743- Regression and Time Series

Mamikon S. Ginovyan

Inherently Linear Models

The Polynomial & Interaction Models

Inherently Linear Models

The multiple linear regression model can be used to make inferences about some **non-linear** models, called **inherently (essentially) linear** models.

Definition.

- A non-linear regression model is called **inherently linear** if it can be transformed into a linear model, that is, if it can be **linearized by suitable transformations** of the underlying variables.
- A non-linear model is called **inherently nonlinear** if it **cannot be transformed** into linear model.

Inherently Linear Models

Assume that our underlying **nonlinear model** is given by equation

$$Y = F(x_1, x_2, \ldots, x_k; e). \qquad (1)$$

If there exist functions ***g*(.)** and ***g_i*(.)**, such that (1) can be written in the form

$$g(Y) = b_0 + b_1 g_1(x_1, \ldots, x_k) + \ldots + b_k g_k(x_1, \ldots, x_k) + e$$

or

$$Y^* = b_0 + b_1 x_1^* + \ldots + b_k x_k^* + e,$$

where $Y^* = g(Y), \quad x_i^* = g_i(x_1, \ldots, x_k), \quad i = \overline{1, k},$

then (1) is an **inherently linear** model.

Examples of inherently linear models

1. Polynomial model

$$Y = b_0 + b_1 x + b_2 x^2 + \cdots + b_k x^k + e$$

$$= b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + e, \quad (x_i = x^i)$$

$$Y = b_0 + b_1 x + b_2 x^2 + e$$

$$= b_0 + b_1 x_1 + b_2 x_2 + e, \quad (x_1 = x;\ x_2 = x^2)..$$

2. Interaction model

$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 + e$$

$$= b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + e \qquad (x_3 = x_1 x_2).$$

Examples of inherently linear models

3. Exponential model

$$Y = \exp\{b_0 + b_1 x_1 + b_2 x_2\} \cdot e \quad \Leftrightarrow$$

$$\ln Y = b_0 + b_1 x_1 + b_2 x_2 + e_1, \quad e_1 = \ln e, \quad (Y^* = \ln Y).$$

4. Reciprocal model

$$Y = \frac{1}{b_0 + b_1 x_1 + b_2 x_2 + e}, \quad \Leftrightarrow$$

$$Y^* = b_0 + b_1 x_1 + b_2 x_2 + e, \quad (Y^* = Y^{-1}).$$

Examples of inherently linear models

5. Logarithmic model (Multiplicative model)

$$Y = b_0 x_1^{b_1} x_2^{b_2} e$$

$$\Leftrightarrow \ln Y = \ln b_0 + b_1 \ln x_1 + b_2 \ln x_2 + \ln e$$

$$\Leftrightarrow Y^* = b'_0 + b_1 x_1^* + b_2 x_2^* + e_1.$$

Remark 1.

In models 3 and 5 should be assumed that the random term e has **log-normal distribution**, that is,

$e_1 = \ln e$ has normal distribution.

Examples of inherently linear models

Remark 2.

The model

$$Y = b_0 x_1^{b_1} x_2^{b_2} + e \qquad (1)$$

is quite similar to model 5, but the similarity is **deceptive** because **no transformation of model (1)** will provide a new model that is **linear** in the parameters.

So (1) is **inherently non linear** model.

Remark 3.

We will consider only **polynomial** and **interaction** models.

Polynomial Regression Model

The general two-variable ***k***-order polynomial regression model is given by equation

$$Y = g_k(x) + e, \qquad (1)$$

where

$$g_k(x) = \sum_{i=0}^{k} b_i x^i = b_0 + b_1 x + b_2 x^2 + \mathbf{L} + b_k x^k \qquad (2)$$

is a polynomial of degree ***k***.

Polynomial Regression Model

This model can be reduced to the **multiple linear** regression model by using the transformations

$$x_i = x^i, \ i = \overline{1, k}. \qquad (3)$$

Thus, the model (1) is **inherently linear**, and is equivalent to the model.

$$Y = b_0 + b_1 x_1 + \mathbf{L} + b_k x_k + e. \qquad (4)$$

For applications, the most interesting case is when ***k* = 2**, that is, the **quadratic** regression model

$$Y = b_0 + b_1 x + b_2 x^2 + e$$

$$= b_0 + b_1 x_1 + b_2 x_2 + e, \quad (x_1 = x; x_2 = x^2) \qquad (5)$$

Example 1.

Suppose we have ***n* = 5** data points given in the following table

x	-2	-1	0	1	2
Y	0	0	1	1	3

Fit a parabola to the given data using the quadratic model (5).

Solution.

Recall, first, that the given data points we have fitted by a straight line, using the two-variable linear model, and got the **prediction (regression line)**

$$\hat{Y} = .7x + 1.$$

Example 1.

Observe that the ***X*** matrix is **different** from that of in the linear case, and has the form (the ***Y*** matrix is the same).

$$X = \left[\begin{array}{ccc} x_0 & x & x^2 \\ 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{array}\right], \quad Y = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{bmatrix}.$$

Example 1.

The matrix products, ***X'X*** and ***X'Y***, are

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \\ 4 & 1 & 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{bmatrix}'$$

$$X'Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \\ 4 & 1 & 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \\ 13 \end{bmatrix}.$$

Example 1.

We omit the process of inverting and simply state that the **inverse matrix** is equal to

$$(X'X)^{-1} = \begin{bmatrix} 17/35 & 0 & -1/7 \\ 0 & 1/10 & 0 \\ -1/7 & 0 & 1/14 \end{bmatrix}.$$

[You may verify that $(X'X)^{-1}X'X = I$.]

Finally,

$$\hat{b} = (X'X)^{-1}X'Y = \begin{bmatrix} 17/35 & 0 & -1/7 \\ 0 & 1/10 & 0 \\ -1/7 & 0 & 1/14 \end{bmatrix} \begin{bmatrix} 5 \\ 7 \\ 13 \end{bmatrix} = \begin{bmatrix} 4/7 \\ 7/10 \\ 3/14 \end{bmatrix} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix}.$$

Example 1.

Hence $\hat{b}_0 = 4/7 \approx .571$, $\hat{b}_1 = 7/10 = .7$, and $\hat{b}_2 = 3/14 \approx .214$,

and the prediction equation is

$$\hat{y} = .571 + .7x + .214x^2.$$

A graph this parabola on Figure 1 will indicate a **good fit** to the data points.

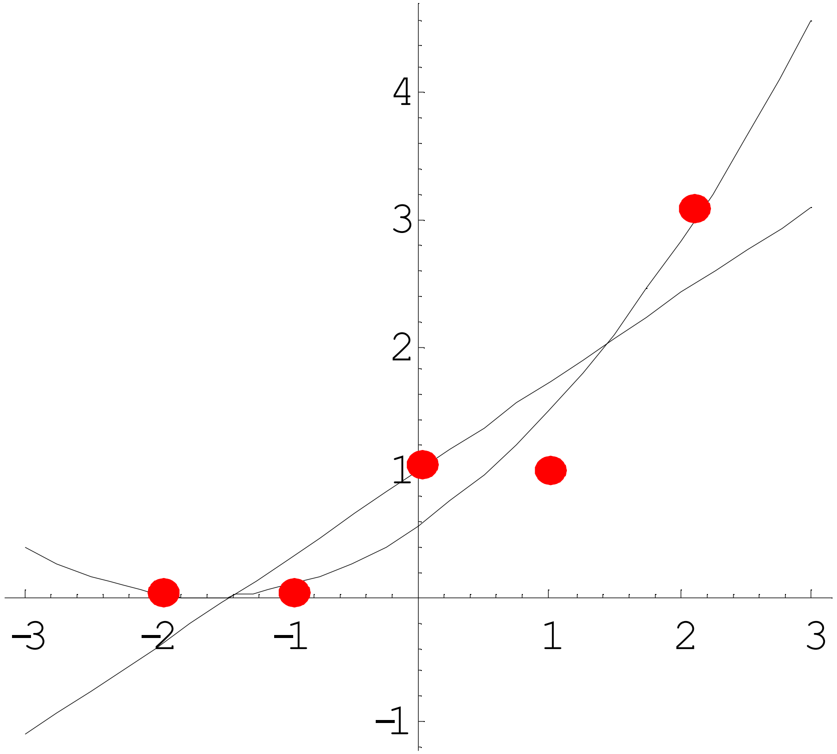


Figure 1

Example 1.

