HW7

Dima Mikhaylow

10/20/2021

Predictors from the dataset:

• 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

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.

```
result <- lm(hipcenter ~ ., data=seatpos)
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
```

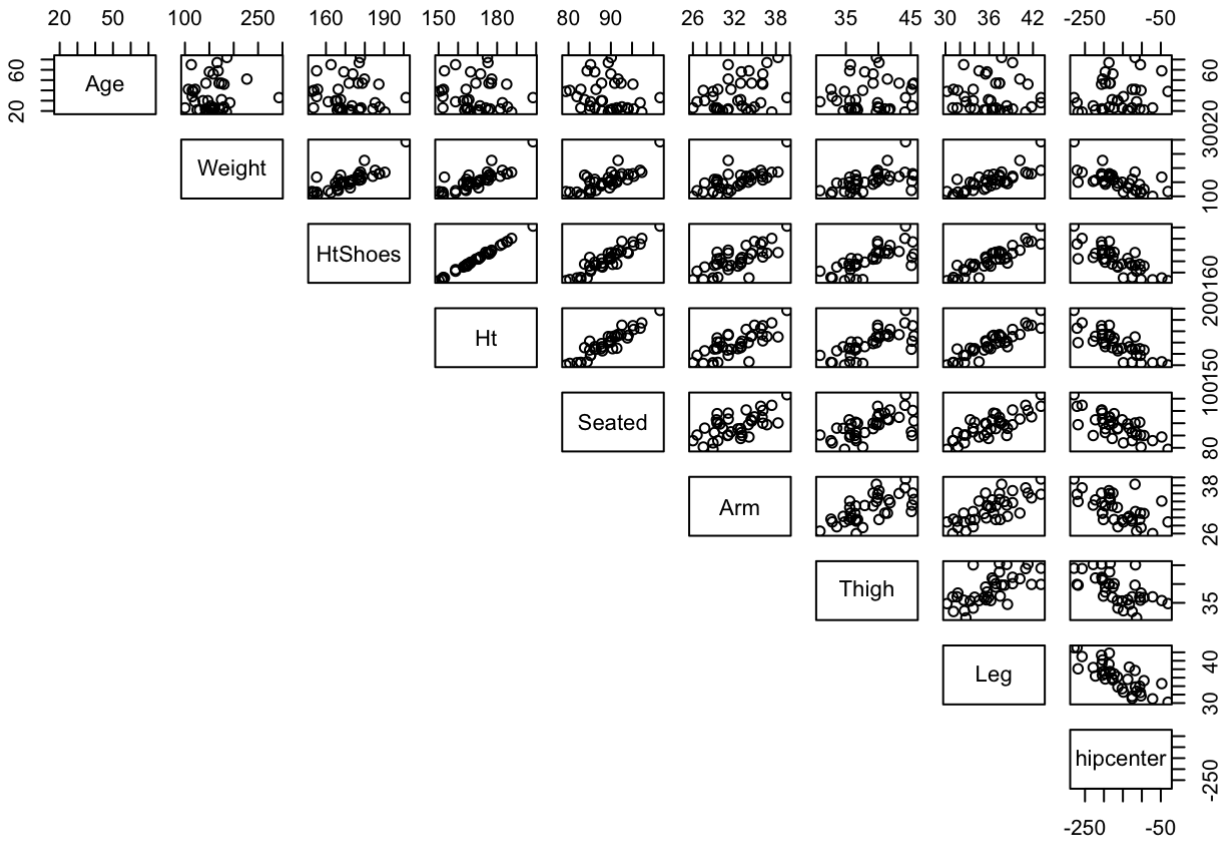
All slope coefficients are not significant, but F test appears highly significant with very small p-value of 1.3e-0.5.

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

There is a possibility of multicollinearity becuase all individual predictor variables came out as statistically insignificant, although the overall F-statistic is highly significant.

