Chapter 14

Multicollinearity

Chanwoo Yoo, Division of Advanced Engineering,
Korea National Open University

Contents

1. Multicollinearity

1. Multicollinearity

- Multicollinearity exists when two or more of the predictors in a regression model are moderately or highly correlated.

2. Multicollinearity Pitfalls

- When multicollinearity exists, any of the following pitfalls can be exacerbated:
  - The estimated regression coefficient of any one variable depends on which other predictors are included in the model
  - The precision of the estimated regression coefficients decreases as more predictors are added to the model

2. Multicollinearity Pitfalls

- When multicollinearity exists, any of the following pitfalls can be exacerbated:
  - The marginal contribution of any one predictor variable in reducing the error sum of squares depends on which other predictors are already in the model
  - Hypothesis tests for $\beta_k = 0$ may yield different conclusions depending on which predictors are in the model

3. 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** 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.

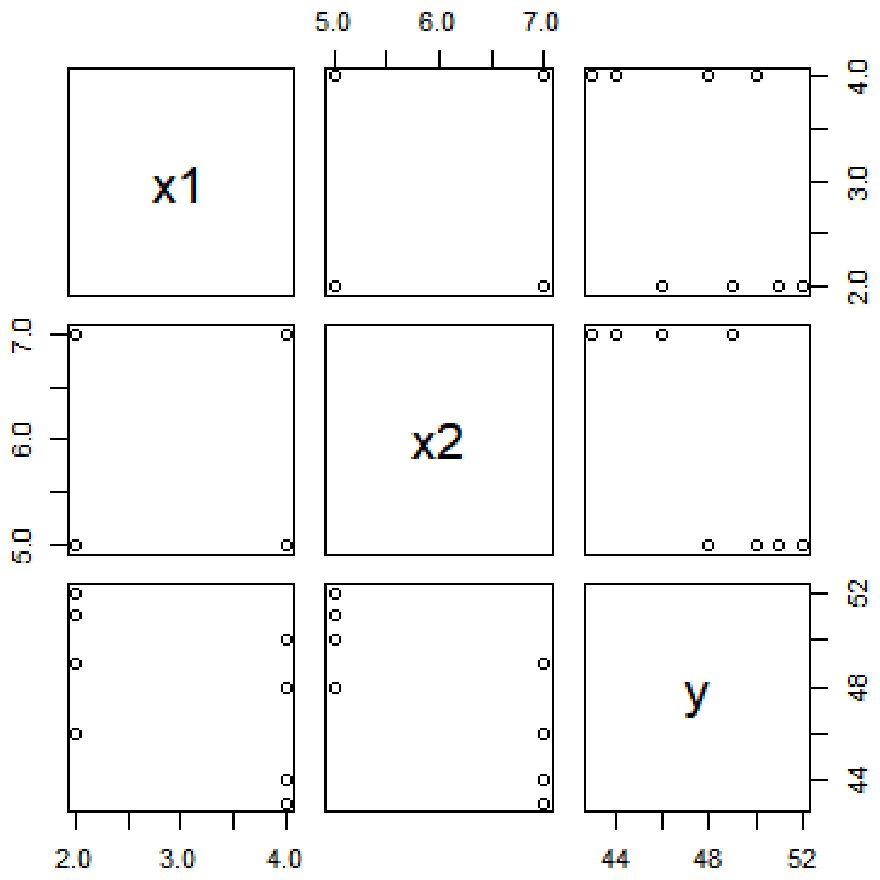
2. Uncorrelated Predictors

1. Uncorrelated Predictors

```
# Uncorrelated Predictors data set
uncorrpreds <- read.table("uncorrpreds.txt", header=T)
attach(uncorrpreds)

pairs(uncorrpreds)
```

1. Uncorrelated Predictors



- There is no apparent relationship at all between the predictors x_1 and x_2. That is, the correlation between x_1 and x_2 is zero.

```
> cor(x1,x2)
[1] 0
```

1. Uncorrelated Predictors

```
model.1 <- lm(y ~ x1)
summary(model.1)
anova(model.1)

model.2 <- lm(y ~ x2)
summary(model.2)
anova(model.2)
```

