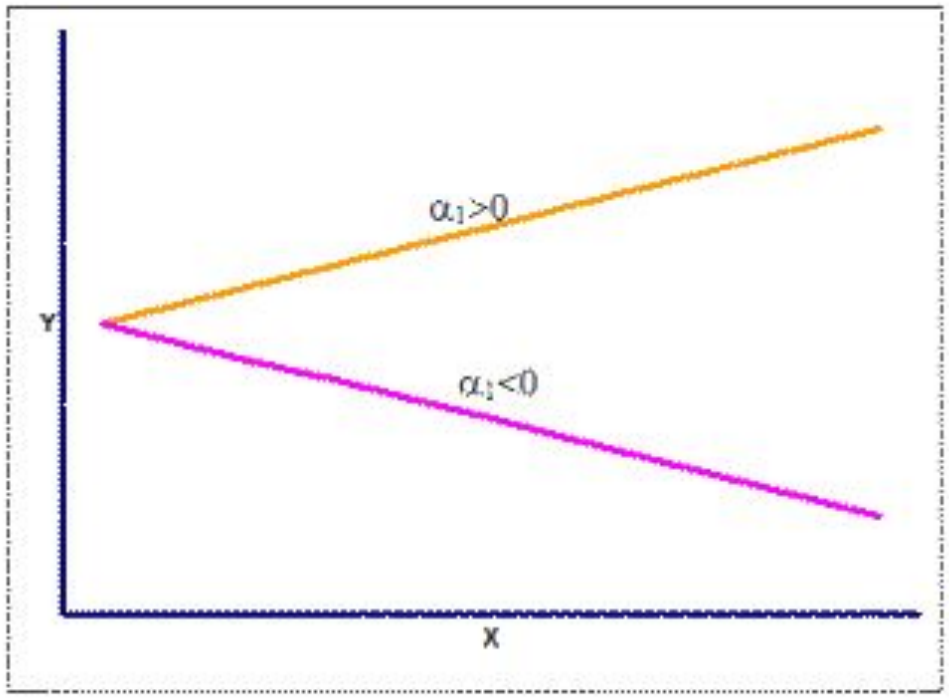
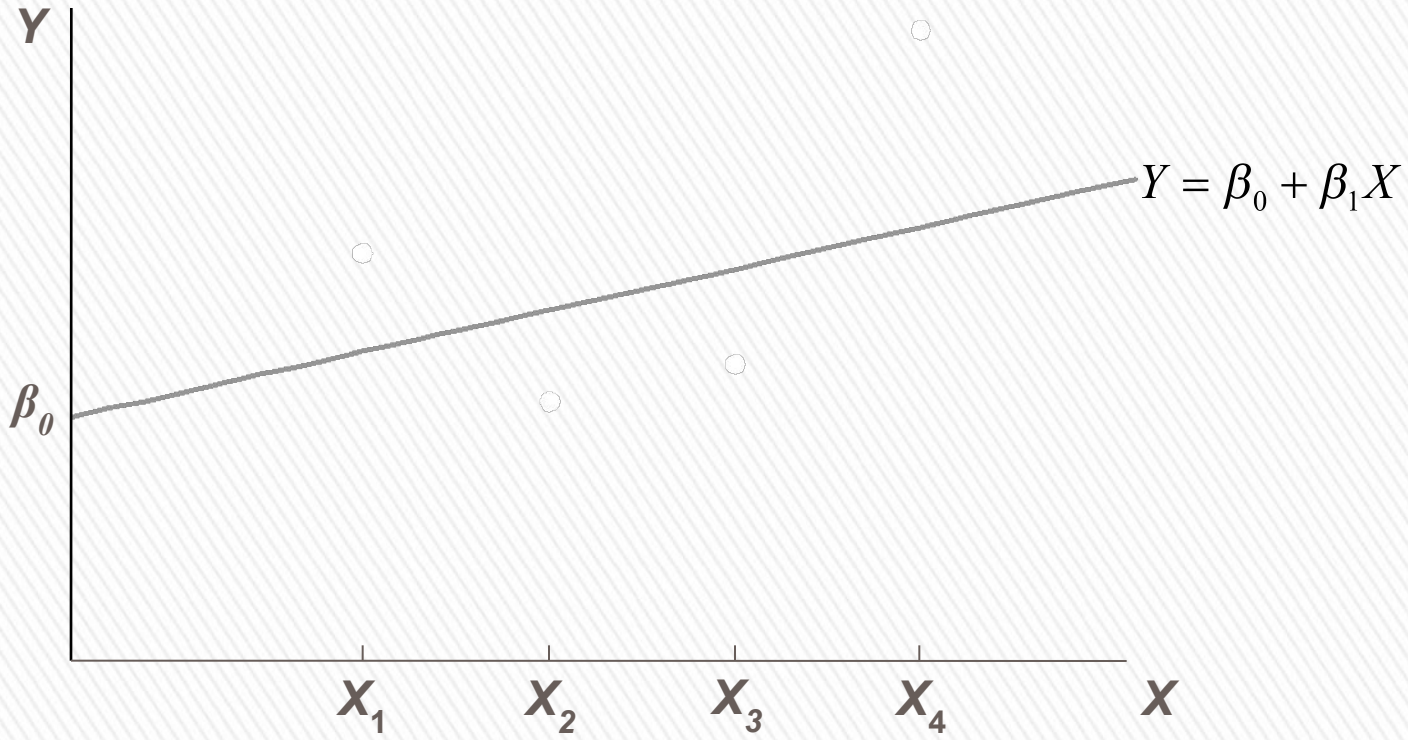
SIMPLE REGRESSION MODEL

The nature of the relationship between variables can take many forms, ranging from simple mathematical functions to extremely complicated ones. The simplest relationship consists of a straight-line or linear relationship (linear function).

This is an example plot of linear function:



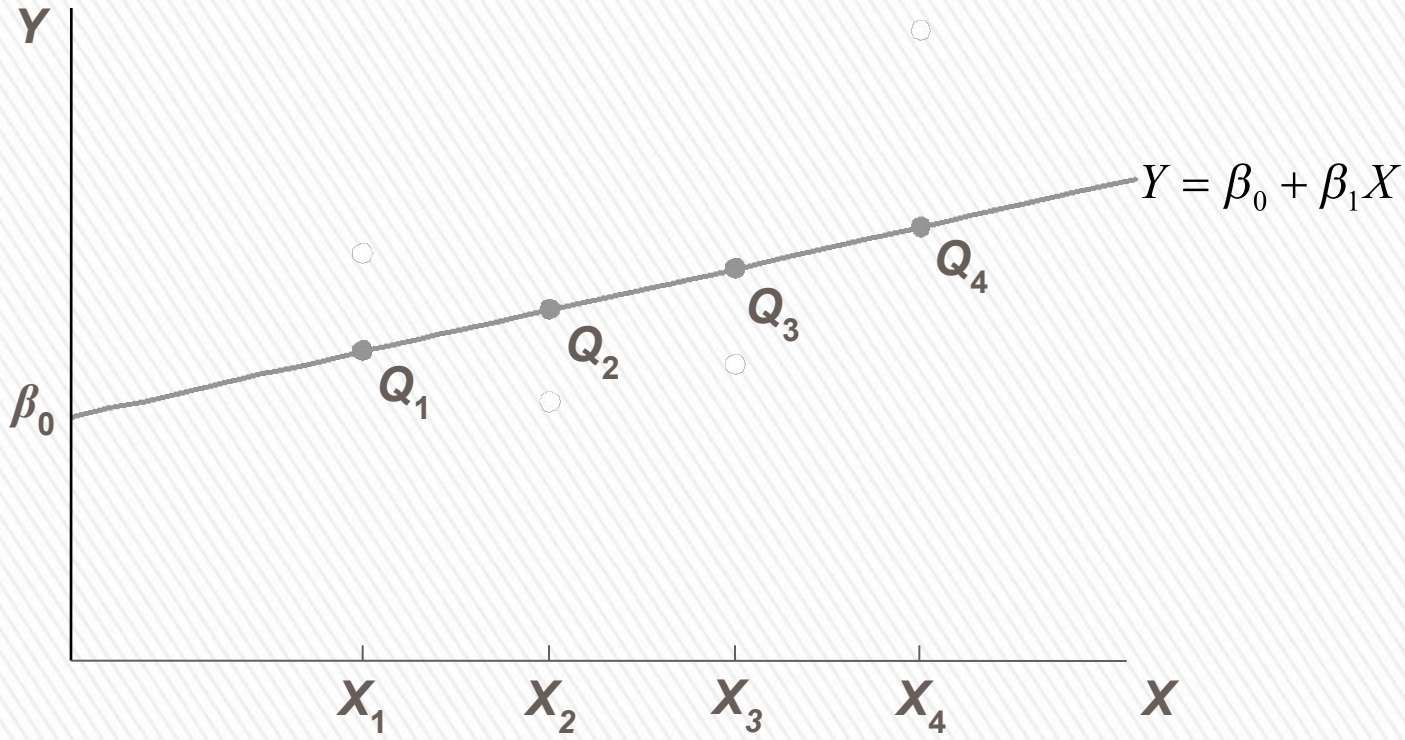
SIMPLE REGRESSION MODEL



Suppose that a variable *Y* is a linear function of another variable *X*, with unknown parameters β_0 and β_1 that we wish to estimate.

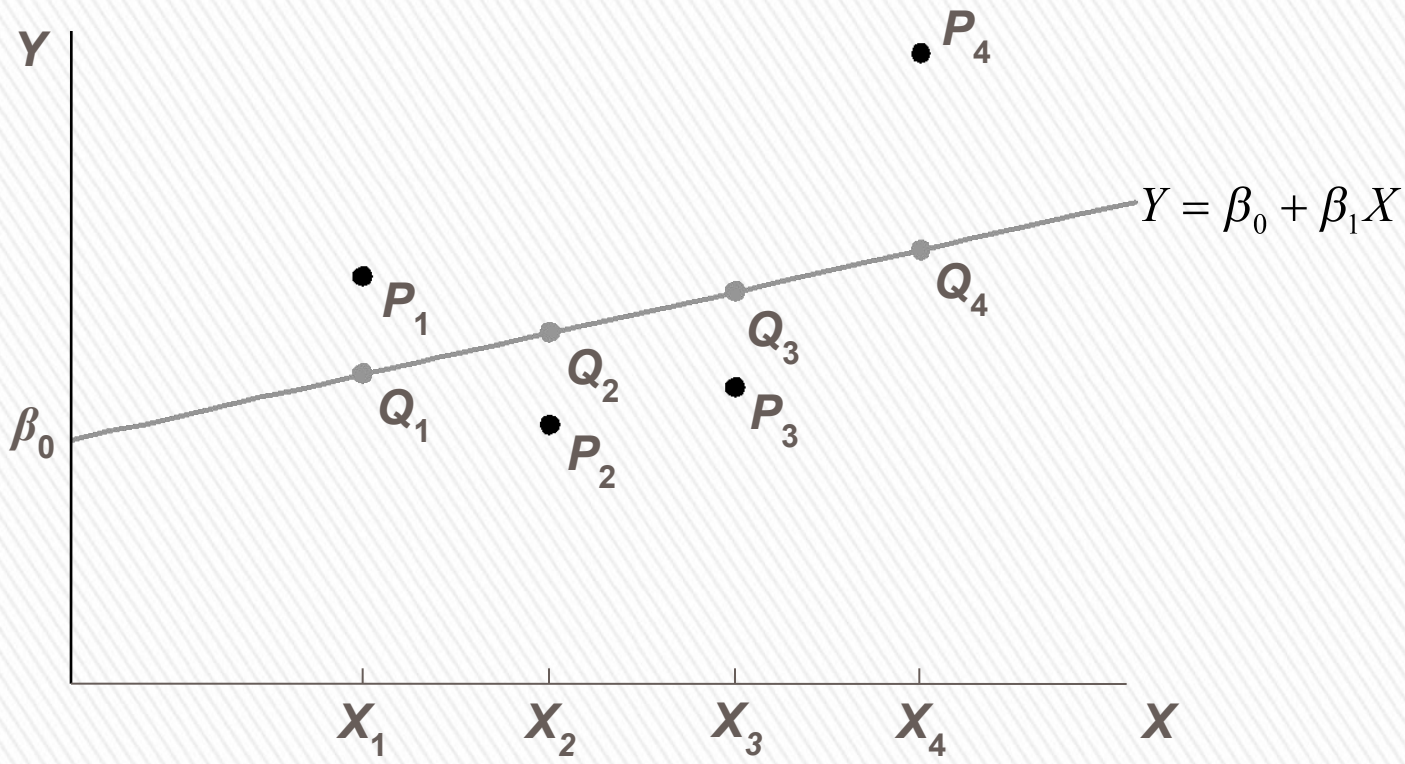
Suppose that we have a sample of 4 observations with *X* values as shown.

