Qualitative indicator variables -

An indicator variable is a distinguishing between qualitative categories.

The easiest way creating an indicator variable is to

1) choose the category to be singled out
2) In a separate column of the X matrix,
   put a 1 wherever the chosen category is correct
   put a 0 otherwise

3) This could be repeated once for each category of the qualitative variable

We have also seen the value 1 used as the first column in an X matrix to fit the mean. This is one use of an indicator variable, but they can be used to fit other means.

Take for example the data set

Category	Value
A	3
A	4
A	5
B	2
B	3
B	4
C	5
C	6
C	7

There are three groups here, with means of 4, 3 and 6 respectively. Suppose we wish to distinguish between these in an X matrix.

$$X=\begin{bmatrix}1&1&0&0\\1&1&0&0\\1&1&0&0\\1&0&1&0\\1&0&1&0\\1&0&1&0\\1&0&0&1\\1&0&0&1\\1&0&0&1\end{bmatrix} \qquad X'X=\begin{bmatrix}9&3&3&3\\3&3&0&0\\3&0&3&0\\3&0&0&3\end{bmatrix}$$

, but this matrix is singular, 4 cols for 3 groups

There are 3 groups, so we can use 2 degrees of freedom after the mean, 3 all together. How about the following options?

SAS drop last   Drop 1st Means   Contrasts   Orth Poly

$$X=\begin{bmatrix}1&1&0\\1&1&0\\1&1&0\\1&0&1\\1&0&1\\1&0&1\\1&0&0\\1&0&0\\1&0&0\end{bmatrix}\quad X=\begin{bmatrix}1&0&0\\1&0&0\\1&0&0\\0&1&0\\0&1&0\\0&1&0\\0&0&1\\0&0&1\\0&0&1\end{bmatrix}\quad X=\begin{bmatrix}1&1&0\\1&1&0\\1&1&0\\1&-1&1\\1&-1&1\\1&-1&1\\1&0&-1\\1&0&-1\\1&0&-1\end{bmatrix}\quad X=\begin{bmatrix}1&-1&1\\1&-1&1\\1&-1&1\\1&0&-2\\1&0&-2\\1&0&-2\\1&1&1\\1&1&1\\1&1&1\end{bmatrix}$$

So how do these means fit in with regression?

In regression we have quantitative variables as well as the qualitative.

$$X=\begin{bmatrix}1&X_1&1&0\\1&X_2&1&0\\1&X_3&1&0\\1&X_4&0&1\\1&X_5&0&1\\1&X_6&0&1\\1&X_7&0&0\\1&X_8&0&0\\1&X_9&0&0\end{bmatrix}\qquad X'X=\begin{bmatrix}9&\sum_{i=1}^{9}X_{i1}&3&3\\\sum_{i=1}^{9}X_{i1}&\sum_{i=1}^{9}X_{i1}^2&\sum_{i=1}^{3}X_{i1}&\sum_{i=4}^{6}X_{i1}\\3&\sum_{i=1}^{3}X_{i1}&3&0\\3&\sum_{i=4}^{6}X_{i1}&0&3\end{bmatrix}$$

Without the quantitative variable, the indicator variables fit means, which are level adjustments.

with the quantitative variable, the indicator variables fit intercepts, which are also level adjustments.

SLR: $Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$

Multiple regression with indicator variable

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

where X_{i2} is an indicator variable

When X_{i2} = 0, then $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 0 + \epsilon_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$

which is the SLR

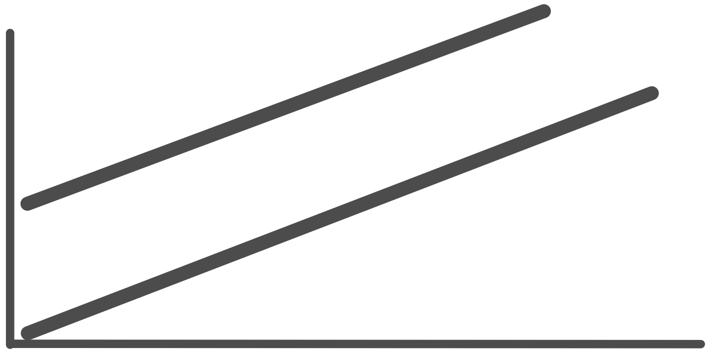
When X_{i2} = 1, then $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 1 + \epsilon_i$

where both β_0 and β_2 are constants, so let $\beta_0' = \beta_0 + \beta_1$

$$Y_i = (\beta_0+\beta_1) + \beta_1 X_{i1} + \beta_2 1 + \epsilon_i = \beta_0' + \beta_1 X_{i1} + \epsilon_i$$

which is another SLR with a different intercept

First line	β_0	$E(Y_i) = \beta_0 + \beta_1 X_{i1}$
Second Line	β_0'	$E(Y_i) = (\beta_0+\beta_1) + \beta_1 X_{i1}$



Note that there is only 1 value for the slope, so both lines have the same slope and are parallel

The effect of interactions with the indicator variable.