1. Uncorrelated Predictors

```
model.12 <- lm(y ~ x1 + x2)
summary(model.12)
anova(model.12)

model.21 <- lm(y ~ x2 + x1)
summary(model.21)
anova(model.21)
```

1. Uncorrelated Predictors

```
> model.1 <- lm(y ~ x1)
> summary(model.1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   52.750      3.346  15.764 4.13e-06 ***
x1            -1.625      1.058  -1.536    0.176
---
Residual standard error: 2.993 on 6 degrees of freedom
Multiple R-squared:  0.2821,  Adjusted R-squared:  0.1625
F-statistic: 2.358 on 1 and 6 DF,  p-value: 0.1755
```

1. Uncorrelated Predictors

```
> anova(model.1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
x1         1 21.125 21.1250  2.3581 0.1755
Residuals  6 53.750  8.9583
```

1. Uncorrelated Predictors

```
> model.2 <- lm(y ~ x2)
> summary(model.2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.1250     4.7888  12.973 1.29e-05 ***
x2           -2.3750     0.7873  -3.017   0.0235 *
---
Residual standard error: 2.227 on 6 degrees of freedom
Multiple R-squared:  0.6027,  Adjusted R-squared:  0.5364
F-statistic: 9.101 on 1 and 6 DF,  p-value: 0.02349
```

1. Uncorrelated Predictors

```
> anova(model.2)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x2         1 45.125  45.125  9.1008 0.02349 *
Residuals  6 29.750   4.958
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Uncorrelated Predictors

```
> model.12 <- lm(y ~ x1 + x2)
> summary(model.12)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.0000     3.1494  21.274 4.25e-06 ***
x1           -1.6250     0.4644  -3.499  0.01729 *
x2           -2.3750     0.4644  -5.115  0.00372 **
---
Residual standard error: 1.313 on 5 degrees of freedom
Multiple R-squared:  0.8848,   Adjusted R-squared:  0.8387
F-statistic:  19.2 on 2 and 5 DF,  p-value: 0.004504
```

1. Uncorrelated Predictors

```
> anova(model.12)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x1         1 21.125  21.125  12.246 0.017294 *
x2         1 45.125  45.125  26.159 0.003724 **
Residuals  5  8.625   1.725
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Uncorrelated Predictors

```
> model.21 <- lm(y ~ x2 + x1)
> summary(model.21)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.0000     3.1494  21.274 4.25e-06 ***
x2           -2.3750     0.4644  -5.115  0.00372 **
x1           -1.6250     0.4644  -3.499  0.01729 *
---
Residual standard error: 1.313 on 5 degrees of freedom
Multiple R-squared:  0.8848,  Adjusted R-squared:  0.8387
F-statistic:  19.2 on 2 and 5 DF,  p-value: 0.004504
```

