Multicollinearity

Tuan V. Nguyen
Professor and NHMRC Senior Research Fellow
Garvan Institute of Medical Research
University of New South Wales
Sydney, Australia

What we are going to learn

- Introduction to MLR
- Interaction
- Polynomial regression

What is multicollinearity?

- Multicollinearity exists whenever two or more of the predictors in a regression model are moderately or highly correlated.
- Multicollinearity happens more often than not in such observational studies.

Types of multicollinearity

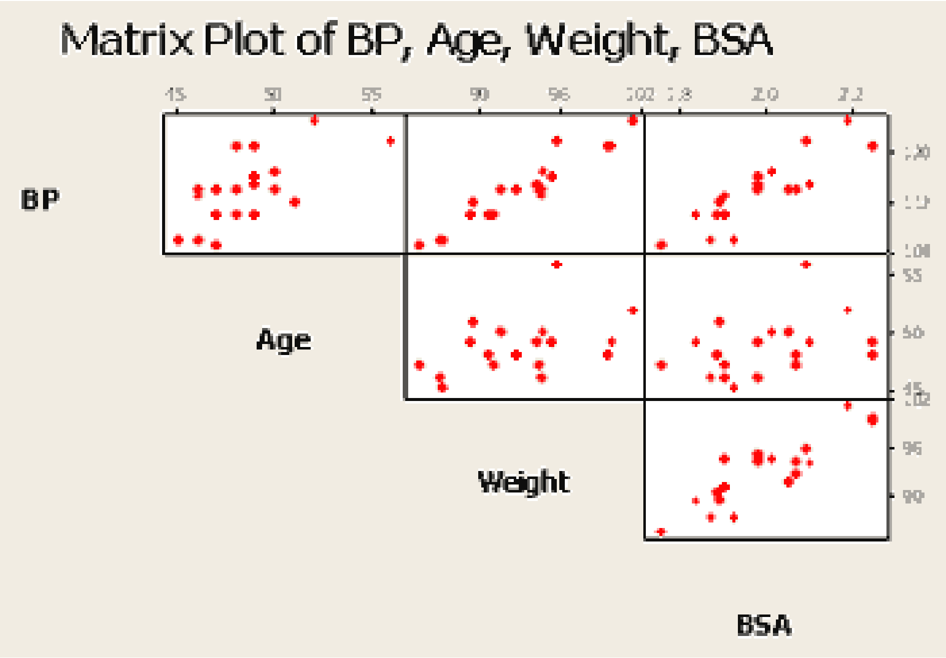
- **Structural multicollinearity** is a mathematical artifact caused by creating new predictors from other predictors — such as, creating the predictor *x*2 from the predictor *x*.
- **Data-based multicollinearity**, on the other hand, is a result of a poorly designed experiment, reliance on purely observational data, or the inability to manipulate the system on which the data are collected.