Thus, we have two prediction equations

$\hat{Y} = 1 + .7x$ (using **linear model**)

$\hat{y} = .571 + .7x + .214x^2.$ (using **quadratic model**).

The following question naturally rises.

Which model better fits given data points?

Example 2.

Do the data of Example 1 present sufficient evidence to indicate **curvature** in the response function?

(a) Test the claim using $\boldsymbol{\alpha = .05}$ and

(b) Give bounds to the attained significance level.

Example 2.

Solution.

Remark. The preceding question assumes that the probabilistic model is a realistic description of the true response and implies a test of the hypothesis

$$H_0 : b_2 = 0 \text{ versus } H_a : b_2 \neq 0$$

in the non-linear model $Y = b_0 + b_1 x + b_2 x^2 + e$ that was fit to the data in Example 1.

(If $b_2 = 0,$ the quadratic term will not appear and the expected value $\boldsymbol{Y}$ will represent a straight-line function of $\boldsymbol{x}$.)

Example 2.

Solution.

The first step in the solution is to calculate ***SSE*** and ***s***^{2} :

$$SSE = Y'Y - \hat{b}'X'Y$$

$$= 11 - [.571 \quad .700 \quad .214]\begin{bmatrix} 5 \\ 7 \\ 13 \end{bmatrix} = 11 - 10.537 = .463,$$

so then

$$s^2 = \frac{SSE}{n-3} = \frac{.463}{2} = .232 \qquad \text{and} \quad \boldsymbol{s = .48}.$$

Notice that the model contains **three parameters** and hence **SSE** is based upon ***n*** **– 3 = 2 df**.

Example 2.

The parameter $\boldsymbol{\beta_2}$ is a linear combination of $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ with $\boldsymbol{a_0} = \mathbf{0}$, $\boldsymbol{a_1} = \mathbf{0}$, and $\boldsymbol{a_2} = \mathbf{1}$. For this choice of $\boldsymbol{a}$, we have

$$\hat{b}_2 = a'b \quad \text{and} \quad a'(X'X)^{-1}a = c_{22}.$$

The calculations in Example 1 yielded

$$\hat{b}_2 = 3/14 \approx .214 \quad \text{and} \quad c_{22} = 1/14.$$

The appropriate test statistic can therefore be written as

$$T = \frac{\hat{b}_2 - 0}{S\sqrt{c_{22}}}.$$

Example 2.

The **observed value** of the test statistic is

$$t_0 = T(obs) = \frac{\hat{b}_2 - 0}{s\sqrt{c_{22}}} = \frac{.214}{.48\sqrt{1/14}} = 1.67.$$

For $\alpha = .05$, from **T** -Table we find

$$t_{a/2,(n-3)} = t_{.025,2} = 4.303.$$

Hence the **rejection region** is:

Reject $H_0 : b_2 = 0$ if $|t| \geq 4.303.$

Decision. Because $|T(obs)| = 1.67 < 4.303,$
we **cannot reject** the null hypothesis $H_0 : b_2 = 0.$

Example 2.

Remark.

We do not accept $H_0 : b_2 = 0,$ because we would need to know the probability of making a Type II error .

(b) Give bounds to the attained significance level.

Because the test is two-tailed,

$$\boldsymbol{P}\text{-value} = 2\boldsymbol{P}(\text{T} > \ \mathbf{T(obs) = 1.67}) \ ,$$

where $\boldsymbol{T}$ has a $\boldsymbol{t}$ -distribution with 2 degrees of freedom.
Using $\boldsymbol{T}$ -Table, we find that

$$\boldsymbol{P}(\mathbf{T} \ \mathbf{> 1.67) > .10}.$$

Thus, we conclude that ***P***-value >.2.
So, again we **cannot reject** $\boldsymbol{H_0}$ at **α = .05.**

Example 2.

Remark.

As a further step in the analysis, we could look at the **width of a confidence** interval for $\boldsymbol{\beta_2}$ to see whether it is short enough to detect a departure from zero that would be of practical significance.

The resulting 95% confidence interval for $\boldsymbol{\beta_2}$ is

$$\hat{b}_2 \pm t_{.025} S\sqrt{c_{22}} = .214 \pm (4.303)(.48)\sqrt{1/14} \Rightarrow .214 \pm .552.$$

Thus the CI for $\boldsymbol{\beta_2}$ is **quite wide**, suggesting that the researcher n eeds to collect **more data** before reaching a decision.

To be sure, that the quadratic regression model is informative, we need to answer the following question.

Example 3.

(a) Do the data of Example 1 provide sufficient evidence to indicate that the second-order model

$$Y = b_0 + b_1 x + b_2 x^2 + e$$

contributes information for the prediction of $\boldsymbol{Y}$?

That is, test the hypothesis

$$\boldsymbol{H_0} : \boldsymbol{\beta_1} = \boldsymbol{\beta_2} = \boldsymbol{0}$$

against the alternative hypothesis:

$\boldsymbol{H_a}$: **at least one** of the parameters $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}$ differ from 0.

Use $\boldsymbol{\alpha} = .05$.

(b) Give bounds for the attained significance level.

Example 3.

Solution.

For the **complete model**, we determined in Example 2 that **SSE$_c$ = .463.**

Because we want to test $\boldsymbol{H_0} : \boldsymbol{\beta_1} = \boldsymbol{\beta_2} = \mathbf{0}$, the appropriate **reduced model** is

$$\boldsymbol{Y} = \boldsymbol{\beta_0} + \boldsymbol{\varepsilon} \, ,$$

for which

$$Y' = [0 \quad 0 \quad 1 \quad 1 \quad 3] \quad \text{and} \quad X' = [1 \quad 1 \quad 1 \quad 1 \quad 1] \rightarrow x_0 .$$

Because $X'X = 5,$ we have $(X'X)^{-1} = 1/5,$ and

$$\hat{b} = (X'X)^{-1} X'Y = (1/5)\sum_{i=1}^{5} y_i = \overline{y} = 5/5 = 1.$$

Example 3.

Thus,

$$SSE_R = Y'Y - \hat{b}'X'Y = \sum_{i=1}^{5} y_i^2 - \overline{y}(\sum_{i=1}^{n} y_i) =$$

$$= \sum_{i=1}^{5} y_i^2 - \frac{1}{n}(\sum_{i=1}^{5} y_i)^2 = 11/(1/5)(5)^2 = 11 - 5 = 6.$$

Notice that in this case the number of independent variables in the complete model is ***k* = 2**, whereas the number of independent variables in the reduced model is ***m* = 0**.

Example 3.

Thus,

$$F(obs) = \frac{(SSE_R - SSE_C)/(k-m)}{(SSE_C)/(n-[k+1])} = \frac{(6-.643)(2-0)}{.463(5-3)} = 11.959.$$

For $\boldsymbol{\alpha} = .05$, $\boldsymbol{df_1} = \mathbf{k - m = 2}$ and $\boldsymbol{df_2} = \mathbf{n - (\ k + 1\ ) = 2,}$ from ***F***-Table we find that

$$F_{.05,2,2} = 19.00. \qquad RR: F > 19.00.$$

Hence the observed value of the test statistic $\boldsymbol{F}(\ obs\ ) = 11.959$ **does not fall in the rejectionregion**, and we conclude that, at the $\boldsymbol{\alpha} = .05$ level, there is **not enough evidence to support** a claim that either $\boldsymbol{\beta_1}$ or $\boldsymbol{\beta_2}$ differs from zero.