SIMPLE REGRESSION MODEL



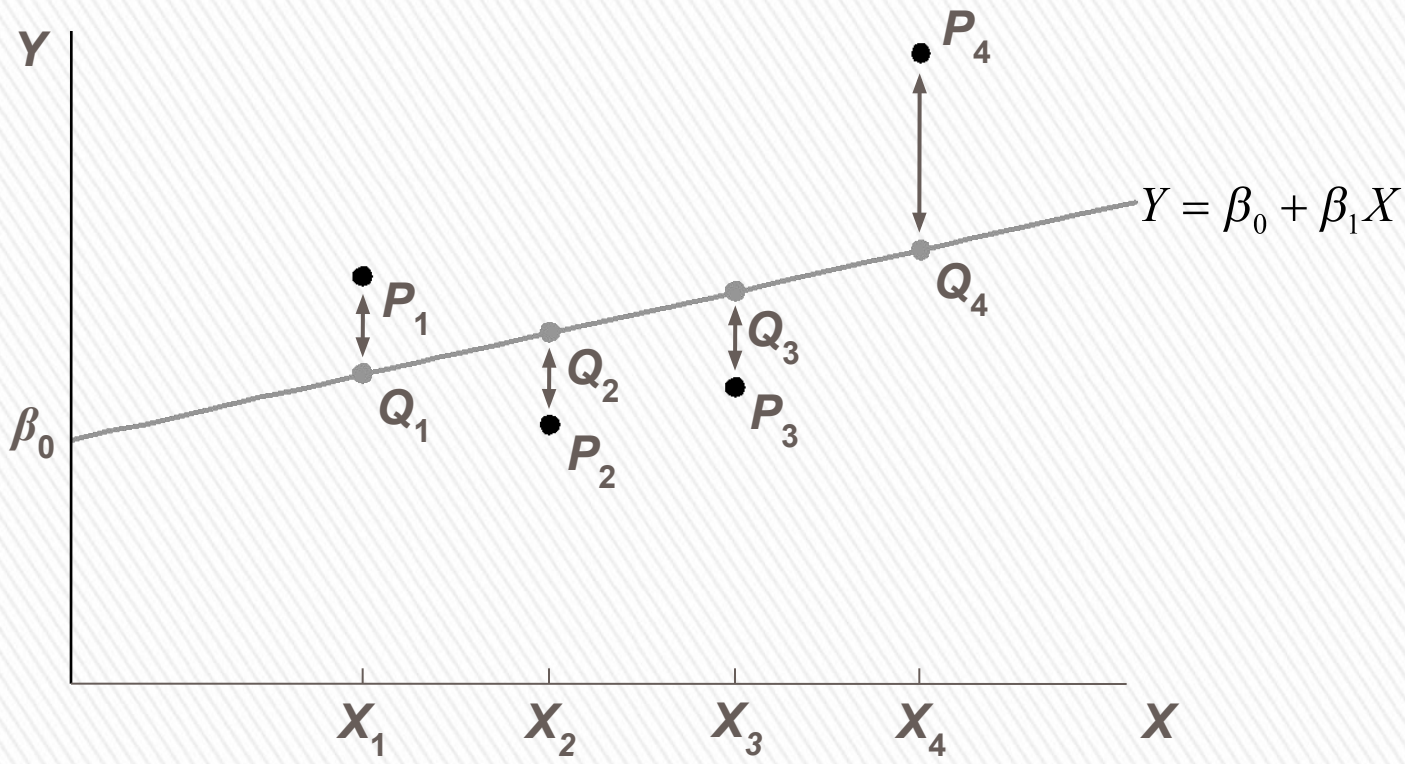
If the relationship were an exact one, the observations would lie on a straight line and we would have no trouble obtaining accurate estimates of β_0 and β_1. When all empirical pairs of X-Y points lie on a straight line – it is called a functional or deterministic relationship.

SIMPLE REGRESSION MODEL



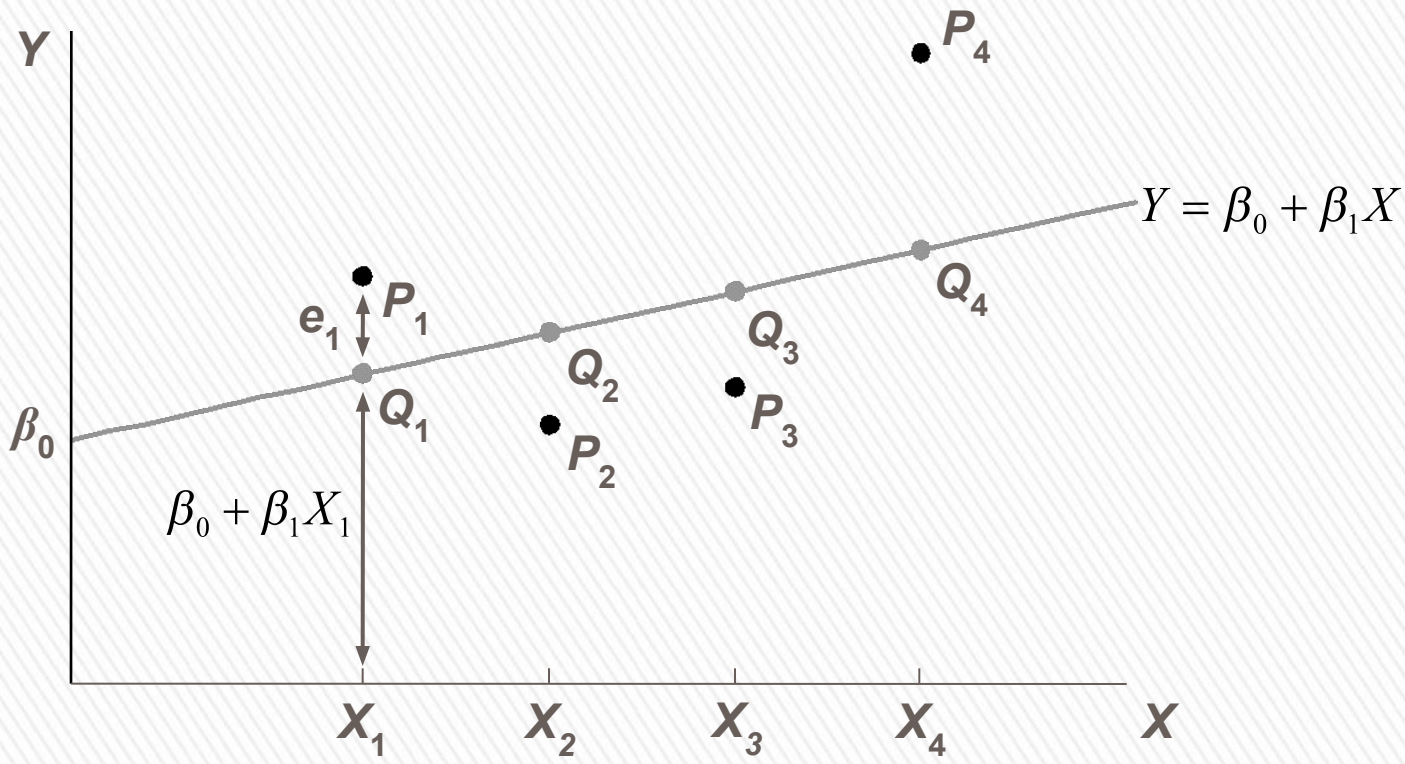
In practice, most economic relationships are not exact and the actual values of *Y* are different from those corresponding to the straight line.

SIMPLE REGRESSION MODEL



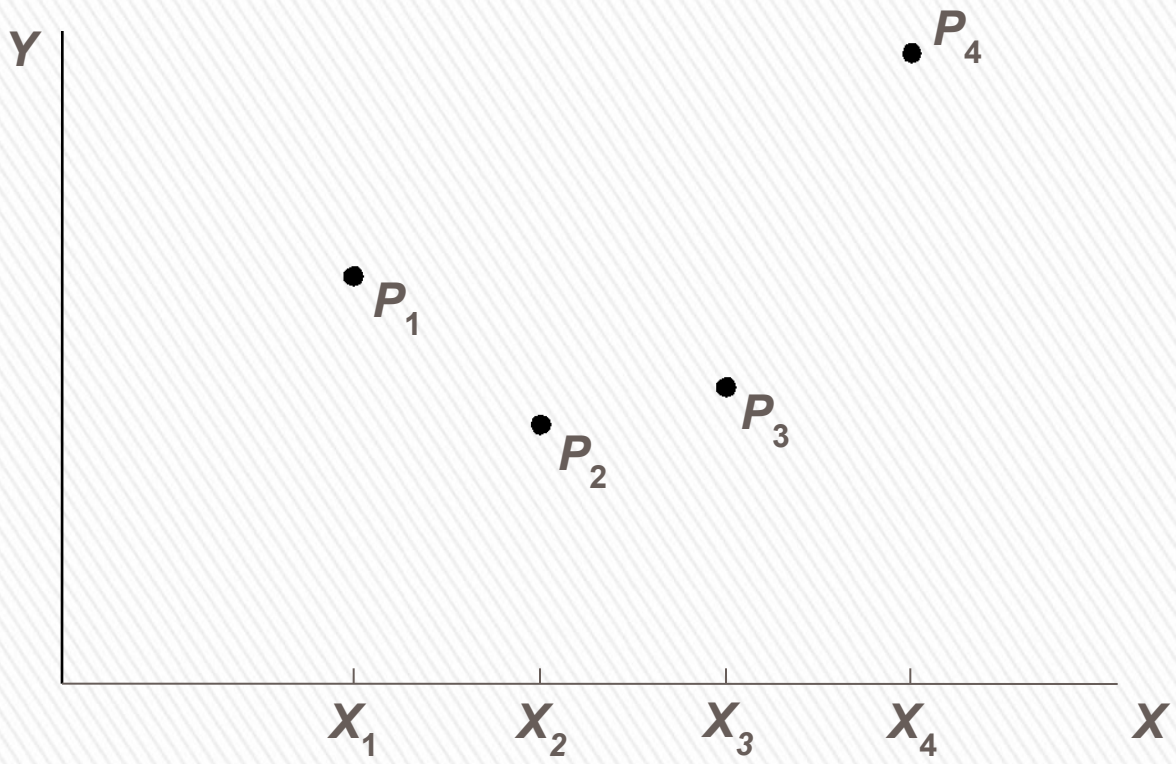
To allow for such divergences, we will write the model as $Y = \beta_0 + \beta_1 X + e$, where e is a disturbance term.

