$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \varepsilon_i$$

Response function:

$$E[Y] = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2$$

This is a parabola and is frequently called a quadratic response function.

Parameters:

β_0 – the mean response when $x = 0$ (ie $X = \bar{X}$)

β_1 – the linear effect coefficient.

β_{11} – the quadratic effect coefficient.

Third order model with one predictor variable

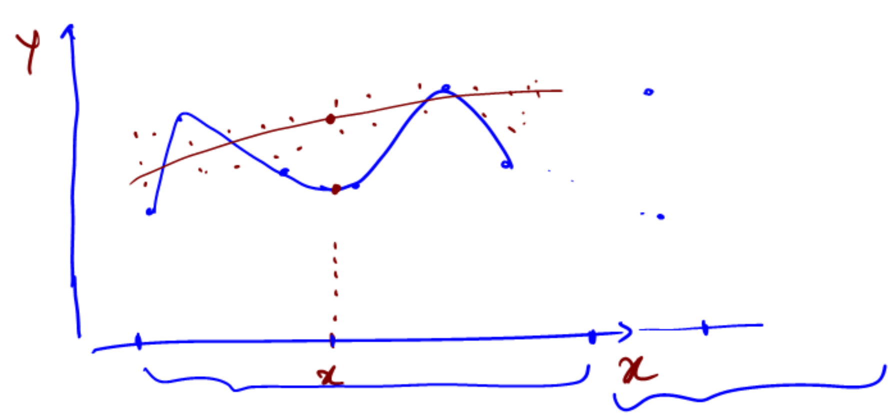
$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{111} x_i^3 + \varepsilon_i,$$

where $x_i = X_i - \bar{X}$

The response function is

$$E[Y] = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{111} x_i^3.$$

Higher orders with one predictor variable



There are some drawbacks of using higher order models.

* Interpretation of coefficient becom difficult.
* Wrong interpolations and extrapolations
* poor predictions.

The Second order model with two predictors

Model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \underline{\beta_{12} x_{i1}' x_{i2}'} + \varepsilon_i,$$

where

$$x_{i1} = X_{i1} - \bar{X}_1 \quad \text{and}$$

$$x_{i2} = X_{i2} - \bar{X}_2$$

The response function

$$E[Y] = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2}_{\text{linear component}} + \underbrace{\beta_{11} x_1^2 + \beta_{22} x_2^2}_{\text{Quadratic component}} + \underbrace{\beta_{12} x_1 x_2}_{\text{cross-product}},$$

β_{12} – Interaction effect coefficient.

The second order model with three predictors is Similar.

$$Y_i = \underbrace{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}_{\text{linear component}} + \underbrace{\beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{33} x_{i3}^2}_{\text{Quadratic component}}$$

$$+ \underbrace{\beta_{12} x_{i1} x_{i2} + \beta_{13} x_{i1} x_{i3} + \beta_{23} x_{i2} x_{i3}}_{\text{cross product.}} + \varepsilon_i$$

__Fitting of Polynomial models__

Since polynomial regression is a Special case of general linear regression model, fitting and making inferences are same to the previous cases.

Hierarchical Approach to fitting

Here we start a higher order (Second or third order) models, and then test whether higher order terms can be dropped.

$$\text{Eg:} \quad Y_i = \beta_0 + \beta_1 x_i + \underbrace{\beta_{11} x_i^2 + \beta_{111} x_i^3} + \varepsilon_i$$

Test : * $\beta_{111} = 0$ or not

* $\beta_{11} = 0$ and $\beta_{111} = 0$ or not (But we do not test "$\beta_{11} = 0$ or not" only).

For these test extra Sum of Squares can be used.

Here
$$SSR = SSR(x_1) + SSR(x^2/x) + SSR(x^3/x, x^2)$$

Test "$\beta_{111} = 0$ or not" $\longrightarrow$ $SSR(x^3/x, x^2)$ can be used.