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

```
pairs(seatpos, lower.panel=NULL)
```



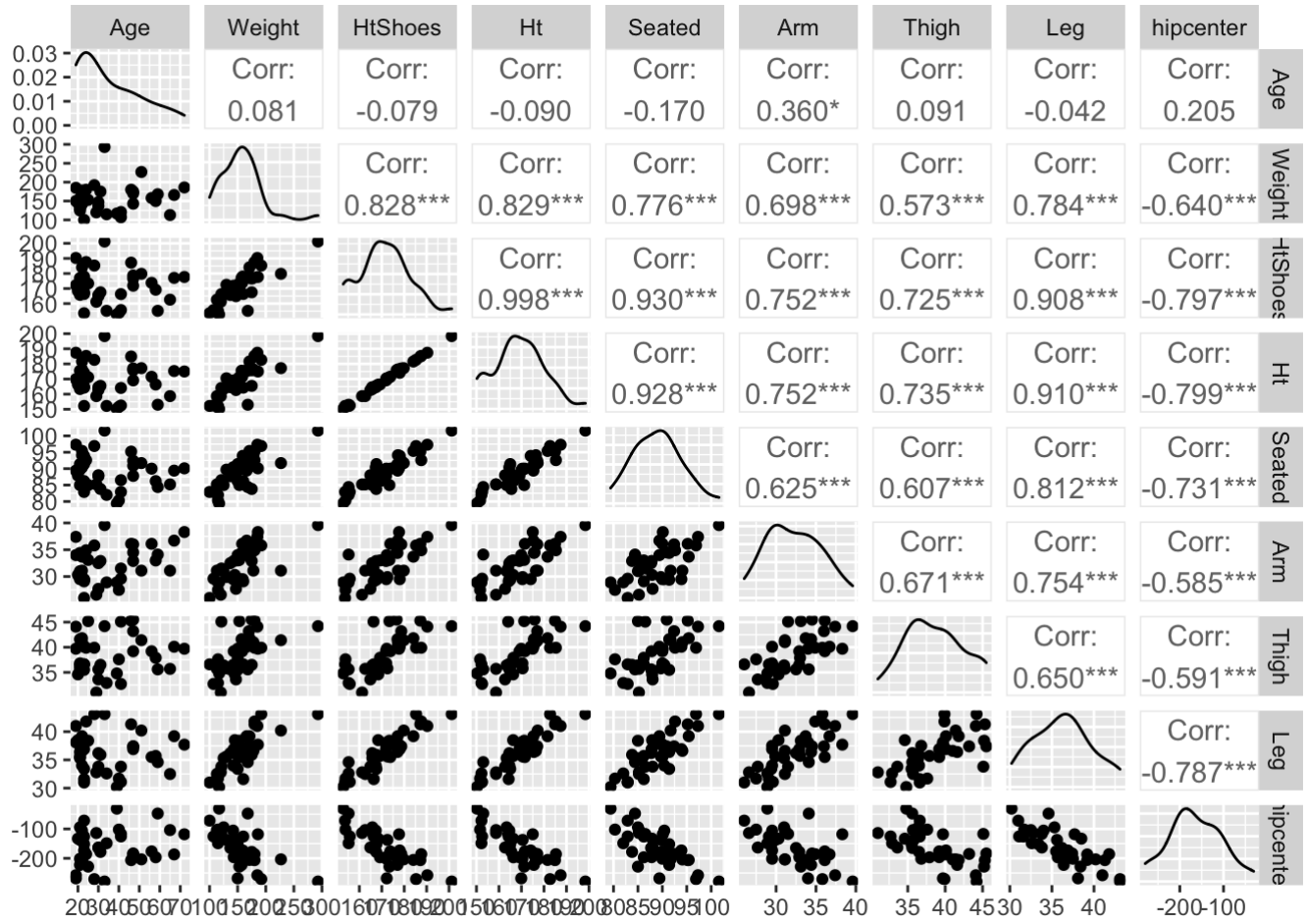

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:faraway':
##
##     happy
```