Example 3.

(b) Give bounds for the attained significance level.

The P-value is given by **P (F > 11.959)** where $F \sim F_{2,2}$. Using *F*-Table, we can see that

$$.05 < P\text{-value} < .10.$$

Thus, if we chose $\alpha = .05$ (in agreement with the previous discussion), there is not enough evidence to support a claim that either β_1 or β_2 differs from zero.

However, if an $\alpha = .10$ were selected, we could claim that either $\beta_1 \neq 0$ or $\beta_2 \neq 0$.
Notice that the little additional effort required to place bounds on the *P* -value provides a considerable amount of additional information.

Example 3.

Remark.

$$.05 < P\text{-value} < .10.$$

Thus, if we chose $\alpha = .05$ there is not enough evidence to support a claim that either β_1 or β_2 differs from zero.

However, if instead of $\alpha = .05$, an $\alpha = .10$ were selected, we could claim that either $\beta_1 \neq 0$ or $\beta_2 \neq 0$.

Notice that the little additional effort required to place bounds on the P -value provides a considerable amount of additional information.

Interaction Models

Consider the following regression model with two independent variables $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ and

$$Y = b_0 + b_1 x_1 + b_2 x_2 + e, \qquad (1)$$

and suppose we are interested in **impact of a change** in $\boldsymbol{x_2}$ on $\boldsymbol{Y}$ (or in $\boldsymbol{x_1}$ on $\boldsymbol{Y}$).

For the mean of $\boldsymbol{Y}$ we have

$$E(Y) = b_0 + b_1 x_1 + b_2 x_2 = (b_0 + b_1 x_1) + b_2 x_2, \text{ or}$$

$$= (b_0 + b_2 x_2) + b_1 x_1 \qquad (2)$$

Interaction Models

This implies that for any particular value of x_1 (respectively x_2), the **slope** of the straight line relating the mean values of Y to x_2 (respectively x_1) will always be the **same** = β_2 (respectively β_1).

That is, no matter what the value of x_1 (respectively x_2) is, the effect is always the same, and would be measured by β_2 (respectively β_1).

In such cases we say that the model **assumes no interaction between the independent variables** x_1 and x_2 .

Thus, (1) is a **free from interaction** model.

Interaction Models

In order to model **interaction** between $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ we use the **cross-product (or interaction)** term $\boldsymbol{x_1 x_2}$.

Therefore, we consider the model

$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 + e. \qquad (3)$$

For this model, the mean of $\boldsymbol{Y}$ is given by

$$E(Y) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 = b_0 + b_1 x_1 + (b_2 + b_3 x_1) x_2. \qquad (4)$$

Interaction Models

This implies that the **slope** of the line relating $\boldsymbol{Y}$ to $\boldsymbol{x_2}$, which is $(\boldsymbol{\beta_2} + \boldsymbol{\beta_3 x_1})$ will be **different for different values** of $\boldsymbol{x_1}$.

That is, the effect now is $(\boldsymbol{\beta_2} + \boldsymbol{\beta_3 x_1})$.
In such cases we say that the **model assumes interaction between the independent variables** $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$.

Thus, we can give the following definition.

Interaction Models

Definition.

If the change in the mean ***Y*** value associated with **one-unit increase** in on independent variable ($\boldsymbol{x_2}$) depends on the value of second independent variable ($\boldsymbol{x_1}$),
then there is **interaction** between these two variables, and this interaction is described by an additional predictor $\boldsymbol{x_3 = x_1 x_2}$.

The corresponding model is given by equation:

$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 \cdot x_2 + e \ = \ b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + e. \qquad (4')$$

Interaction Models

Remark 1.

In applied work, quadratic predictors $\boldsymbol{x_1^2}$ and $\boldsymbol{x_2^2}$ are often included to model a **curved relationship**. This leads to the **full quadratic** or **complete second-order** model

$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 + b_4 x_1^2 + b_5 x_2^2 + e. \qquad (5)$$

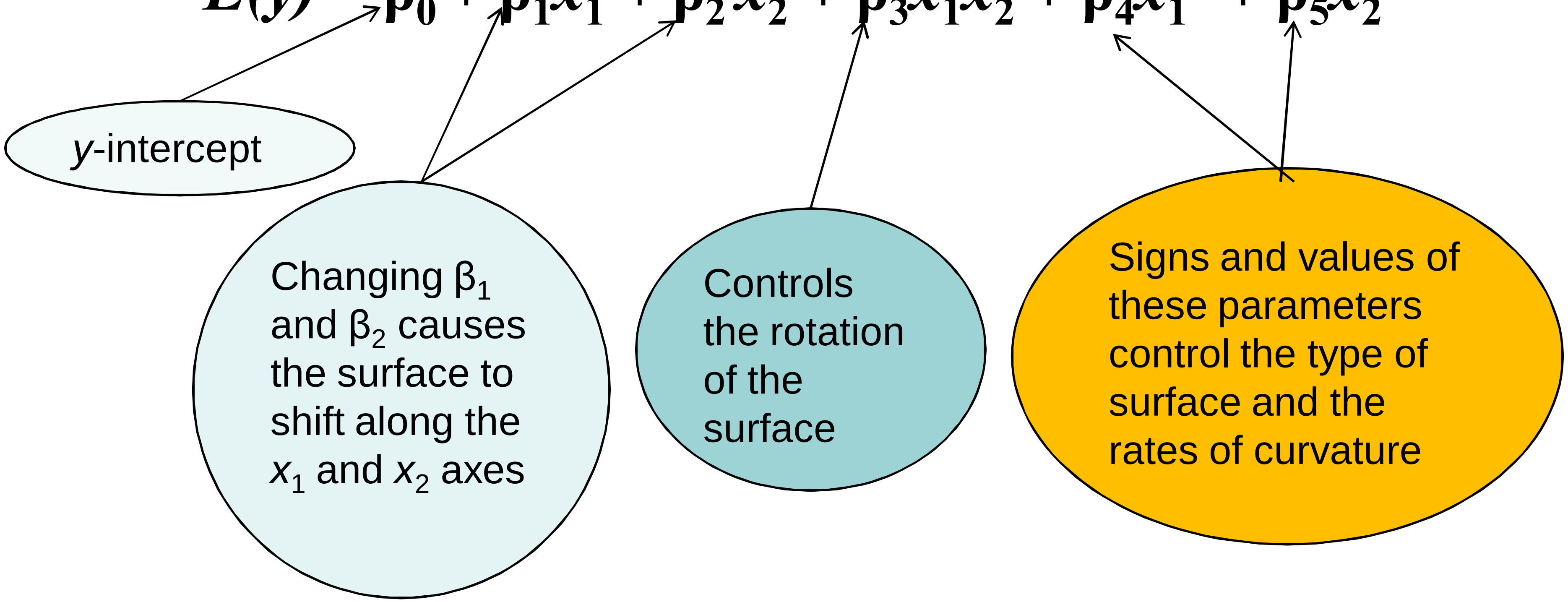
Remark 2.

Testing the null hypotheses $\boldsymbol{H_0 : \beta_3 = 0}$ provides a test for **interaction**, and testing the null hypothesis $\boldsymbol{H_0 : \beta_3 = \beta_4 = \beta_5 = 0}$ provides a test for presence of **nonlinearity**.

