflation Factor for predictor x_j measures the combined effect of linear dependencies among predictors on the variance of predictor x_j. Since the Variance Inflation Factors for $HtShoes$ and Ht are far greater than 4, 5, or even 10, our linear model is seriously multicollinear. The estimated regression coefficients for $HtShoes$ and Ht are poorly estimated because of multicollinearity. Multicollinearity tends to produce estimated regression coefficients that are too large in magnitude. We are unable to interpret the relationships between predictors, and may be unable to use our model for prediction.

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.

```
data_frame_of_predictor_values <-
    seatpos %>% select(HtShoes, Ht, Seated, Arm, Thigh, Leg)
correlation_matrix <- round(cor(data_frame_of_predictor_values), 3)
correlation_matrix
```

```
##         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
```

```
analyze_correlation_matrix(correlation_matrix)
```

```
## HtShoes
##     V:  HtShoes, Ht, Seated, Leg
##     H:  Arm, Thigh
##     M:
##     L:
##     N:
## Ht
##     V:  HtShoes, Ht, Seated, Leg
##     H:  Arm, Thigh
##     M:
##     L:
##     N:
## Seated
##     V:  HtShoes, Ht, Seated
##     H:  Leg
##     M:  Arm, Thigh
##     L:
##     N:
## Arm
##     V:  Arm
##     H:  HtShoes, Ht, Leg
##     M:  Seated, Thigh
##     L:
##     N:
## Thigh
##     V:  Thigh
##     H:  HtShoes, Ht
##     M:  Seated, Arm, Leg
##     L:
##     N:
## Leg
##     V:  HtShoes, Ht, Leg
##     H:  Seated, Arm
##     M:  Thigh
##     L:
##     N:
```

6. Since all 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 will use $HtShoes$, as $HtShoes$ may be thought of as an object with the other variables as components.

7. Based on your choice in part 6, fit a multiple linear regression model with your choice of predictor to keep, along with the predictors $x_1 = Age$ and $x_2 = Weight$. Check the Variable Inflation Factors for this model. Comment on whether we still have an issue with multicollinearity.

```
library(TomLeversRPackage)
reduced_model <- lm(hipcenter ~ HtShoes + Age + Weight, data = seatpos)
summarize_linear_model(reduced_model)
```

```
##
## Call:
## lm(formula = hipcenter ~ HtShoes + Age + Weight, data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.029 -25.249   0.294  25.200  54.717
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 532.877125 137.104715   3.887 0.000448 ***
## HtShoes      -4.178042   0.996501  -4.193 0.000186 ***
## Age           0.557597   0.406456   1.372 0.179094
## Weight       -0.008688   0.310512  -0.028 0.977842
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.55 on 34 degrees of freedom
## Multiple R-squared:  0.6549, Adjusted R-squared:  0.6244
## F-statistic:  21.5 on 3 and 34 DF,  p-value: 5.46e-08
##
## E(y | x) =
##     B_0 +
##     B_HtShoes * HtShoes +
##     B_Age * Age +
##     B_Weight * Weight
## E(y | x) =
##     532.877124547553 +
##     -4.17804245194146 * HtShoes +
##     0.55759700862984 * Age +
##     -0.00868800192199593 * Weight
## Number of observations: 38
## Estimated variance of errors: 1336.26480994201
## Multiple R:  0.809238322705622   Adjusted R:  0.790198532920499
## Critical value t(alpha/2 = 0.05/2, DFRes = 34): 2.03224450931772
## Critical value F(alpha = 0.05, DFR = 3, DFRes = 34): 2.88260420426123
```

```
vif(reduced_model)
```

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

Since the Variance Inflation Factors for all predictors in the reduced model are less than 4 and are drastically reduced relative to the VIF's of the full model, the reduced model is less multicollinear than the full model. Since the VIF's are non-zero, we still have an issue with multicollinearity. The estimated regression coefficients for the predictors in the reduced model are well estimated relative to the full model.