SLR: $Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$

Multiple regression with the added indicator variable

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

Multiple regression with an indicator variable and interaction term

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} * X_{i2} + \epsilon_i$$

When X_{i2} = 0, then $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 *0 + \beta_3 X_{i1} *0 + \epsilon_i$

$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$

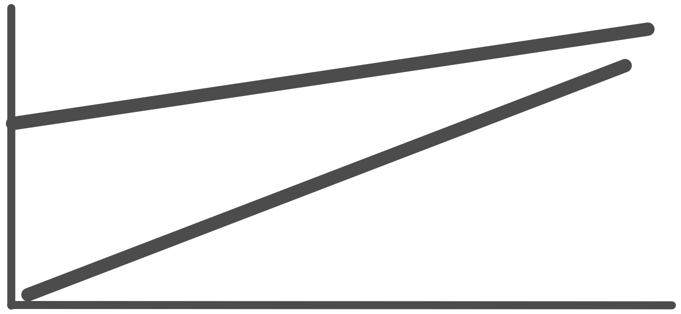
When X_{i2} = 1, then $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 *1 + \beta_3 X_{i1} *1 + \epsilon_i$

$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{i1} + \epsilon_i$

$Y_i = \beta_0' + \beta_1' X_{i1} + \epsilon_i$

NOTE that both the intercept and the slope are different (though not necessarily significantly so). This essentially fits two entirely different regression lines (two slopes and 2 intercepts).

First line	β_0 & β_1	$E(Y_i) = \beta_0 + \beta_1 X_{i1}$
Second line	β_0' & β_1'	$E(Y_i) = (\beta_0 + \beta_1) + (\beta_1 + \beta_3) X_{i1}$



Now there are 2 slopes, so each line has its own slope and intercept. These are two separate lines.

Interpreting the estimated coefficients

When the indicator is 0, the coefficients β_0 and β_1 represent the slope and intercept of the group which is allocated the “0" indication. The model reduces down to a SLR

When the indicator is 1, the coefficients β_0 and β_1 are recombined with β_2 and β_3 (respectively) to create new slopes and intercepts

$$Y_i = (\beta_0+\beta_2) + (\beta_1+\beta_3)X_{i1} + \epsilon_i$$

Therefore, β_2 is a value which shows how much GREATER (or less if $-$) the intercept for the second group (1) is than the first (0).

A test of this value (aganinst 0) is actually a test of the difference in the intercepts.

(ie. if $\beta_2 = 0$ then the intercepts are not different)

Likewise, β_3 is a value which shows how different the slope for the second group (1) is from the first (0). A test of this value is actually a test of the difference in the slopes.

HANDOUTS

1) Raw data (see coding) and Raw data Plot

2) Overall SLR, see coefficients & overall fit
Residual plot : not very good

3) Separate Intercepts : improved fit, different coefficients
Residual plot : better, but ...

4) Separate slopes and intercepts : improved fit, different coefficients
Residual plot : pretty good

5) Separate fits : note sameness of reg coeff
note differences in se, and WHAT IS TESTED

NOTES

1) It can be extended to as many categories as necessary, each getting its its own intercept and/or slope.

The textbook examples are restricted primarily to two categories. However, in practice we can adjust for as many separate slopes and intercepts as desired.

2) A model fitted to indicator variables only, with no quantitative variables, fits an ANOVA model.

3) In using this approach to test for a constant relationship between two populations, we make all of the usual assumptions (NIDrv(0,σ^2).

We must also assume that the two populations have the same variance.

Fitting the functions in this fashion produces the same regression coefficients as if the models were fitted separately. However, we have the advantage that all observations from both models contribute to the estimation of the variances.

4) The usual diagnostics apply to these approaches. However, one must have sufficient foresight to include some notation for the indicator variable in residual plots.

Residual plots of diverging lines may appear as nonhomogeneous variance when fitted as an SLR.