```
ggpairs(seatpos)
```



Obviously, most predictors are highly correlated to each other. Specifically, "HtShoes", "Ht", "Seated", and "Leg" are very strongly (above absolute 70%) correlated with the dependent variable "hipcenter" as well as with each other.

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

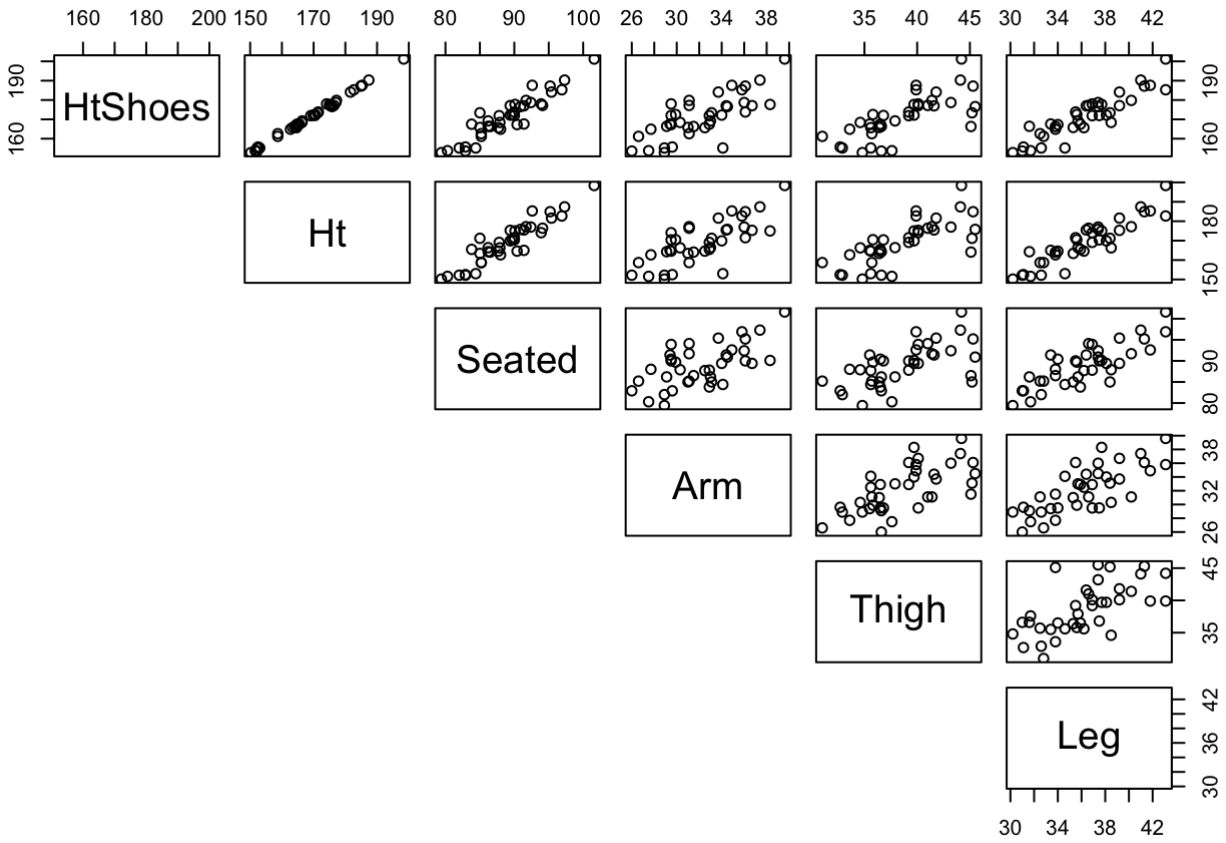
```
vif(result)
```

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