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.

```
full_model <- lm(
    hipcenter ~ HtShoes + Age + Weight + Ht + Seated + Arm + Thigh + Leg,
    data = seatpos
)
```

```
analyze_variance_for_reduced_and_full_linear_models(reduced_model, linear_model)
```

```
## Analysis of Variance Table
##
## Model 1: hipcenter ~ HtShoes + Age + Weight
## 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
##
## test statistic F0 for Partial F Test: 0.586331325647288
## Fc(alpha = 0.05, predictors_dropped = 5, DFRes(full) = 29) = 2.54538648794854
## P(F > F0) for Partial F Test: 0.710270048806473
## significance level: 0.05
```

The test statistic for a partial F test $F_0 = 0.5863$. Since the test statistic is less than a critical value $F_c = 2.545$, we have insufficient evidence to reject a null hypothesis that the regression coefficients for the dropped predictors are 0. We have insufficient evidence to support an alternate hypothesis that a regression coefficient for a dropped predictor is not 0. The dropped predictors are jointly insignificant. The p-value for a partial F test $P(F > F_0) = 0.710$. Since this p-value is greater than a significance level $\alpha = 0.05$, we have insufficient evidence to reject a null hypothesis that the regression coefficients for the dropped predictors are 0. We have insufficient evidence to support an alternate hypothesis that a regression coefficient for a dropped predictor is not 0. The dropped predictors are jointly insignificant.

```
analyze_variance_for_one_linear_model(full_model)
```

```
## Analysis of Variance Table
##
## Response: hipcenter
##           Df Sum Sq Mean Sq F value    Pr(>F)
## HtShoes    1  83534   83534 58.7099 1.897e-08 ***
## Age        1   2671    2671  1.8775    0.1811
## Weight     1      1       1  0.0007    0.9786
## Ht         1    183     183  0.1288    0.7223
## Seated     1    538     538  0.3779    0.5435
## Arm        1    726     726  0.5105    0.4806
## Thigh      1     69      69  0.0485    0.8273
## Leg        1   2655    2655  1.8659    0.1824
## Residuals 29  41262    1423
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## DFR: 8, SSR: 90377.2054319607, MSR: 11297.1506789951
## F0: 7.93997112819486, F(alpha = 0.05, DFR = 8, DFRes = 29): 2.27825084905155
## p: 1.3057731397195e-05
## DFT: 37, SST: 131638.988932342
## R2: 0.686553476025338, Adjusted R2: 0.600085469411638
## Number of observations: 38
```

The regression sum of squares for the dropped predictors given that the kept predictors were already in the model and the sum of regression sum of squares for predictors dropped from the full model given that previous predictors were already in the model

$$SS_R\left(\boldsymbol{x}_d|\boldsymbol{x}_k\right) = SS_R^{Ht} + SS_R^{Seated} + SS_R^{Arm} + SS_R^{Thigh} + SS_R^{Leg} = 183 + 538 + 726 + 69 + 2655 = 4171$$

The number of predictors dropped $d = 5$. The regression mean square for the dropped predictors given

that the kept predictors were already in the model

$$MS_R\left(\boldsymbol{x}_d|\boldsymbol{x}_k\right) = \frac{SS_R\left(\boldsymbol{x}_d|\boldsymbol{x}_k\right)}{d} = 834.2$$

The residual mean square $MS_{Res} = 1423$. The test statistic F for the Partial F Test

$$F_0 = \frac{MS_R\left(\boldsymbol{x}_d|\boldsymbol{x}_k\right)}{MS_{Res}} = \frac{834.2}{1423} = 0.586$$