Inferences about regression lines

Conceptually, there are several approaches to testing inferences about the various regression lines fitted. All can be expressed in the form of “Full versus reduced models" and as EXTRA SS

1) Can fit directly as full and reduced models.

2) Can use tests of b_i provided by PROC REG for some tests

3) Can use tests of TYPE I SS provided by PROC GLM for most tests, and TYPE III for others.

There are many hypotheses which may be of interest, and which can be tested about the regression lines.

Test of interest; where $X_0=1$,
X_1=quantitative,
X_2=indicator(s),
$X_3=X_1*X_2$ interaction(s)

if there are more than one column for indicator variables, the column should be tested simultaneously (as a group).

Possible series of models in Analysis of Covariance
or Multisource Regression

Given X_0 has been fitted,   Test for a slope - $H_0{:}\beta_1=0$
a) $SSR(X_1|X_0)$ **simple linear regression**

Given X_0 has been fitted,   Test for a separate levels - $H_0{:}\beta_2=0$
b) $SSR(X_2|X_0)$ **ANOVA**

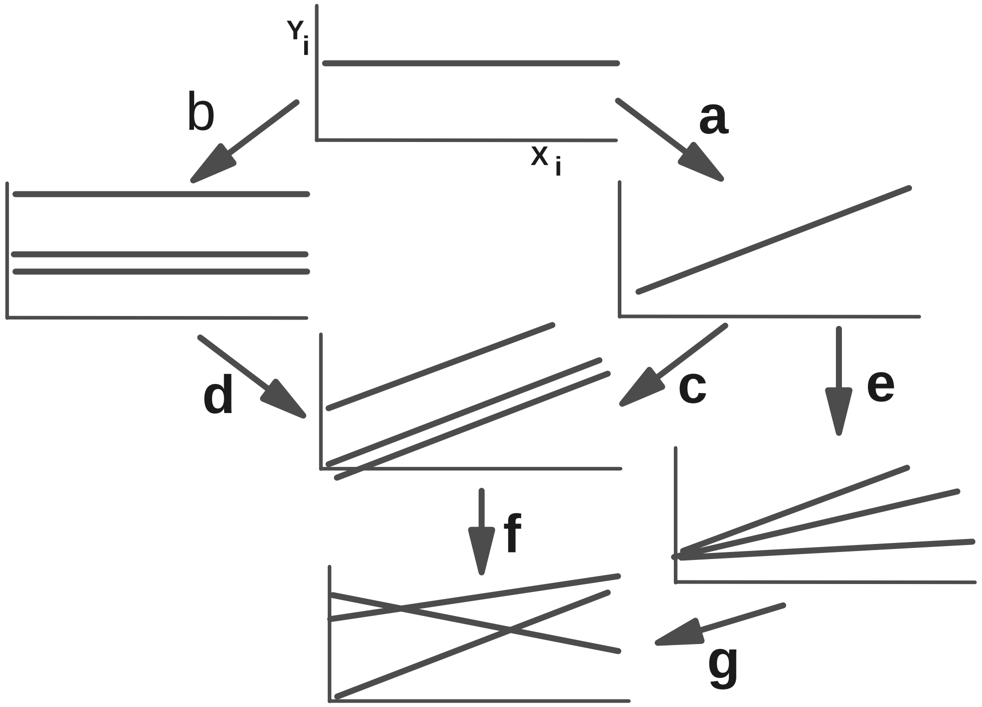
Given X_0 and X_1 have been fitted,   Test for a common intercept - $H_0{:}\beta_2=0$
c) $SSR(X_2|X_0,X_1)$ **this model has a single slope**

Given X_0 and X_2 have been fitted,   Test for a common intercept - $H_0{:}\beta_2=0$
d) $SSR(X_1|X_0,X_2)$ **this model has a single slope, same as above**

Given X_0 and X_1 have been fitted,   Test for common slopes - $H_0{:}\beta_3=0$
e) $SSR(X_3|X_0,X_1)$ **this model has a single intercept**

Given X_0 and X_1 and X_2 have been fitted,   Test for common slopes - $H_0{:}\beta_3=0$
f) $SSR(X_3|X_0,X_1,X_2)$ **two separate lines**

Given X_0 and X_1 and X_3 have been fitted,   Test for common slopes - $H_0{:}\beta_2=0$
g) $SSR(X_2|X_0,X_1,X_3)$ **two separate lines, same as above**



The Bottom line is SLR, do the slopes come from different sources?