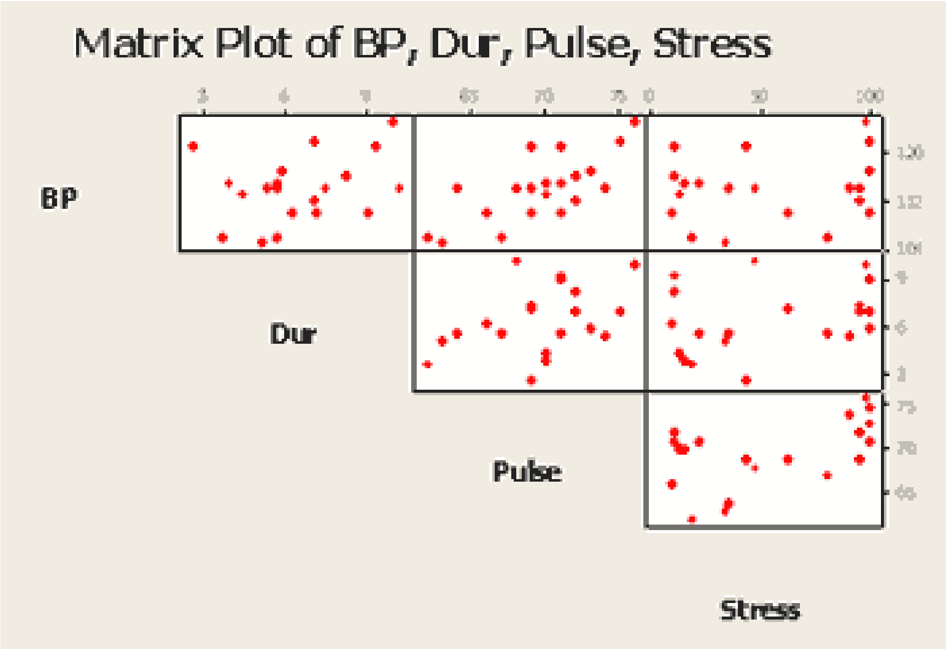
Consider the following study

Pt	BP	Age	Weight	BSA	Dur	Pulse	Stress
1	**105**	47	85.4	1.75	5.1	63	33
2	**115**	49	94.2	2.1	3.8	70	14
3	**116**	49	95.3	1.98	8.2	72	10
4	**117**	50	94.7	2.01	5.8	73	99
5	**112**	51	89.4	1.89	7	72	95
6	**121**	48	99.5	2.25	9.3	71	10
7	**121**	49	99.8	2.25	2.5	69	42
8	**110**	47	90.9	1.9	6.2	66	8
9	**110**	49	89.2	1.83	7.1	69	62
10	**114**	48	92.7	2.07	5.6	64	35
11	**114**	47	94.4	2.07	5.3	74	90
12	**115**	49	94.1	1.98	5.6	71	21
13	**114**	50	91.6	2.05	10.2	68	47
14	**106**	45	87.1	1.92	5.6	67	80
15	**125**	52	101.3	2.19	10	76	98
16	**114**	46	94.5	1.98	7.4	69	95
17	**106**	46	87	1.87	3.6	62	18
18	**113**	46	94.5	1.9	4.3	70	12
19	**110**	48	90.5	1.88	9	71	99
20	**122**	56	95.7	2.09	7	75	99

blood pressure (y = *BP*, in mm Hg); **age** ($x1$ = *Age*, in years); **weight** ($x2$ = *Weight*, in kg); **body surface area** ($x3$ = *BSA*, in sq m); **duration of hypertension** ($x4$ = *Dur*, in years); **basal pulse** ($x5$ = *Pulse*, in beats per minute); **stress index** ($x6$ = *Stress*)

Inter-correlations among variables

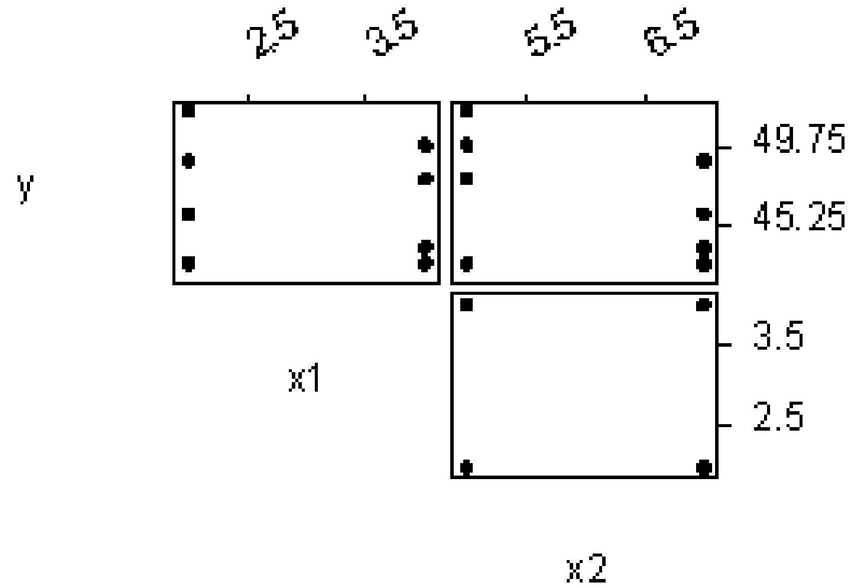




	BP	Age	Weight	BSA	Duration	Pulse
Age	0.659					
Weight	0.950	0.407				
BSA	0.866	0.378	0.875			
Duration	0.293	0.344	0.201	0.131		
Pulse	0.721	0.619	0.659	0.465	0.402	
Stress	0.164	0.368	0.034	0.018	0.312	0.506

Uncorrelated predictors

- **What is the effect on regression analyses if the predictors are perfectly uncorrelated ?**




```
Pearson correlation of
x1 and x2 = 0.000
```