1. Uncorrelated Predictors

```
> anova(model.21)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x2         1 45.125  45.125  26.159 0.003724 **
x1         1 21.125  21.125  12.246 0.017294 *
Residuals  5  8.625   1.725
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Uncorrelated Predictors

Model	b_1	$se(b_1)$	b_2	$se(b_2)$	Seq SS
x_1 only	-1.625	1.058			$SSR(x_1) = 21.125$
x_2 only			-2.375	0.7873	$SSR(x_2) = 45.125$
x_1, x_2 (in order)	-1.625	0.4644	-2.375	0.4644	$SSR(x_2 \mid x_1) = 45.125$
x_2, x_1 (in order)	-1.625	0.4644	-2.375	0.4644	$SSR(x_1 \mid x_2) = 21.125$

2. Observations

- The estimated slope coefficients b_1 and b_2 are the same regardless of the model used.
- The sum of squares $SSR(x_1)$ is the same as the sequential sum of squares $SSR(x_1|x_2)$.
- The sum of squares $SSR(x_2)$ is the same as the sequential sum of squares $SSR(x_2|x_1)$.

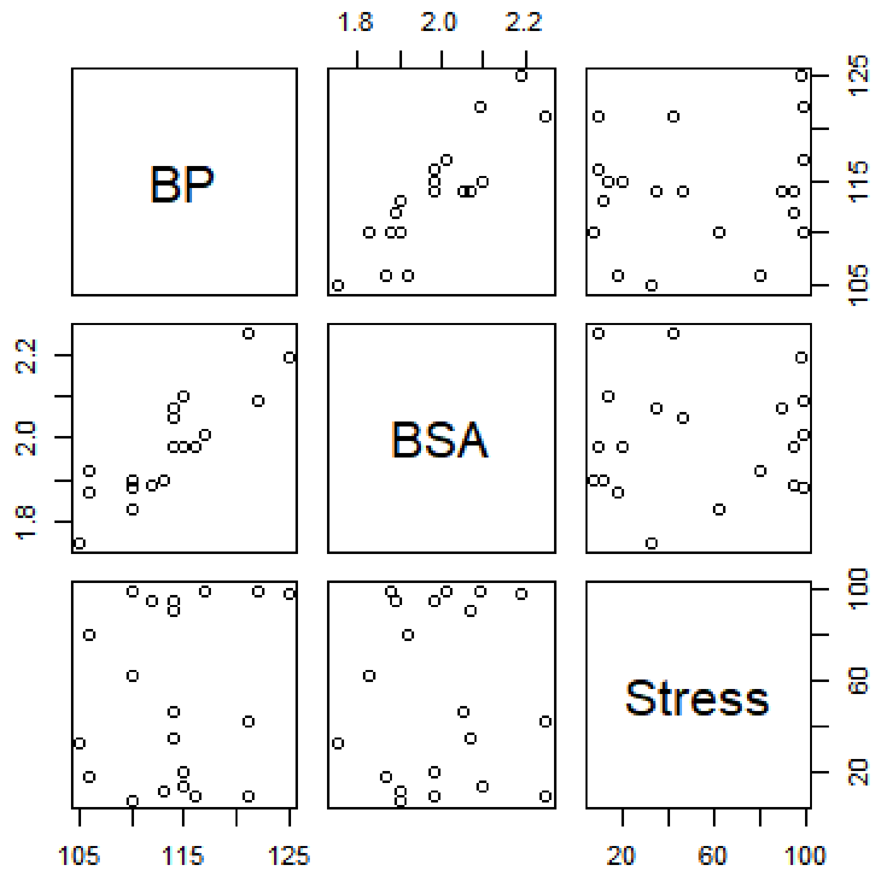
3. Data: Blood Pressure

- Data: Blood Pressure
  - y (BP): blood pressure in mm Hg
  - x_1 (Age): age in years
  - x_2 (Weight): weight in kg
  - x_3 (BSA): body surface area in sq m
  - x_4 (Dur): duration of hypertension in years

3. Data: Blood Pressure

- x_5 (Pulse) : basal pulse in beats per minute
- x_6 (Stress) : stress index

4. Nearly Uncorrelated Predictors



- There appears to be a strong relationship between BP and the predictor = BSA, a weak relationship between BP and Stress, and an almost non-existent relationship between BSA and Stress.

4. Nearly Uncorrelated Predictors

```
> round(cor(bloodpress[,c(2:8)]),3)
          BP   Age Weight   BSA   Dur Pulse Stress
BP     1.000 0.659  0.950 0.866 0.293 0.721  0.164
Age    0.659 1.000  0.407 0.378 0.344 0.619  0.368
Weight 0.950 0.407  1.000 0.875 0.201 0.659  0.034
BSA    0.866 0.378  0.875 1.000 0.131 0.465  0.018
Dur    0.293 0.344  0.201 0.131 1.000 0.402  0.312
Pulse  0.721 0.619  0.659 0.465 0.402 1.000  0.506
Stress 0.164 0.368  0.034 0.018 0.312 0.506  1.000
```

4. Nearly Uncorrelated Predictors

```
model.1 <- lm(y ~ x1)
summary(model.1)
anova(model.1)

model.2 <- lm(y ~ x2)
summary(model.2)
anova(model.2)
```

4. Nearly Uncorrelated Predictors