Test "$\beta_{11} = 0$ and $\beta_{111} = 0$ or not" $\Rightarrow$ $SSR(x^2, x^3/x) = SSR(x^2/x) + SSR(x^3/x, x^2)$ can be used.

The model in terms of the original variables:

$$\hat{Y} = b_0 + b_1 x + b_{11} x^2 \quad \text{(Second order model)}$$

$$= b_0 + b_1(X - \bar{X}) + b_{11}(X - \bar{X})^2$$

$$= \underbrace{(b_0 - b_1\bar{X} - b_{11}\bar{X}^2)} + \underbrace{(b_1 - 2b_{11}\bar{X})}_{b_1'} X + \underbrace{b_{11}}_{b_{11}'} X^2$$

$$= b_0' + b_1' X + b_{11}' X^2$$

Note:

* Fitted values and the residuals in terms of x and X are the same.
* Centered observations (x) reduce multicollinearity and calculation difficulties.

Interaction Regression Models

A regression model with $p-1$ predictors contains additive effect if

$$E[Y] = f_1(X_1) + f_2(X_2) + \dots + f_{p-1}(X_{p-1}) \longrightarrow (*)$$

Eg: 1. $E[Y] = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_1^2}_{f(X_1)} + \underbrace{\beta_3 X_2}_{f(X_2)}$

$\Rightarrow$ effect of X_1 and X_2 are additive

2. $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$ (not in the form of $(*)$)

$\Rightarrow$ effect are not additive. cross-product term.

* Cross-product terms model the interaction effect of two predictor variables. This is also called as an interaction term or bi-linear interaction term.

* The meaning of β_1 and β_2 is not same as that given earlier.

* The change in mean response with a unit increase in X_1 when X_2 is held constant is

$$\frac{\partial[E[Y]]}{\partial X_1} = \beta_1 + \beta_3 X_2 \quad (\text{depends on } X_2)$$

Similarly

$$\frac{\partial[E[Y]]}{\partial X_2} = \beta_2 + \beta_3 X_1 \quad (\text{depends on } X_1)$$

Consider three Response functions:

a) $E[Y] = 10 + 2X_1 + 5X_2$ b) $E[Y] = 10 + 2X_1 + 5X_2 + 0.5X_1X_2$ c) $E[Y] = 10 + 2X_1 + 5X_2 - 0.5X_1X_2$

Calculate the mean response with unit increase of X_1 when $X_2 = 1$ or $X_2 = 3$.

* When $X_2 = 1$

a) $\dfrac{\partial[E[Y]]}{\partial X_1} = 2$

b) $\dfrac{\partial E[Y]}{\partial X_1} = 2 + 0.5(1) = 2.5$

c) $\dfrac{\partial(E[Y])}{\partial X_1} = 2 - 0.5(1) = 1.5$

* when $X_2 = 3$

a) $\dfrac{\partial[E[Y]]}{\partial X_1} = 2$

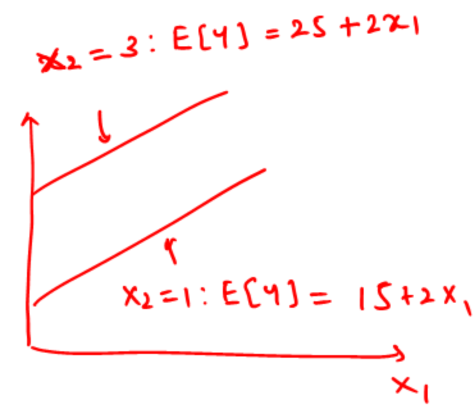
$\dfrac{\partial E(Y)}{\partial X_1} = 2 + 0.5(3) = 3.5$