Uncorrelated predictors

- **What is the effect on regression analyses if the predictors are perfectly uncorrelated ?**
- Regresion model 1: x1 is the predictor

The regression equation is y = 48.8 - 0.63 x1

Predictor	Coef	SE Coef	T	P
Constant	48.750	4.025	12.11	0.000
x1	-0.625	1.273	-0.49	0.641

- Regresion model 1: x2 is the predictor

The regression equation is y = 55.1 - 1.38 x2

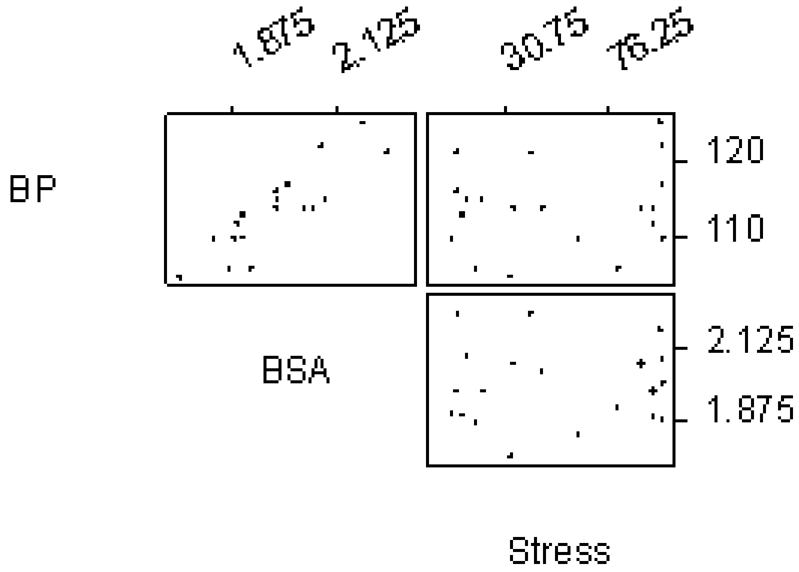
Predictor	Coef	SE Coef	T	P
Constant	55.125	7.119	7.74	0.000
x2	-1.375	1.170	-1.17	0.285

- Regresion model 1: x1 and x2 are predictors

The regression equation is y = 57.0 - 0.63 x1 - 1.38 x2

Predictor	Coef	SE Coef	T	P
Constant	57.000	8.486	6.72	0.001
x1	-0.625	1.251	-0.50	0.639
x2	-1.375	1.251	-1.10	0.322

What is the effect on regression analyses if the predictors are *nearly* uncorrelated?



	BP	Age	Weight	BSA	Duration	Pulse
Age	0.659					
Weight	0.950	0.407				
BSA	0.866	0.378	0.875			
Duration	0.293	0.344	0.201	0.131		
Pulse	0.721	0.619	0.659	0.465	0.402	
Stress	0.164	0.368	0.034	0.018	0.312	0.506

- The regression of the response *y = BP* on the predictor *x6 = Stress*:

What is the effect on regression analyses if the predictors are *nearly* uncorrelated?

- **The regression of the response *y = BP* on the predictor *x6 = Stress*:**

```
The regression equation is
BP = 113 + 0.0240 Stress

Predictor        Coef      SE Coef         T        P
Constant      112.720        2.193     51.39    0.000
Stress        0.02399      0.03404      0.70    0.490

S = 5.502        R-Sq = 2.7%       R-Sq(adj) = 0.0%
```

- The regression of the **response *y = BP*** on the **predictor *x3 = BSA***:

```
The regression equation is
BP = 45.2 + 34.4 BSA

Predictor        Coef      SE Coef         T        P
Constant       45.183        9.392      4.81    0.000
BSA            34.443        4.690      7.34    0.000

S = 2.790        R-Sq = 75.0%      R-Sq(adj) = 73.6%
```

What is the effect on regression analyses if the predictors are *nearly* uncorrelated?

