

# Multiple Regression Part I

STAT315, 19-20/3/2014

## Regression problem

Predictors/independent variables/features  $X_i$

$$\mathbf{X} = (X_1, X_2, \dots, X_p)$$

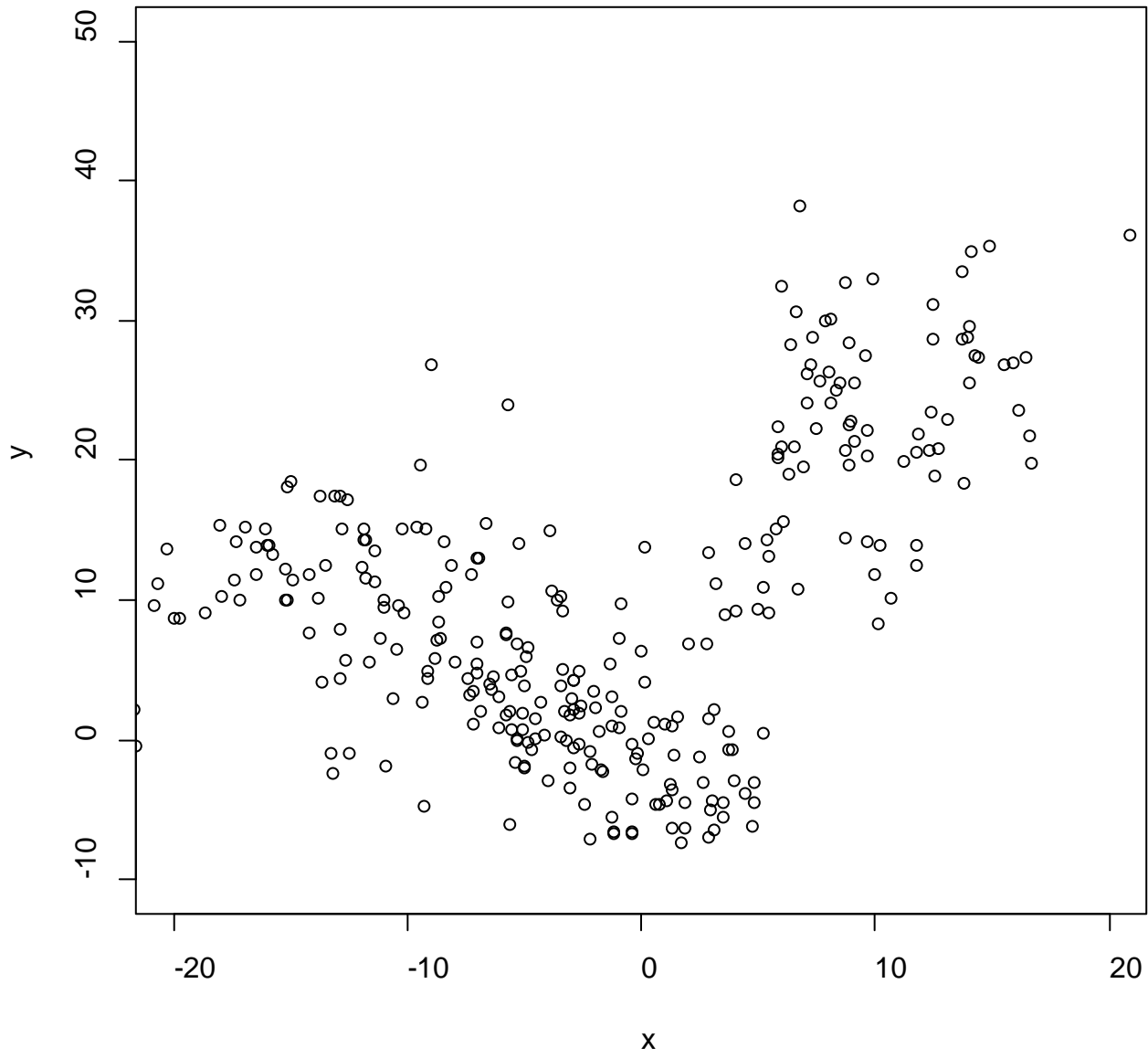
$$E[Y|\mathbf{X}] = f(\mathbf{X})$$

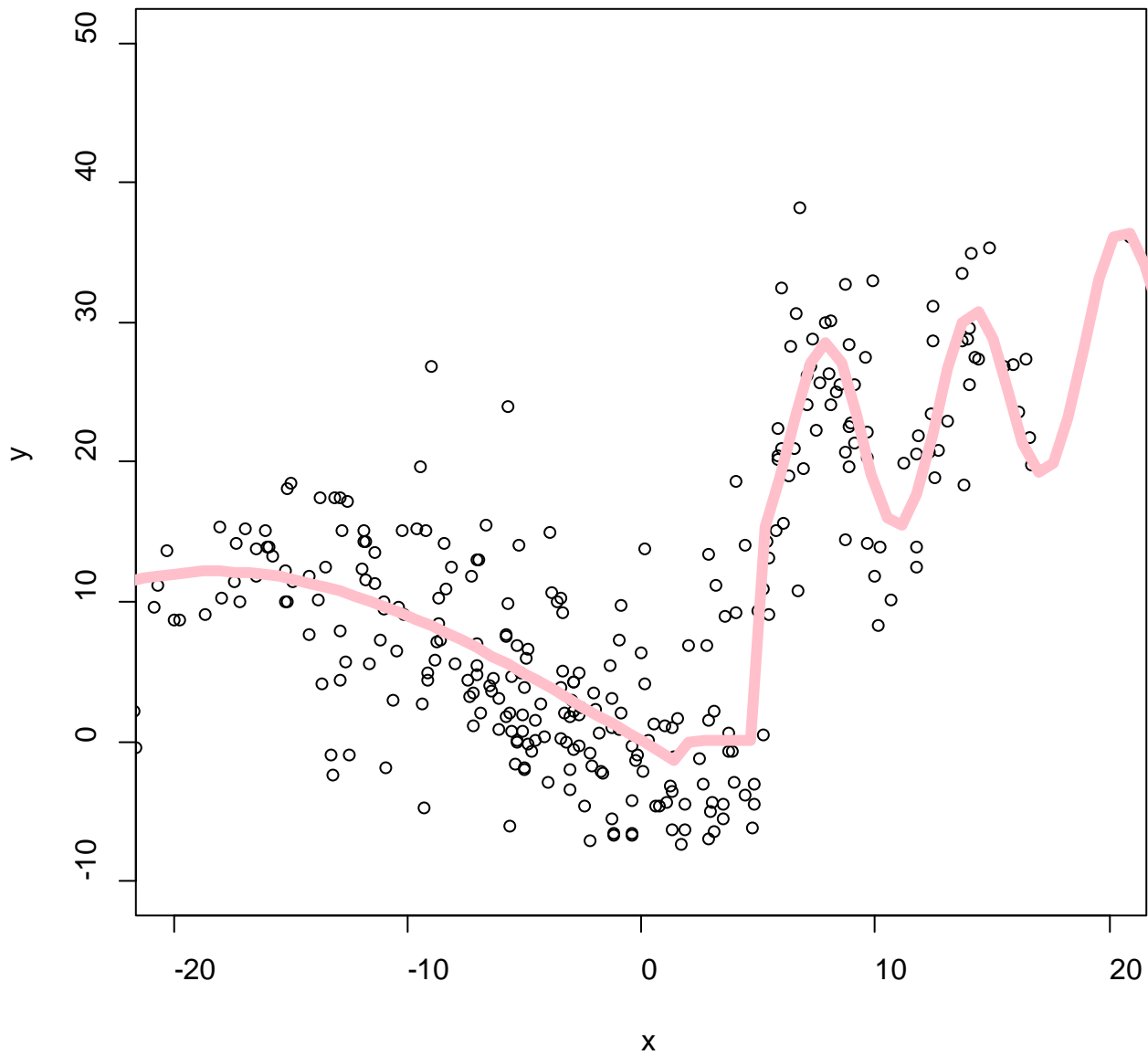
Or:

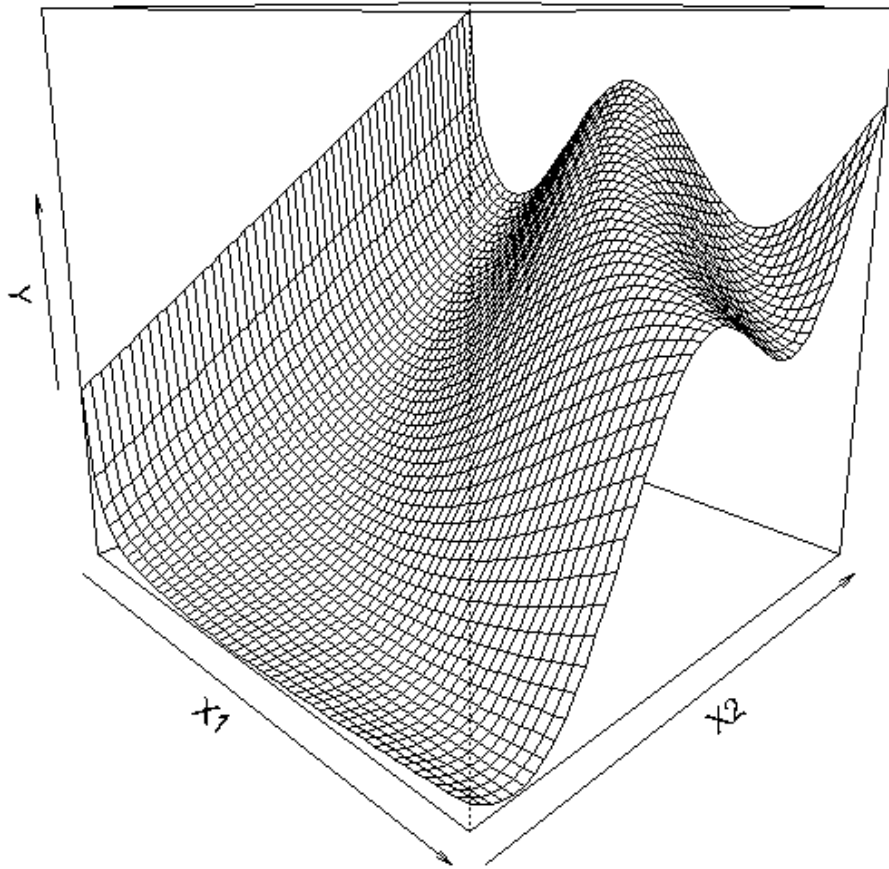
$$Y = f(\mathbf{X}) + \varepsilon$$

← **Error** which can never be eliminated. Our task is to estimate the regression function  $f$ .

**Regression function**, describing how the expected value of  $y$  is related to the  $x$ 's.







Since the regression function might be very complicated, we usually want to make some simplifying assumptions. The simplest kind of function which involves all the independent variables is a linear function.

## Linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Fitted by minimising the sum of squared errors. This leads to:

Estimate of  
beta

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Fitted values  
of Y

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

“Hat Matrix”

