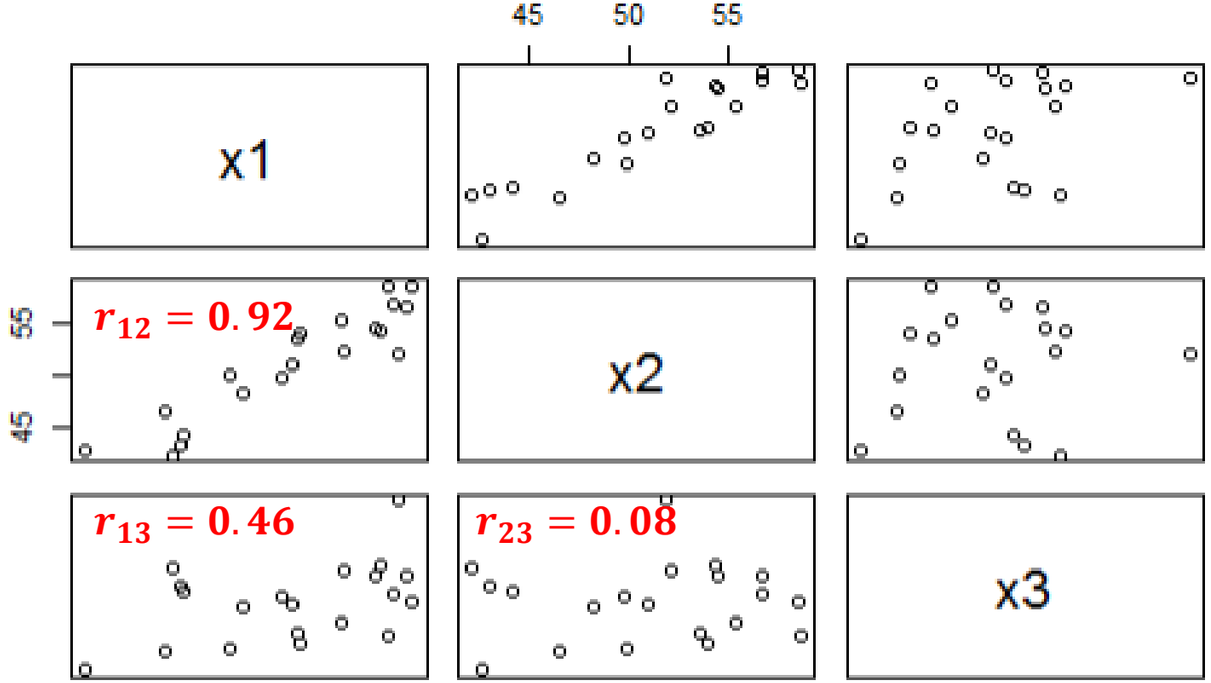
Effects of Multicollinearity

The body fat example

- The body fat example: a study of the relation of amount of body fat (Y) to several possible predictor variables, based on a sample of n=20 healthy females 25-34 years old. The possible predictors are
  - X1: The triceps skinfold thickness;
  - X2: The thigh circumference;
  - X3: midarm circumference.



X1 and X2 are highly correlated;
X3 is not so related to X1 and X2 individually

Effects of multicollinearity on regression coefficients, b_k

- The regression coefficient of one variable (eg. For X1) varies markedly depending on other variables in the model.

- If predictors are correlated, the regression coefficient cannot capture the true effect of the individual predictor variable, and instead, only represents a marginal or partial effect.

As a result, the coefficients in the MLR model do not accurately reflect the linear impact of the variable on Y.

- If intercorrelated predictor variables are left out of the model, they can still affect the coefficients of the remaining variables in the model.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4961     3.3192  -0.451    0.658
x1            0.8572     0.1288   6.656 3.02e-06 ***
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.6345     5.6574  -4.178 0.000566 ***
x2            0.8565     0.1100   7.786  3.6e-07 ***
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.1742     8.3606  -2.293   0.0348 *
x1            0.2224     0.3034   0.733   0.4737
x2            0.6594     0.2912   2.265   0.0369 *
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
x1             4.334      3.016   1.437    0.170
x2            -2.857      2.582  -1.106    0.285
x3            -2.186      1.595  -1.370    0.190
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.7916     4.4883   1.513   0.1486
x1            1.0006     0.1282   7.803 5.12e-07 ***
x3           -0.4314     0.1766  -2.443   0.0258 *
```

Effects of multicollinearity on the standard error of the coefficients, $s\{b_k\}$

When only X1 in the model $s\{b1\} = 0.1288$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4961     3.3192  -0.451    0.658
x1            0.8572     0.1288   6.656 3.02e-06 ***
```