- The regression of the **response *y = BP*** on the **predictors *x6= Stress*** and ***x3 = BSA*** (in that order)

```
The regression equation is
BP = 44.2 + 0.0217 Stress + 34.3 BSA

Predictor        Coef     SE Coef       T       P
Constant       44.245       9.261    4.78   0.000
Stress        0.02166     0.01697    1.28   0.219
BSA            34.334       4.611    7.45   0.000
```

- Finally, the regression of the **response *y = BP*** on the **predictors *x3 = BSA*** and ***x6= Stress*** (in that order)

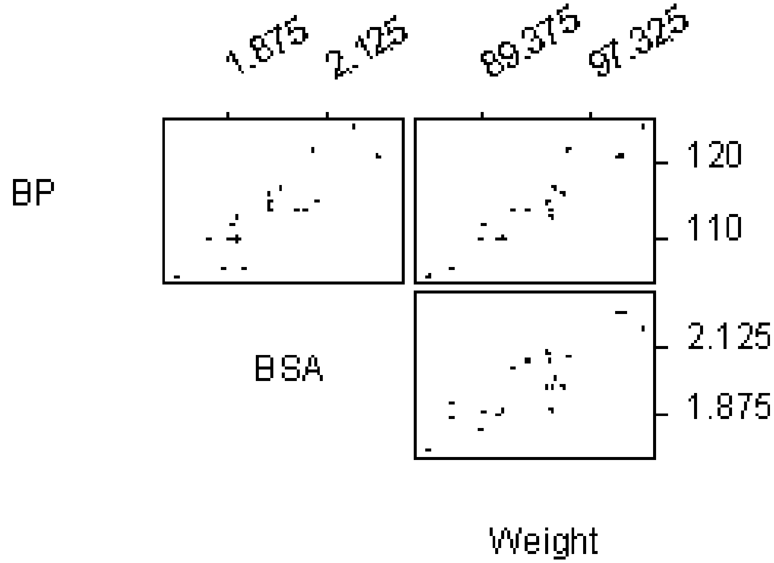
```
The regression equation is
BP = 44.2 + 34.3 BSA + 0.0217 Stress

Predictor        Coef     SE Coef       T       P
Constant       44.245       9.261    4.78   0.000
BSA            34.334       4.611    7.45   0.000
Stress        0.02166     0.01697    1.28   0.219
```

What have we seen?

Model	b_6	$se(b_6)$	b_3	$se(b_3)$	Seq SS
x_6 only	0.0240	0.0340	---	---	$SSR(x_6)$ 15.04
x_3 only	---	---	34.443	4.690	$SSR(x_3)$ 419.86
x_6, x_3 (in order)	0.0217	0.0170	34.334	4.611	$SSR(x_3 \vert x_6)$ 417.07
x_3, x_6 (in order)	0.0217	0.0170	34.334	4.611	$SSR(x_6 \vert x_3)$ 12.26

- We don't get identical, but very *similar* slope estimates *b*3 and *b*6, regardless of the predictors in the model.
- The sum of squares *SSR*(*x*3) is not the same, but very *similar* to the sequential sum of squares *SSR*(*x*3|*x*6).
- The sum of squares *SSR*(*x*6) is not the same, but very *similar* to the sequential sum of squares *SSR*(*x*6|*x*3).

What happens if the predictor variables are highly correlated?



	BP	Age	Weight	BSA	Duration	Pulse
Age	0.659					
Weight	0.950	0.407				
BSA	0.866	0.378	0.875			
Duration	0.293	0.344	0.201	0.131		
Pulse	0.721	0.619	0.659	0.465	0.402	
Stress	0.164	0.368	0.034	0.018	0.312	0.506

What happens if the predictor variables are highly correlated?

- **The regression of the response *y = BP* on the predictor *x2= Weight*:**

```
The regression equation is
BP = 2.21 + 1.20 Weight

Predictor        Coef     SE Coef        T        P
Constant        2.205       8.663     0.25    0.802
Weight        1.20093     0.09297    12.92    0.000

S = 1.740        R-Sq = 90.3%      R-Sq(adj) = 89.7%
```

- The regression of the response ***y = BP*** on the predictor ***x3= BSA***

```
The regression equation is
BP = 45.2 + 34.4 BSA

Predictor        Coef     SE Coef        T        P
Constant       45.183       9.392     4.81    0.000
BSA            34.443       4.690     7.34    0.000

S = 2.790        R-Sq = 75.0%      R-Sq(adj) = 73.6%
```

What happens if the predictor variables are highly correlated?

