CHAPTER 9: MODELS WITH QUALITATIVE EXPLANATORY VARIABLES

- In simple linear regression problem: $y_i = \alpha + \beta X_i + \epsilon_i$
  $y_i \sim N(\alpha + \beta X_i, \sigma^2)$
  $\epsilon_i \sim N(0, \sigma^2)$
  X_i is non-stochastic.

- A dummy variable can be thought as a binary variable that takes values 0 or 1 to indicate the presence/absence of some categorical effect which may be expected to shift the outcome. e.g., employment/marital status.

- <u>Dummy Variable Trap:</u> $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$, where

$$x_{i1} = \begin{cases} 1 & \text{if the highest degree of the candidate is BSc.} \\ 0 & \text{" " " " " " NOT BSc.} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if the highest degree of the candidate is MSc.} \\ 0 & \text{otherwise.} \end{cases}$$

$$x_{i3} = \begin{cases} 1 & \text{if the highest degree of the candidate is PhD.} \\ 0 & \text{otherwise.} \end{cases}$$

Due to multicollinearity; calculations of $\beta_0, \beta_1, \beta_2$, and β_3 would be indeterminate. Since $x_{i1} = 1 - x_{i2} - x_{i3}$ and the OLS-based normal equations are NOT independent / $X'X$ is a singular matrix. This is called "<u>Dummy Variable Trap</u>".

- <u>Dummy Variables to seperate Blocks of Data:</u> Suppose we wish to introduce into a model the idea that there are two types of machines (Type A and Type B) that produces different levels of response, in addition to the variation that occurs due to other regressors. One way to do this is to add a dummy variable Z ($Z = 0,1$). Consider the simple model with one regression variable X and one dummy variable Z.

$$y = \beta_0 + \beta_1 x + \alpha Z + \epsilon$$

$$Z = \begin{cases} 0 & \text{if the observation is from machine A.} \\ 1 & \text{if the observation is from machine B.} \end{cases}$$

Let $\hat{\beta}_0, \hat{\beta}_1, \hat{\alpha}$ be LSEs of β_0, β_1 and α, respectively. Then the fitted model is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\alpha} Z.$$

Machine A data are estimated by setting $Z = 0$: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Machine B data are estimated by setting $Z = 1$: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\alpha}$

Both are straight lines with different intercepts

Thus, $\hat{\alpha}$ simply estimates the difference in response level between machine A and machine B.

MODEL: $y = \beta_0 + \beta_1 x + \alpha z + \epsilon$.

MLR: $Y = X\beta + \epsilon$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_1+n_2} \end{pmatrix}, \quad X = \begin{array}{c} \begin{array}{ccc} x_0 & \text{obs } X\text{'s} & z \end{array} \\ \left[\begin{array}{ccc} 1 & & 0 \\ 1 & & 0 \\ 1 & & 0 \\ \vdots & & \vdots \\ 1 & & 0 \\ \hline 1 & & 1 \\ 1 & & 1 \\ \vdots & & \vdots \\ 1 & & 1 \end{array}\right] \end{array} \begin{array}{l} \text{Machine A} \\ (n_1 \text{ obsn}) \\ \text{Mac. B} \\ (n_2 \text{ obsn}) \end{array}$$

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

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \alpha \end{pmatrix}$$

Two Blocks require two dummy variables including x_0.

Three Blocks, three dummy variables :-

$$(z_1, z_2) = \begin{cases} (1,0) & \text{for Machine A} \\ (0,1) & \text{for Machine B} \\ (0,0) & \text{for Machine C} \end{cases}$$

$$X = \begin{array}{c} \begin{array}{cccc} x_0 & \text{other } X\text{'s} & z_1 & z_2 \end{array} \\ \left[\begin{array}{cccc} 1 & & 1 & 0 \\ 1 & & 1 & 0 \\ 1 & & 1 & 0 \\ \hline 1 & & 0 & 1 \\ 1 & & \vdots & \vdots \\ 1 & & 0 & 1 \\ \hline 1 & & 0 & 0 \\ 1 & & \vdots & \vdots \\ 1 & & 0 & 0 \end{array}\right] \end{array} \begin{array}{l} \text{M/c A} \\ \\ \text{M/c B} \\ \\ \text{M/c C} \end{array}$$

The model would be $Y = \beta_0 x_0 + \beta X + \alpha_1 z_1 + \alpha_2 z_2 + \epsilon$.

$$Y = X\beta + \epsilon \xrightarrow{\text{OLS}} \hat{\beta} = (X'X)^{-1}(X'Y)$$

Suppose the fitted equation is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}X + \hat{\alpha}_1 z_1 + \hat{\alpha}_2 z_2$$

Machine A data are estimated by setting $(z_1, z_2) = (1,0)$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}X + \hat{\alpha}_1$$

" B " " " " " $(z_1, z_2) = (0,1)$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}X + \hat{\alpha}_2$$

" C " " " " " $(z_1, z_2) = (0,0)$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}X$$

$\hat{\alpha}_1$ estimates the diff. in response level between A & C.

$\hat{\alpha}_2$ " " " " " " B & C.