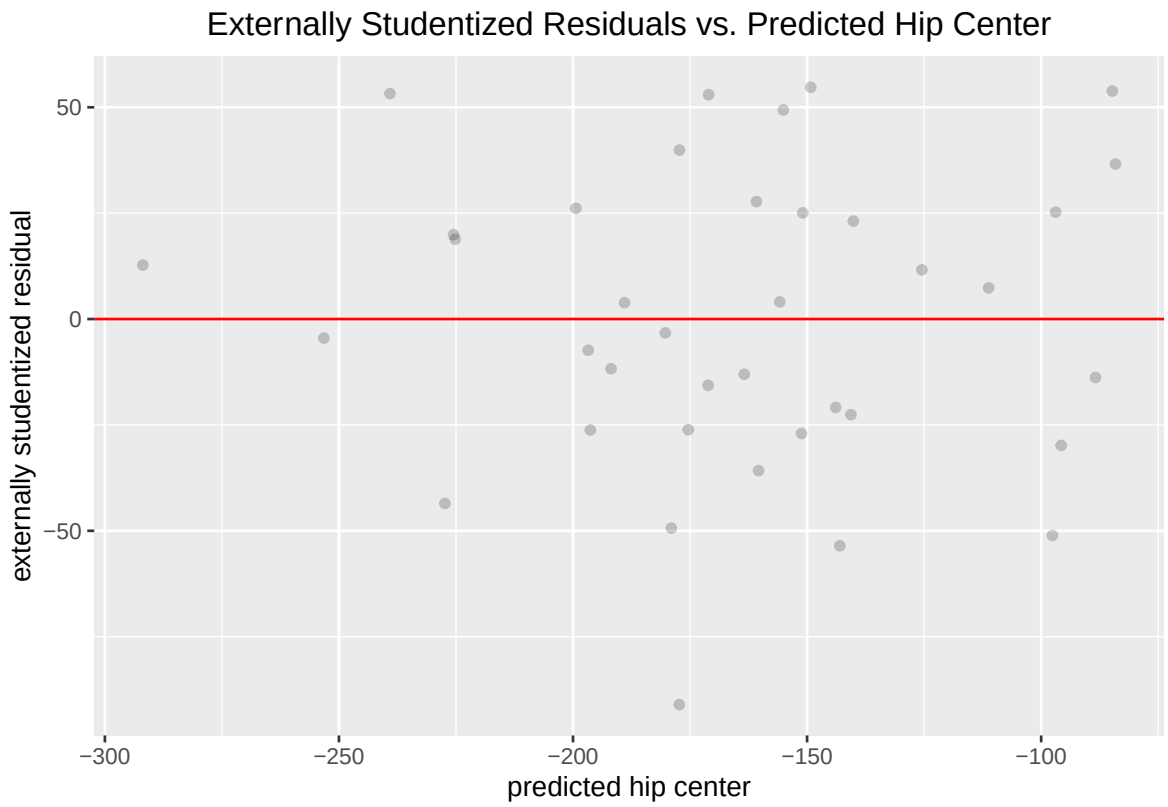
The corresponding p-value

```
test_statistic_F0_for_Partial_F_Test <- 0.586
number_of_predictors_dropped <- 5
number_of_observations <- 38
number_of_variables <- 9
residual_degrees_of_freedom <- number_of_observations - number_of_variables
calculate_p_value_from_F_statistic_and_regression_and_residual_degrees_of_freedom(
    test_statistic_F0_for_Partial_F_Test,
    number_of_predictors_dropped,
    residual_degrees_of_freedom
)
```