Complete Second-Order Model

Complete Second-Order Model with Two Quantitative Independent Variables:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$



y-intercept

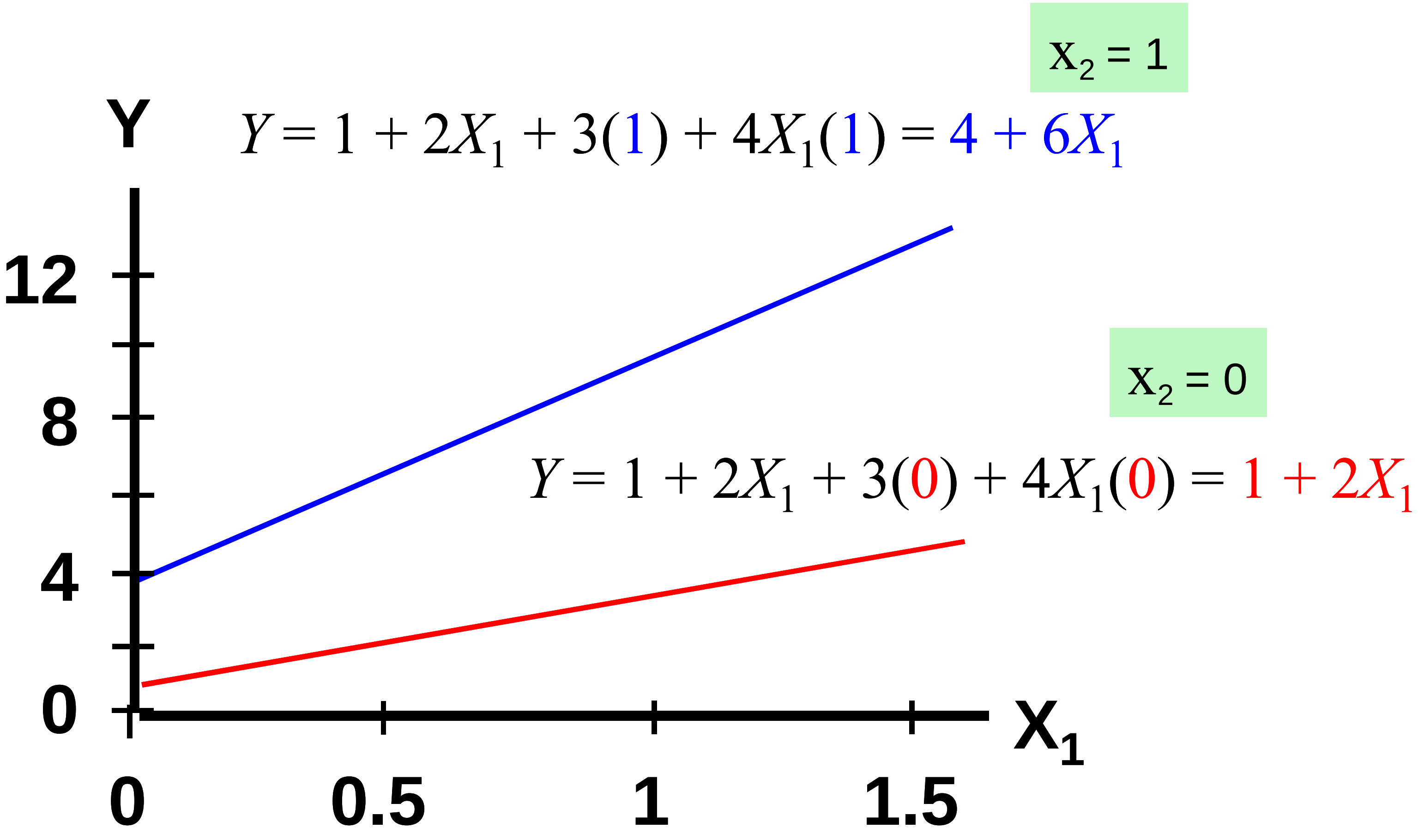
Changing β$_1$ and β$_2$ causes the surface to shift along the x_1 and x_2 axes

Controls the rotation of the surface

Signs and values of these parameters control the type of surface and the rates of curvature

Interaction Example 1

Given the model: $\boldsymbol{Y = 1 + 2X_1 + 3X_2 + 4X_1X_2}$



$x_2 = 1$

$Y = 1 + 2X_1 + 3(1) + 4X_1(1) = 4 + 6X_1$

$x_2 = 0$

$Y = 1 + 2X_1 + 3(0) + 4X_1(0) = 1 + 2X_1$

Effect (slope) of X_1 on Y does depend on X_2 value.

Regression models Involving Dummy Variables

Thus far we have considered regression models involving only **quantitative (numerical)** predictor variables.

However, in some applied problems, it is important to include the model **qualitative (categorical)** predictor variables, such as

type of college (private or state), or
type of wood (pine, oak, or walnut), and so on.

Using simple **numerical coding**, with any such variable,
we associate a **dummy (or indicator, or binary)** variable $\boldsymbol{x}$
whose possible values **0** and **1** indicate which category
is relevant for any particular observation.

A simple dummy variable model

Suppose a firm uses **two types** of production process **A** and **B**. Assuming that the output obtained from each process is **normally** distributed with **different means** but **identical variances**, we can represent the production process by a regression equation

$$Y_i = b_1 + b_2 x_i + e_i, \quad i = 1, 2, \ldots, n, \qquad (1)$$

where $\boldsymbol{Y_i}$ is the output associated with $\boldsymbol{i}$-th input process and $\boldsymbol{x_i}$ is a **dummy variable**, defined by

$$x_i = \begin{cases} 1, & \text{if output obtained from } A \\ 0, & \text{if output obtained from } B. \end{cases}$$

A simple dummy variable model

For model (1) we have

$$E[Y_i] = \begin{cases} b_1, & \text{if } x_i = 0, \\ b_1 + b_2, & \text{if } x_i = 1. \end{cases}$$

A test of the hypothesis $\boldsymbol{H_0 : \beta_2 = 0}$ is a test of the hypothesis that there is **no difference** in the output associated with processes **A** and **B**.

A model involving two dummy variables

Suppose now that a firm uses **three types** of production process **A, B** and **C**.

To describe the production process as a regression model we introduce **two dummy** predictor variables (**not three!**) ***x*₁** and ***x*₂**:

$$x_1 = \begin{cases} 1, & \text{if output obtained from } A \\ 0, & \text{otherwise.} \end{cases}$$

$$x_2 = \begin{cases} 1, & \text{if output obtained from } B \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the three production process are represented by the following combination of values of dummy variables.

A model involving two dummy variables

Process	x_1	x_2
A	1	0
B	0	1
C	0	0

Therefore the model is described by equation

$$Y = b_0 + b_1 x_1 + b_2 x_2 + e$$

with expected value ***E*[*Y*]**:

$$E[Y] = \begin{cases} b_0 + b_1, & \text{if } x_1 = 1, \quad x_2 = 0 \qquad (A) \\ b_0 + b_2, & \text{if } x_1 = 1, \quad x_2 = 0 \qquad (B) \\ b_0, & \text{if } x_1 = x_2 = 0 \qquad (C). \end{cases}$$

A model involving two dummy variables

- β_0 represents the expected value of output associated with process **C**.
- β_1 represents the difference in output associated with a change process **C** to process **A**, and
- β_2 measures the average change in output associated with a change from process **C** to presses **B**.

- A test of $H_0 : \beta_1 = 0$ is a test of the hypothesis that there is **no difference between A and C**, while
- a test of $H_0 : \beta_2 = 0$ provide a test of **no difference between B and C**.

A model involving two dummy variables

Remark 1.

Do not make the mistake of representing the dummy-variable process by using three indicator variables $\boldsymbol{x_1}$, $\boldsymbol{x_2}$ and $\boldsymbol{x_3}$, where

$$x_3 = \begin{cases} 1, & \text{if output obtained from } C \\ 0, & \text{otherwise.} \end{cases}$$

The introduction of $\boldsymbol{x_3}$ **adds no further information but does add a non-independent equation in the derivation of the least-squares estimators**.