- The regression of the response ***y = BP*** on the predictors ***x2= Weight*** and ***x3 = BSA*** (in that order)

The regression equation is
BP = 5.65 + 1.04 Weight + 5.83 BSA

Predictor	Coef	SE Coef	T	P
Constant	5.653	9.392	0.60	0.555
Weight	1.0387	0.1927	5.39	0.000
BSA	5.831	6.063	0.96	0.350

- the regression of the response ***y = BP*** on the predictors ***x3= BSA*** and ***x2= Weight*** (in that order):

The regression equation is
BP = 5.65 + 5.83 BSA + 1.04 Weight

Predictor	Coef	SE Coef	T	P
Constant	5.653	9.392	0.60	0.555
BSA	5.831	6.063	0.96	0.350
Weight	1.0387	0.1927	5.39	0.000

What happens if the predictor variables are highly correlated?

- Let's summarize the results in a table:

Model	b_2	$se(b_2)$	b_3	$se(b_3)$	Seq SS
x_2 only	1.2009	0.0930	---	---	$SSR(x_2)$ 505.47
x_3 only	---	---	34.443	4.690	$SSR(x_3)$ 419.86
x_2, x_3 (in order)	1.0387	0.1927	5.831	6.063	$SSR(x_3 \mid x_2)$ 2.81
x_3, x_2 (in order)	1.0387	0.1927	5.831	6.063	$SSR(x_2 \mid x_3)$ 88.43

Effects of multicollinearity

- **Effect #1. When predictor variables are correlated, the estimated regression coefficient of any one variable depends on which other predictor variables are included in the model**

Variables in model	b_2	b_3
x_2	1.20	---
x_3	---	34.4
x_2, x_3	1.04	5.83

- If ***x3*** = ***BSA*** is the only predictor included in our model, we claim that for every additional one square meter increase in body surface area (*BSA*), bloodpressure (*BP*) increases by 34.4 mm Hg.
- On the other hand, if ***x2*** = ***Weight*** and ***x3*** = ***BSA*** are both included in our model, we claim that for every additional one square meter increase in body surface area (*BSA*), bloodpressure (*BP*) increases by only 5.83 mm Hg.

Effects of multicollinearity

- **Effect #2. When predictor variables are correlated, the precision of the estimated regression coefficients decreases as more predictor variables are added to the model**

Variables in model	$se(b_2)$	$se(b_3)$
x_2	0.093	---
x_3	---	4.69
x_2, x_3	0.193	6.06

- The standard error for the estimated slope *b*2 obtained from the model including both ***x2*** = ***Weight*** and ***x3*** = ***BSA*** is about double the standard error for the estimated slope *b*2 obtained from the model including only ***x2*** = ***Weight***.
- the standard error for the estimated slope *b*3 obtained from the model including both ***x2*** = ***Weight*** and ***x3*** = ***BSA*** is about 30% larger than the standard error for the estimated slope *b*3 obtained from the model including only ***x3*** = ***BSA***.

Effects of multicollinearity

- **Effect #3.** When predictor variables are correlated, the marginal contribution of any one predictor variable in reducing the error sum of squares varies depending on which other variables are already in the model.
- regressing the response y = *BP* on the predictor $x2$ = *Weight*, we obtain $SSR(x2)$ = 505.47.
- regressing the response y = *BP* on the two predictors $x3$ = *BSA* and $x2$ = *Weight* (in that order), we obtain $SSR(x2|x3)$ = 88.43.

Effects of multicollinearity

- **Effect #4. When predictor variables are correlated, hypothesis tests for βk = 0 may yield different conclusions depending on which predictor variables are in the model. (This effect is a direct consequence of the three previous effects.)**
- **The regression of the response *y = BP* on the predictor *x3 = BSA*:**

The regression equation is
BP = 45.2 + 34.4 BSA

Predictor	Coef	SE Coef	T	P
Constant	45.183	9.392	4.81	0.000
BSA	34.443	4.690	7.34	0.000

- The regression of the response ***y* = *BP*** on the predictor ***x2* = *Weight***:

The regression equation is
BP = 2.21 + 1.20 Weight

Predictor	Coef	SE Coef	T	P
Constant	2.205	8.663	0.25	0.802
Weight	1.20093	0.09297	12.92	0.000

And, the regression of the response ***y* = *BP*** on the predictors ***x2* = *Weight*** and ***x3* = *BSA***