SIMPLE REGRESSION MODEL



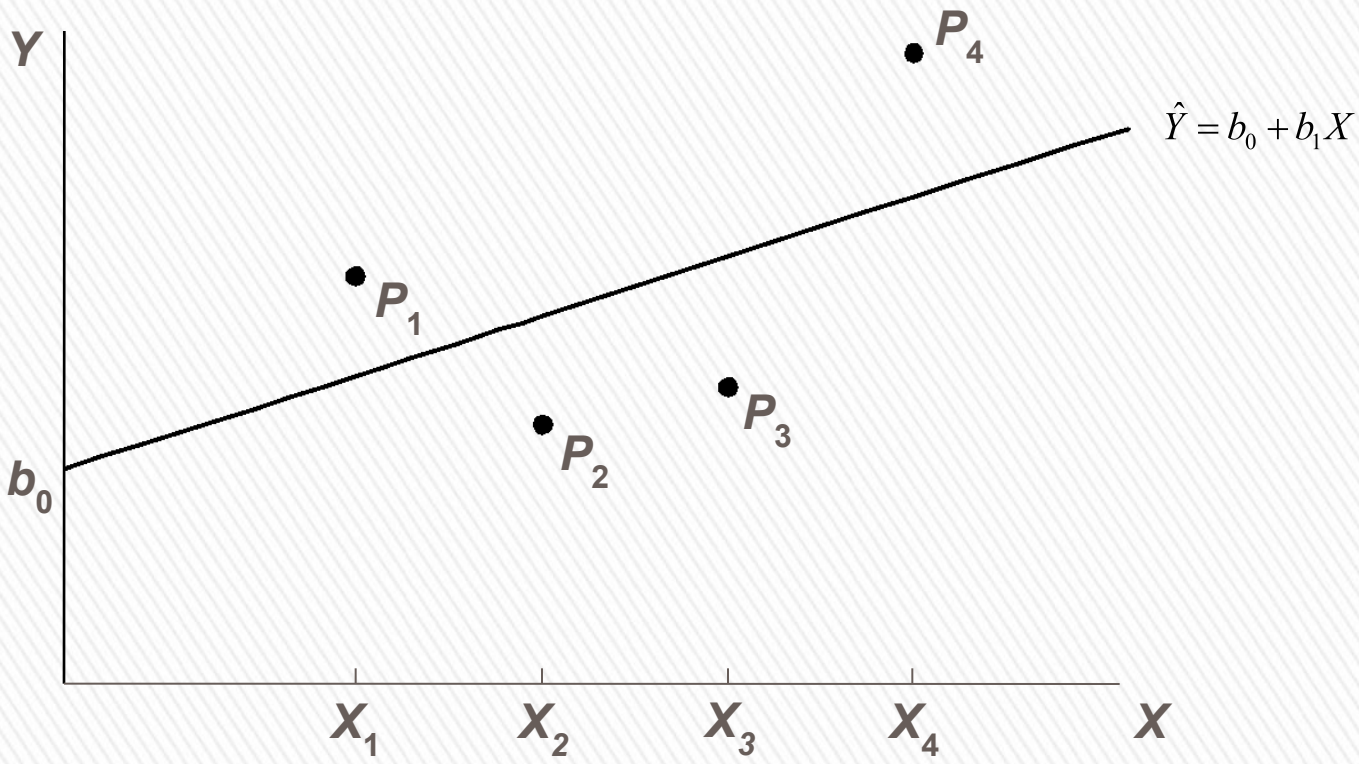
Each value of *Y* thus has a nonrandom component, $\beta_0 + \beta_1 X$, and a random component, *e*. The first observation has been decomposed into these two components.

SIMPLE REGRESSION MODEL



In practice we can see only the *P* points.

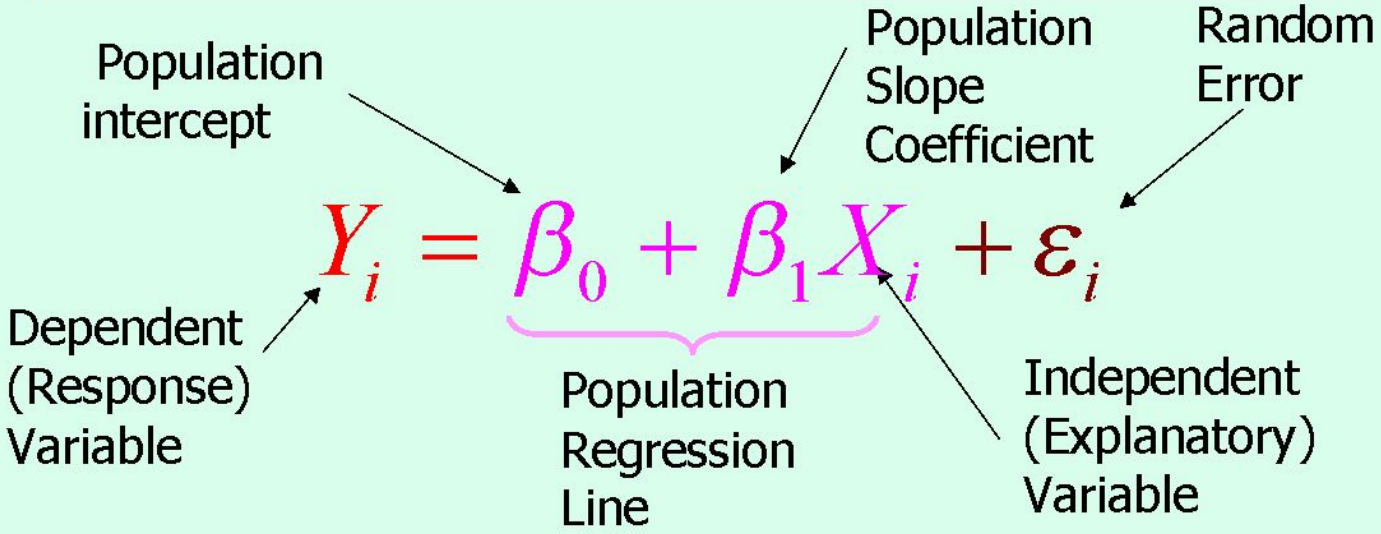
SIMPLE REGRESSION MODEL



Obviously, we can use the *P* points to draw a line which is an approximation to the line $Y = \beta_0 + \beta_1 X$. If we write this line $\hat{Y} = b_0 + b_1 X$, b_0 is an estimate of β_0 and b_1 is an estimate of β_1.

Population Linear Regression

Population regression line is a straight line that describes the dependence of the average value (conditional mean) of one variable on the other



SIMPLE REGRESSION MODEL

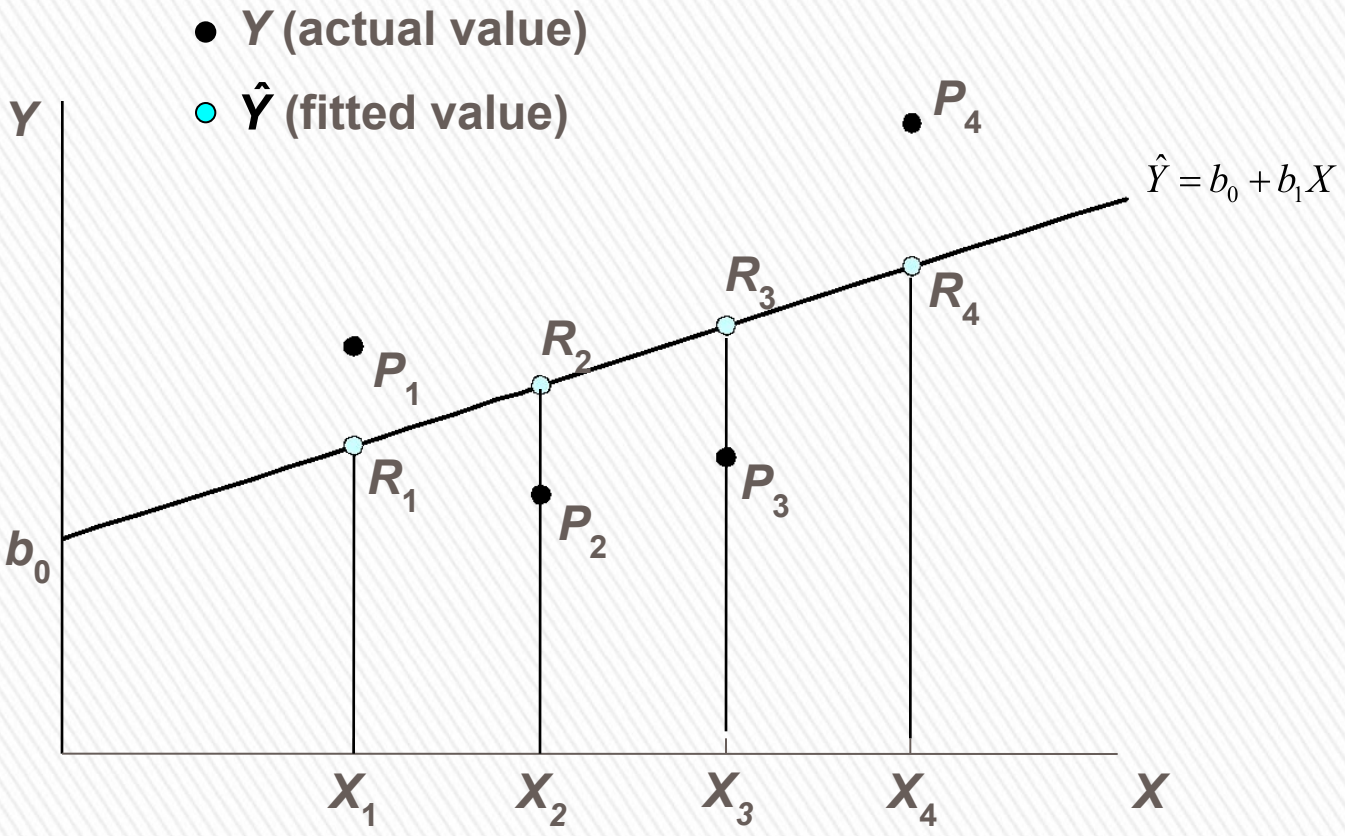
However, we have obtained data from only a random sample of the population. For a sample, b_0 and b_1 can be used as estimates (estimators) of the respective population parameters β_0 and β_1

$$\hat{y}_i = b_0 + b_1 x_i + e_i$$

The intercept b_0 and the slope b_1 are the coefficients of the regression line. The slope b_1 is the change in Y (increase, if >0, and decrease, if <0) associated with a unit change in X. The intercept is the value of Y when X=0; it's the point at which the population regression line intersects the Y axis. In some cases the intercept has no real-world meaning (for example when X is the class size, Y is the test score – the intercept is the predicted value of test scores when there are no students in the class!).

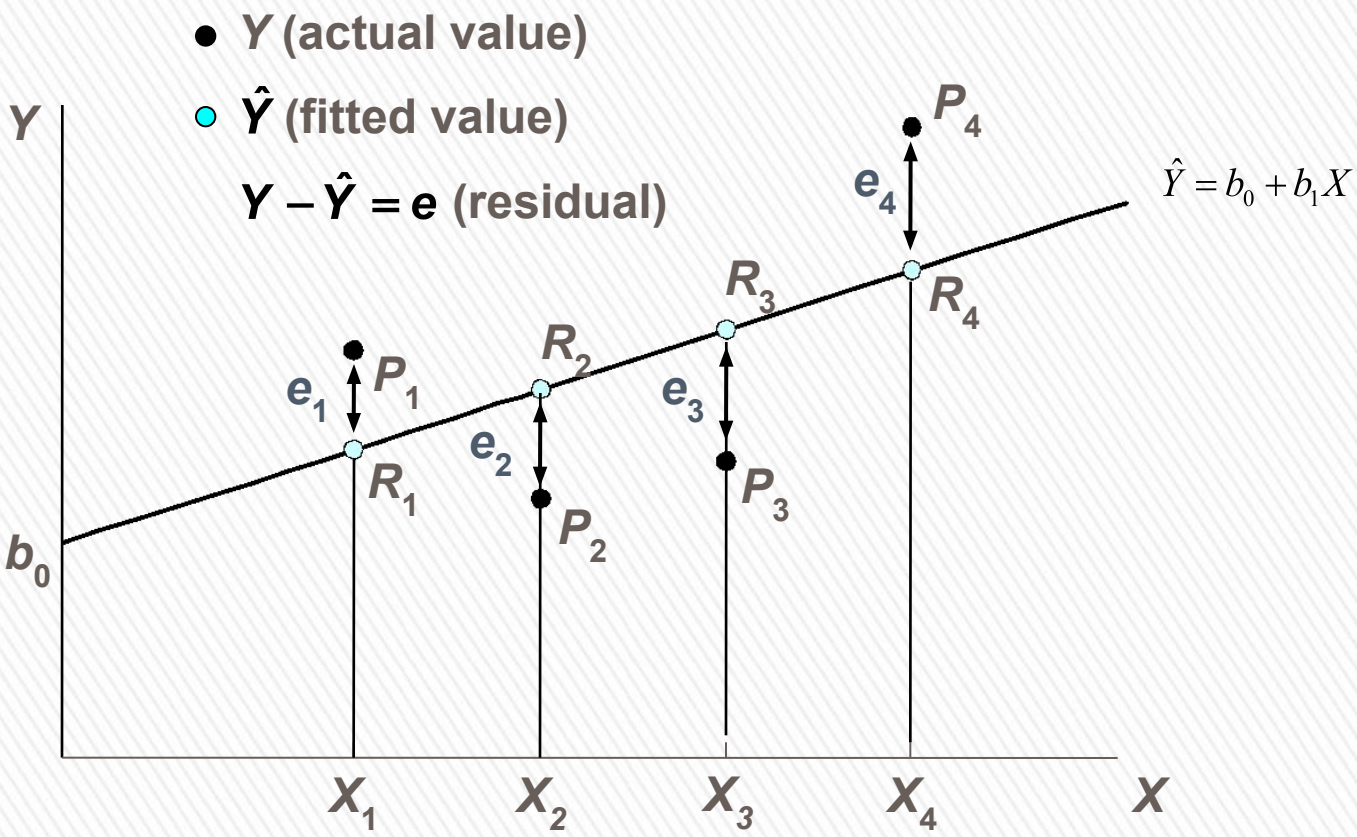
Random error contains all the other factors besides X that determine the value of the dependent variable Y, for a specific observation.