```
model.12 <- lm(BP ~ Stress + BSA)
summary(model.12)
anova(model.12)

model.21 <- lm(BP ~ BSA + Stress)
summary(model.21)
anova(model.21)
```

4. Nearly Uncorrelated Predictors

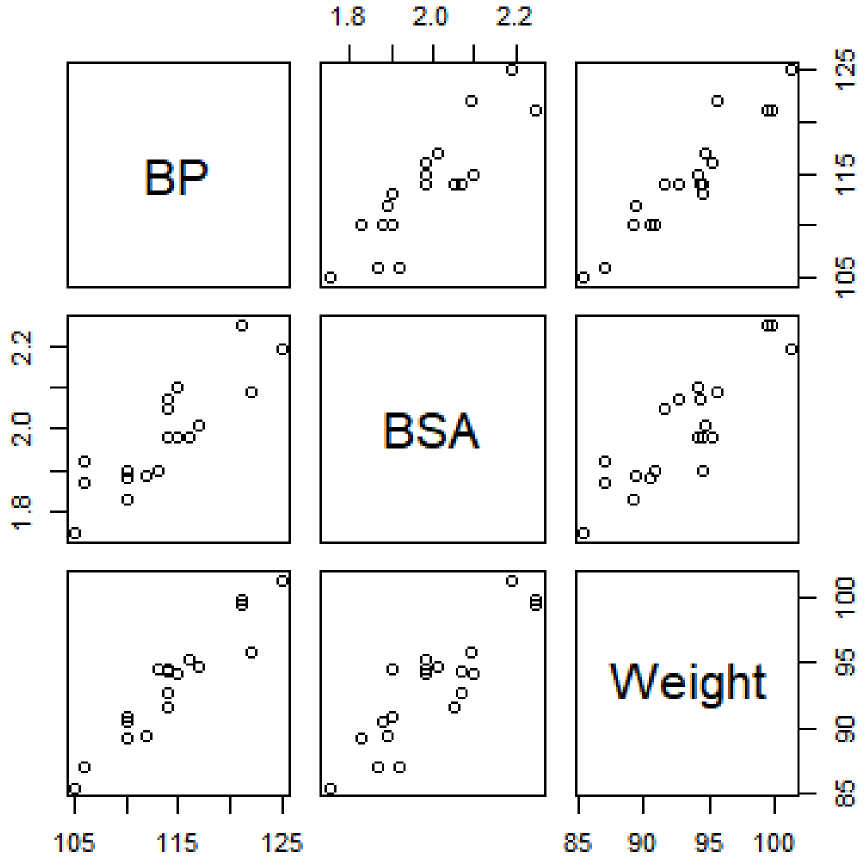
Model	b_6	$se(b_6)$	b_3	$se(b_3)$	Seq SS
x_6 only	0.02399	0.03404			$SSR(x_6) = 15.04$
x_3 only			34.443	4.690	$SSR(x_3) = 419.86$
x_6, x_3 (in order)	0.02166	0.01697	34.334	4.611	$SSR(x_3 \| x_6) = 417.07$
x_3, x_6 (in order)	0.02166	0.01697	34.334	4.611	$SSR(x_6 \| x_3) = 12.26$

5. Observations

- 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 $SSR(x_6|x_3)$.

3. Highly Correlated Predictors

1. Highly Correlated Predictors



- There appears to be not only a strong relationship between BP and Weight and a strong relationship between BP and the predictor BSA, but also a strong relationship between the two predictors Weight and BSA.

1. Highly Correlated Predictors

```
> round(cor(bloodpress[,c(2,5,4)]),3)
          BP   BSA Weight
BP     1.000 0.866  0.950
BSA    0.866 1.000  0.875
Weight 0.950 0.875  1.000
```

1. Highly Correlated Predictors

```
model.1 <- lm(BP ~ Weight)
summary(model.1)
anova(model.1)

model.2 <- lm(BP ~ BSA)
summary(model.2)
anova(model.2)
```

1. Highly Correlated Predictors

```
model.12 <- lm(BP ~ Weight + BSA)
summary(model.12)
anova(model.12)