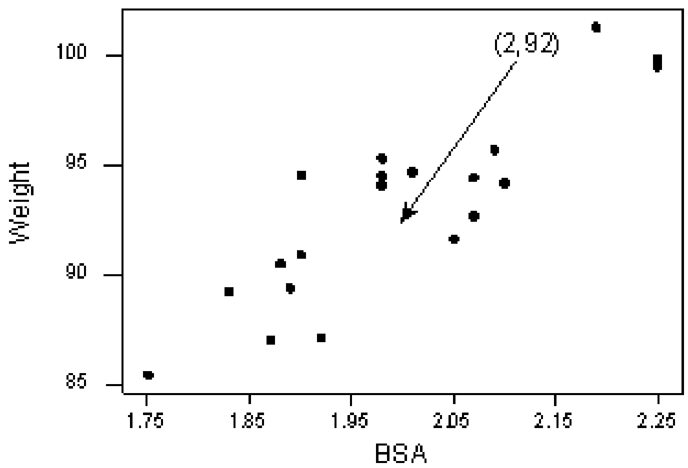
Effects of multicollinearity

- And, the regression of the response ***y = BP*** on the predictors ***x2 = Weight*** and ***x3 = BSA***

```
The regression equation is
BP = 5.65 + 1.04 Weight + 5.83 BSA

Predictor     Coef     SE Coef          T        P
Constant     5.653       9.392       0.60    0.555
Weight      1.0387      0.1927       5.39    0.000
BSA          5.831       6.063       0.96    0.350
```

Effects of multicollinearity

- **Effect #5.** High multicollinearity among predictor variables does not prevent good, precise predictions of the response within the scope of the model.



Weight	Fit	SE Fit	95.0% CI	95.0% PI
92	112.7	0.402	(111.85,113.54)	(108.94,116.44)

BSA	Fit	SE Fit	95.0% CI	95.0% PI
2	114.1	0.624	(112.76,115.38)	(108.06,120.08)

BSA	Weight	Fit	SE Fit	95.0% CI	95.0% PI
2	92	112.8	0.448	(111.93,113.83)	(109.08, 116.68)

Detection of collinearity

- **Variance inflation factor (VIF)**
- **For the model in which x_k is the only predictor**

$$y_i = \beta_0 + \beta_k x_{ik} + \varepsilon_i$$

- it can be shown that the variance of the estimated coefficient b_k is:

$$Var(b_k)_{\min} = \frac{\sigma^2}{\sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2}$$

- Let's consider such a model with correlated predictors:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$$

- It can be shown that the variance of b_k is

$$Var(b_k) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2} \times \frac{1}{1 - R_k^2}$$

Variance inflation factor (VIF)

- **How much larger? To answer this question, all we need to do is take the ratio of the two variances. Doing so, we obtain:**

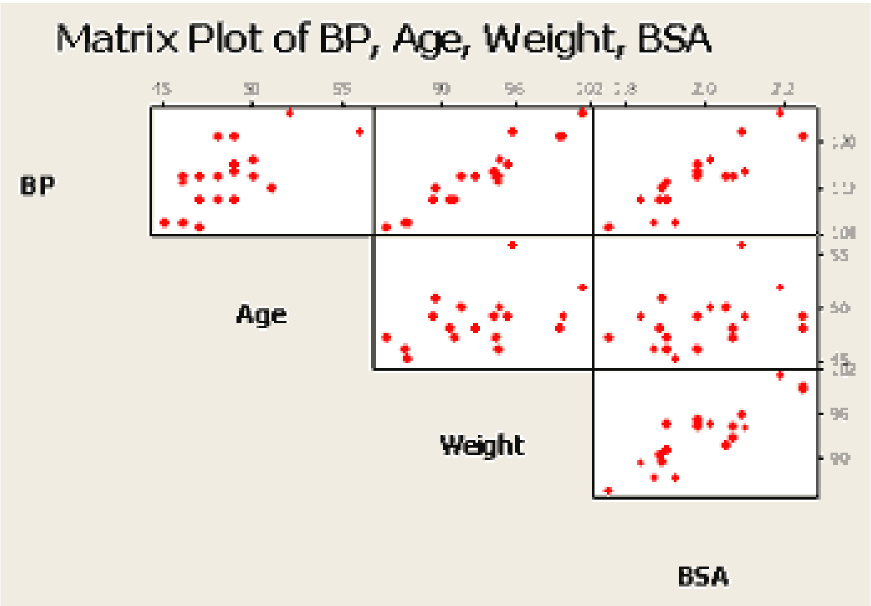
$$\frac{Var(b_k)}{Var(b_k)_{min}} = \frac{\left(\frac{\sigma^2}{\sum (x_{ik} - \bar{x}_k)^2} \times \frac{1}{1 - R_k^2}\right)}{\left(\frac{\sigma^2}{\sum (x_{ik} - \bar{x}_k)^2}\right)} = \frac{1}{1 - R_k^2}$$