In fact, there is **prefect collinearity** in the model because

$$x_3 = 1 - x_1 - x_2 .$$

Mixed dummy-continuous models

We denote by

$\boldsymbol{x_i}$ the **continuous (quantitative)** predictor variables, and

$\mathbf{z}_{\boldsymbol{i}}$ the **dummy** predictor variables.

1. Pure continuous model

$$Y = b_0 + b_1 x_1 + \mathbf{L} + b_k x_k + e .$$

2. Pure dummy model

$$Y = b_0 + b_1 z_1 + \mathbf{L} + b_m z_m + e$$

Mixed dummy-continuous models

3. Mixed continuous-dummy model without interaction

$$Y = b_0 + b_1 x + b_2 z + e$$

$$\begin{aligned} E(Y) &= b_0 + b_2 + b_1 x \quad \text{if } z = 1 \\ &= b_0 + b_1 x \quad \text{if } z = 0. \end{aligned}$$

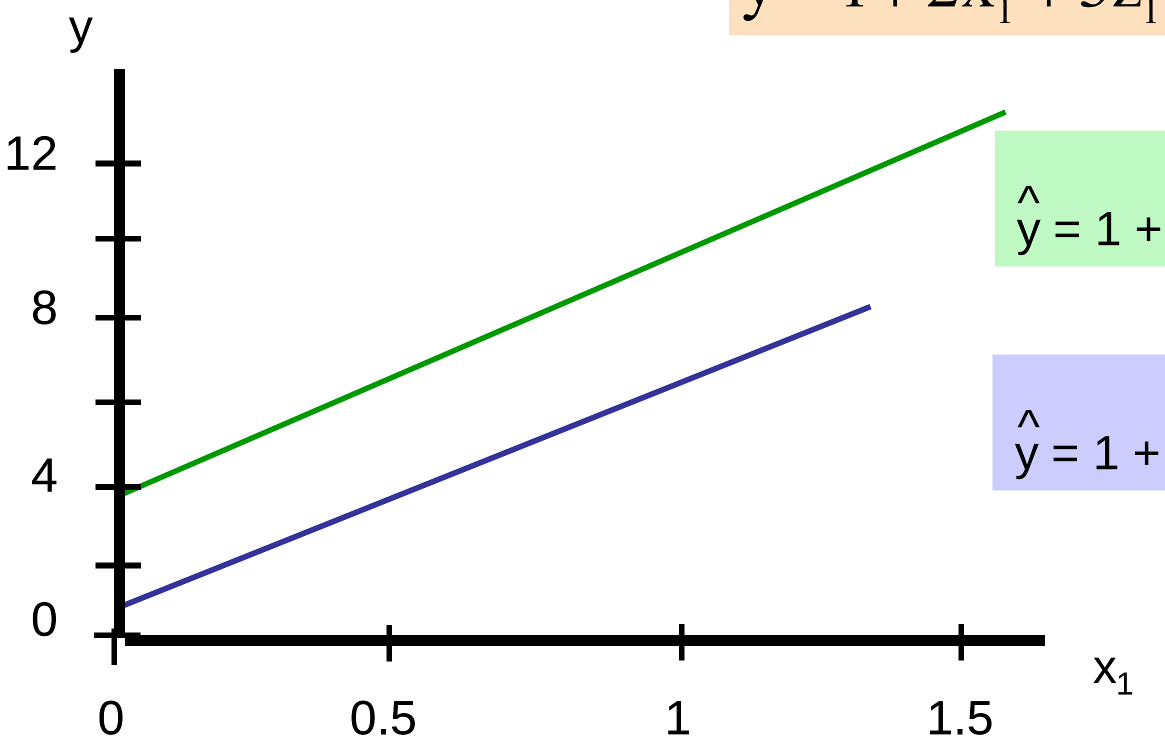
That is, in this case, the regression lines are **parallel** and can **differ only** by ***y*** - **intercepts.**

A test of whether the ***y*** -intercept change is statistically significant provides testing the null hypothesis $\boldsymbol{H_0 : \beta_2 = 0}$.

Mixed continuous-dummy model without interaction

Let $\boldsymbol{x_1}$ be a continuous predictor variable and $\boldsymbol{z_1}$ be a dummy predictor variable, and let the estimated regression equation is:

$$\hat{y} = 1 + 2x_1 + 3z_1$$



$z_1 = 1$:
$\hat{y} = 1 + 2x_1 + 3(1) = 4 + 2x_1$

$z_1 = 0$:
$\hat{y} = 1 + 2x_1 + 3(0) = 1 + 2x_1$

Slopes are equal if the effect of $\mathbf{x_1}$ on **y** does not depend on $\mathbf{z_1}$ value.

Mixed continuous-dummy model with interaction

4. Mixed continuous-dummy model with interaction

$$
\begin{aligned}
Y &= b_0 + b_1 x + b_2 z + b_3 xz + e \\
E(Y) &= b_0 + b_2 + (b_1 + b_2)x \quad \text{if } z = 1 \\
&= b_0 + b_1 x \quad \text{if } z = 0.
\end{aligned}
$$

That is, in this case, the regression lines can have **both different slopes** and ***y*** -**intercepts**.

First test for interaction:

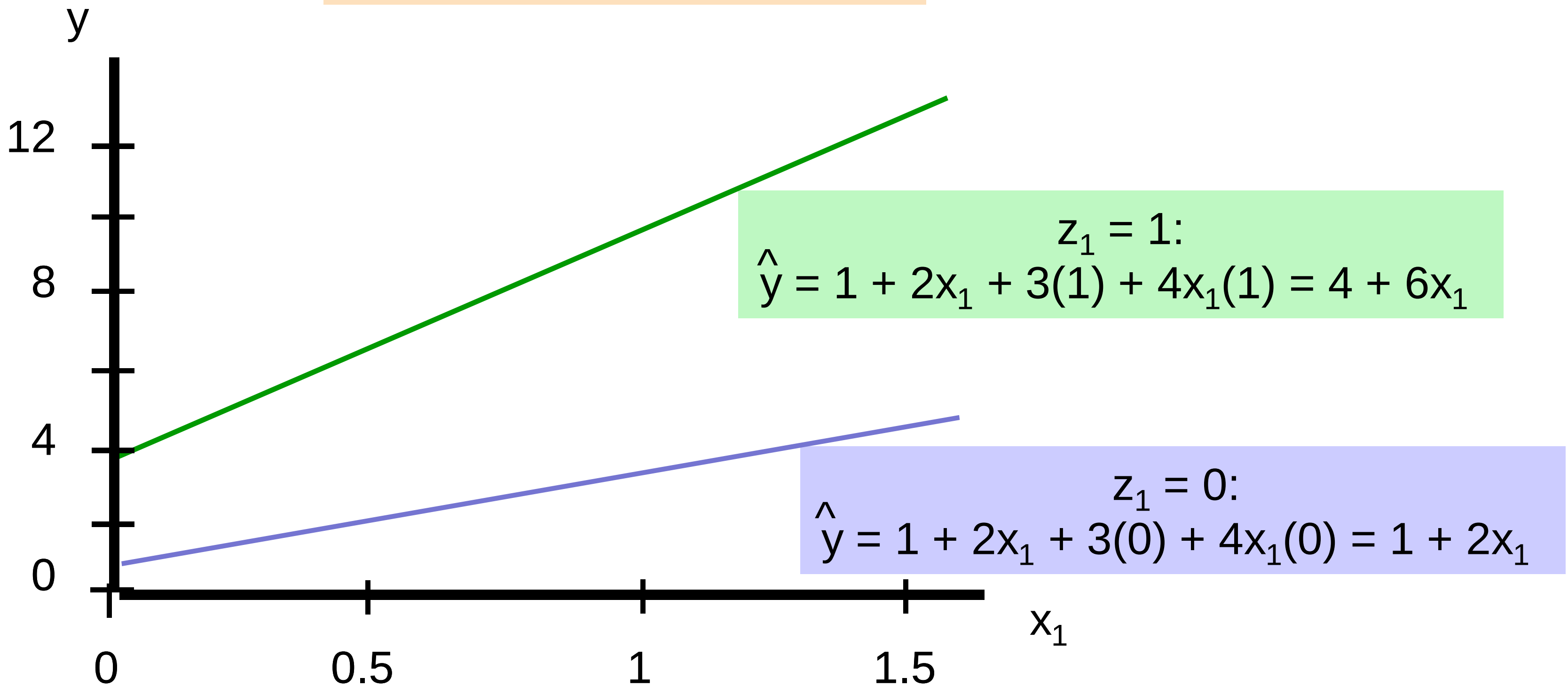
$$\boldsymbol{H_0 : \beta_3 = 0} \text{ vs. } \boldsymbol{H_a : \beta_3 \neq 0}$$

If we do not reject $\boldsymbol{H_0}$, then we can use model 3 (parallel regression lines) to make statistical inferences.