- Start with SLR
  - Test for separate intercepts
  - Test for separate slopes (given separate intercepts)
  - Interpret from the bottom up

Analysis of Covariance approach - Bottom line is a designed experiment, do we need to adjust for slopes (a covariable).

- Start with ANOVA
  - Test for a slope
  - Test for separate slopes
  - Interpret from the bottom up;
  - If separate slopes, PUNT
  - If a covariable is required, include it in the model

Uses of the indicator variable technique

1) Fitting and testing for separate regression functions between groups

Analysis of Covariance and Multisource regression

2) Fitting a change in slope (and/or level) at some value of X.

Piecewise Linear Regression

3) Adjusting for corrections between groups over time (eg quarters).

Piecewise Linear Regression

If as some **KNOWN** point X_Δ we have a change in the slope,

BUT THE LINE IS CONTINUOUS, we can fit a model such that at that point the regression assumes a new slope.

This is fitted as $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 (X_{1i} - X_\Delta) X_{2i} + \epsilon_i$

where

X_{i2} is 0 if X_{i1} is less than X_Δ, and 1 otherwise

When X_{i1} is less than X_Δ, the model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 (X_{i1} - X_\Delta)*0 + \epsilon_i$$

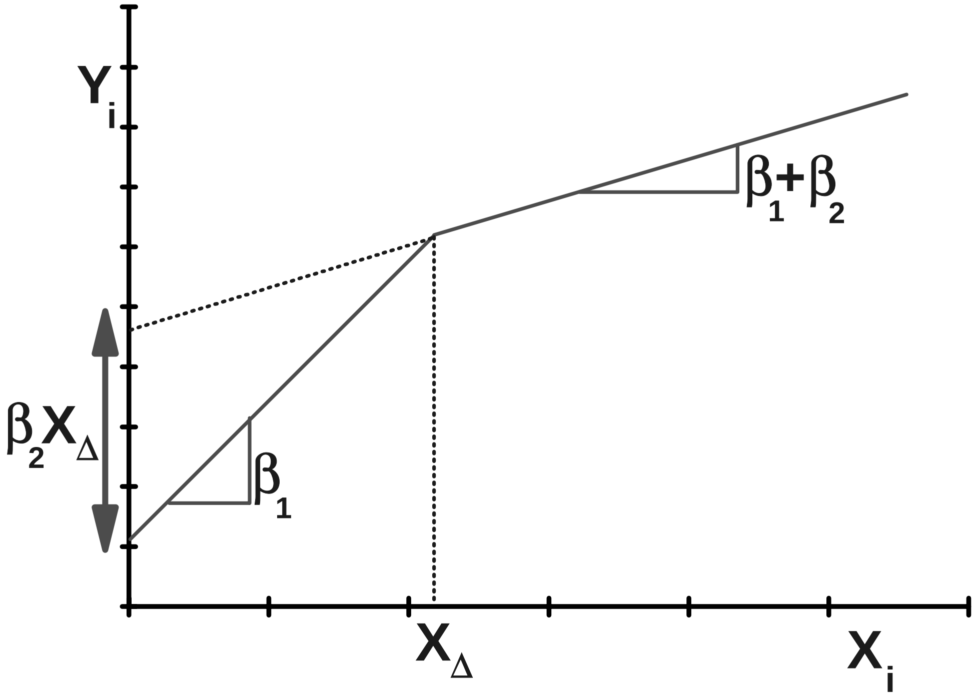
and when X_{i1} is greater than X_Δ,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 (X_{i1} - X_\Delta)*1 + \epsilon_i$$

$$Y_i = (\beta_0 - \beta_2 X_\Delta) + \beta_1 X_{i1} + \beta_2 X_{i1} + \epsilon_i$$

$$Y_i = (\beta_0 - \beta_2 X_\Delta) + (\beta_1 + \beta_2) X_{i1} + \epsilon_i$$

this forces the intercept of the second line to be whatever it must so that the two lines cross at $X_{i1} = X_\Delta$