VIF suggests multicollinearity is present as there are several factors greater than 10, for example "HtShoes" and "Ht" appear exceptionally high. Others, "Seated" and "Leg" have large values as well.

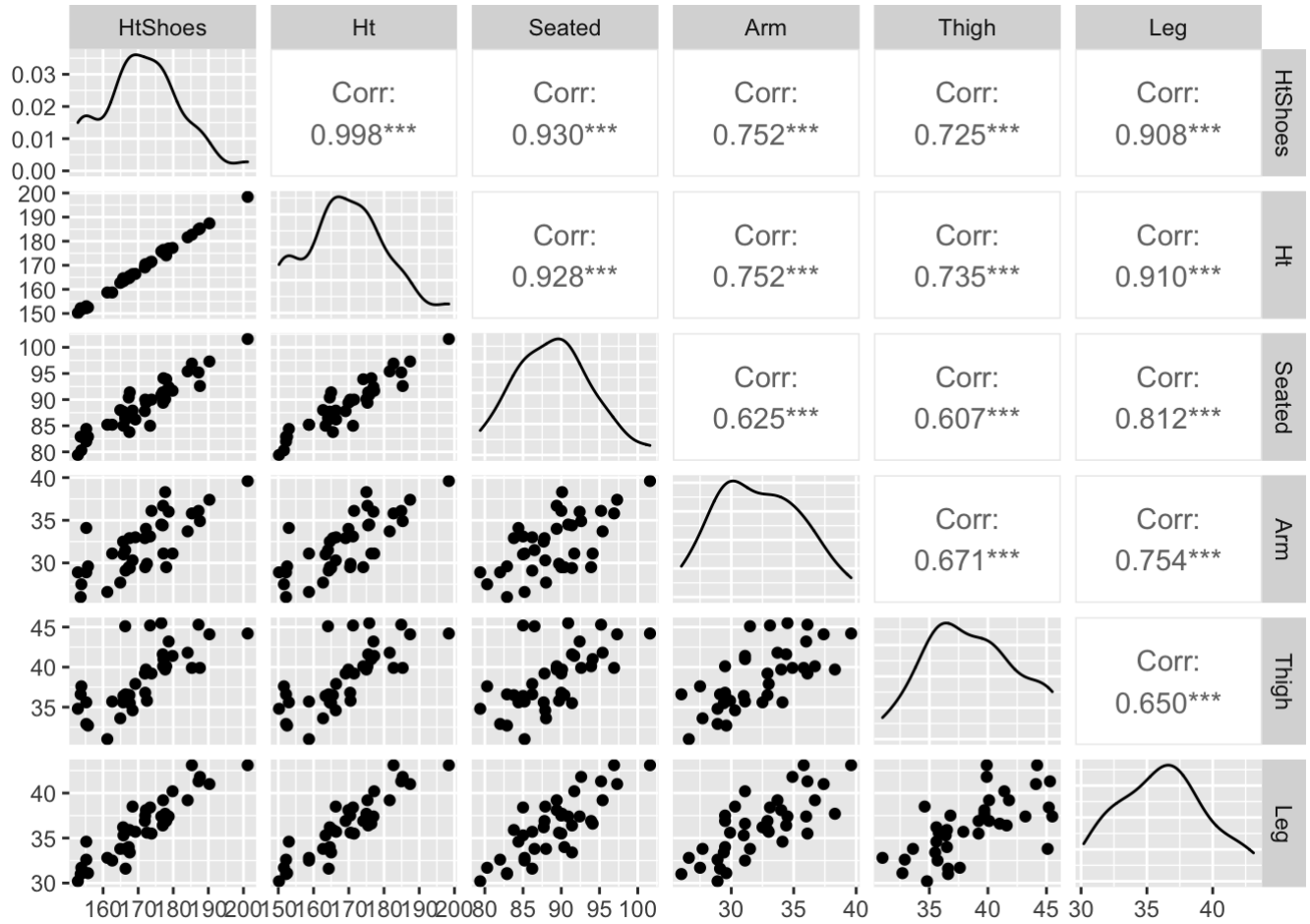
5. Looking at the data, we may want to look at the correlations for the variables that describe length of body parts: HtShoes, Ht, Seated, Arm, Thigh, and Leg. Comment on the correlations of these six predictors.