model.21 <- lm(BP ~ BSA + Weight)
summary(model.21)
anova(model.21)
```

1. Highly Correlated Predictors

Model	b_2	$se(b_2)$	b_3	$se(b_3)$	Seq SS
x_2 only	1.20093	0.09297			$SSR(x_2) = 505.47$
x_3 only			34.443	4.690	$SSR(x_3) = 419.86$
x_2, x_3 (in order)	1.0387	0.1927	5.8313	6.0627	$SSR(x_3 \mid x_2) = 2.81$
x_3, x_2 (in order)	1.0387	0.1927	5.8313	6.0627	$SSR(x_2 \mid x_3) = 88.43$

2. Observation1

- We get wildly different estimates of the slope parameters b_2 and b_3.
- If BSA is the only predictor included in our model, we claim that for every additional one square meter increase in body surface area (BSA), blood pressure (BP) increases by 34.4 mm Hg.

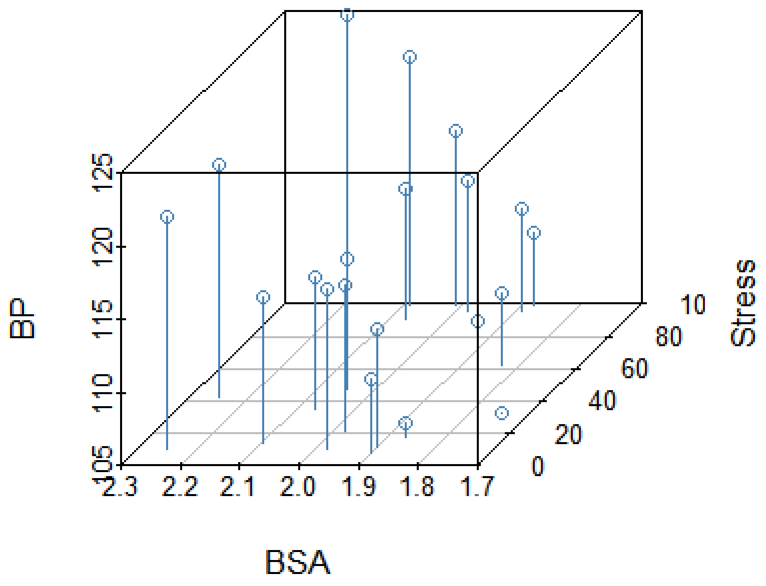
2. Observation1

- On the other hand, if Weight and BSA are both included in our model, we claim that for every additional one square meter increase in body surface area (BSA), holding weight constant, blood pressure (BP) increases by only 5.83 mm Hg.

3. Observation 2

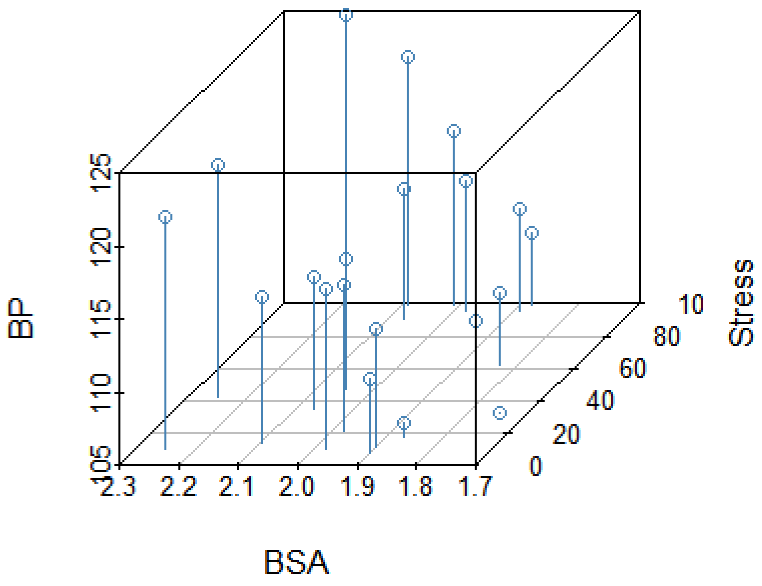
- The standard error for the estimated slope b_2 obtained from the model including both Weight and BSA is about double the standard error for the estimated slope b_2 obtained from the model including only Weight. And, the standard error for the estimated slope b_3 obtained from the model including both Weight and BSA is about 30% larger than the standard error for the estimated slope b_3 obtained from the model including only BSA.

3. Observation 2



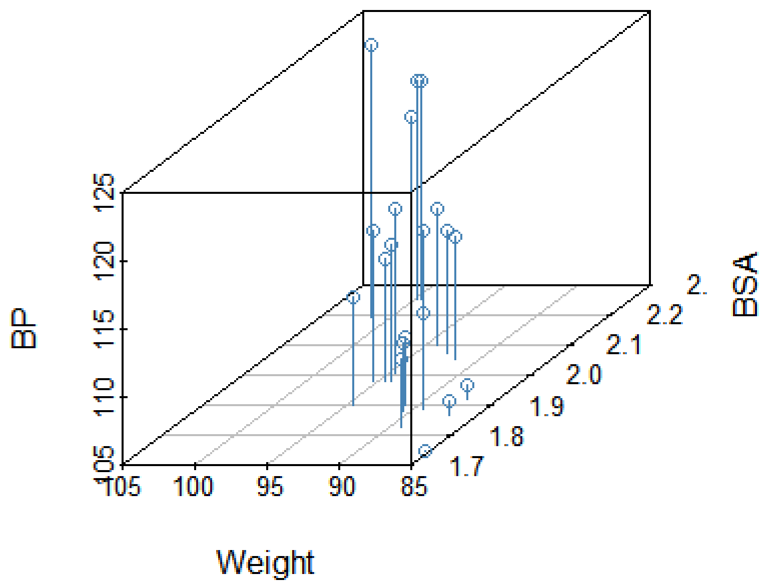