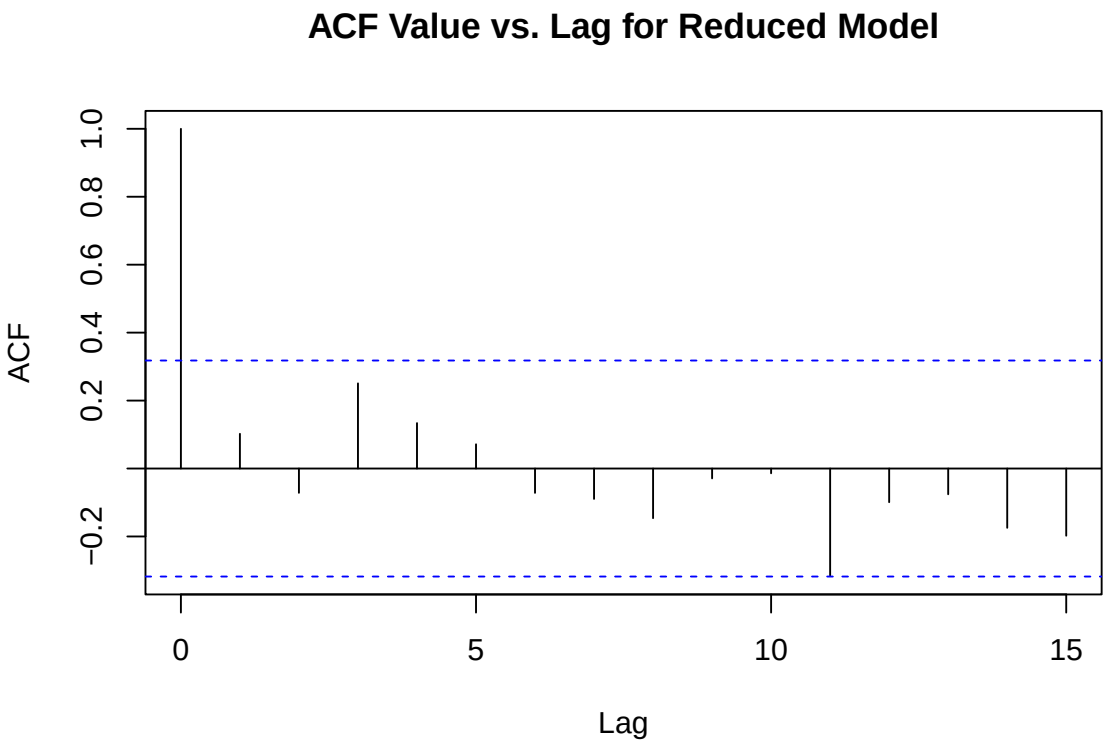
```
## [1] 0.7105135
```

9. 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 linear regression model. Also produce an ACF plot and a QQ plot of the residuals, and comment on the plots.

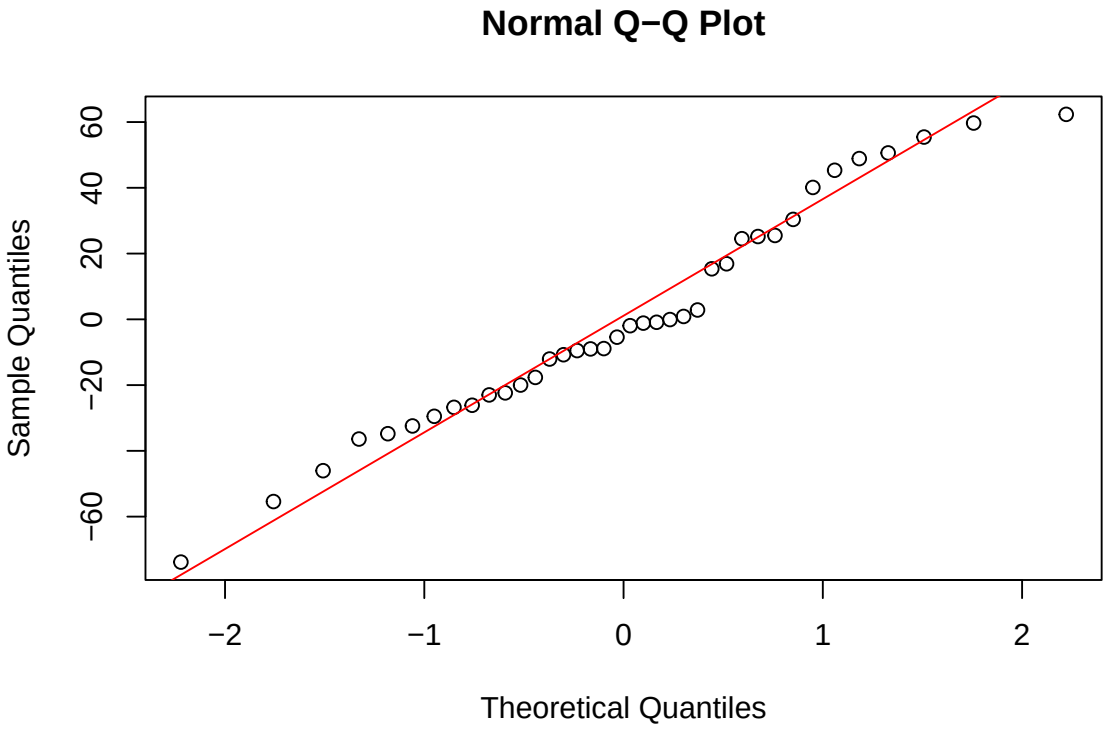
```
library(ggplot2)
ggplot(
    data.frame(
        externally_studentized_residual = reduced_model$residuals,
        predicted_hip_center = reduced_model$fitted.values
    ),
    aes(x = predicted_hip_center, y = externally_studentized_residual)
) +
    geom_point(alpha = 0.2) +
    geom_hline(yintercept = 0, color = "red") +
    labs(
        x = "predicted hip center",
        y = "externally studentized residual",
        title = "Externally Studentized Residuals vs. Predicted Hip Center"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
```




```
acf(linear_model$residuals, main = "ACF Value vs. Lag for Reduced Model")
```



```
qqnorm(linear_model$residuals)
qqline(linear_model$residuals, col = "red")
```