```
selected_vars = seatpos[, c("HtShoes", "Ht", "Seated", "Arm", "Thigh", "Leg")]
pairs(selected_vars, lower.panel=NULL)
```



All selected variables appear to be very correlated.

```
ggpairs(selected_vars)
```



Better visual above shows stromg correlations between body parts. Below is correlation matrix with exact values:

```
cor(selected_vars)
```

```
##           HtShoes        Ht    Seated       Arm     Thigh       Leg
## HtShoes 1.0000000 0.9981475 0.9296751 0.7519530 0.7248622 0.9084334
## Ht      0.9981475 1.0000000 0.9282281 0.7521416 0.7349604 0.9097524
## Seated  0.9296751 0.9282281 1.0000000 0.6251964 0.6070907 0.8119143
## Arm     0.7519530 0.7521416 0.6251964 1.0000000 0.6710985 0.7538140
## Thigh   0.7248622 0.7349604 0.6070907 0.6710985 1.0000000 0.6495412
## Leg     0.9084334 0.9097524 0.8119143 0.7538140 0.6495412 1.0000000
```

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.

It seems that “Ht” is a better single predictor as it appears to have the strongest correlation with all other variables at the same time, at the same time “Thigh” has smaller VIF and better business case for predicting seat position.

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.

```
subset_result <- lm(hipcenter ~ Age + Weight + Thigh, data=seatpos)
summary(subset_result)
```

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

All the slopes are significant, the overall F test is also significant.

```
vif(subset_result)
```

```
##      Age   Weight    Thigh
## 1.009626 1.489650 1.492388
```

All the values are less than 10 – not sure how to interpret (?)

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.

H0: B3 = B4 = B5 = B6 = B8 = 0 and Ha: B3, B4, B5, B6, B8 not 0.

In words, null hypothesis is going with the reduced subset model, and alternative supports the full model not dropping any of the predictors.

First approach - compare two F-statistics from the ANOVA. Below are the outputs of ANOVA for the reduced, subset model and original, full model.

```
anova(subset_result, result)
```

```
## 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-statistic is not too small abd p-value is high, greater than 0.05 – two-sided shoudl be used here, question (?). Not enought evidence to reject H0 – question (?)

```
# change in residuals over number of variables to drop:
F_stat <- ((57963-41262) / 5) / (41262 / 29)
F_stat
```

```
## [1] 2.347579
```

Second approach - sequantial SSR: use ANOVA on the full test model and look for the predictors to drop being listed last

```
# Other possibilities were considered here, for example "Leg"..., the problem seems to c
ome from "Age"
test_model <- lm(hipcenter ~ Age + Weight + Thigh + ., data=seatpos)
anova(test_model)
```

```
## 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 statistic is the same as above from the first approach:
F0_stat = ((12900+54+419+674+2655)/5)/(41262/29)
F0_stat
```

```
## [1] 2.347719
```