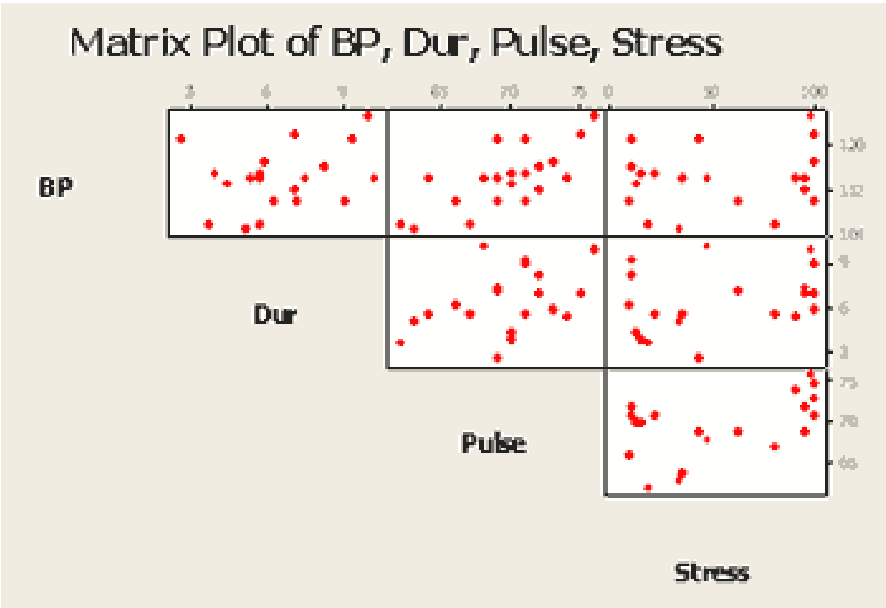
- The above quantity is what is deemed the variance inflation factor for the *kth* predictor. That is:

$$VIF_k = \frac{1}{1 - R_k^2}$$

Where R_k^2 is the R^2-value obtained by regressing the k^{th} predictor on the remaining predictors.

VIF - Example





	BP	Age	Weight	BSA	Duration	Pulse
Age	0.659					
Weight	0.950	0.407				
BSA	0.866	0.378	0.875			
Duration	0.293	0.344	0.201	0.131		
Pulse	0.721	0.619	0.659	0.465	0.402	
Stress	0.164	0.368	0.034	0.018	0.312	0.506

VIF - Example

- Regressing *y* = BP on all six of the predictors and asking Minitab to report the variance inflation factors, we obtain

Predictor	Coef	SE Coef	T	P	VIF
Constant	-12.870	2.557	-5.03	0.000	
Age	0.70326	0.04961	14.18	0.000	1.8
Weight	0.96992	0.06311	15.37	0.000	8.4
BSA	3.776	1.580	2.39	0.033	5.3
Dur	0.06838	0.04844	1.41	0.182	1.2
Pulse	-0.08448	0.05161	-1.64	0.126	4.4
Stress	0.005572	0.003412	1.63	0.126	1.8

S = 0.4072   R-Sq = 99.6%   R-Sq(adj) = 99.4%

VIF - Example

- Now, let's verify the calculation of the VIF for the predictor *Weight*. Regressing the predictor *x2* = *Weight* on the remaining five predictors

Predictor	Coef	SE Coef	T	P	VIF
Constant	19.674	9.465	2.08	0.057	
Age	-0.1446	0.2065	-0.70	0.495	1.7
BSA	21.422	3.465	6.18	0.000	1.4
Dur	0.0087	0.2051	0.04	0.967	1.2
Pulse	0.5577	0.1599	3.49	0.004	2.4
Stress	-0.02300	0.01308	-1.76	0.101	1.5

S = 1.725 R-Sq = 88.1% R-Sq(adj) = 83.9%

$$VIF_{Weight} = \frac{Var(b_{Weight})}{Var(b_{Weight})_{\min}} = \frac{1}{1 - R^2_{Weight}} = \frac{1}{1 - 0.881} = 8.4$$

What to do with multicollinearity

- Reducing data-based multicollinearity
- Reducing structural multicollinearity
- The hierarchical approach to model fitting