Multicollinearity

1. Introduction

Multicollinearity refers to a situation in which two or more **explanatory variables** in a multiple regression model are highly linearly related.

2. Data

```r
library(readr)
bloodpressure <- read_csv("bloodpressure.csv")
bloodpressure
```

```
# A tibble: 20 x 9
      X1    Pt    BP   Age Weight   BSA   Dur Pulse Stress
   <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl>
 1     1     1   105    47   85.4  1.75   5.1    63     33
 2     2     2   115    49   94.2  2.1    3.8    70     14
 3     3     3   116    49   95.3  1.98   8.2    72     10
 4     4     4   117    50   94.7  2.01   5.8    73     99
 5     5     5   112    51   89.4  1.89   7      72     95
 6     6     6   121    48   99.5  2.25   9.3    71     10
 7     7     7   121    49   99.8  2.25   2.5    69     42
 8     8     8   110    47   90.9  1.9    6.2    66      8
 9     9     9   110    49   89.2  1.83   7.1    69     62
10    10    10   114    48   92.7  2.07   5.6    64     35
11    11    11   114    47   94.4  2.07   5.3    74     90
12    12    12   115    49   94.1  1.98   5.6    71     21
13    13    13   114    50   91.6  2.05  10.2    68     47
14    14    14   106    45   87.1  1.92   5.6    67     80
15    15    15   125    52  101.   2.19  10      76     98
16    16    16   114    46   94.5  1.98   7.4    69     95
17    17    17   106    46   87    1.87   3.6    62     18
18    18    18   113    46   94.5  1.9    4.3    70     12
19    19    19   110    48   90.5  1.88   9      71     99
20    20    20   122    56   95.7  2.09   7      75     99
```

2.1 Variable description

1. Y: BP (blood pressure, in mmHg)
2. X_1: Age (in years)
3. X_2: Weight (in kg)
4. X_3: BSA (body surface area, in sq m)

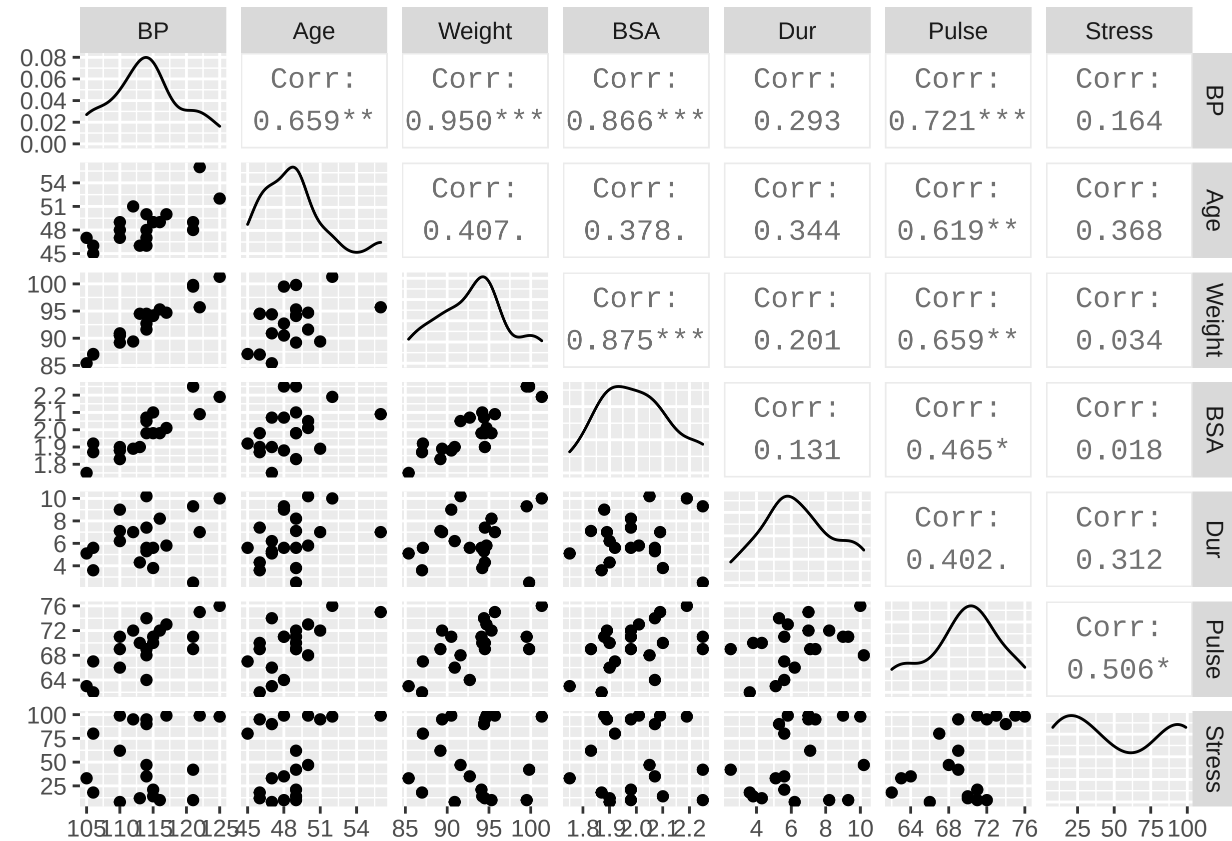
5. X_4: Dur (duration of hypertension)
6. X_5: Pulse (basal pulse)
7. X_6: Stress (stress index)