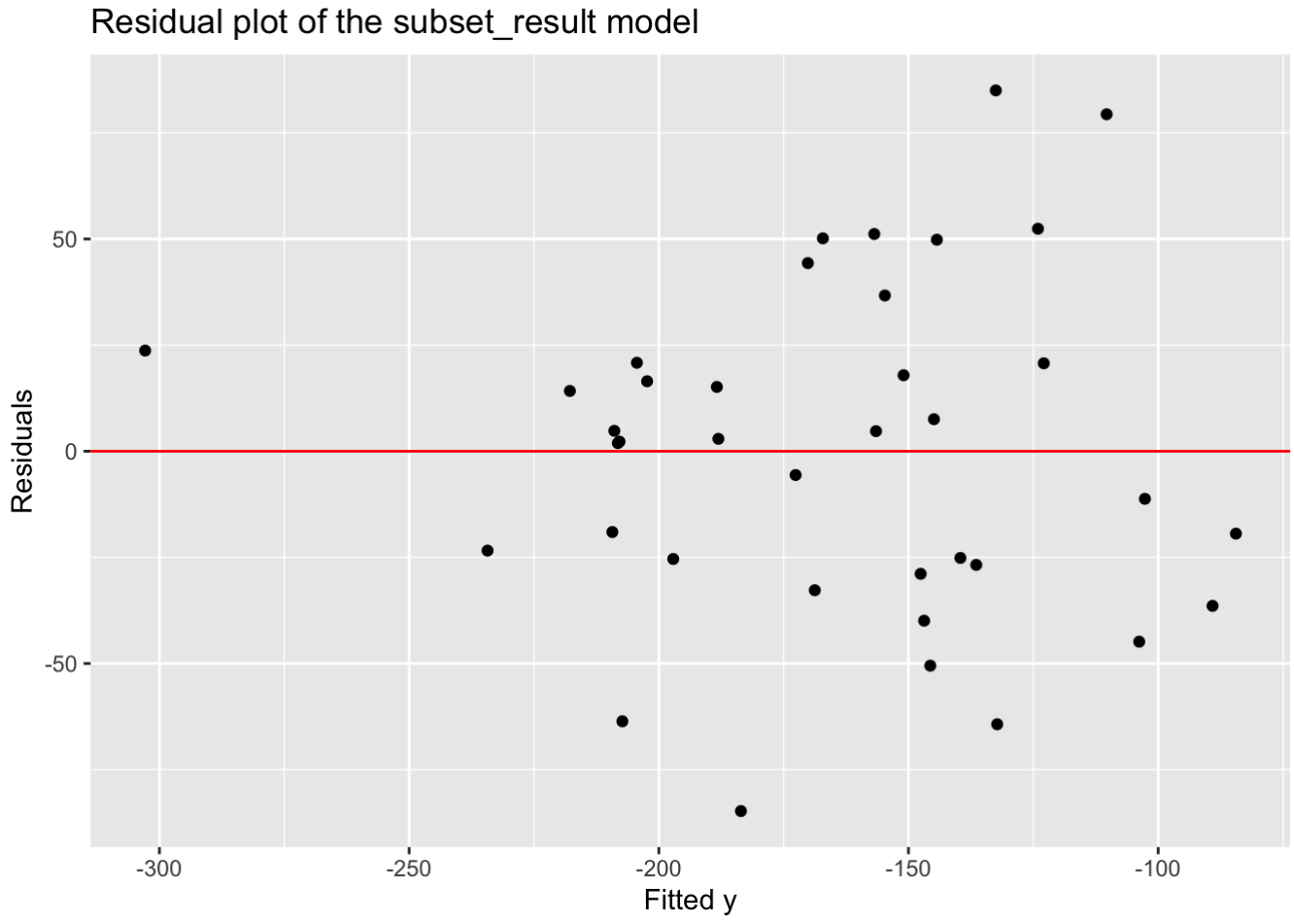
```
# Check the p-value for this F statistic
1-pf(F0_stat, 5, 29)
```

```
## [1] 0.06610044
```