1. The assumption that the relationship between response / hip center and predictors is linear, at least approximately, is met cannot be addressed.
2. The assumption that the residuals of the linear model of hip center versus predictors have mean 0 is met. Residuals are evenly scattered around $e = 0$ at random.
3. The assumption that the distributions of residuals of the linear model for different predictors have constant variance is met. Residuals are evenly scattered around $e = 0$ with constant vertical variance.
4. The assumption that the residuals of the linear model are uncorrelated is met. The ACF value for lag 0 is always 1; the correlation of the vector of residuals with itself is always 1. Since all ACF value are insignificant (the ACF value for lag 11 barely), we have insufficient evidence to reject a null hypothesis that the residuals of the linear model are uncorrelated. We have insufficient evidence to conclude that the residuals of the linear model are correlated. We have insufficient evidence to conclude that the assumption that the residuals of the linear model are uncorrelated is not met.
5. The assumption that the residuals of the linear model are normally distributed is met. A linear model is robust to these assumptions. Considering a plot of sample quantiles versus theoretical quantiles for the residuals of the linear model, since observations lie near the line of best fit / their theoretical values, a probability vs. externally studentized residuals plot / distribution is normal.

10. Based on your results, write your estimated regression equation from part 7. Also report the R^2 of this model, and compare with the R^2 you reported in part 1, for the model with all predictors. Also comment on the adjusted R^2 for both models.

    The R^2 of the linear model from part 1 is 0.6866. The adjusted R^2 for the linear model from part 1 is 0.6001. This coefficient of determination is the proportion of variation in hip center that is explained

by the multiple linear relationship / predictors. The adjusted R^2 penalizes us for adding terms to the model that are not helpful. R^2 tends to optimistically estimate the fit of the linear regression. It always increases as the number of predictors in the model increases. Adjusted R^2 attempts to correct for this overestimation. Adjusted R^2 might decrease if a specific predictor does not improve the model. Adjusted R^2 is always less than or equal to R^2. A value of 1 indicates a model that perfectly predicts values in the target field. A value that is less than or equal to 0.09 indicates a model that has no predictive value. In the real world, adjusted R^2 lies between 0 and 1.

See part 7 for the estimated regression equation of the reduced model. The R^2 of the reduced model is 0.6549. The adjusted R^2 of the reduced model is 0.6244. The R^2 of the reduced model is less than the R^2 of the linear model from part 1, though the R^2 of the reduced model is closer to the adjusted R^2 from both the linear model from part 1 and the reduced model. The adjusted R^2 for the reduced model is greater than the adjusted R^2 for the linear model from part 1.