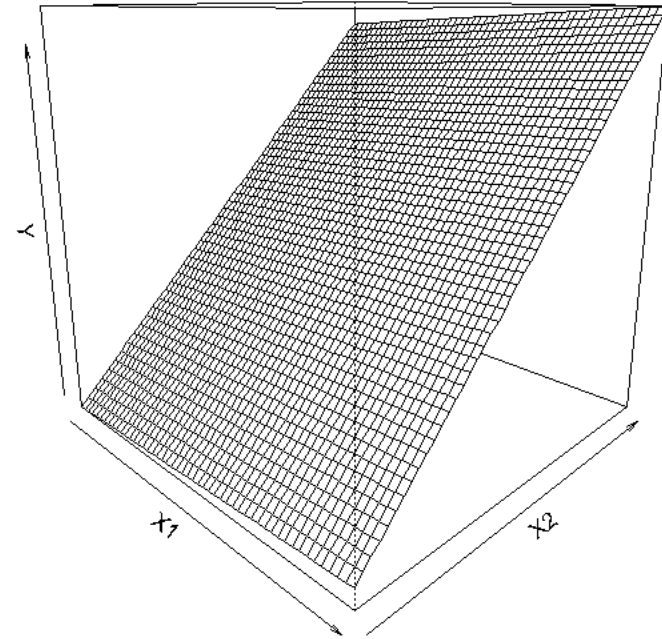
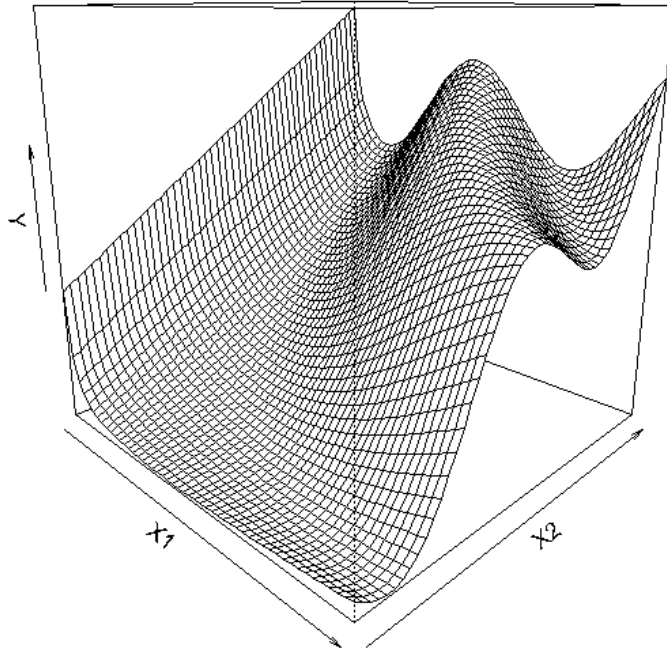
Interpretation: an increase of one unit in  $X_i$  is associated with an increase of  $\hat{\beta}_j$  units in  $Y$  provided the other independent variables are held constant (sometimes called *ceteris paribus* condition)

e.g.

Y = salary of a cricketer

X1 = length of career so far

X2 = total number of runs scored in career



If  $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_{n \times n})$  i.i.d errors

then the method of least squares is also the method of maximum likelihood, and we are able to make all sorts of inferences, such as confidence intervals. These can be calculated from the output given by software.



$$\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

Multivariate normal

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{N - (p + 1)}$$

Estimate of the variance  $\sigma^2$  of the error terms

$$\frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)} \sim t_{N-(p+1)}$$

Used to calculate confidence intervals for the regression coefficients

↑  
Calculated by software

↑  
t distribution

$$\hat{\beta}_i \pm t_{\alpha/2} SE(\hat{\beta}_i)$$

100(1 -  $\alpha$ )% *confidence interval*

```
> data(stackloss)
> model <- lm(stack.loss ~ ., data=stackloss)
> summary(model)
```

```
Call:
lm(formula = stack.loss ~ ., data = stackloss)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-7.2377 -1.7117 -0.4551  2.3614  5.6978
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.9197    11.8960  -3.356  0.00375 **
Air.Flow      0.7156     0.1349   5.307  5.8e-05 ***
Water.Temp    1.2953     0.3680   3.520  0.00263 **
Acid.Conc.   -0.1521     0.1563  -0.973  0.34405
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-squared:  0.9136,    Adjusted R-squared:  0.8983
F-statistic: 59.9 on 3 and 17 DF,  p-value: 3.016e-09
```

```
> 1.2953 + c(-1,1)*0.3680*qt(0.025, df=17)
[1] 2.0717121 0.5188879
```