3. How to detect multicollinearity?

1. Correlation matrix and scatterplot matrix

This is limiting. It is possible that the pairwise correlations between variables are small, but a linear dependence exists among three or even more variables in the dataset. Hence, we use **variance inflation factors (VIF)** to detect multicollinearity.

```
library(GGally)
ggpairs(bloodpressure[, -c(1, 2)])
```



2. Variance Inflation Factors (VIF)

Variance inflation factor for j^{th} variable

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 value obtained by regressing the j^{th} predictor on the remaining predictors.

```r
library(broom)
bp <- lm(BP ~ Age + Weight + BSA + Dur + Pulse + Stress, data=bloodpressure)
bp
```

```
Call:
lm(formula = BP ~ Age + Weight + BSA + Dur + Pulse + Stress, 
    data = bloodpressure)

Coefficients:
(Intercept)          Age       Weight          BSA          Dur        Pulse  
 -12.870476     0.703259     0.969920     3.776491     0.068383    -0.084485  
     Stress  
   0.005572  
```

```r
summary(bp)
```

```
Call:
lm(formula = BP ~ Age + Weight + BSA + Dur + Pulse + Stress, 
    data = bloodpressure)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.93213 -0.11314  0.03064  0.21834  0.48454 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12.870476   2.556650  -5.034 0.000229 ***
Age           0.703259   0.049606  14.177 2.76e-09 ***
Weight        0.969920   0.063108  15.369 1.02e-09 ***
BSA           3.776491   1.580151   2.390 0.032694 *  
Dur           0.068383   0.048441   1.412 0.181534    
Pulse        -0.084485   0.051609  -1.637 0.125594    
Stress        0.005572   0.003412   1.633 0.126491    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4072 on 13 degrees of freedom
Multiple R-squared:  0.9962,	Adjusted R-squared:  0.9944 
F-statistic: 560.6 on 6 and 13 DF,  p-value: 6.395e-15
```

4. Calculate VIF

```r
library(car)
vif(bp)
```

```
     Age   Weight      BSA      Dur    Pulse   Stress 
1.762807 8.417035 5.328751 1.237309 4.413575 1.834845 
```

5. Illustration of the output for weight variable

Build a regression model taking *weight* as the dependent variable and remaining `x` variables as the independent variables.

```
weight <- lm(Weight ~ Age  + BSA + Dur + Pulse + Stress, data=bloodpressure)
weight
```

```
Call:
lm(formula = Weight ~ Age + BSA + Dur + Pulse + Stress, data = bloodpressure)

Coefficients:
(Intercept)          Age          BSA          Dur        Pulse       Stress
  19.674438    -0.144643    21.421654     0.008696     0.557697    -0.022997
```

```
summary(weight)
```

```
Call:
lm(formula = Weight ~ Age + BSA + Dur + Pulse + Stress, data = bloodpressure)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7697 -1.0120  0.1960  0.6955  2.7035

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.674438   9.464742   2.079  0.05651 .
Age         -0.144643   0.206491  -0.700  0.49510
BSA         21.421654   3.464586   6.183 2.38e-05 ***
Dur          0.008696   0.205134   0.042  0.96678
Pulse        0.557697   0.159853   3.489  0.00361 **
Stress      -0.022997   0.013079  -1.758  0.10052
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.725 on 14 degrees of freedom
Multiple R-squared:  0.8812,    Adjusted R-squared:  0.8388
F-statistic: 20.77 on 5 and 14 DF,  p-value: 5.046e-06
```