- The predictor values are spread out and just about anchored in each of four corners, providing a solid base over which to draw the response plane.
- Even if the responses varied somewhat from sample to sample, the plane couldn't change all that much because of the solid base.

3. Observation 2



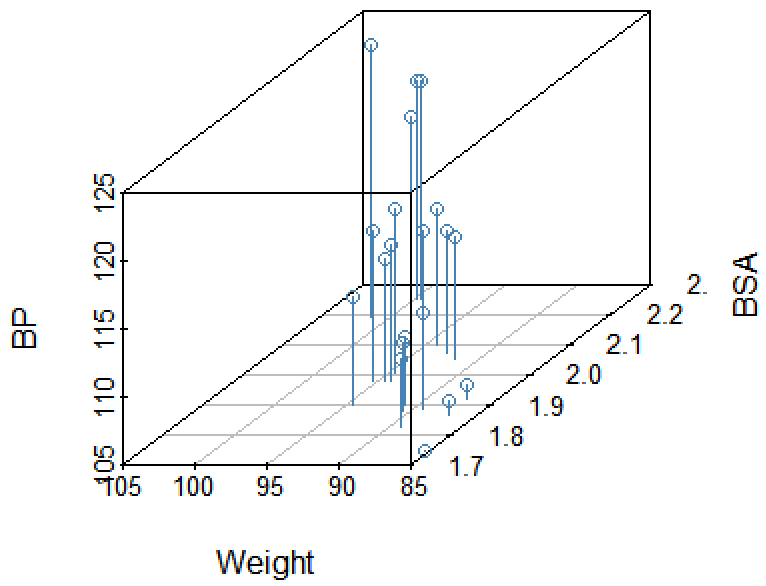
- That is, the estimated coefficients couldn't change all that much.
- The standard errors of the estimated coefficients will necessarily be small.

3. Observation 2



- The predictor values tend to fall in a straight line. That is, there are no anchors in two of the four corners.
- Therefore, the base over which the response plane is drawn is not very solid.

3. Observation 2



- If the responses varied somewhat from sample to sample, the position of the plane could change significantly.
- That is, the estimated coefficients could change substantially.
- The standard errors of the estimated coefficients will be necessarily larger.

4. Observation 3

- Because weight and body surface area are highly correlated, most of the variation in blood pressure explained by weight could just have easily been explained by body surface area. Therefore, once you take into account a person's body surface area, there's not much variation left in the blood pressure for weight to explain.

4. Observation 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.

5. Observation 4

- When predictor variables are correlated, hypothesis tests for coefficients may yield different conclusions depending on which predictor variables are in the model.

5. Observation 4