95% CI for  $\beta_1$  is [0.52, 2.07] (“we are 95% confident that an increase of 1 unit in *Water.Temp* is associated with an increase of between 0.52 and 2.07 units in *stack.loss*”)

One of the most useful things about linear regression is that it can handle categorical variables as well as numerical ones. For example:

$$\text{Salary} = \beta_0 + \beta_1(\text{Years of education}) + \beta_2\text{Gender}$$

where gender is coded as 0/1. Fitting the same slope for both genders, but different intercepts; parallel regression lines.

Question: what term or terms do we need to add if we want separate slopes for the different genders as well?

- Interpretation of  $\beta_2$  ?
- What if you have a categorical variable with several categories? e.g. Apple/Pear/Orange

$$\text{FruitWeight} = \beta_0 + \beta_1\text{Apple} + \beta_2\text{Orange}$$


One category has to be treated as the baseline, in this case Pear. Interpretation: weight of Pear is  $\beta_0$ , weight of Apple is  $\beta_0 + \beta_1$ , weight of orange is  $\beta_0 + \beta_2$

## SAS example:

```
data fruit;
  input apple orange weight;
  cards;
  0 0 29
  0 0 26
  0 0 25
  1 0 22
  1 0 23
  1 0 26
  0 1 50
  0 1 34
  0 1 30
  ;
```

```
proc reg data=fruit;
  model weight = apple orange;
run;
quit;
```

Note: this “significant” p-value is completely irrelevant. Why? What is it telling us?



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	26.66667	3.66161	7.28	0.0003
apple	1	-3.00000	5.17830	-0.58	0.5834
orange	1	11.33333	5.17830	2.19	0.0712

# R<sup>2</sup>

$$\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Variance of predicted y-values /  
variance of actual y-values

- Not really the square of anything interesting, except in the one-variable case.
- Often quoted, but not necessarily a good way of measuring the quality of the fit.
- The more x-variables you add, the bigger R-squared will be.
- The more spread out the x-values, the bigger R-squared will be.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-39.9197	11.8960	-3.356	0.00375	**
Air.Flow	0.7156	0.1349	5.307	5.8e-05	***
Water.Temp	1.2953	0.3680	3.520	0.00263	**
Acid.Conc.	-0.1521	0.1563	-0.973	0.34405	

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.243 on 17 degrees of freedom  
Multiple R-squared: 0.9136, Adjusted R-squared: 0.8983  
F-statistic: 59.9 on 3 and 17 DF, p-value: 3.016e-09

## ANOVA

Nested models:

$$M1: Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

$$M2: Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \beta_{k+1} X_{k+1} + \dots + \beta_p X_p$$

$$H0 : \beta_{k+1} = \dots = \beta_p = 0$$

H1 : at least one of these  $\beta_i$  is not zero.

Under the null hypothesis, the quantity

$$\frac{(RSS(M_1) - RSS(M_2))/(p - k)}{RSS(M_2)/(n - (p + 1))}$$

has an  $F(p-k, n-(p+1))$  distribution.

Special case: M1:  $Y = \beta_0$  is the **null model**. The p-value for the F-test is usually reported in regression output, e. g.

