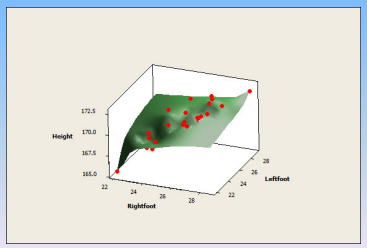
MULTICOLLINEARITY IN MULTIPLE LINEAR REGRESSION



After studying this lecture, you should be able to:

• determine if your fitted model suffers from multicollinearity or not,

• compute Variance Inflation Factors (VIF), and discuss how VIF is related to multicollinearity,

• fit a linear model avoiding the problem of multicollinearity,

•discuss other ways in which the problem of multicollinearity can be addressed.

Multicollinearity

•The predictors in an MLR model ideally should be uncorrelated with each other.

•Low correlations among predictors do not cause problems in estimating coefficients of MLR models.

•High Low correlations among predictors cause serious problems in estimating coefficients of MLR models. This is referred to as multicollinearity.

Example 1 (Height vs left-foot and right-foot lengths)
The heights, left-foot and right-foot lengths of 25 subjects are given. Find a linear regression model for Height as a function of the left-foot and right-foot lengths.

Use data file Height vs leftfoot rightfoot.csv

Leftfoot	Rightfoot	Height
23.5	23.6	167.7
25.2	25.7	169.7
24.1	23.5	169
27.1	26.7	172.3
29	28.6	172.1
25.8	26.9	171.1
27.4	26.6	171.7
22	22.1	165.5
25.8	26.3	170.2
26.3	25.4	169.5
23.2	23.9	169.2
23.4	24	167.7
24	24	168.2
25.6	24.3	171
26.4	26.8	171.9
26	26.4	170.5
25.3	24.4	169.4
27.1	27.4	171.3
25.6	25.3	169.5
25.8	25.4	170.8
25.8	25.3	169.7
27.5	26.8	170.5
26.1	26.1	170.2
25.1	24.9	168.6
25.7	25.8	172.7

Regression Analysis: Height versus Leftfoot, Rightfoot

```
x <- read.csv("Height vs leftfoot rightfoot.csv",
header=TRUE)
lm1 <- lm(x$Height~x$Leftfoot+x$Rightfoot)
library(HH)
vif(lm1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 145.6668     3.2759  44.466  <2e-16 ***
x$Leftfoot    0.4978     0.3210   1.551   0.135
x$Rightfoot   0.4563     0.3372   1.353   0.190
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9275 on 22 degrees of freedom
Multiple R-squared: 0.7189,     Adjusted R-squared: 0.6933
F-statistic: 28.13 on 2 and 22 DF,  p-value: 8.666e-07
```


Note that each of the β's are not significant at 5% level, but the P-value for overall F-test is much smaller than .05, indicating the fitted model is significant. This is one indication of multicollinearity. Some other indicators of multicollinearity are:

•Unusually large SE's of estimated coefficients,

•Counterintuitive signs of estimated coefficients,

What causes multicollinearity?

Recall the OLS estimation of the linear model $\underline{y} = X\beta$

OLS estimate:

$\hat{\underline{\beta}} = (X^T X)^{-1} X^T \underline{y}$

$\text{Var}(\hat{\beta}_i) = \hat{\sigma}^2 \times C_{ii}$

where

C_{ii} = i-th diagonal of $(X^T X)^{-1}$

When two of the predictors are highly correlated (or one of the predictors is highly correlated with some linear combination of the other predictors), the determinant of the matrix X^TX is very small, and its inversion (involves division by determinant which is close to 0) becomes inaccurate or impossible.

Variance Inflation Factor (VIF)

$VIF(\hat{\beta}_i) = \frac{1}{1 - R_i^2}$

where

$R_i^2 = R^2$ – value when a model is fitted for X_i as a function of all other predictors.

Clearly high R_i^2 value is associated with high VIF-values.

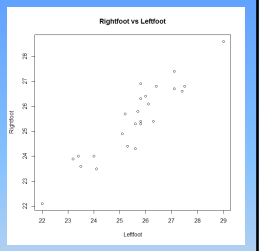
Computing VIF in R:
Install package HH
library(HH)
vif(lm1) # see slide 5 for the linear model lm1

x$Leftfoot x$Rightfoot
6.929271 6.929271

Values of VIF much >1 indicate multicollinearity, and VIF much larger than 10 indicate serious problems due to multicollinearity.

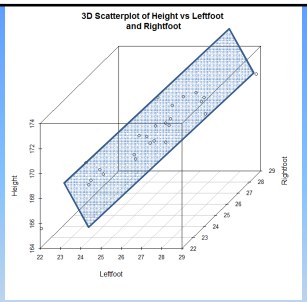
Let us plot Rightfoot vs Leftfoot in R:

```
plot(x$Leftfoot,x$Rightfoot,
xlab = "Leftfoot",
ylab="Rightfoot",
main = "Rightfoot vs Leftfoot")
```



Above graph shows that Leftfoot and Rightfoot measurements are highly correlated.

Example 1



Since the Leftfoot and Rightfoot measurements tend to fall along a line, the least squares regression equation (a plane in 3D for 2 predictors) can easily be rotated without deviating from the points– which is equivalent to high uncertainty is β-estimates.

Simplest way to deal with multicollinearity is to remove one (or more) correlated variables. For Example 1, this method yields two very similar equations:

Regression Analysis: Height versus Leftfoot

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	147.0112	3.1775	46.266	0.000***
x$Leftfoot	0.8997	0.1241	7.248	0.000

Multiple R-squared: 0.6955, Adjusted R-squared: 0.6822
F-statistic: 52.53 on 1 and 23 DF, p-value: 2.236e-07

Regression Analysis: Height versus Rightfoot

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	146.0775	3.3634	43.431	0.000
x$Rightfoot	0.9401	0.1320	7.124	0.000

Multiple R-squared: 0.6881, Adjusted R-squared: 0.6746
F-statistic: 50.75 on 1 and 23 DF, p-value: 2.954e-07