$\hat{\alpha}_1 - \hat{\alpha}_2$ " " " " " " A & B.

If desired, t test can be performed to test the diff. in response level between A & C.

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \alpha_1 \\ \alpha_2 \end{pmatrix}$$

$H_0: \alpha_1 = 0$ ag. $H_1: \alpha_1 \neq 0$

↳ diff. in response model

Test statistic $\quad t = \dfrac{\hat{\alpha}_1}{\sqrt{(X'X)^{-1}_{33}\, MS_{Res}}}$

Critical region: $|t| > t_{\alpha/2,\ \text{Res df}}$.

$H_0: \alpha_2 = 0$ ag. $H_1: \alpha_2 \neq 0$

↳ diff. in response level between B and C.

Test statistic: $t = \dfrac{\hat{\alpha}_2}{\sqrt{(X'X)^{-1}_{44} MS_{Res}}}$

Critical region: $|t| > t_{\alpha/2, \text{Res d.f.}}$

$H_0: \alpha_1 - \alpha_2 = 0$ Vs. $H_1: \alpha_1 - \alpha_2 \neq 0$

↳ diff. in response level between A & B.

$$t = \frac{\hat{\alpha}_1 - \hat{\alpha}_2}{\sqrt{V(\hat{\alpha}_1 - \hat{\alpha}_2)}} \;;\; V(\hat{\alpha}_1 - \hat{\alpha}_2) = V(\hat{\alpha}_1) + V(\hat{\alpha}_2) - 2\,\mathrm{Cov}(\hat{\alpha}_1, \hat{\alpha}_2)$$

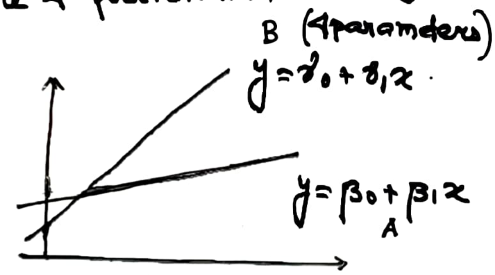
Critical region: $|t| > t_{\alpha/2, \text{Res df}}$. <u>Example</u>: See Pg. 6.

Interaction Terms Involving Dummy Variables

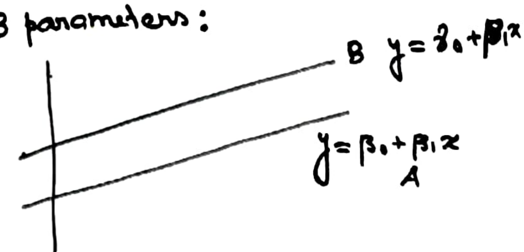
Two sets of data, straight line models.

Suppose A & B denote two sets of data and we are considering fits involving straight lines. There are 4 possibilities:

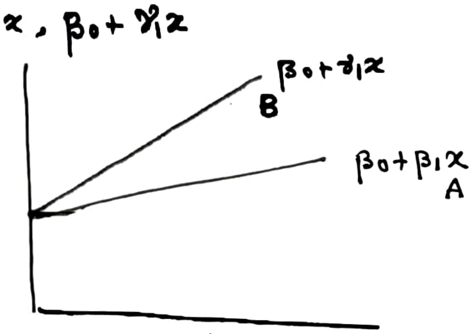
(a) Two distinct lines $\beta_0 + \beta_1 x$, $\gamma_0 + \gamma_1 x$



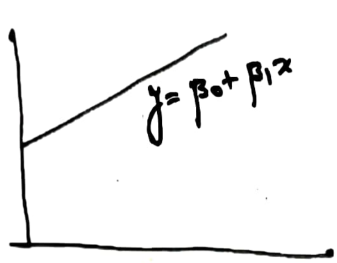
(b) Two parallel lines $\beta_0 + \beta_1 x$, $\gamma_0 + \beta_1 x$, 3 parameters:



(c) Two lines with the same intercepts $\beta_0 + \beta_1 x$, $\beta_0 + \gamma_1 x$
3 parameters



(d) One line $\beta_0 + \beta_1 x$



<u>NOTE</u>: For r blocks and r dummies. In general, we can also deal with r blocks by introducing $(r-1)$ dummies in addition to X_0.

We can take care of 4 possibilities at once by choosing two dummies, including X_0.

X_0	Z	
1	0	for A (Block A)
1	1	for B (Block B)

Then the model could be

$$Y = X_0(\beta_0 + \beta_1 x) + Z(\alpha_0 + \alpha_1 x) + \epsilon$$

$$= \beta_0 + \beta_1 x + \alpha_0 Z + \alpha_1 Z x + \epsilon \quad \text{———} \quad (*)$$

This model contains not only Z but an interaction term involving Z.
The seperate models for A & B are given by setting $Z=0$ & $Z=1$.

$$Y = \beta_0 + \beta_1 x \quad \text{for } A$$
$$= (\beta_0 + \alpha_0) + (\beta_1 + \alpha_1)x \quad \text{for } B$$
$$= \gamma_0 + \gamma_1 x$$

To test whether two parallel lines will do, i.e., to test the appropriateness of case (b) we would fit $(*)$ & then test.

$$H_0: \alpha_1 = 0 \quad Vs. \quad H_1: \alpha_1 \neq 0$$

To test the appropriateness of the case (c) we would fit $(*)$ & then test

$$H_0: \alpha_0 = 0 \quad Vs. \quad H_1: \alpha_0 \neq 0$$

To test the appropriateness of the case (d), we would test $H_0: \alpha_0 = \alpha_1 = 0$ Vs. H_1: H_0 is not true.

