Lecture 14:
Multicollinearity
STAT 632, Spring 2020

# Multicollinearity

When predictors in a regression model are strongly correlated there can be a number of issues:

- The signs of the coefficients can be the opposite of what we expect.
- The standard errors are inflated so the t-tests may fail to reveal significant predictors.
- The predictor variables are not significant when the overall F-test is highly significant.

# Multicollinearity

Multicollinearity can be detected in several ways:

- Examining the relationships between the predictor variables in the scatter plot matrix.
- Examining the correlation matrix of the predictor variables.
- Variance inflation factors (VIFs).

# Variance Inflation Factor (VIF)

Consider the multiple linear regression model with two predictors

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

It can be shown that

$$Var(\hat{\beta}_j) = \frac{1}{1 - r_{12}^2} \cdot \frac{\sigma^2}{(n-1)S_{x_j}^2}$$

- $r_{12}$ denotes the correlation between $x_1$ and $x_2$
- $S_{x_j}$ denotes the standard deviation of $x_j$

Notice that the variance of $\hat{\beta}_j$ increases as $r_{12}^2$ increases. Thus, the correlation between the predictors increases the variance of the estimated regression coefficient.

# Variance Inflation Factor (VIF)

Consider the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$$

It can be shown that

$$Var(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \cdot \frac{\sigma^2}{(n-1)S_{x_j}^2}$$

where $R_j^2$ is obtained from the regression of $x_j$ on all other predictors (i.e., the percent of variation in $x_j$ explained by the other predictors).

The term $\frac{1}{1-R_j^2}$ is called the **variance inflation factor** (VIF). A commonly used rule is that a VIF greater than 5 indicates that there are issues with multicollinearity.

# Variance Inflation Factor (VIF)

For example, suppose that $R_j^2 = 0.99$, then

$$VIF_j = \frac{1}{1 - 0.99} = 100$$

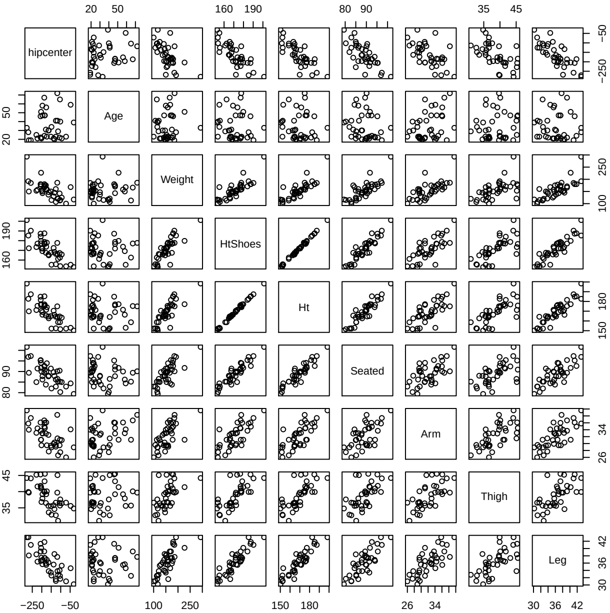
The interpretation is that the standard error of the estimated regression coefficient, $se(\hat{\beta}_j)$, is $\sqrt{100} = 10$ times larger than it would be if there was no collinearity ($R_j^2 = 0$).

# Example[1]

- Car drivers like to adjust the seat position for their own comfort. Car designers would find it helpful to know where different drivers will position the seat depending on their size and age.
- Researchers at the 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.
- The predictors of interest are age, weight, height with and without shoes, seated height, arm length, thigh length, and lower leg length.

[1] Example from Julian Faraway, *Linear Models in R*, 1st edition, Ch. 5, pp. 83–87

```
> library(faraway)
> head(seatpos)
  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
4  19    185   190.3 187.4   97.3 37.4  44.1 41.0  -257.720
5  23    159   178.0 174.1   93.9 29.5  40.1 36.9  -173.230
6  47    170   178.7 177.0   92.4 36.0  43.2 37.4  -185.150
```


hipcenter
Age
Weight
HtShoes
Ht
Seated
Arm
Thigh
Leg
20
50
160
190
80
90
35
45
−50
−250
250
100
180
150
34
26
42
36
30
−250
−50
100
250
150
180
26
34
30
36
42