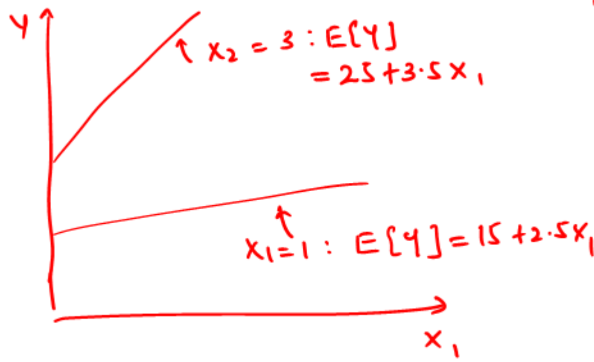
$\dfrac{\partial(E[Y])}{\partial X_1} = 2 - 0.5(3) = 0.5$

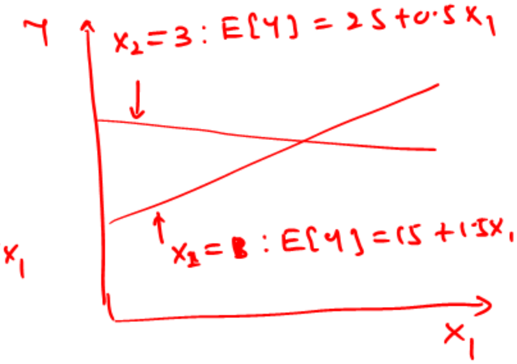
doesn't depend on the level of X_2

unit increase in X_1 has larger effect on mean response when X_2 is at a higher level.

unit increase in X_1 has smaller effect on mean response when X_2 is at a higher level.







Reinforcement type effects

Interference type effects.

Qualitative predictors

When there are qualitative predictors in a regression model, indicator variables are used to denote the classes of the qualitative variable:

Eg:- i) gender $\rightarrow$ male / female — 2 levels.

$$X = \begin{cases} 1 & : \text{male} \\ 0 & : \text{female} \end{cases}$$

ii) Disability Status : $\rightarrow$ not disable / partially disable / fully disable. } 3- levels.

$$X_1 = \begin{cases} 1 : & \text{if not disable} \\ 0 : & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 : & \text{if partially disable} \\ 0 : & \text{otherwise} \end{cases}$$

Status	X_1	X_2
not disable	1	0
partially	0	1
fully	0	0

* Qualitative variable with "c" classes is represented by "c-1" indicator variables.

Eg:- Y – Number of month elapsed (time for a particular insurance innovation)

Predictors:

* the size of the insurance firm (X_1)
* the type of the firm (Stock company, mutual company)

Define $X_2 = \begin{cases} 0 : & \text{if Stock company} \\ 1 : & \text{if mutual company} \end{cases}$

Regression model:

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

Response function:

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

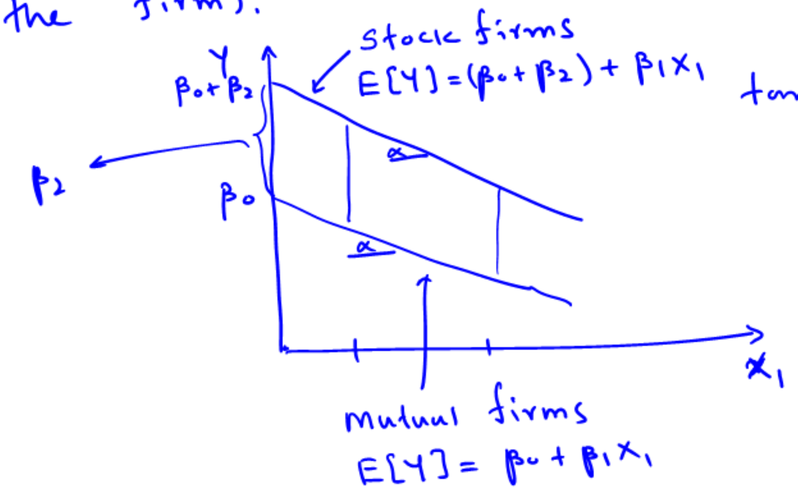
Meaning of the regression Coefficients

* When $X_2 = 0$ $\quad E[Y] = \beta_0 + \beta_1 X_1 \longrightarrow$ mutual firms
* when $X_2 = 1$ $\quad E[Y] = \beta_0 + \beta_1 X_1 + \beta_2(1) = \beta_0 + \beta_2 + \beta_1 X_1 \rightarrow$ Stock firms

Both response functions are straight lines with same slope.