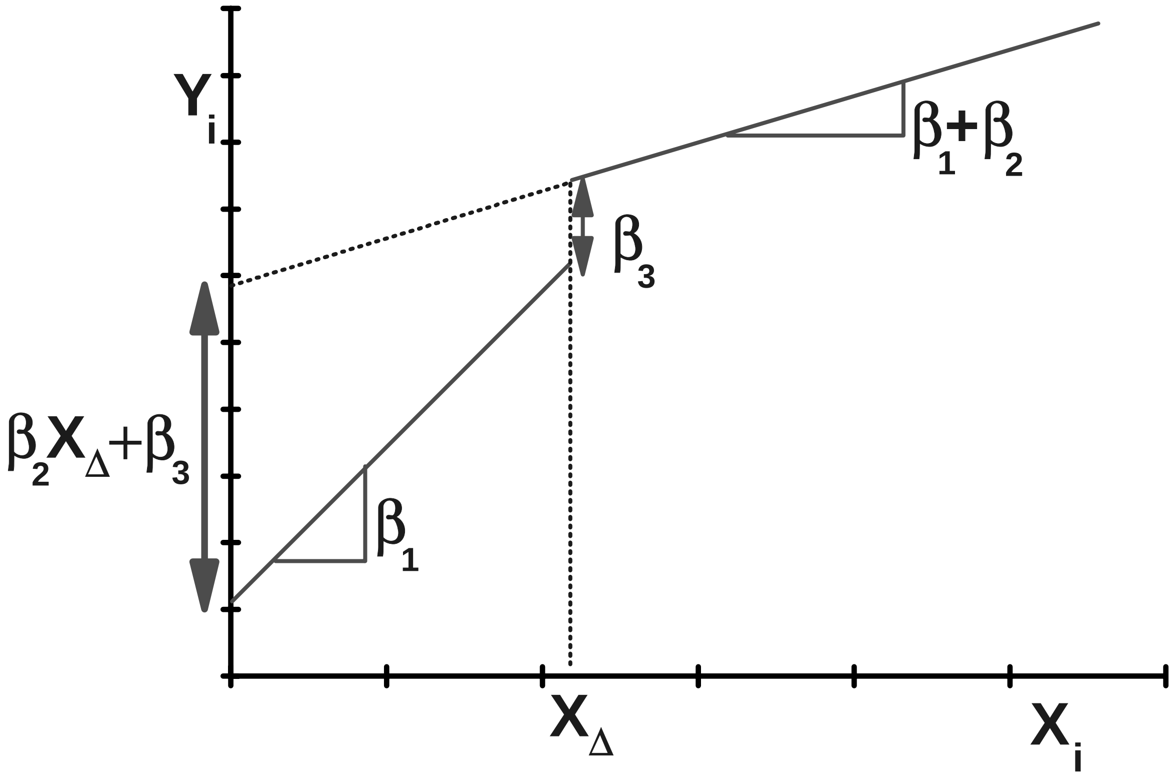


If as some point X_Δ we have a change in the slope, and the IS DISCONTINUOUS, we can fit a model such that at that point the regression assumes a new slope and intercept.

This is fitted as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 (X_{i1} - X_\Delta) X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

eg



Comparison of models : AnaCov and Piecewise. Both have an SLR as the reduced model.

ANACOV first fits a multiple regression with indicator variable (X_2)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \boldsymbol{\beta_2 X_{i2}} + \epsilon_i$$

When $X_{i2} = 0$, then

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 {*} 0 + \epsilon_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$$

When $X_{i2} = 1$, then

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 {*} 1 + \epsilon_i = (\beta_0 + \beta_2) + \beta_1 X_{i1} + \epsilon_i$$

ANACOV then adds an interaction (of the quantative term and indicator variable)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \boldsymbol{\beta_3 X_{i1} {*} X_{i2}} + \epsilon_i$$

When $X_{i2} = 0$, then

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 {*} 0 + \beta_3 X_{i1} {*} 0 + \epsilon_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$$

When $X_{i2} = 1$, then

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 {*} 1 + \beta_3 X_{i1} {*} 1 + \epsilon_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{i1} + \epsilon_i$$

Piecewise regression first fits

$$Y_i = \beta_0 + \beta_1 X_{i1} + \boldsymbol{\beta_2 (X_{i1} - X_\Delta) X_{i2}} + \epsilon_i$$

Which when X_{i1} is less than X_Δ reduces to ,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 (X_{i1} - X_\Delta) {*} 0 + \epsilon_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$$

and when X_{i1} is greater or equal to X_Δ the model is,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 (X_{i1} - X_\Delta) {*} 1 + \epsilon_i = (\beta_0 - \beta_2 X_\Delta) + (\beta_1 + \beta_2) X_{i1} + \epsilon_i$$

The second step in piecewise regression fits the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 (X_{i1} - X_\Delta) X_{i2} + \boldsymbol{\beta_3 X_{i3}} + \epsilon_i$$