SIMPLE REGRESSION MODEL



The line is called the fitted model and the values of *Y* predicted by it are called the fitted values of *Y*. They are given by the heights of the *R* points.

SIMPLE REGRESSION MODEL



The discrepancies between the actual and fitted values of *Y* are known as the residuals.

Least squares criterion:

Minimize *SSE* (residual sum of squares), where

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = e_1^2 + \ldots + e_n^2$$

To begin with, we will draw the fitted line so as to minimize the sum of the squares of the residuals, *SSE*. This is described as the least squares criterion.

Least squares criterion:

Minimize *SSE* (residual sum of squares), where

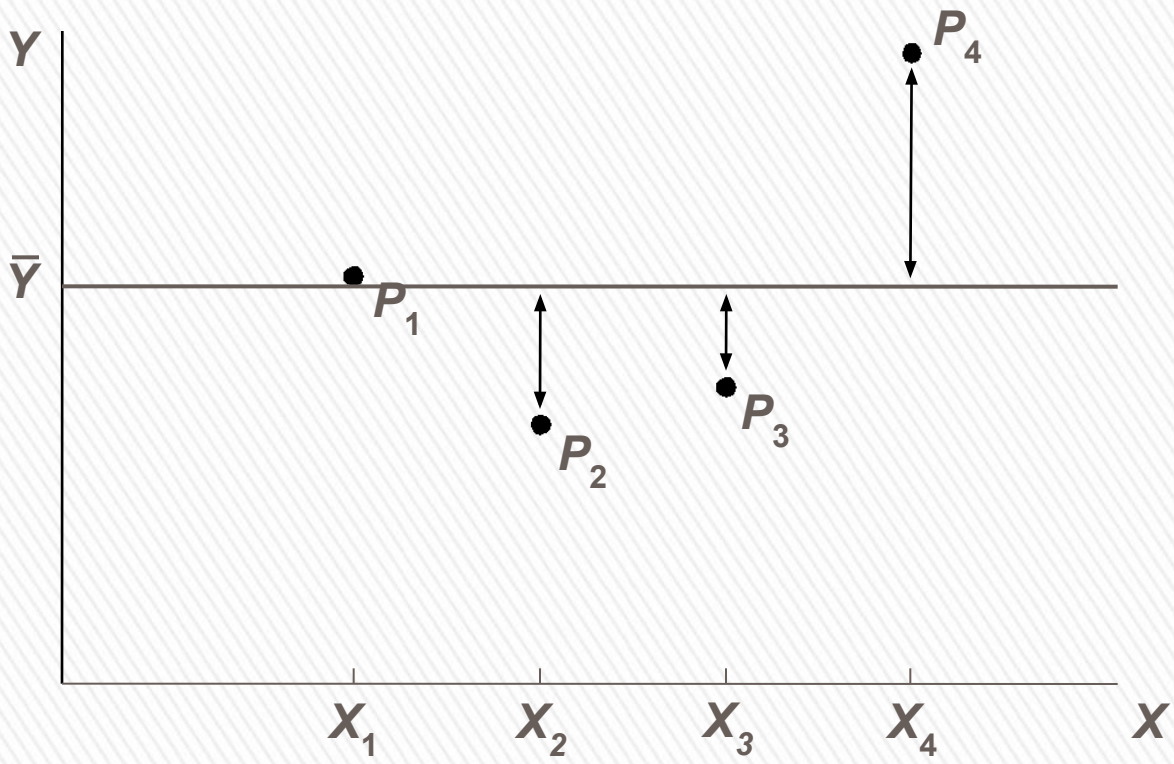
$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = e_1^2 + \ldots + e_n^2$$

Why not minimize

$$\sum_{i=1}^{n} e_i = e_1 + \ldots + e_n$$

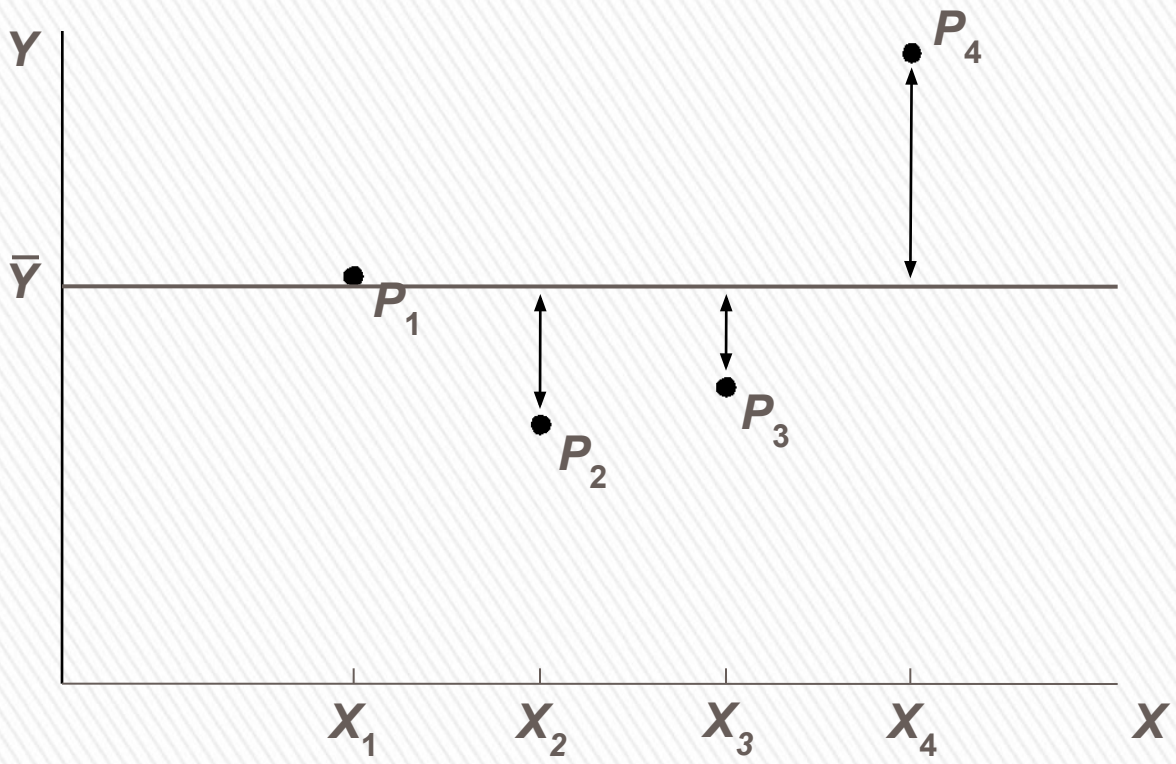
Why the squares of the residuals? Why not just minimize the sum of the residuals?

SIMPLE REGRESSION MODEL



The answer is that you would get an apparently perfect fit by drawing a horizontal line through the mean value of *Y*. The sum of the residuals would be zero.

SIMPLE REGRESSION MODEL



You must prevent negative residuals from cancelling positive ones, and one way to do this is to use the squares of the residuals.

SIMPLE REGRESSION MODEL

Since $\hat{y}_i = b_0 + b_1 x_i$ **we are minimizing** $SSE = \sum e_i^2 = \sum (y_i - (b_0 + b_1 x_i))^2$

which has two unknowns, b_0 and b_1. A mathematical technique which determines the values of b_0 and b_1 that best fit the observed data is known as the Ordinary Least Squares method (OLS).

Ordinary Least Squares is a procedure that selects the best fit line given a set of data points, by minimizing the sum of the squared deviations of the points from a line. That is, if $\hat{y}_i = b_0 + b_1 x_i$ is the equation of the best line to fit through the data then in order to get this best line, using the least squares criteria, for each value data point (x_i, y_i) if $e_i = y_i - \hat{y}_i$ where $\hat{y} = b_0 + b_1 x$, then e_i is the amount of deviation of the data point from the line. The least squares criteria minimizes, finds the slope b_1 and the y-intercept b_0 from the data, that minimizes the sum of the square deviations, $\sum_{i=1}^{n} e_i^2$

SIMPLE REGRESSION MODEL

For the mathematically curious , I provide a condensed derivation of the coefficients.

To minimize $SSE = \sum e_i^2 = \sum (y_i - (b_0 + b_1 x_i))^2$ **determine the partial derivatives with respect to** b_0 **and with respect to** b_1. **These are:**

$$f_{b0} = 2\sum (y - b_0 - b_1 x)(-1)$$

$$f_{b1} = 2\sum (y - b_0 - b_1 x)(-x)$$

Setting $f_{b0} = f_{b1} = 0$ **and solving for** b_0 **and** b_1 **results in equations given below.**

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

SIMPLE REGRESSION MODEL

Since there are two equations with two unknown, we can solve these equations simultaneously for b_0 and b_1 as follows:

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i-1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

ONLY FOR REGRESSION MODELS WITH ONE INDEPENDENT VARIABLE!

We also note that the regression line always goes through the mean ($\overline{X}, \overline{Y}$ **).**

SIMPLE REGRESSION MODEL

In matrix notation OLS may be written as:
$Y = Xb + e$
The normal equations in matrix form are now

$$X^T Y = X^T X b$$

And when we solve it for *b* we get:

$$b = (X^T X)^{-1} X^T Y$$

where Y is a column vector of the Y values and X is a matrix containing a column of ones (to pick up the intercept) followed by a column of the X variable containing the observations on it and *b* is a vector containing the estimators of regression parameters.

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} \qquad b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

SIMPLE REGRESSION MODEL

We can state as follows:

$$X^T X = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \qquad X^T Y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_i \end{bmatrix}$$

How to inverse $X^T X$?

1. **matrix determinant** $deX^T X = n \cdot \sum x^2 - (\sum x)^2$
2. **minor matrix**

$$\min X^T X = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} = \begin{bmatrix} \sum x^2 & \sum x \\ \sum x & n \end{bmatrix}$$

3. **cofactor matrix**

$$(X^T X)D = \begin{bmatrix} \sum x^2 \cdot (-1)^{1+1} & \sum x \cdot (-1)^{1+2} \\ \sum x \cdot (-1)^{2+1} & n \cdot (-1)^{2+2} \end{bmatrix}$$

4. **inverse matrix**

$$(X^T X)^{-1} = \frac{1}{\det X^T X} \begin{bmatrix} \sum x^2 & -\sum x \\ -\sum x & n \end{bmatrix}$$

SIMPLE REGRESSION MODEL

EXAMPLE

In this problem we were looking at the way home size is effected by the family income. We will use this model to try to predict the value of the dependent variable based on the independent variable. Also, the slope will help us to understand how the Y variable changes for each unit change in the X variable.