* **Meaning of β_2**

β_2 indicates how much higher (or lower) the response function for 2^{nd} level (stock firms) than that for the first level (mutual firms) for any given value of the first predictor (size of the firm).



$\tan\alpha = $ Slope

β_2 : effect of type of firm (differential effect).

Qualitative Predictors with more than two classes

Eg:- consider the regression of tool wear (Y) on tool speed X_1 and tool model (qualitative : m_1, m_2, m_3, m_4).

Define three indicator variables:

$$X_2 = \begin{cases} 1 & : \text{ if model is } m_1 \\ 0 & : \text{ otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & : \text{ if model is } m_2 \\ 0 & : \text{ otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & : \text{ if model is } m_3 \\ 0 & : \text{ otherwise.} \end{cases}$$

model	X_1	X_2	X_3
m_1	1	0	0
m_2	0	1	0
m_3	0	0	1
m_4	0	0	0

Model

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

Response function:

$E[Y] = \beta_0 + \beta_1 X_1$: model m_4

$E[Y] = (\beta_0 + \beta_2) + \beta_1 X_1$: m_1

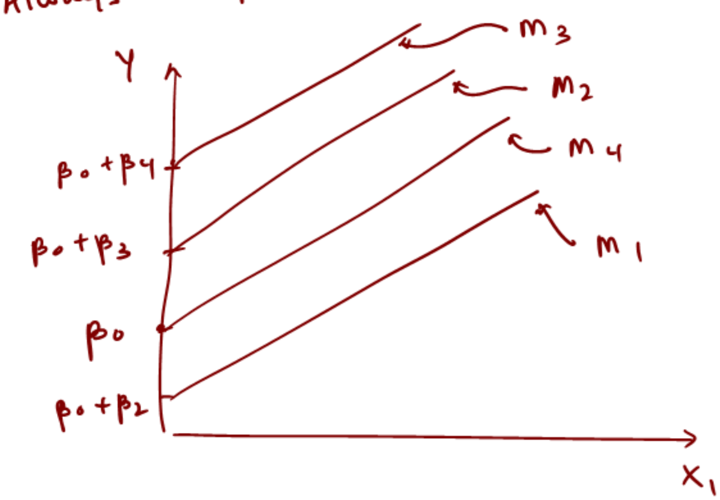
$E[Y] = (\beta_0 + \beta_3) + \beta_1 X_1$: m_2

$E[Y] = (\beta_0 + \beta_4) + \beta_1 X_1$: m_3

Interpretation:

$\beta_2, \beta_3, \beta_4$: how much higher (or lower) the response function for tool models m_1, m_2, m_3 than that for respectively for tool model m_4.

* Always compared with the class for which $X_2 = X_3 = X_4 = 0$,



To estimate differential effect other than against tool model m_4
(ie estimate differences between regression coefficients)

Eg: $\beta_4 - \beta_3$

: how much higher (or lower) the response function for m_3 than that for m_2.

$Var(X-Y) = Var(X) + Var(Y) - 2\,cov(X,Y)$

Point estimator = $b_4 - b_3$

$S^2\{b_4 - b_3\} = S^2\{b_4\} + S^2\{b_3\} - 2S\{b_4, b_3\}$.

These can be obtained from the variance covariance matrix $S^2\{b\}$.

Chapter – 9 : Building a Regression Model : Model Selection and Validation.

In this chapter we discuss the idea of model selection (ie how to choose the model which is good for data) and validation (ie how to choose the model which is appropriate for the application).

Model building can be thought of as a 4 step process.

1) Data collection and preperation
2) Reduction of explanatory (or predictor) variables
3) Model refinement and Selection
4) Model validation.

Data collection and Preperation

There are basically four types of research designs.

1. controlled experiments
2. controlled experiments with co-variates
3. Confirmatory observational studies
4. Exploratory observational studies.