There are several strong correlations between the predictors.

```
> round(cor(seatpos[, -9]), 2)
          Age Weight HtShoes    Ht Seated  Arm Thigh   Leg
Age      1.00   0.08   -0.08 -0.09  -0.17 0.36  0.09 -0.04
Weight   0.08   1.00    0.83  0.83   0.78 0.70  0.57  0.78
HtShoes -0.08   0.83    1.00  1.00   0.93 0.75  0.72  0.91
Ht      -0.09   0.83    1.00  1.00   0.93 0.75  0.73  0.91
Seated  -0.17   0.78    0.93  0.93   1.00 0.63  0.61  0.81
Arm      0.36   0.70    0.75  0.75   0.63 1.00  0.67  0.75
Thigh    0.09   0.57    0.72  0.73   0.61 0.67  1.00  0.65
Leg     -0.04   0.78    0.91  0.91   0.81 0.75  0.65  1.00
```

The model shows signs of multicollinearity. The overall F statistic is large and $R^2 = 0.6$, but none of the individual predictors are significant.

```
> lm1 <- lm(hipcenter ~ ., data=seatpos)
> summary(lm1)
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
```

Several of the variance inflation factors are large and exceed the 5 cut-off.

```
> library(faraway) # to use vif() function
> round(vif(lm1), 2)
    Age  Weight HtShoes      Ht  Seated     Arm   Thigh     Leg
   2.00    3.65  307.43  333.14    8.95    4.50    2.76    6.69
```

For HtShoes the interpretation is that the standard error for this predictor is $\sqrt{307.4} = 17.5$ times larger than it would be without collinearity.

We can also compute the VIFs manually.

```
# create data frame only containing predictors
> x <- seatpos[, -9]

> summary(lm(Ht ~., data=x))$r.squared
[1] 0.9969982

> 1/(1 - 0.9969982)
[1] 333.1335
```

# Your Turn

Manually compute the VIF for the predictor `Seated`, which is seated height in cm.

- Many of the variables in the full model are redundant, and do the same job at predicting the response.
- For example, the following predictors all measure the length of the driver in some way: `HtShoes`, height in shoes; `Ht`, height bare foot; `Seated`, seated height; `Arm`, arm length; `Thigh`, thigh length; and `Leg`, leg length.
- Instead of using all of these length predictors, we can just select one to include in the model, and drop the others.
- However, because of collinearity, we should not conclude that the variables we drop have nothing to do with the response.

Consider the correlation matrix with just the length variables. All of these predictor variables are strongly correlated with each other. We pick `Ht` since it is the simplest measure, and more strongly correlated with the response than the other predictors.

```
> round(cor(seatpos[, 3:9]), 2)
          HtShoes    Ht Seated   Arm Thigh   Leg hipcenter
HtShoes      1.00  1.00   0.93  0.75  0.72  0.91     -0.80
Ht           1.00  1.00   0.93  0.75  0.73  0.91     -0.80
Seated       0.93  0.93   1.00  0.63  0.61  0.81     -0.73
Arm          0.75  0.75   0.63  1.00  0.67  0.75     -0.59
Thigh        0.72  0.73   0.61  0.67  1.00  0.65     -0.59
Leg          0.91  0.91   0.81  0.75  0.65  1.00     -0.79
hipcenter   -0.80 -0.80  -0.73 -0.59 -0.59 -0.79      1.00
```

Removing some correlated predictors fixes many of the issues caused by multicollinearity. The predictor `Ht` is now highly significant in the model. Further simplification is clearly possible.

```
> lm2 <- lm(hipcenter ~ Age + Weight + Ht, data=seatpos)
> summary(lm2)

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

> vif(lm2)
     Age   Weight       Ht
1.093018 3.457681 3.463303
```

The $R^2$ of the reduced model with just 3 predictors (`Age`, `Weight`, and `Ht`) is close to the $R^2$ of the full model with all the strongly correlated predictors. In fact, the adjusted $R^2$ for the reduced model is slightly higher than the full model.

```
> summary(lm1)$r.squared
[1] 0.6865535
> summary(lm2)$r.squared
[1] 0.6561654

> summary(lm1)$adj.r.squared
[1] 0.6000855
> summary(lm2)$adj.r.squared
[1] 0.6258271
```

# Concluding Remarks

Some ways to deal with multicollinearity:

- If several predictors are strongly correlated with each other, pick one predictor out of the bunch to use in the reduced model. The $R^2$ should not change much after removing some correlated predictors.
- You can also combine predictors. For instance, by taking the sum or average of two correlated predictors.
- Automated variable selection techniques (e.g., stepwise selection, LASSO) can also be used (see ISLR, Ch. 6).