F-statistic: 59.9 on 3 and 17 DF, p-value: 3.016e-09

Often laid out as an ANOVA table

$$\begin{array}{ccc}
 p & \sum(\hat{y}_i - \bar{y})^2 & \sum(\hat{y}_i - \bar{y})^2 / p & F \\
 (n - p - 1) & \sum(\hat{y}_i - y_i)^2 & \sum(\hat{y}_i - y_i)^2 / (n - p - 1) &
 \end{array}$$


---

$$\begin{array}{cc}
 n - 1 & \sum(\hat{y}_i - \bar{y})^2
 \end{array}$$

SAS output example:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	31638	2433.65468	108.08	<.0001
Error	492	11079	22.51785		
Corrected Total	505	42716			

## Model comparison in R

```
> lm(stack.loss ~ Air.Flow, data=stackloss) -> model1
> lm(stack.loss ~ Water.Temp + Air.Flow, data=stackloss)
-> model2
➤ anova(model1, model2)
```

### Analysis of Variance Table

Model 1: stack.loss ~ Air.Flow

Model 2: stack.loss ~ Water.Temp + Air.Flow

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	19	319.12				
2	18	188.80	1	130.32	12.425	0.002419 **

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1  
' ' 1

Used in deciding whether to add or drop variables from the model. Leads to the idea of stepwise selection to choose the best variables to be in the regression model.



## Multicollinearity

A common problem, especially if too many variables are included in the regression. Often manifests itself in questions like:

- How come the F-test is significant, but none of the t-tests for the individual coefficients are significant?
- How come the standard error of one of my coefficients is 10000 times larger than the coefficient itself?
- Why did the regression fail with an error message about something being singular?

Happens if some linear combination of the x-variables adds up a constant (or close to a constant) or alternatively, some subset of the centred x-variables is linearly dependent, which happens if and only if the variance-covariance matrix of the X's is **singular** (or nearly singular.)

e.g.  $X_1 - X_2 = 0$

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\ &= \beta_0 + (\beta_1 + 1)X_1 + (\beta_2 - 1)X_2 \\ &= \beta_0 + (\beta_1 + 2)X_1 + (\beta_2 - 2)X_2 \end{aligned}$$

and so on.

One diagnostic for multicollinearity is the **variance inflation factor** or VIF.

$VIF(X_i) = 1/(1 - R^2)$  for a regression of  $X_i$  on the other  $X_j, j \neq i$

If there is no linear combination which adds up to zero, then VIF will be approx. 1.

```
proc reg data = boston;  
  model MEDV = CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX  
PTRATIO B LSTAT / vif;  
  run;
```

1. CRIM per capita crime rate by town
2. ZN proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS proportion of non-retail business acres per town
4. CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX nitric oxides concentration (parts per 10 million)
6. RM average number of rooms per dwelling
7. AGE proportion of owner-occupied units built prior to 1940
8. DIS weighted distances to five Boston employment centres
9. RAD index of accessibility to radial highways
10. TAX full-value property-tax rate per \$10,000
11. PTRATIO pupil-teacher ratio by town
12. B  $1000(B_k - 0.63)^2$  where  $B_k$  is the proportion of African-Americans by town
13. LSTAT % lower status of the population
14. MEDV Median value of owner-occupied homes in \$1000's

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	1	36.45949	5.10346	7.14	<.0001	0
CRIM	1	-0.10801	0.03286	-3.29	0.0011	1.79219
ZN	1	0.04642	0.01373	3.38	0.0008	2.29876
INDUS	1	0.02056	0.06150	0.33	0.7383	3.99160
CHAS	1	2.68673	0.86158	3.12	0.0019	1.07400
NOX	1	-17.76661	3.81974	-4.65	<.0001	4.39372
RM	1	3.80987	0.41793	9.12	<.0001	1.93374
AGE	1	0.00069222	0.01321	0.05	0.9582	3.10083
DIS	1	-1.47557	0.19945	-7.40	<.0001	3.95594
RAD	1	0.30605	0.06635	4.61	<.0001	<b>7.48450</b>
TAX	1	-0.01233	0.00376	-3.28	0.0011	<b>9.00855</b>
PTRATIO	1	-0.95275	0.13083	-7.28	<.0001	1.79908
B	1	0.00931	0.00269	3.47	0.0006	1.34852
LSTAT	1	-0.52476	0.05072	-10.35	<.0001	2.94149

In fact,  
corr(TAX,RAD) = 0.9