```
> model.1 <- lm(BP ~ Weight)
> summary(model.1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.20531    8.66333   0.255    0.802
Weight       1.20093    0.09297  12.917 1.53e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. Observation 4

```
> model.2 <- lm(BP ~ BSA)
> summary(model.2)

Coefficients:
            Estimate Std. Error t value Pr(>¦t¦)
(Intercept)   45.183      9.392   4.811  0.00014 ***
BSA           34.443      4.690   7.343 8.11e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5. Observation 4

```
> model.12 <- lm(BP ~ Weight + BSA)
> summary(model.12)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.6534     9.3925   0.602    0.555
Weight        1.0387     0.1927   5.392 4.87e-05 ***
BSA           5.8313     6.0627   0.962    0.350
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. Observation 5

- High multicollinearity among predictor variables does not prevent good, precise predictions of the response within the scope of the model.
- Geometrically, the best fitting plane through the responses may tilt from side to side from sample to sample (because of the correlation), but the center of the plane (in the scope of the model) won't change all that much.

6. Observation 5

```
# The following output illustrates how the predictions don't
change all that much from model to model:

> predict(model.1, interval="prediction",
+         newdata=data.frame(Weight=92))
      fit     lwr     upr
1 112.691 108.938 116.444
> predict(model.12, interval="prediction",
+         newdata=data.frame(Weight=92, BSA=2))
       fit      lwr      upr
1 112.8794 109.0801 116.6787
```

6. Observation 5

```
# The following output illustrates how the predictions don't
change all that much from model to model:

> predict(model.2, interval="prediction",
+         newdata=data.frame(BSA=2))
       fit      lwr      upr
1 114.0689 108.0619 120.0758
> predict(model.12, interval="prediction",
+         newdata=data.frame(Weight=92, BSA=2))
       fit      lwr      upr
1 112.8794 109.0801 116.6787
```

7. Conclusion

- In the presence of multicollinearity:
  - It is okay to use an estimated regression model to predict y or estimate μ_Y as long as you do so within the scope of the model.

7. Conclusion

- In the presence of multicollinearity:
  - We can no longer make much sense of the usual interpretation of a slope coefficient as the change in the mean response for each additional unit increase in the predictor x_k, when all the other predictors are held constant, since changing one predictor necessarily would change the values of the others.

4. Detecting Multicollinearity

1. Variance Inflation Factor (VIF)

- $VIF_k = \frac{1}{1-R_k^2}$
  - R_k^2: R^2 value obtained by regressing the kth predictor on the remaining predictors.

2. Calculating VIF

```
install.packages('car')
library(car)
model.1 <- lm(BP ~ Age + Weight + BSA + Dur + Pulse + Stress)
vif(model.1)
```

```
> vif(model.1)
     Age   Weight      BSA      Dur    Pulse   Stress
1.762807 8.417035 5.328751 1.237309 4.413575 1.834845
```

2. Calculating VIF

```
> model.2 <- lm(Weight ~ Age + BSA + Dur + Pulse + Stress)
> summary(model.2)

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
```

$$VIF_{Weight} = \frac{1}{1 - 0.8812} = 8.417$$

3. Dealing with Multicollinearity

- One solution to dealing with multicollinearity is to remove some of the violating predictors from the model.

3. Dealing with Multicollinearity

```
install.packages("corrplot")
library(corrplot)
corrplot(round(cor(bloodpress[,c(2:8)]),2), method="number")
```

3. Dealing with Multicollinearity

	BP	Age	Weight	BSA	Dur	Pulse	Stress
BP	1.00	0.66	0.95	0.87	0.29	0.72	0.16
Age	0.66	1.00	0.41	0.38	0.34	0.62	0.37
Weight	0.95	0.41	1.00	0.88	0.20	0.66	0.03
BSA	0.87	0.38	0.88	1.00	0.13	0.46	0.02
Dur	0.29	0.34	0.20	0.13	1.00	0.40	0.31
Pulse	0.72	0.62	0.66	0.46	0.40	1.00	0.51
Stress	0.16	0.37	0.03	0.02	0.31	0.51	1.00

- We see that the predictors Weight and BSA are highly correlated (r = 0.88).
- We can choose to remove either predictor from the model. The decision of which one to remove is often a scientific or practical one.

3. Dealing with Multicollinearity

1. Choose the two predictors which show the largest absolute value of pairwise correlation.
2. Remove the predictor with the smaller correlation with y among the two predictors.
3. Repeat steps 1–2 until no absolute correlations are above the threshold.

3. Dealing with Multicollinearity

```
> model.3 <- lm(BP ~ Age + Weight + Dur + Stress)
> vif(model.3)
     Age   Weight      Dur   Stress
1.468245 1.234653 1.200060 1.241117
```

3. Dealing with Multicollinearity

```
> summary(model.3)
...
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

> summary(model.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
```

Next

Chapter 15
Generalized Linear Model