Which when X_{i1} is less than X_Δ and X_2 is 0 reduces to ,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 (X_{i1} - X_\Delta) {*} 0 + \beta_3 {*} 0 + \epsilon_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$$

and when X_{i1} is greater or equal to X_Δ and X_2 and X_3 are 1 the model is,

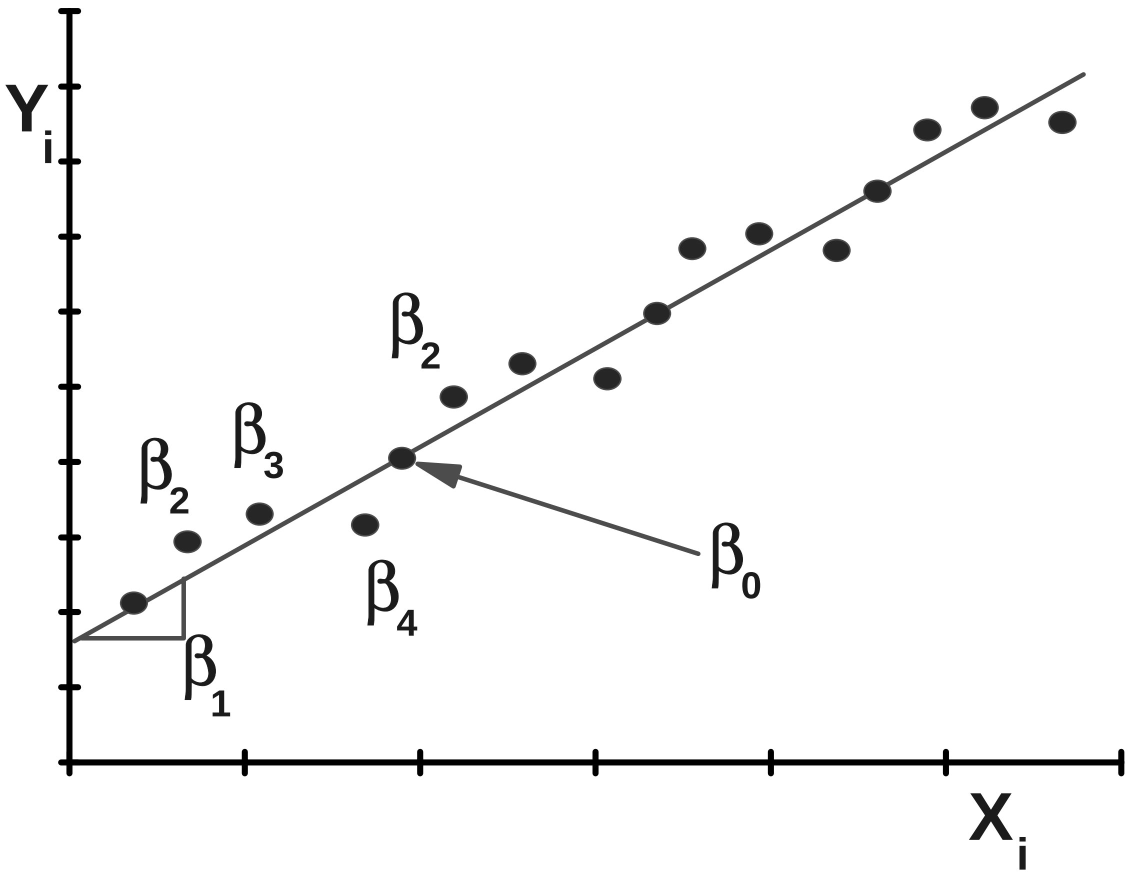
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 (X_{i1} - X_\Delta) {*} 1 + \beta_3 {*} 1 + \epsilon_i =$$
$$(\beta_0 - \beta_2 X_\Delta + \beta_3) + (\beta_1 + \beta_2) X_{i1} + \epsilon_i$$

Adjustments in time series, or for a recurring adjustment
Suppose we are regression for a trend over years, and are using quarterly data
Frequently there is some recurrent quarterly pattern
eg. in Employment, summer jobs and Christmas jobs cause higher seasonal employment.
In biology, growth also follows seasonal patterns.

Then we may wish to adjust for this recurrent pattern. Create indicator variables for quarters, and obtain the following variables

X_1 = Year
X_2 = 1 for first quarter, 0 otherwise
X_3 = 1 for second quarter, 0 otherwise
X_4 = 1 for third quarter, 0 otherwise

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i$$



eg