$$VIF_weight = \frac{1}{1 - R_w^2 eight} = \frac{1}{1 - 0.8812} = 8.42$$

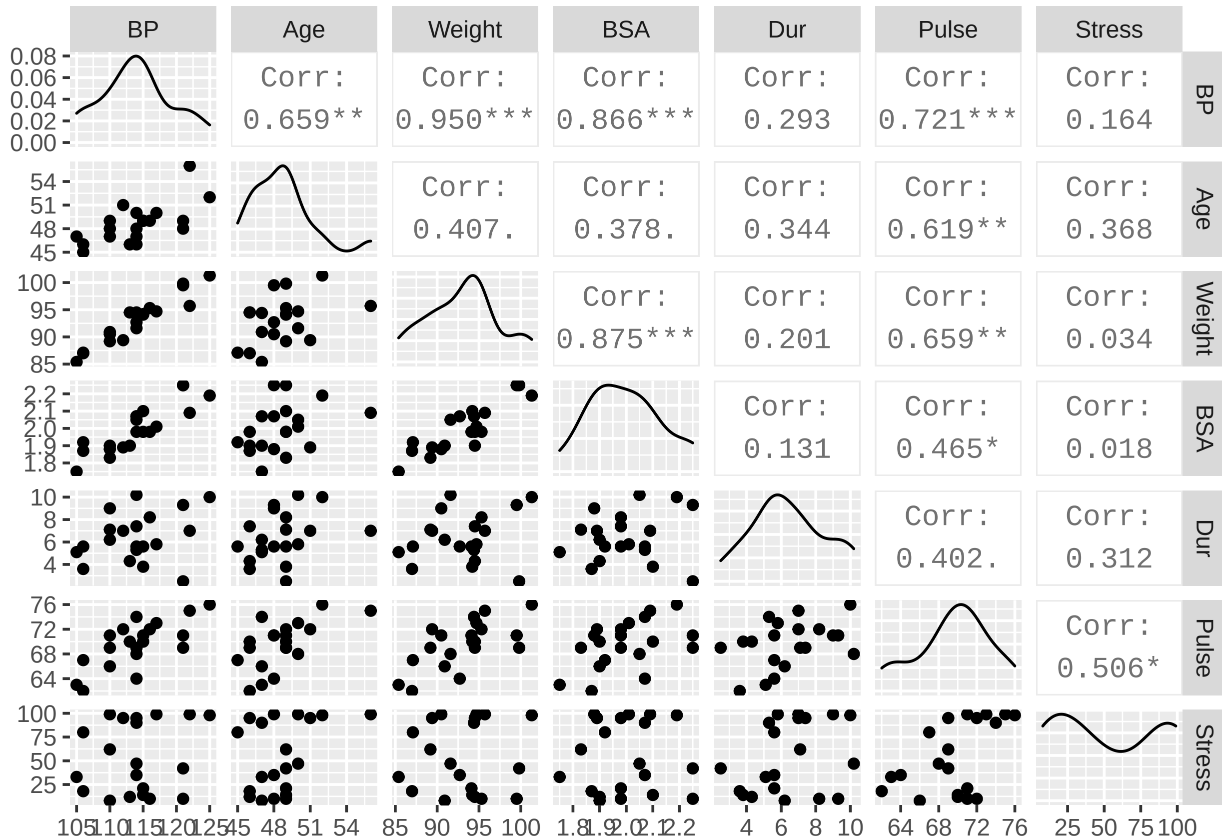
VIFs exceeding 4 indicates high multicollinearity while VIFs exceeding 10 are considered evidence of serious multicollinearity requiring correction.

6. What to do now?

One solution is to remove some of the variables with high VIF. Variables `Weight`, `BSA` and `Pulse` have high VIF values. If we review the pairwise correlations again, we can see `Weight` and `BSA` are highly correlated. We can choose to remove either predictor from the model.

Which one to remove? In-class discussion.

```r
library(GGally)
ggpairs(bloodpressure[, -c(1, 2)])
```



New model without Pulse and BSA

```r
library(broom)
bp2 <- lm(BP ~ Age + Weight  + Dur  + Stress, data=bloodpressure)
bp2
```

```
Call:
lm(formula = BP ~ Age + Weight + Dur + Stress, data = bloodpressure)

Coefficients:
(Intercept)          Age       Weight          Dur       Stress
 -15.869829     0.683741     1.034128     0.039889     0.002184
```

```r
summary(bp2)
```

```
Call:
lm(formula = BP ~ Age + Weight + Dur + Stress, data = bloodpressure)

Residuals:
     Min       1Q   Median       3Q      Max
-1.11359 -0.29586  0.01515  0.27506  0.88674

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.869829   3.195296  -4.967 0.000169 ***
Age           0.683741   0.061195  11.173 1.14e-08 ***
Weight        1.034128   0.032672  31.652 3.76e-15 ***
Dur           0.039889   0.064486   0.619 0.545485
Stress        0.002184   0.003794   0.576 0.573304
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5505 on 15 degrees of freedom
Multiple R-squared:  0.9919,    Adjusted R-squared:  0.9897
F-statistic: 458.3 on 4 and 15 DF,  p-value: 1.764e-15
```

```r
vif(bp2)
```

```
     Age   Weight      Dur   Stress
1.468245 1.234653 1.200060 1.241117
```

Ackowledgement

Data: The Pennsylvania State University.