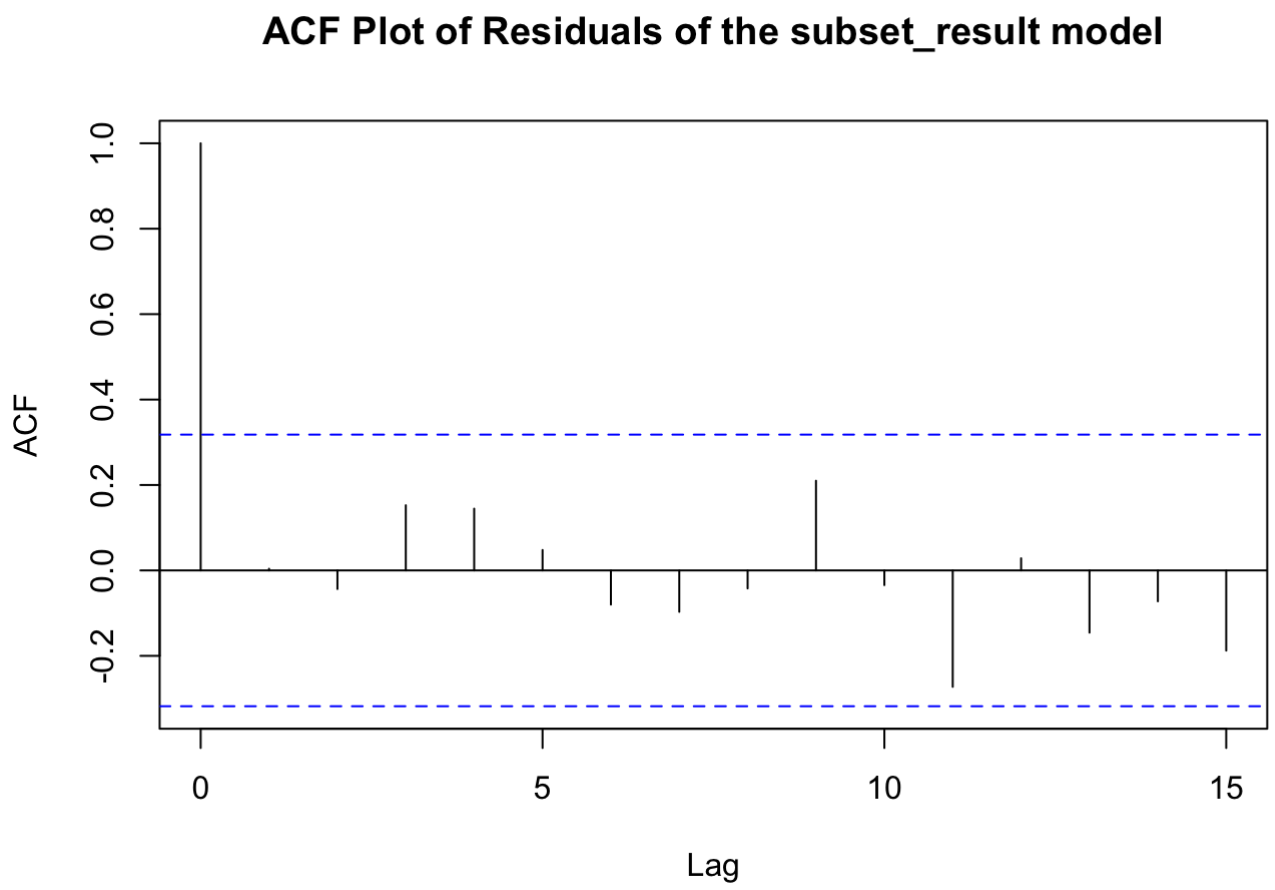
Conclusion: Fail to reject H0 => go with the reduced subset model. Question: “Age” seems to be the biggest problem, dropping height related variables helps somewhat, but still the model is not good.