Mixed continuous-dummy model with interaction

Let $\boldsymbol{x_1}$ be a continuous predictor variable and $\mathbf{z_1}$ be a dummy predictor variable, and let the estimated regression equation is:

$$\hat{y} = 1 + 2x_1 + 3z_1 + 4x_1z_1$$



$z_1 = 1$:
$\hat{y} = 1 + 2x_1 + 3(1) + 4x_1(1) = 4 + 6x_1$

$z_1 = 0$:
$\hat{y} = 1 + 2x_1 + 3(0) + 4x_1(0) = 1 + 2x_1$

Slopes are different if the effect of $\mathbf{x_1}$ on $\mathbf{y}$ depends on $\mathbf{z_1}$ value.

Multicollinearity

Multicollinearity comes in two forms:

- **Extreme (perfect)** and
- **Non- extreme.**

1. Extreme (Perfect) Collinearity

One of the assumptions of the multiple regression model is that **there is no exact linear relationship between any of the independent variables** in the model.

If such a **linear relationship does exist**, we say that the independent variables are **perfectly collinear**.

Multicollinearity

Assume, for example, that we have a multiple regression model described by equation

$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 + e,$$

and assume that

$$\boldsymbol{X_4 = c_1 + c_2 X_3},$$

where $\boldsymbol{c_1}$ and $\boldsymbol{c_2}$ are some non-zero constant, then

$$\boldsymbol{Corr}\ (\boldsymbol{X_3}, \boldsymbol{X_4}) = 1,$$

and the variables $\boldsymbol{X_3}$ and $\boldsymbol{X_4}$ are **perfectly collinear**.

Multicollinearity

Perfect collinearity is easy to discover because it will be **impossible** to calculate **least-squares estimates** of the parameters.

With collinearity, the system of equations to be solved contains two or more equations which are **non independent.**

In such cases

$$\det(X'X) = 0,$$

and hence the **inverse** $(X'X)^{-1}$ does not exist.

Non-Extreme Collinearity

Non-Extreme Collinearity arises when
two or more variables are **highly (but not perfectly)**
correlated with each other, that is,
the correlation between these variables is **close to 1 or -1.**

Suppose two variables are related in this manner.
Then it **will be possible to obtain least-squares estimates**
of the regression coefficients, but
interpretation of the coefficients will be quite **difficult**.

Non-Extreme Collinearity

Thus, the **presence of multicollinearity** implies that there will be **very little data** in the sample to give one confidence about such an interpretation.

The distributions of the estimated regression parameters are **quite sensitive** to the

- **correlation between independent variables**, and
- **magnitude of the standard error** of the regression.

Indications of Multicollinearity

How Can Multicollinearity be Diagnosed?

The easiest ways to tell whether multicollinearity is causing problems are:

1) To examine the standard errors of the coefficients.

If several coefficients have high standard errors, and dropping one or more variables from the equation lowers the standard errors of the remaining variables,
could be indicative of multicollinearity in the model.

Indications of Multicollinearity

How Can Multicollinearity be Diagnosed?
The easiest ways to tell whether multicollinearity is causing problems are:

2) To examine the covariance between estimated parameters.

A high degree of collinearity will be associated with a relatively high (in absolute value) covariance between estimated parameters.

3) To examine the value of the test statistic:

An estimated model with **high standard errors** and **low *t*- test** could be indicative of multicollinearity.

Standardized Coefficients and Elasticities

Standardized Coefficients

describe the **relative importance of the independent variables** in a multiple regression model.

To calculate **standardized coefficients**, we simply perform a linear regression in which
each variable is **normalized** by subtracting its **mean** and dividing by its estimated **standard deviation**.

Then the **normalized** (**standardized**) regression looks as follows:

Standardized Coefficients and Elasticities

Then the **normalized** (**standardized**) regression looks as follows:

$$\frac{Y-\bar{Y}}{s_Y}=b_1^*\frac{X_{1i}-\bar{X}_1}{s_{X_1}}+b_2^*\frac{X_{2i}-\bar{X}_2}{s_{X_2}}+\mathbf{L}+b_k^*\frac{X_{ki}-\bar{X}_k}{s_{X_k}}+e_i.$$

or

$$Y^*=b_1^*X_{1i}^*+b_2^*X_{2i}^*+\mathbf{L}+b_k^*X_{ki}^*+e_i.$$

- The standardized coefficients bear a **close relationship to the estimated coefficients** of the original non-normalized multiple regression model.
- It is not difficult to show that

$$\hat{b}_j^*=\hat{b}_j\frac{s_{X_j}}{s_Y}, \quad j=1,2,3,\mathbf{L},k.$$

Standardized Coefficients and Elasticities

Definition: An **Elasticity** measures the **effect on the dependent variable (*Y*) of a 1 percent change in an independent variable (*X*).**

For example, the **elasticity** of ***Y*** with respect to $\boldsymbol{X_3}$ is the percentage change in ***Y*** divided by the percentage change in $\boldsymbol{X_3}$.

In general, the **elasticities** are **not constants** but change when measured at different points along the regression line.

For the **j-th** coefficient the **elasticity** is evaluated by formula:

$$E_j = \hat{b}_j \cdot \frac{\bar{X}_j}{\bar{Y}} \approx \frac{\partial Y}{\bar{Y}} / \frac{X_j}{\bar{X}_j} = \frac{\partial Y}{\bar{Y}} \cdot \frac{\bar{X}_j}{X_j}, \quad j = 1, 2, 3, \ldots, k.$$

Partial Correlation

In the multiple regression models, it is natural to extend the **simple correlation concept** to see how much the dependent variable (Y) and one independent variable (X) are related after **netting out the effect** of other independent variables in the model.

To do so, we consider the model,

$$Y_i = b_1 + b_2 X_{2i} + b_3 X_{3i} + e_i, \quad i = 1, 2, \ldots, n.$$

The **partial correlation coefficient (PCC)** between $\boldsymbol{Y}$ and $\boldsymbol{X_2}$ must be defined in such a way that it measures the effect of $\boldsymbol{X_2}$ on $\boldsymbol{Y}$ which is **not accounted** for by the other variables in the model.

Partial Correlation

More specifically, the PCC is calculated by **eliminating** the linear effect of $\boldsymbol{X_3}$ on $\boldsymbol{Y}$ (as well as the linear effect of $\boldsymbol{X_3}$ on $\boldsymbol{X_2}$) and then running the appropriate regression.

The steps are as follows:

1. Run the regression of $\boldsymbol{Y}$ on $\boldsymbol{X_3}$ and obtain fitted values

$$\hat{Y} = \hat{a}_1 + \hat{a}_2 X_3.$$

2. Run the regression of $\boldsymbol{X_2}$ on $\boldsymbol{X_3}$ and obtain fitted values

$$\hat{X}_2 = \hat{g}_1 + \hat{g}_2 X_3.$$

3. Remove the influence of $\boldsymbol{X_3}$ on the both $\boldsymbol{Y}$ and $\boldsymbol{X_2}$.

Let $Y^* = Y - \hat{Y}$ and $X_2^* = X_2 - \hat{X}_2$.