When only X2 in the model $s\{b2\} = 0.11$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.6345     5.6574  -4.178 0.000566 ***
x2            0.8565     0.1100   7.786  3.6e-07 ***
```

When only X1, x2 in the model $s\{b1\} = 0.3034$
$s\{b2\} = 0.2912$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.1742     8.3606  -2.293   0.0348 *
x1            0.2224     0.3034   0.733   0.4737
x2            0.6594     0.2912   2.265   0.0369 *
```

When X1, x2, and x3 in the model $s\{b1\} = 3.016$
$s\{b2\} = 2.582$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
x1             4.334      3.016   1.437    0.170
x2            -2.857      2.582  -1.106    0.285
x3            -2.186      1.595  -1.370    0.190
```

- High degree of multicollinearity among the predictor variables is responsible for the inflated variability of the estimated regression coefficients, resulting in an increased standard error of the estimates.

Effects of multicollinearity on sums of squares

- In the presence of correlated predictors, the impact of a single predictor variable on reducing the error sum of squares can differ depending on the other variables included in the model. Therefore, when evaluating the reduction in total variance attributed to a particular predictor, it must be considered in the context of the other correlated predictors.

SSR(X1)=352.27 SSR(X1|X2)=3.47

SSR(X1) > SSR(X1|X2)

$$R_1^2 = 0.72 > R_{1|2}^2 = \frac{3.47}{3.47 + 109.95} = 0.03$$

- When SSR(X1) < SSR(X1|X2), X2 is called ***suppressor*** variable.
- In other words, a suppressor variable enhances the relationship between the Y variable and another predictor variable by removing the influence of extraneous or confounding variables, which allows for a more accurate prediction of the outcome variable. The evidence of suppressor variable can be verified via the corresponding ESS terms or partial correlation coefficients values.

Suppressor Variable Example:

- The dependent variable, Y (Chance to cancel a streaming service subscription)
- The Predictors variables X1(age), X2(subscription plan cost), X3(frequency of watching)
- The confounding variable Xc (the availability of alternative service in the location)
- The suppressor variable Xs (location)

```
anova(lm(y~x1, bodyfat))

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 352.27  352.27  44.305 3.024e-06 ***
Residuals 18 143.12    7.95
```

```
anova(lm(y~x2+x1, bodyfat))

Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x2         1 381.97  381.97  59.057 6.281e-07 ***
x1         1   3.47    3.47   0.537    0.4737
Residuals 17 109.95    6.47
```

```
anova(lm(y~x2, bodyfat))

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x2         1 381.97  381.97  60.617 3.6e-07 ***
Residuals 18 113.42    6.30
```

Effects of multicollinearity on mean response estimate

When only X1 in the model MSE= 7.95

When $X1 = 25$, $\hat{Y}_h$= -1.4961+0.8572(25)=19.93

$$s\{\hat{Y}_h\} = \sqrt{\boldsymbol{X_h'}\boldsymbol{\Sigma}\{\boldsymbol{b}\}\boldsymbol{X_h}} = \sqrt{s^2(\frac{1}{n} + \frac{(25-\bar{X})^2}{SSX})} = 0.63$$

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 352.27  352.27  44.305 3.024e-06 ***
Residuals 18 143.12    7.95
```

When only X1, x2 in the model MSE=6.47

When $X1 = 25$, $X2 = 50$, $\hat{Y}_h$= -19.1742+0.2224(25)+0.6594(50)=19.39

$$s\{\hat{Y}_h\} = \sqrt{\boldsymbol{X_h'}\boldsymbol{\Sigma}\{\boldsymbol{b}\}\boldsymbol{X_h}} = 0.624$$

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x2         1 381.97  381.97  59.057 6.281e-07 ***
x1         1   3.47    3.47   0.537    0.4737
Residuals 17 109.95    6.47
```

When X1, x2, and x3 in the model $MSE = 6.15$

When $X1 = 25$, $X2 = 50$, $X3 = 29$, $\hat{Y}_h$= 19.19 $s\{\hat{Y}_h\} = 0.619$

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 352.27  352.27 57.2768 1.131e-06 ***
x2         1  33.17   33.17  5.3931   0.03373 *
x3         1  11.55   11.55  1.8773   0.18956
Residuals 16  98.40    6.15
```

When more variables are added to the model, the high degree of multicollinearity

- does not prevent the SSE from being steadily reduced.
- could prevent the MSE from being steadily reduced.
- will increase the standard error of the mean response estimate.

Need for more powerful diagnostics for multicollinearity

- Multicollinearity in predictor variables can significantly impact the interpretation and utilization of a regression model.

- The correlation coefficient and partial correlation coefficient are common diagnostic tools used to identify multicollinearity and can be useful in detecting the issue.

- However, it is possible for serious multicollinearity to exist without being detected by these methods. Later, we will discuss some remedial measures that can be taken to address this issue.