Assume a real-estate developer is interested in determining the relationship between family income (X, in thousand of dollars) of the local resident and the square footage of their homes (Y, in hundreds of square feet). A random sample of ten families is obtained with the following results:

X	22	26	45	37	28	50	56	34	60	40
Y	16	17	26	24	22	21	32	18	30	20

SIMPLE REGRESSION MODEL

For our example (X-family income; Y-home size)

Family	x	y
1	22	16
2	26	17
3	45	26
4	37	24
5	28	22
6	50	21
7	56	32
8	34	18
9	60	30
10	40	20

X

	1	22
	1	26
	1	45
	1	37
	1	28
	1	50
	1	56
	1	34
	1	60
	1	40

Y

16
17
26
24
22
21
32
18
30
20

X^T

1	1	1	1	1	1	1	1	1	1
22	26	45	37	28	50	56	34	60	40

X^TX

10	398
398	17330

X^TY

226
9522

SIMPLE REGRESSION MODEL

det X^TX		14896
so it's nonsingular		
Matrix of minors		
Min X^TX	17330	398
	398	10
Cofactor matrix		
$(X^TX)D$	17330	-398
	-398	10
Inverse matrix		
$(X^TX)^{-1}$	1,163	-0,03
	-0,03	0,0007

X –family income ($ 000)

Y – home size (hundreds of square feet)

Estimators	
b_0	8,51
b_1	0,35

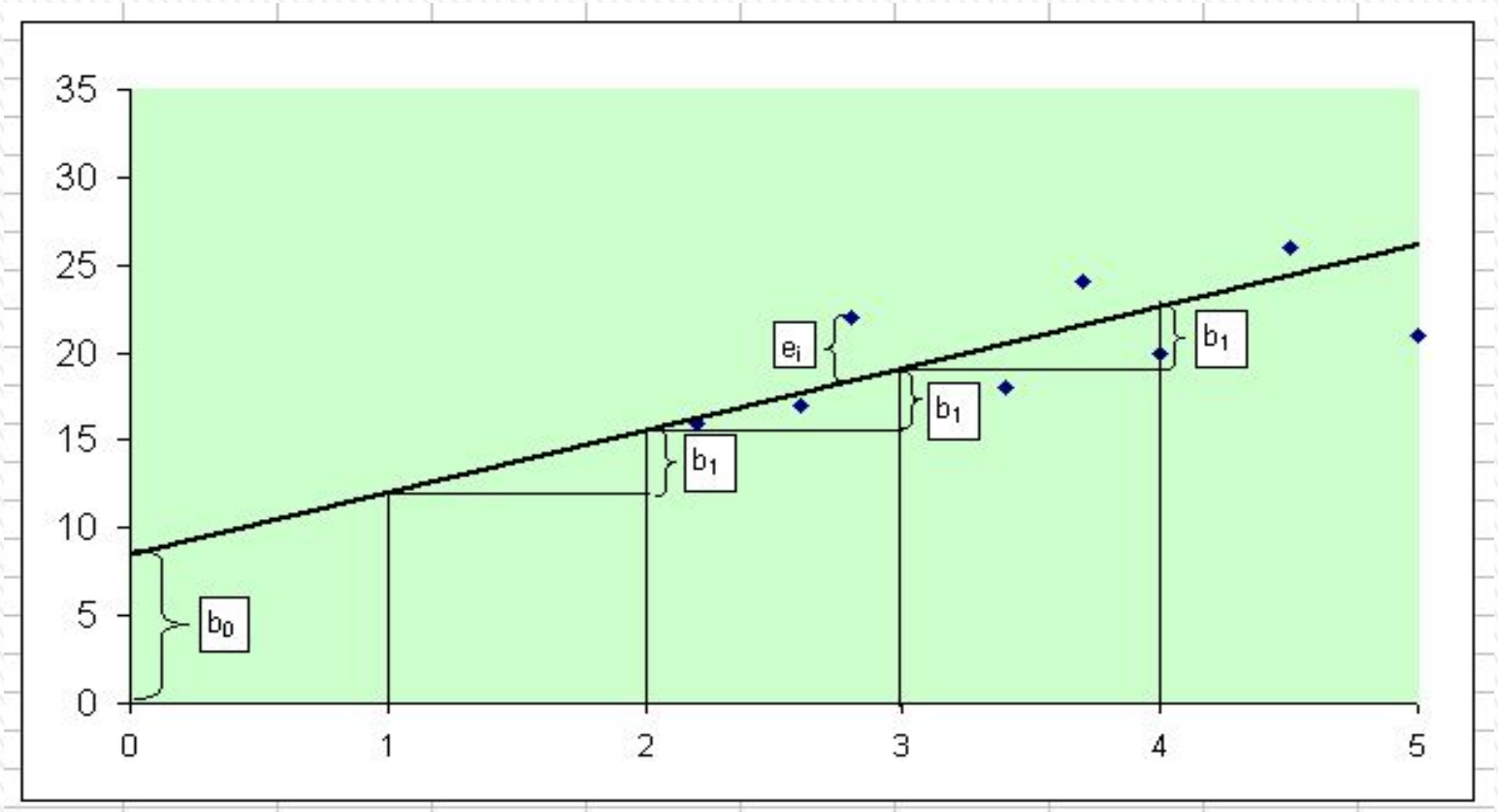
If the family has no income, the home size is equal to 851 square feet

If the family income increase 1000 $, the home size will increase 35 square feet, on average.

The regression equation can be stated as follows:

$$\hat{y}_i = 8{,}51 + 0{,}35x_i + e_i$$

SIMPLE REGRESSION MODEL



Let's try another example:
X – commercial time (minutes) Y – sales ($ hundred thousand)

x	y
1	9
5	20
6	22
5	15
5	17
9	30
3	18
7	25
3	10
6	20

$X =$

1	1
1	5
1	6
1	5
1	5
1	9
1	3
1	7
1	3
1	6

$Y =$

9
20
22
15
17
30
18
25
10
20

$X^T =$

1	1	1	1	1	1	1	1	1	1
1	5	6	5	5	9	3	7	3	6

$X^TX =$

10	50
50	296

$X^TY =$

186
1050

Estimators

b_0	5,56
b_1	2,61

$\hat{y} = 5.56 + 2.61x$

How are the slope and intercept interpreted for this example?

X – commercial time (minutes)

Y – sales ($ hundred thousand)

Intercept: $b_0 = 5.56$ When we have no commercials, we would expect $556,000 in sales.

Slope: $b_1 = 2.61$ For each minute increase in commercial time we have a $261,000 increase in sales.

Notice that when $x = \bar{x} = 5$ $\hat{y} = 5.56 + 2.61(5) = 18.61 = \bar{y}$. As we said, the least squares regression line always goes through the points $(\bar{x}, \bar{y})$

REGRESSION MODEL WITH TWO EXPLANATORY VARIABLES

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

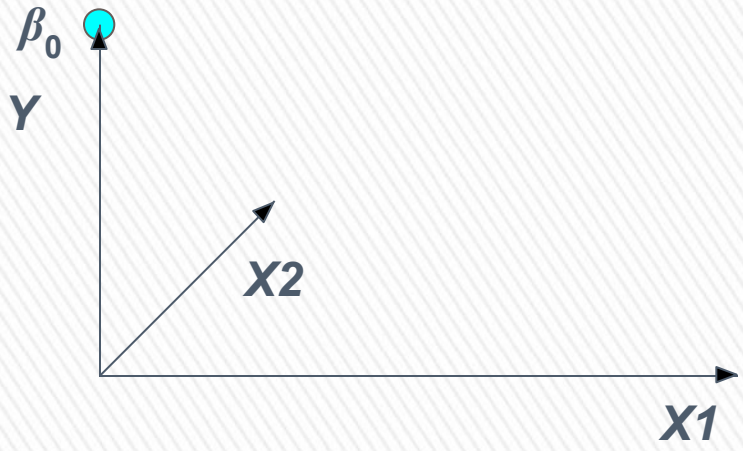
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i$$

This sequence provides a geometrical interpretation of a multiple regression model with two explanatory variables.

Y – weekly salary ($)

X1 – length of employment (in months)

X2 – age (in years)



Specifically, we will look at weekly salary function model where weekly salary, *Y*, depend on length of employment X1, and age, X2.

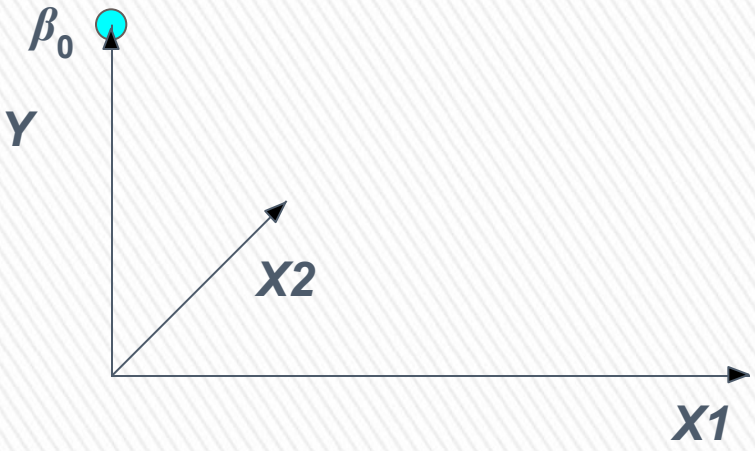
MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i$$

Y – weekly salary ($)

X1 – length of employment (in months)

X2 – age (in years)



The model has three dimensions, one each for *Y, X1*, and X2. The starting point for investigating the determination of *Y* is the intercept, β_0.

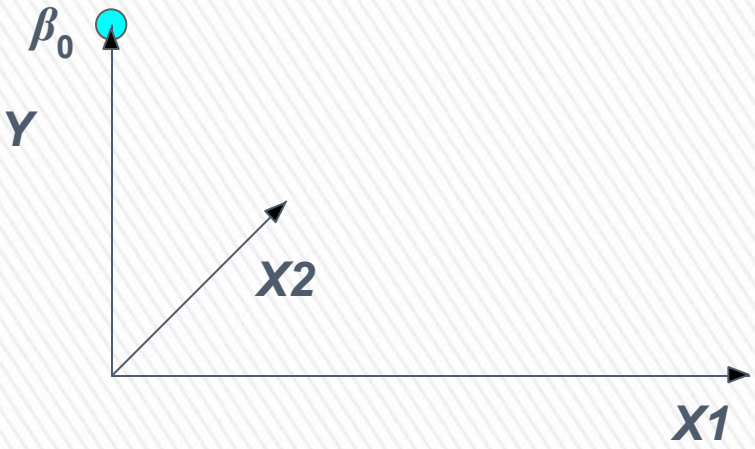
MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i$$

Y – weekly salary ($)

X1 – length of employment (in months)

X2 – age (in years)



Literally the intercept gives *weekly salary* for those respondents who have no age (??) and no length of employment (??). Hence a literal interpretation of β_0 would be unwise.

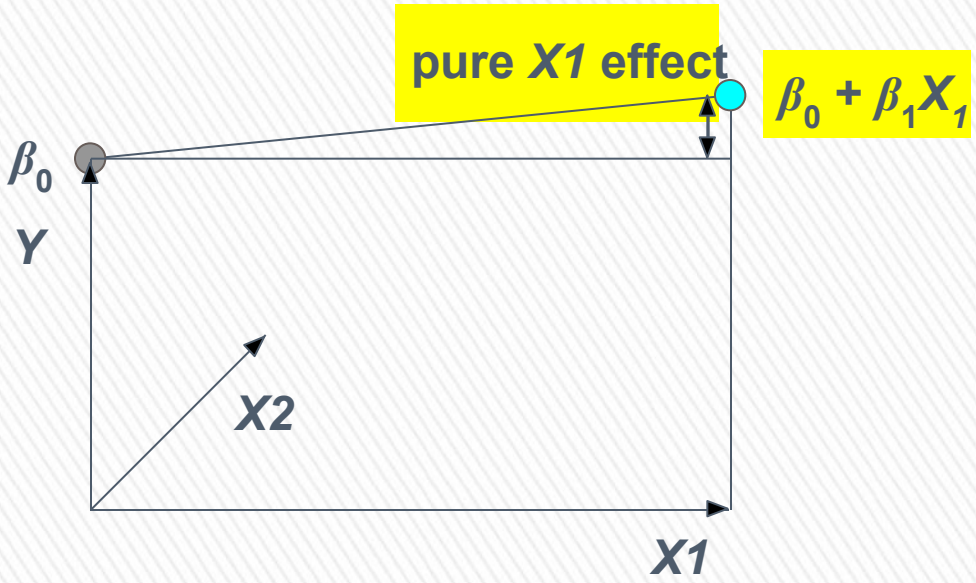
MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i$$

Y – weekly salary ($)

X1 – length of employment (in months)

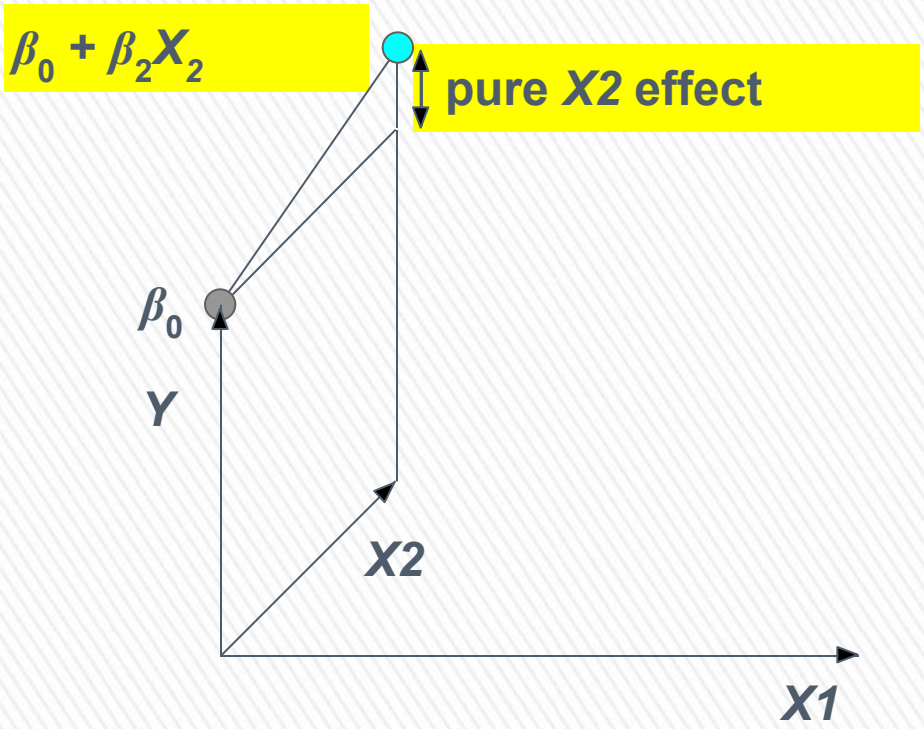
X2 – age (in years)