Partial Correlation

4. The **partial correlation** between X_2 and Y is then the **simple correlation** between Y^* and $X_2{}^*$:

$$\mathbf{PCC}(X_2\, ,\, Y\,) = \mathbf{Corr}\ (Y^*, X_2{}^*).$$

To see why the regression of Y^* on $X_2{}^*$ will give us the **PCC**, **note that**

Y^* and $X_2{}^*$ are both **uncorrelated** with X_3 (by construction).

Then the regression of Y^* and $X_2{}^*$ relates

the part of Y which is uncorrelated with X_3
to the part of X_2 which is uncorrelated with X_3 .

Partial Correlation

We denote the **PCC** and **simple correlations** as follows;

$r_{YX_2 \cdot X_3}$ = partial correlation of $\boldsymbol{Y}$ and $\boldsymbol{X_2}$ (controlling for $\boldsymbol{X_3}$);

r_{YX_2} = simple correlation between $\boldsymbol{Y}$ and $\boldsymbol{X_2}$;

$r_{X_2X_3}$ = simple correlation between $\boldsymbol{X_2}$ and $\boldsymbol{X_3}$.

Partial Correlation

We have the following

relationship between **partial** and **simple correlations.**

We state the result without proof (the details are complicated):

$$r_{YX_2\cdot X_3} = \frac{r_{YX_2} - r_{YX_3} r_{X_2X_3}}{\sqrt{1 - r^2_{X_2X_3}}\sqrt{1 - r^2_{YX_3}}} \qquad (1)$$

$$r_{YX_3\cdot X_2} = \frac{r_{YX_3} - r_{YX_2} r_{X_2X_3}}{\sqrt{1 - r^2_{X_2X_3}}\sqrt{1 - r^2_{YX_2}}} \qquad (2)$$

Partial Correlation

The relationship between PCC and Coefficient of Determination *R^2*.

In the two-variable **linear model** we have proved that

$$R^2 = r_{YX}^2$$

It is also possible to interpret the **PCC** between ***Y*** and ***X_2*** as the square root of the percentage of variance in ***Y*** which is not accounted for by ***X_3*** but which is accounted for by the part of ***X_2*** which is uncorrelated with ***X_3***.

We have the following **relationship** between **multiple and partial correlations**:

$$r^2{}_{YX_2 \cdot X_3} = \frac{R^2 - r^2{}_{YX_3}}{1 - r^2{}_{YX_3}} \Leftrightarrow 1 - R^2 = (1 - r^2{}_{YX_3})(1 - r^2{}_{YX_2 \cdot X_3}) \qquad (3)$$

PCC-Example.

The following table contains data for the winning weights in a weightlifting competition:

X_{2i}	0	0	0	1	1	1	1
X_{3i}	40	50	60	50	60	70	80
Y_i	63	72	81	85	100	108	125

where the independent variables are

$$X_2 = \begin{cases} 1, & \text{if the contestant is male} \\ 0, & \text{if the contestant is female,} \end{cases}$$

X_3 = the contestant's weight (in kilograms), and
Y = winning lift (in kilograms).

PCC-Example.

You are **given** $r_{YX_2} = .8;\; r_{YX_3} = .95;\; r_{X_2X_3} = .6.$

Find $r_{YX_3 \cdot X_2}$, $r_{YX_2 \cdot X_3}$ and $\boldsymbol{R^2}$.

Solution. We have

$$r_{YX_3 \cdot X_2} = \frac{r_{YX_3} - r_{YX_2} r_{X_2X_3}}{\sqrt{1-r^2_{X_2X_3}}\sqrt{1-r^2_{YX_2}}} = \frac{.95-(.8)(.6)}{\sqrt{1-(.6)^2}\sqrt{1-(.8)^2}} = .98.$$

$$r_{YX_2 \cdot X_3} = \frac{r_{YX_2} - r_{YX_3} r_{X_2X_3}}{\sqrt{1-r^2_{X_2X_3}}\sqrt{1-r^2_{YX_3}}} = \frac{.8-(.95)(.6)}{\sqrt{1-(.6)^2}\sqrt{1-(.95)^2}} = .92.$$

$$R^2 = 1-(1-r^2{}_{YX_3})(1-r^2{}_{YX_2 \cdot X_3}) = 1-[1-(.95)^2][1-(.92)^2] = .985.$$

Piecewise Linear Regression

Most of the regression models we have studied have been **continuous**, with small changes in one variable having a measurable effect on another variable.

This framework was modified when we used **dummy variables** to account for **shifts in slope or intercept or both**.

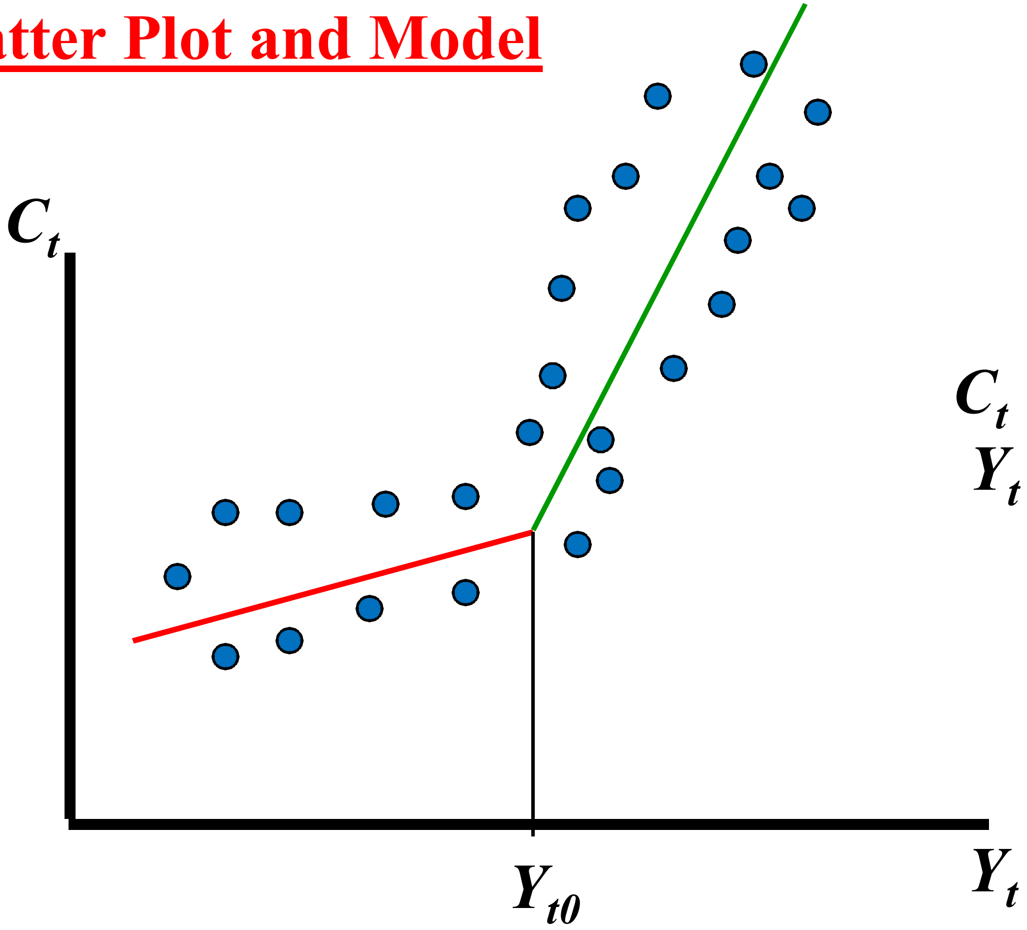
We extend now the analysis **one further step** by allowing for **changes in slope**, assuming that the line being estimated is **continuous**.