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

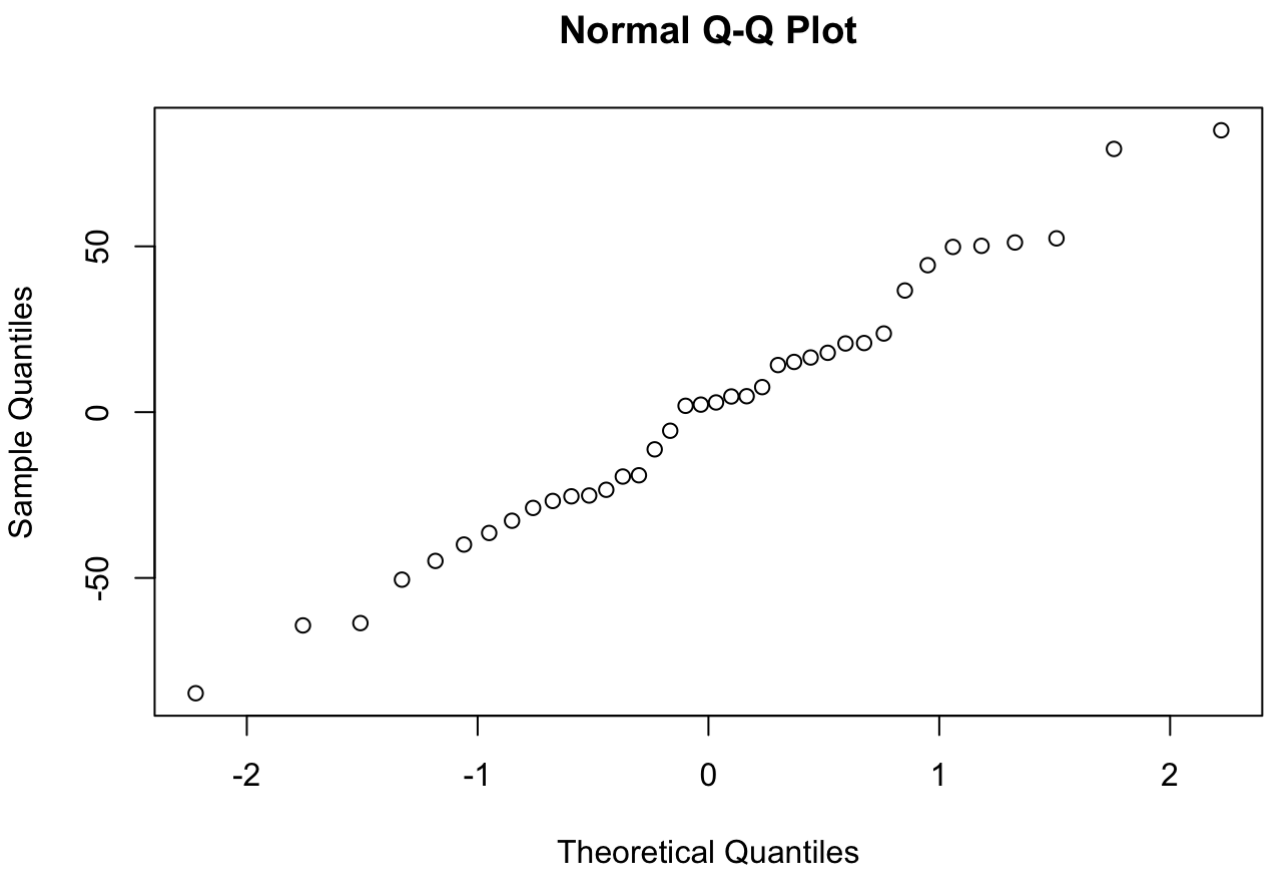
```
y_hat<-subset_result$fitted.values
residuals<-subset_result$residuals
plot_data<-data.frame(y_hat,residuals)


ggplot(plot_data, aes(x=y_hat, y=residuals))+
  geom_point()+
  geom_hline(yintercept=0, color="red")+
  labs(x="Fitted y", y="Residuals", title="Residual plot of the subset_result model")
```




```
acf(residuals, main="ACF Plot of Residuals of the subset_result model")
```




```
qqnorm(residuals)
```



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.

Estimated regression: Hipcenter_hat = Age + Weight + Thigh. This subset model produced R_squared of only 0.5597, this is smaller than the original R_squared of 0.6866. Adjusted R_squared also decreased from 0.6 down to 0.52. Check with class (?)