The next term on the right side of the equation gives the effect of *X1*. A one month of employment increase in *X1* causes weekly salary to increase by β_1 dollars, holding *X2* constant.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

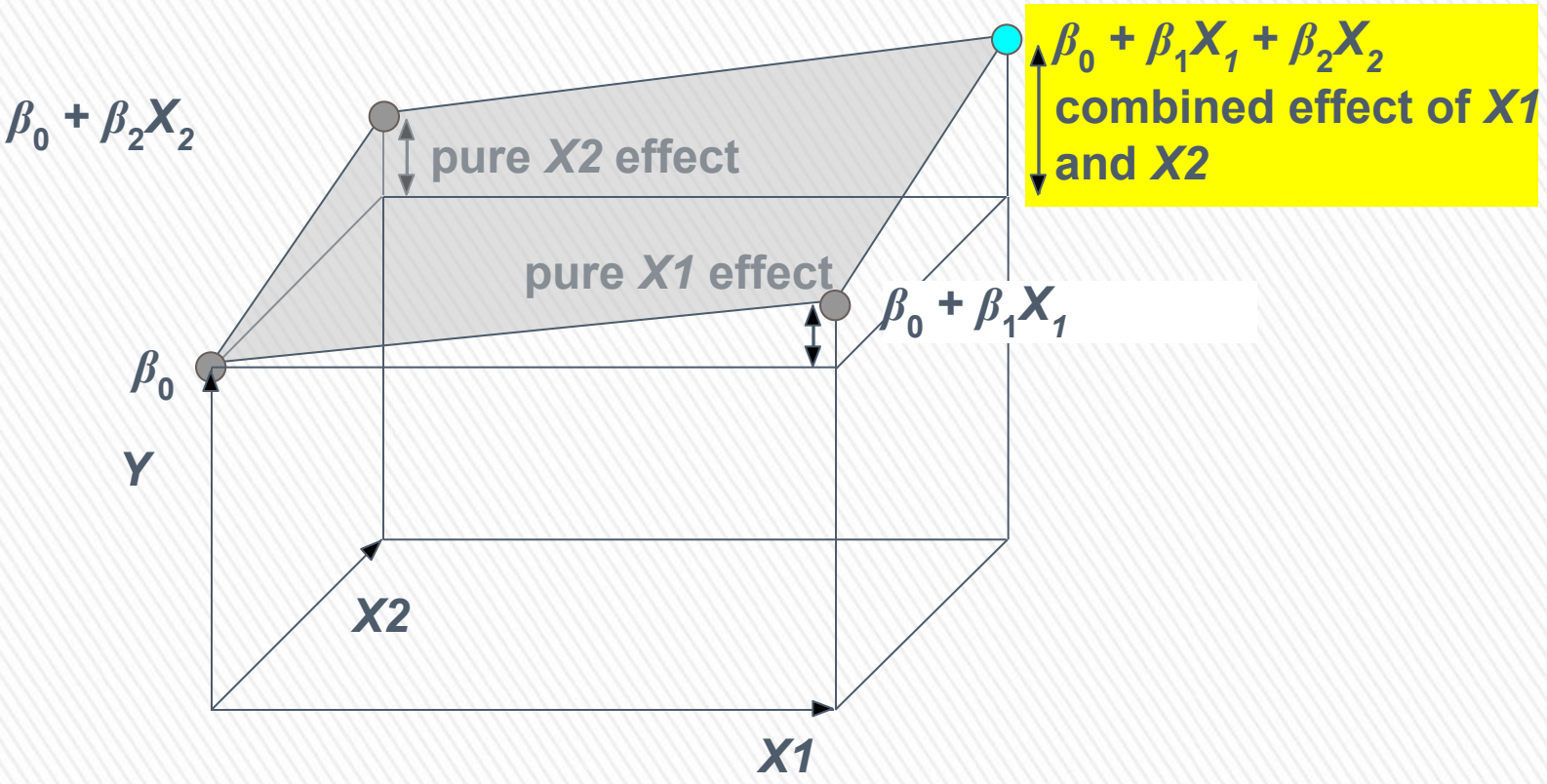
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i$$



Similarly, the third term gives the effect of variations in *X2*. A one year of age increase in *X2* causes weekly salary to increase by β_2 dollars, holding *X1* constant.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

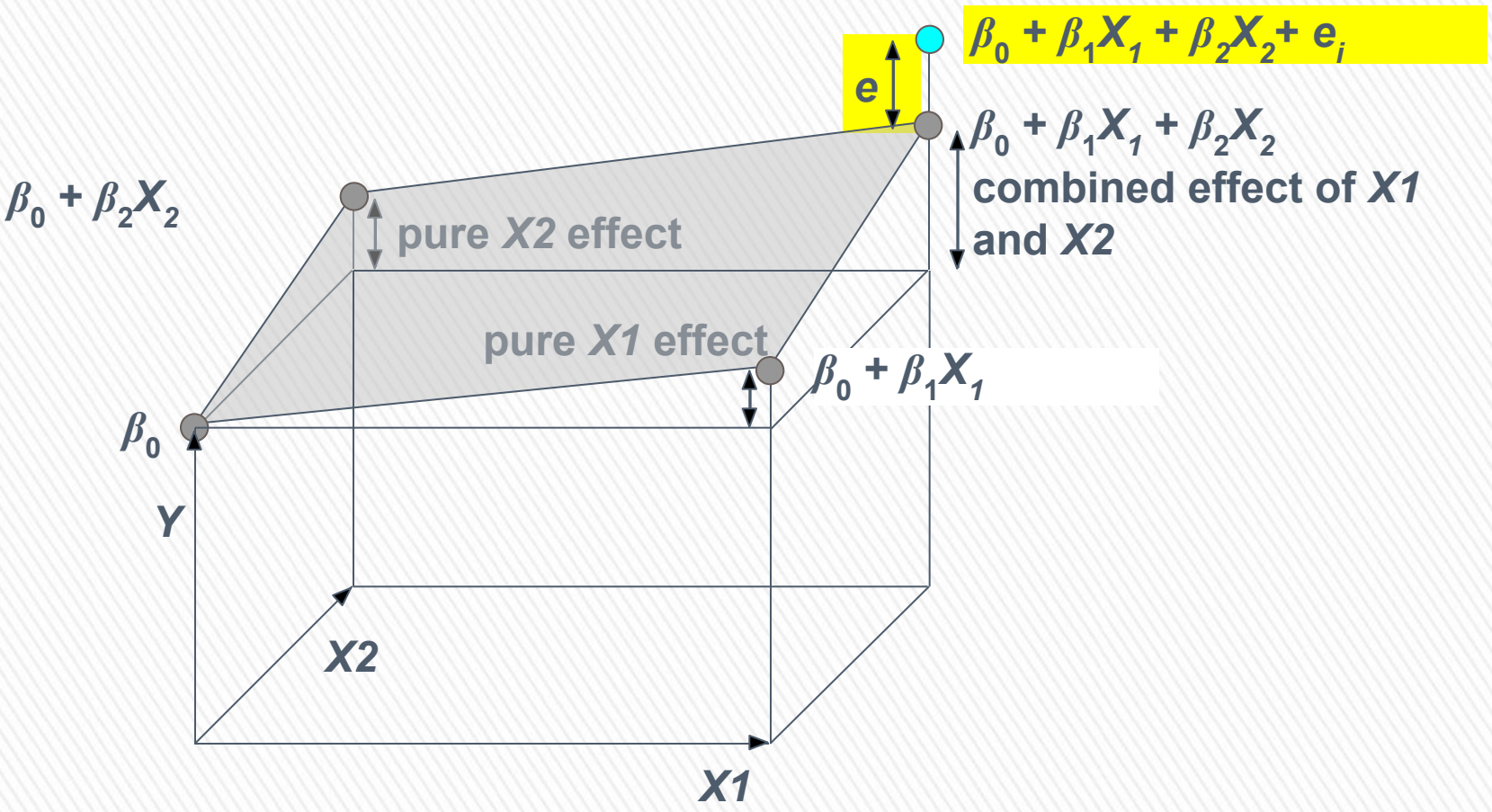
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i$$



Different combinations of *X1* and *X2* give rise to values of *weekly salary* which lie on the plane shown in the diagram, defined by the equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. This is the nonrandom component of the model.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

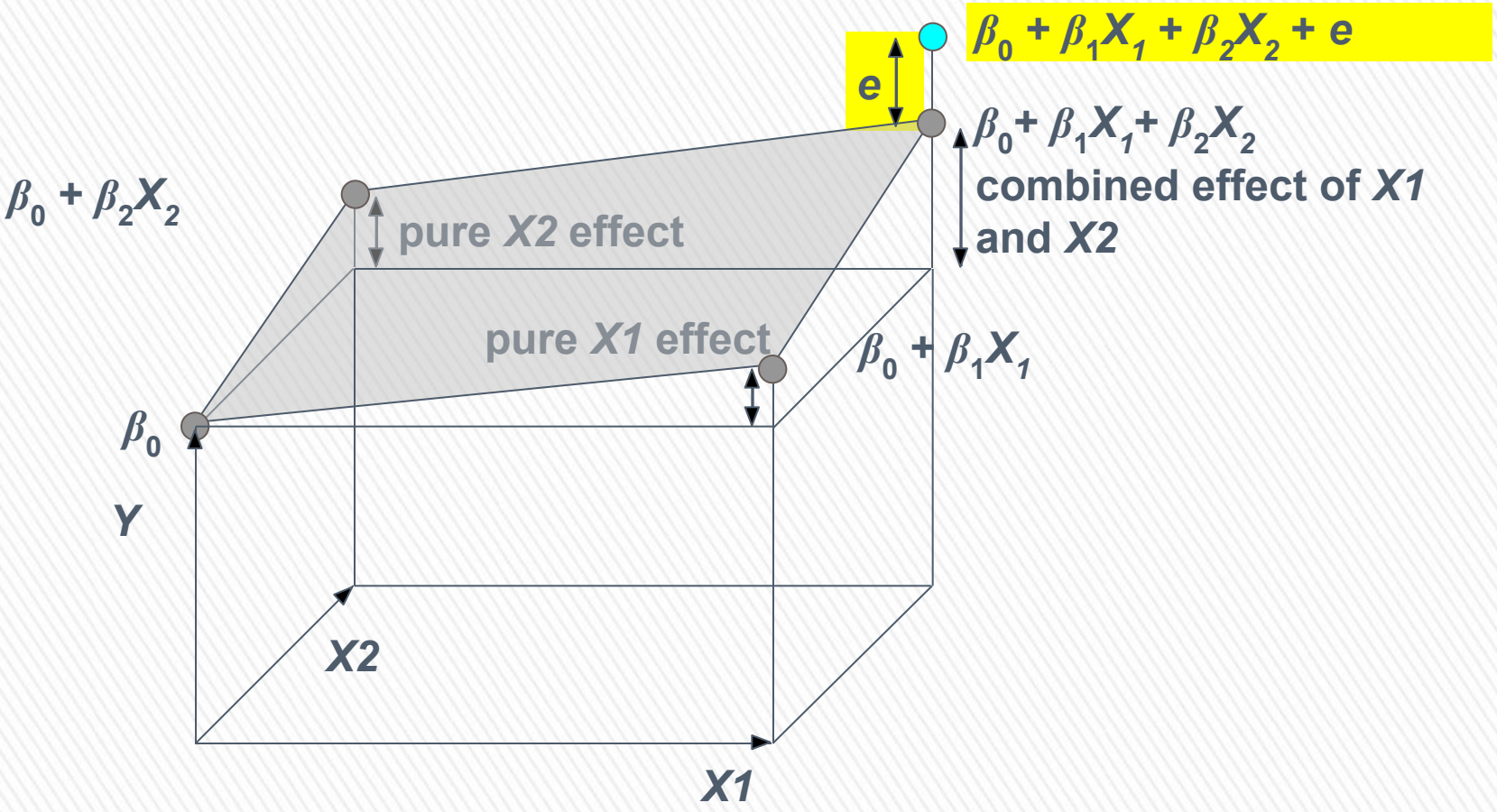
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i$$