A simple example is drawn in Fig. 1.

The true model is **continuous**, with a **structural break**.

Piecewise Linear Regression

Scatter Plot and Model



C_t = consumption
Y_t = income

structural break

Piecewise Linear Regression

For example, if we were explaining **consumption as a function of income**, the **structural break** might occur sometime during **World War II** (or there might be two breaks, one at the beginning and one at the end).

Note that there is **no discontinuity or shift** in the consumption level from year to year.

This piecewise linear model consists of two straight-lines.

Piecewise Linear Regression

Piecewise linear models are special cases of a much larger set of models or relationships called **spline functions**.

Spline functions
have **distinct pieces**, but the **curve** representing each piece is a **continuous function** and **not necessarily a straight line**.

In a typical case, the spline is chosen to be a **polynomial of the third degree** and the procedure guarantees that the first and second derivatives will be **continuous**.

Piecewise Linear Regression

To estimate the model given in Fig. 1, consider the expression

$$C_t = b_1 + b_2 Y_t + b_3 (Y_t - Y_{t_0}) D_t + e_t$$

where

$\boldsymbol{C_t}$ = **consumption**

$\boldsymbol{Y_t}$ = **income**

$\boldsymbol{Y_{t0}}$ = income in year in which **structural break** occurs,

and

$$D_t = \begin{cases} 1, & \text{if} \quad t > t_0 \\ 0, & \text{otherwise.} \end{cases}$$

Piecewise Linear Regression

For years **before** and including the break, $\boldsymbol{D_t} = \mathbf{0}$, so that

$$E(C_t) = b_1 + b_2 Y_t$$

However, after the break, $\boldsymbol{D_t} = \mathbf{1}$, so that

$$E(C_t) = b_1 + b_2 Y_t + b_3 Y_t - b_3 Y_{t_0}$$

$$= (b_1 - b_3 Y_{t_0}) + (b_2 + b_3) Y_t .$$

Before the break the line has slope $\boldsymbol{\beta_2}$, but the slope changes to $\boldsymbol{\beta_2} + \boldsymbol{\beta_3}$ afterward (and the intercept changes as well).

Note, however, that there is **no discontinuity** since

$$E(C_{t_0}) = b_1 + b_2 Y_{t_0}$$

$$= (b_1 - b_3 Y_{t_0}) + (b_2 + b_3) Y_t = b_1 + b_2 Y_{t_0} .$$

Piecewise Linear Regression

Note also that when $\boldsymbol{\beta_3}$=0, the consumption equation reduces to a single straight-line segment, so that a ***t*** - test of $\boldsymbol{H_0}$: $\boldsymbol{\beta_3}$=0 provides a simple test for **structural change**.

What if there were **two structural breaks**, occurring at times $\boldsymbol{t_0}$ and $\boldsymbol{t_1}$?

The appropriate model equation then would be

$$C_t = b_1 + b_2 Y_t + b_3 (Y_t - Y_{t_0}) D + b_4 (Y_t - Y_{t_1}) D' + e_t,$$

where $\boldsymbol{Y_{t1}}$ = income in year in which the **second structural break occurs**, and

Piecewise Linear Regression

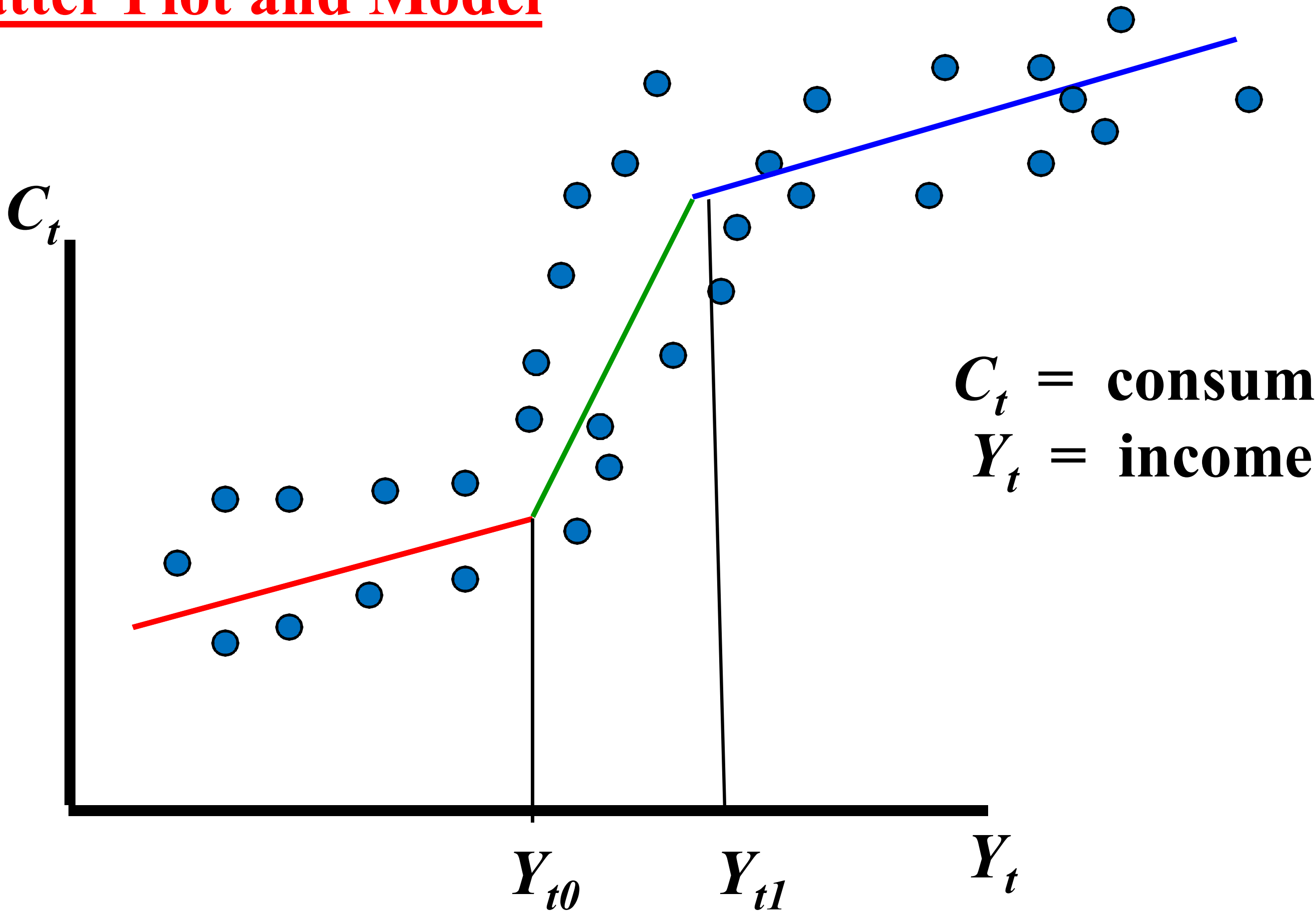
$$D_t^{'} = \begin{cases} 1, & \text{if} \quad t > t_1 \\ 0, & \text{otherwise.} \end{cases}$$

The equations of each of the three line segments are then

$$E(C_t) = \begin{cases} b_1 + b_2 Y_t, & \text{if} \quad 0 < t \leq t_0 \\ (b_1 - b_3 Y_{t_0}) + (b_2 + b_3) Y_t, & \text{if} \quad t_0 \leq t \leq t_1 \\ (b_1 - b_3 Y_{t_0} - b_4 Y_{t_1}) + (b_2 + b_3 + b_4) Y_t, & \text{if} \quad t > t_1. \end{cases}$$

Piecewise Linear Regression

Scatter Plot and Model



C_t = consumption

Y_t = income

structural breaks