The final element of the model is the error term, *e*. This causes the actual values of *Y* to deviate from the plane. In this observation, *e* happens to have a positive value.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

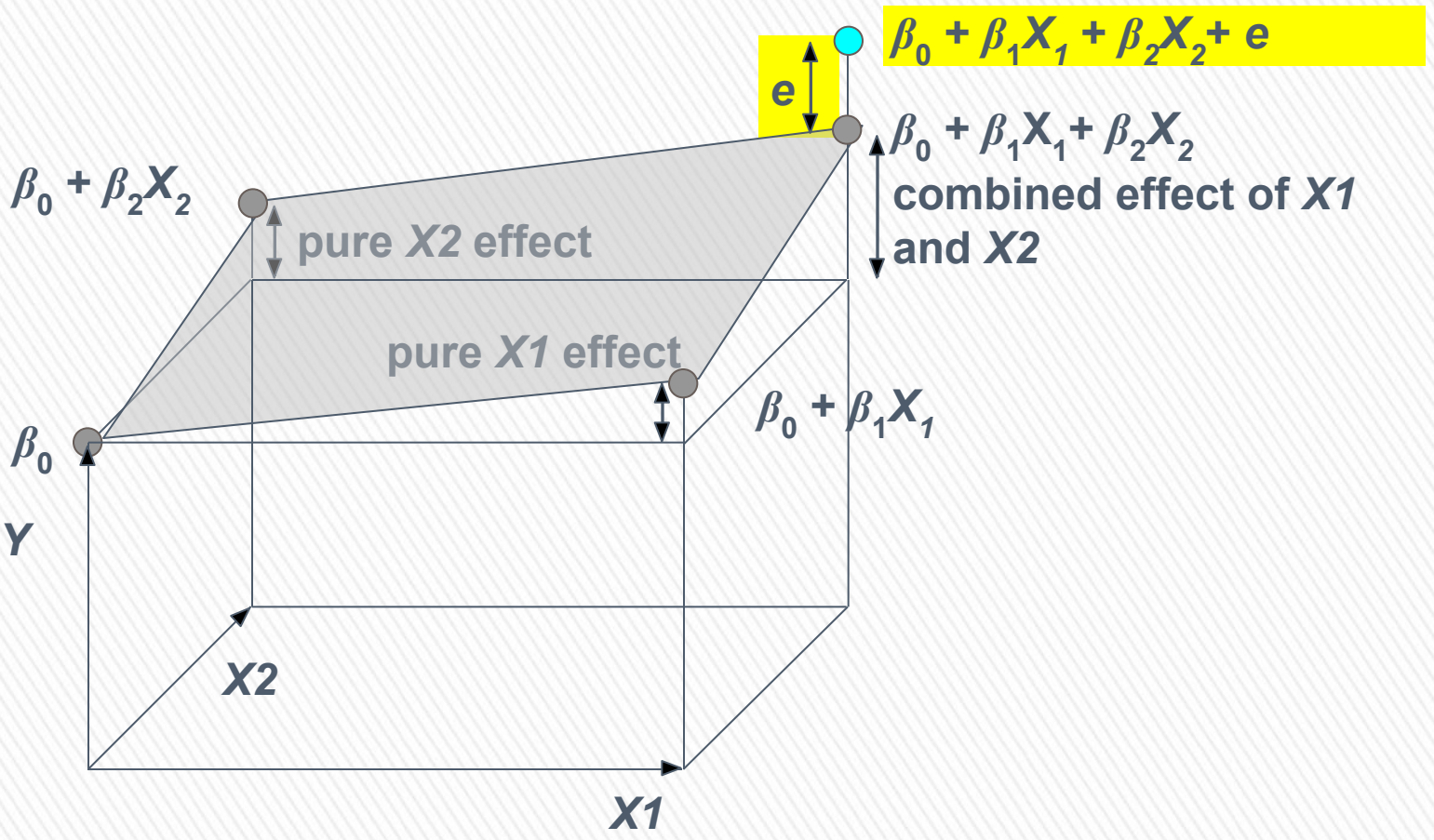
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i$$



A sample consists of a number of observations generated in this way. Note that the interpretation of the model does not depend on whether *X1* and *X2* are correlated or not.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i$$



However we do assume that the effects of *X1* and *X2* on *salary* are additive.
The impact of a difference in *X1* on *salary* is not affected by the value of *X2*, or vice versa.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i}$$

Slope coefficients are interpreted as partial slope/partial regression coefficients:

- **b_1 = average change in Y associated with a unit change in X_1, with the other independent variables held constant (all else equal);**
- **b_2 = average change in Y associated with a unit change in X_2, with the other independent variables held constant (all else equal).**

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i}$$

The regression coefficients are derived using the same least squares principle used in simple regression analysis. The fitted value of *Y* in observation *i* depends on our choice of b_0, b_1, and b_2.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i}$$

$$e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i}$$

The residual e_i in observation *i* is the difference between the actual and fitted values of *Y*.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

$$SSE = \sum e_i^2 = \sum (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})^2$$

We define *SSE*, the sum of the squares of the residuals, and choose b_0, b_1, and b_2 so as to minimize it.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

$$SSE = \sum e_i^2 = \sum (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})^2$$

$$= \sum (Y_i^2 + b_0^2 + b_1^2 X_{1i}^2 + b_2^2 X_{2i}^2 - 2b_0 Y_i - 2b_1 X_{1i} Y_i$$
$$- 2b_2 X_{2i} Y_i + 2b_0 b_1 X_{1i} + 2b_0 b_2 X_{1i} + 2b_1 b_2 X_{1i} X_{2i})$$
$$= \sum Y_i^2 + nb_0^2 + b_1^2 \sum X_{1i}^2 + b_2^2 \sum X_{2i}^2 - 2b_0 \sum Y_i$$
$$- 2b_1 \sum X_{1i} Y_i - 2b_2 \sum X_{2i} Y_i + 2b_0 b_1 \sum X_{1i}$$
$$+ 2b_0 b_2 \sum X_{2i} + 2b_1 b_2 \sum X_{1i} X_{2i}$$

$$\frac{\partial SSE}{\partial b_0} = 0 \qquad \frac{\partial SSE}{\partial b_1} = 0 \qquad \frac{\partial SSE}{\partial b_2} = 0$$

First we expand *SSE* as shown, and then we use the first order conditions for minimizing it.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

$$b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

$$b_1 = \frac{\mathrm{Cov}(X_1, Y)\mathrm{Var}(X_2) - \mathrm{Cov}(X_2, Y)\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_1)\mathrm{Var}(X_2) - [\mathrm{Cov}(X_1, X_2)]^2}$$

$$b_2 = \frac{\mathrm{Cov}(X_2, Y)\mathrm{Var}(X_1) - \mathrm{Cov}(X_1, Y)\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_1)\mathrm{Var}(X_2) - [\mathrm{Cov}(X_1, X_2)]^2}$$

We thus obtain three equations in three unknowns. Solving for b_0, b_1, and b_2, we obtain the expressions shown above.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

$$b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

$$b_1 = \frac{\mathrm{Cov}(X_1, Y)\mathrm{Var}(X_2) - \mathrm{Cov}(X_2, Y)\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_1)\mathrm{Var}(X_2) - [\mathrm{Cov}(X_1, X_2)]^2}$$

$$b_2 = \frac{\mathrm{Cov}(X_2, Y)\mathrm{Var}(X_1) - \mathrm{Cov}(X_1, Y)\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_1)\mathrm{Var}(X_2) - [\mathrm{Cov}(X_1, X_2)]^2}$$

The expression for b_0 is a straightforward extension of the expression for it in simple regression analysis.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

$$b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

$$b_1 = \frac{\mathrm{Cov}(X_1, Y)\mathrm{Var}(X_2) - \mathrm{Cov}(X_2, Y)\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_1)\mathrm{Var}(X_2) - [\mathrm{Cov}(X_1, X_2)]^2}$$

$$b_2 = \frac{\mathrm{Cov}(X_2, Y)\mathrm{Var}(X_1) - \mathrm{Cov}(X_1, Y)\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_1)\mathrm{Var}(X_2) - [\mathrm{Cov}(X_1, X_2)]^2}$$

However, the expressions for the slope coefficients are considerably more complex than that for the slope coefficient in simple regression analysis.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

$$b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

$$b_1 = \frac{\mathrm{Cov}(X_1, Y)\mathrm{Var}(X_2) - \mathrm{Cov}(X_2, Y)\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_1)\mathrm{Var}(X_2) - [\mathrm{Cov}(X_1, X_2)]^2}$$

$$b_2 = \frac{\mathrm{Cov}(X_2, Y)\mathrm{Var}(X_1) - \mathrm{Cov}(X_1, Y)\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_1)\mathrm{Var}(X_2) - [\mathrm{Cov}(X_1, X_2)]^2}$$

For the general case when there are many explanatory variables, ordinary algebra is inadequate. It is necessary to switch to matrix algebra.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES

In matrix notation OLS may be written as:

$$Y = Xb + e$$

The normal equations in matrix form are now

$$X^T Y = X^T X b$$

And when we solve it for *b* we get:

$$b = (X^T X)^{-1} X^T Y$$

where Y is a column vector of the Y values and X is a matrix containing a column of ones (to pick up the intercept) followed by a column of the X variables containing the observations on them and *b* is a vector containing the estimators of regression parameters.

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & X_{11} & X_{21} \\ 1 & X_{12} & X_{22} \\ \dots & \dots & \dots \\ 1 & X_{1n} & X_{2n} \end{bmatrix} \qquad b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

MATRIX ALGEBRA: SUMMARY

A **vector** is a collection of n numbers or elements, collected either in a column (a **column vec**tor) or in a row (a **row vector**).
A **matrix** is a collection, or array, of numbers of elements in which the elements are laid out in columns and rows. The dimension of matrix is $n \times m$ where n is the number of rows and m is the number of columns.

Types of matrices

A matrix is said to be **square** if the number of rows equals the number of columns. A square matrix is said to be **symmetric** if its (i, j) element equals its (j, i) element. A **diagonal** matrix is a square matrix in which all the off-diagonal elements equal zero, that is, if the square matrix A is diagonal, then $a_{ij} = 0$ for $i \neq j$.
The **transpose** of a matrix switches the rows and the columns. That is, the transpose of a matrix turns the $n \times m$ matrix A into the $m \times n$ matrix denoted by A^T, where the (i, j) element of A becomes the (j, i) element of A^T; said differently, the transpose of a matrix A turns the rows of A into the columns of A^T. The **inverse** of the matrix A is defined as the matrix for which $A^{-1}A=1$. If in fact the inverse matrix A^{-1} exists, then A is said to be **invertible** or **nonsingular**.

Vector and matrix multiplication

The matrices A and B can be multiplied together if they are conformable, that is, if the number of columns of A equals the number of rows of B. In general, matrix multiplication does not commute, that is, in general $AB \neq BA$.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES: EXAMPLE

Data for weekly salary based upon the length of employment and age of employees of a large industrial corporation are shown in the table.

Employee	Weekly salary ($)	Length of employment (X1, months)	Age (X2, years)
1	639	330	46
2	746	569	65
3	670	375	57
4	518	113	47
5	602	215	41
6	612	343	59
7	548	252	45
8	591	348	57
9	552	352	55
10	529	256	61
11	456	87	28
12	674	337	51
13	406	42	28
14	529	129	37
15	528	216	46
16	592	327	56

Calculate the OLS estimates for regression coefficients for the available sample. Comment on your results.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES: EXAMPLE

Y-weekly salary ($) **X1 –length of employment (months)** **X2-age (years)**

i	Y	X_1	X_2
1	639	330	46
2	746	569	65
3	670	375	57
4	518	113	47
5	602	215	41
6	612	343	59
7	548	252	45
8	591	348	57
9	552	352	55
10	529	256	61
11	456	87	28
12	674	337	51
13	406	42	28
14	529	129	37
15	528	216	46
16	592	327	56

X

1	330	46
1	569	65
1	375	57
1	113	47
1	215	41
1	343	59
1	252	45
1	348	57
1	352	55
1	256	61
1	87	28
1	337	51
1	42	28
1	129	37
1	216	46
1	327	56

Y

639
746
670
518
602
612
548
591
552
529
456
674
406
529
528
592

X^T

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
330	569	375	113	215	343	252	348	352	256	87	337	42	129	216	327
46	65	57	47	41	59	45	57	55	61	28	51	28	37	46	56

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES: EXAMPLE

	16	4291	779
X^TX	4291	1417105	227875
	779	227875	39771

	9192
X^TY	2617701
	457709

det	2105037674

min11	1417105	227875
	227875	39771

det	4432667330

min12	4291	227875
	779	39771

det	-6857264

min13	4291	1417105
	779	227875

det	-126113170

min21	4291	779
	227875	39771

det	-6857264

min22	16	779
	779	39771

det	29495

min23	16	4291
	779	227875

det	303311

min31	4291	779
	1417105	227875

det	-126113170

min32	16	779
	4291	227875

det	303311

min33	16	4291
	4291	1417105

det	4260999

matrix of minors

cofactor matrix

minors XTX	4432667330	-6857264	-126113170
	-6857264	29495	303311
	-126113170	303311	4260999

	4432667330	6857264	-126113170
$(X^TX)D$	6857264	29495	-303311
	-126113170	-303311	4260999

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES: EXAMPLE

	2,1057	0,0033	-0,0599
$(X^TX)^{-1}$	0,0033	0,00001	-0,0001
	-0,0599	-0,0001	0,002

	9192
X^TY	2617701
	457709

vector of parameters' estimates

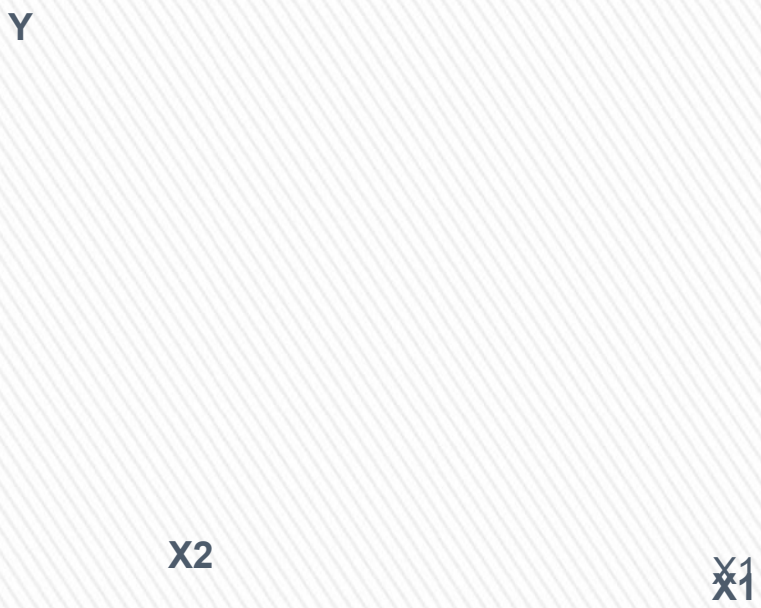
	461,85	=b0
$b=(X^TX)^{-1}X^TY$	0,671	=b1
	-1,383	=b2

Y-weekly salary ($) **X1 –length of employment (months)** **X2-age (years)**

Our regression equation with two predictors (X1, X2):

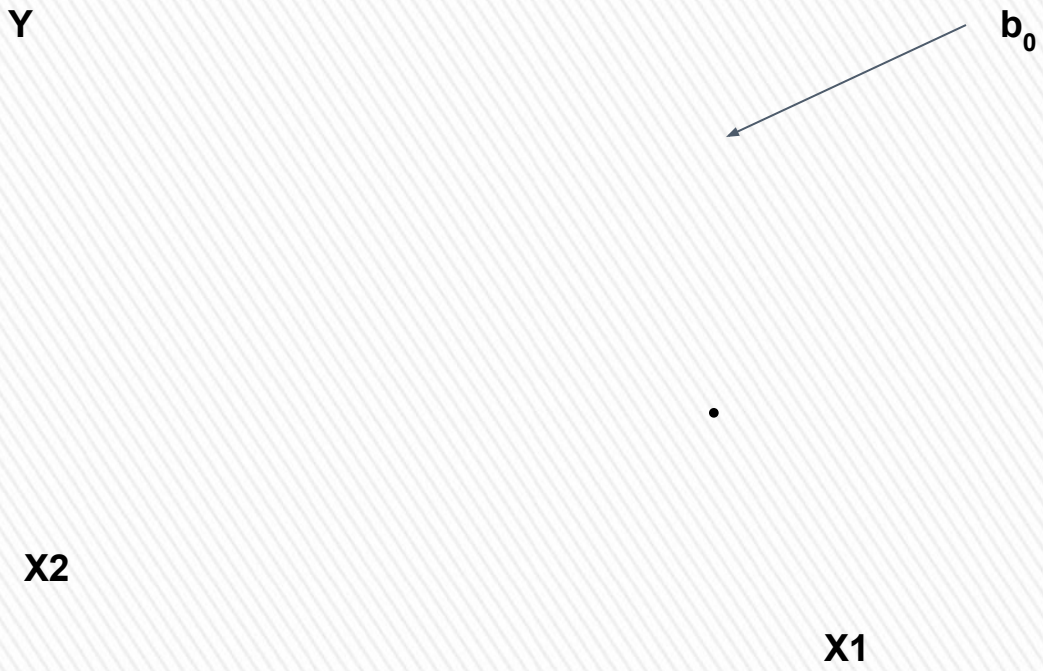
$$\hat{y}_i = 461{,}85 + 0{,}671 \cdot X_1 - 1{,}383 \cdot X_2$$

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES: EXAMPLE



These are our data points in 3dimensional space (graph drawn using *Statistica 6.0*)

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES: EXAMPLE



Data points with the regression surface (*Statistica 6.0*)

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES: EXAMPLE

Data points with the regression surface (*Statistica 6.0*) after rotation.

Dummy Variables
Nonquantitative Variables

There are times when a variable of interest in a regression cannot possibly be considered quantitative. An example is the variable gender.

Although this variable may be considered important in predicting a quantitative dependent variable, it cannot be regarded as quantitative.

The best course of action in such case is to take separate samples of males and females and conduct two separate regression analyses.

The results for the males can be compared with the results for the females to see if the same predictor variables and the same regression coefficients results.

If a large sample size is not possible, a dummy variable can be employed to introduce qualitative variable into the analysis.

A DUMMY VARIABLE
IN A REGRESSION ANALYSIS
IS A QUALITATIVE OR CATEGORICAL VARIABLE
THAT IS USED AS A PREDICTOR VARIABLE.

For example, a male could be designated with the code 0 and the female could be coded as 1.

Each person sampled could then be measured as either a 0 or a 1 for the variable gender, and this variable, along with the quantitative variables for the persons, could be entered into a multiple regression program and analyzed.

Example 1

Returning to real-estate developer, we noticed that all the houses in the population were from three neighborhoods, A, B, and C.

X1 - family income ($ 000)
X2 - family size
X3 - neighborhood

Family	Y	X1	X2	X3
1	16	22	2	B
2	17	26	2	C
3	26	45	3	A
4	24	37	4	C
5	22	28	4	B
6	21	50	3	C
7	32	56	6	B
8	18	34	3	B
9	30	60	5	A
10	20	40	3	A

Using these data, we can construct the necessary dummy variables and determine whether they contribute significantly to the prediction of home size (Y).

One way to code neighborhoods would be to define:

$$X_3 \begin{cases} 0 \ \ if\ neighborhood\ A \\ 1 \ \ if\ neighborhood\ B \\ 2 \ \ if\ neighborhood\ C \end{cases}$$

However, this type of coding has many problems. First, because 0 < 1< 2, the codes imply that neighborhood A is smaller then neighborhood B, which is smaller then neighborhood C. A better procedure is to use the necessary number of dummy variables to represent the neighborhood.

To represent the three neighborhoods, we use two dummy variables, by letting

$$X_3 = \begin{cases} 1 & if\ house\ is\ in\ A \\ 0 & otherwise \end{cases}$$

$$X_4 = \begin{cases} 1 & if\ house\ is\ in\ B \\ 0 & otherwise \end{cases}$$

What happened to neighborhood C? It is not necessary to develop a third dummy variable.

IT IS VERY IMPORTANT

THAT YOU NOT INCLUDE IT!!

If you attempted to use three such dummy variables in your model, you would receive a message in your computer output informing you that no solution exists for this model.

Why?

One predictor variable is a linear combination (including a constant term) of one or more other predictors, then mathematically no solution exists for the least squares coefficients. To arrive at a usable equation, any such predictor variable must not be included. We don't lose any information – this excluded category is the reference system. The coefficients are the measure of the categories included in comparison to this one excluded.

The final array of data is

Family	Y	X1	X2	X3 (A)	X4 (B)
1	16	22	2	0	1
2	17	26	2	0	0
3	26	45	3	1	0
4	24	37	4	0	0
5	22	28	4	0	1
6	21	50	3	0	0
7	32	56	6	0	1
8	18	34	3	0	1
9	30	60	5	1	0
10	20	40	3	1	0

$$\hat{Y} = 7{,}772 + 0{,}082x_1 + 3{,}27x_2 + 1{,}613x_3 - 0{,}9x_4$$
$$(2{,}5573)\ (0{,}1059)\ (0{,}987)\ \ (1{,}8009)\ \ (1{,}8414)$$

- If family income increases 1000$ the average home size will increase about 0,082 hundred of square feet (holding family size constant)
- If family size increases 1 person the average home size will increase about 3,27 hundred of square feet (holding family income constant)

$$\hat{Y} = 7{,}772 + 0{,}082x_1 + 3{,}27x_2 + 1{,}613x_3 - 0{,}9x_4$$
$$(2{,}5573)\ (0{,}1059)\ (0{,}987)\ \ (1{,}8009)\ \ (1{,}8414)$$

- The houses located in neighborhood A are 1,613 hundred of square feet bigger then houses from neighborhood C.

- The houses located in neighborhood B are 0,9 hundred of square feet smaller then houses from neighborhood C.

Example 2

Joanne Herr, an analyst for the Best Foods grocery chain, wanted to know whether three stores have the same average dollar amount per purchase or not. Stores can be thought of a single qualitative variable set at 3 levels – A, B, and C.

A model can be set up to predict the dollar amount per purchase:

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + e_i$$

where

Y^- expected dollar amount per purchase

$$x_1 = \begin{cases} 1 \; if \; the \; purchase \; is \; made \; in \; store \; A \\ 0 \; if \; the \; purchase \; is \; not \; made \; in \; store \; A \end{cases}$$

$$x_2 = \begin{cases} 1 \; if \; the \; purchase \; is \; made \; in \; store \; B \\ 0 \; if \; the \; purchase \; is \; not \; made \; in \; store \; B \end{cases}$$

The data

purchase (dollars)	Store A	Store B
Y	X1	X2
12,05	1	0
23,94	1	0
14,63	1	0
25,78	1	0
17,52	1	0
18,45	1	0
15,17	0	1
18,52	0	1
19,57	0	1
21,4	0	1
13,59	0	1
20,57	0	1
9,48	0	0
6,92	0	0
10,47	0	0
7,63	0	0
11,9	0	0
5,92	0	0
273,51		

The variables X1 and X2 are dummy variables representing purchases in store A or B, respectively. Note that the three levels of the qualitative variable have been described with only two variables.

$$b = (X^T X)^{-1} X^T Y$$

$$b = \begin{bmatrix} 8,72 \\ 10,01 \\ 9,42 \end{bmatrix}$$

The regression equation

$$\hat{Y} = 8,72 + 10,01x_1 + 9,42x_2$$

$$\hat{Y} = 8,72 + 10,01x_1 + 9,42x_2$$

· the average dollar amount per purchase is for store A is 10,01$ higher comparing with store C

· the average dollar amount per purchase is for store B is 9,42$ higher comparing with store C

always compare to the excluded category!!

Store A $\hat{Y} = 8,72 + 10,01 \cdot 1 + 9,42 \cdot 0 = 18,73\,\$$

Store B $\hat{Y} = 8,72 + 10,01 \cdot 0 + 9,42 \cdot 1 = 18,14\,\$$

Store C $\hat{Y} = 8,72 + 10,01 \cdot 0 + 9,42 \cdot 0 = 8,72\,\$$