Three sets of data, straight line models:—

To allow the fitting of three seperate straight lines, we form the model: $Y = X_0(\beta_0 + \beta_1 X) + Z_1(\gamma_0 + \gamma_1 X) + Z_2(\delta_0 + \delta_1 X) + \epsilon$

$X_0 = 1$ & Z_1, Z_2 are two additional dummy variables.

	X_0	Z_1	Z_2
A →	1	1	0
B →	1	0	1
C →	1	0	0

$$Y = \beta_0 + \beta_1 X + \gamma_0 Z_1 + \gamma_1 X Z_1 + \delta_0 Z_2 + \delta_1 X Z_2 + \epsilon$$

Note that we have two interaction terms XZ_1 & XZ_2.

To test whether 3 lines are identical, we test

$$H_0: \gamma_0 = \gamma_1 = \delta_0 = \delta_1 = 0 \text{ Vs. } H_1: H_0 \text{ is not true.}$$

$$Y = (\beta_0 + \beta_1 X) + Z_1(\gamma_0 + \gamma_1 X) + Z_2(\delta_0 + \delta_1 X) + \epsilon.$$

$$F = \frac{\{SS_{Reg}(\text{Full model}) - SS_{Reg}(\text{Restricted Model})\}/4}{SS_{Res}/(n-6)} \sim F_{4,n-6}$$

Restricted Model $\to Y = \beta_0 + \beta_1 X$; $4 \to (6-2)$

Critical region: $F > F_{\alpha,4,n-6}$ (Reject Ho)

To test three lines are parallel,

Ho : $\gamma_1 = \delta_1 = 0$ Vs. H_1: Ho is not true.

$$F = \frac{\{SS_{Reg}(\text{Full model}) - SS_{Reg}(\text{Restricted model})\}/2}{SS_{Res}/(n-6)} \sim F_{2,n-6}$$

Restricted model $\to Y = \beta_0 + \beta_1 X + \gamma_0 Z_1 + \gamma_1 Z_2 + \epsilon$; $2 \to (6-4)$

If $F > F_{\alpha,2,n-6}$, then reject Ho.

<u>EXAMPLE</u>: See Pg. 7.

Two sets of data. Quadratic Model:-

Suppose we have two sets of data on Y and X and we have in mind to model of the form

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \epsilon$$

$\boxed{Z = X^2}$

We fit the model

$$Y = Z_0(\beta_0 + \beta_1 X + \beta_{11} X^2) + Z_1(\alpha_0 + \alpha_1 X + \alpha_{11} X^2) + \epsilon$$

Z_0	Z_1	
1	0	for A
1	1	for B

(1) Test: Ho; $\alpha_0 = \alpha_1 = \alpha_{11} = 0$ Vs H_1: Ho is not true.

If Ho is rejected then we conclude the models are not the same.

(2). If Ho is (1) is rejected, test Ho; $\alpha_1 = \alpha_{11} = 0$ Vs. H_1: Ho is not true.

If Ho is accepted, we conclude that the two sets of data have the same slope & curvature.

(3.) If Ho in (2) is rejected, then test Ho; $\alpha_{11} = 0$ Vs. H_1: $\alpha_{11} \neq 0$

Model differ only in zero & first order term.

Example (TURKEY Data): Ref: Applied Regression Analysis by Draper & Smith.

G: Georgia, V: Virginia, W: Wisconsin

Weights (Y) in pounds	Ages (X) in weeks	origin (Z)
13.3	28	G
8.9	20	G
15.1	32	G
10.4	22	G
13.1	29	V
12.4	27	V
13.2	28	V
11.8	26	V
11.5	21	W
14.2	27	W
15.4	29	W
13.1	23	W
13.8	25	W

We would like to relate Y to X via a simple straight line model, but the different origins of the turkeys may cause a problem. If they do, how do we handle it?

SOLUTIONS:

If we fit a simple regression model (Y against X) without considering origin:

$$\hat{Y} = 1.98 + 0.4167X.$$

Consider dummy variables Z_1 and Z_2 and fit the model

$$Y = \beta_0 + \beta_1 X + \alpha_1 Z_1 + \alpha_2 Z_2 + \epsilon.$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \alpha_1 \\ \alpha_2 \end{pmatrix}; \quad X = \begin{array}{c} \begin{array}{cccc} X_0 & X & Z_1 & Z_2 \end{array} \\ \begin{bmatrix} 1 & 28 & 1 & 0 \\ 1 & 20 & 1 & 0 \\ 1 & 32 & 1 & 0 \\ 1 & 22 & 1 & 0 \\ \hline 1 & 29 & 0 & 1 \\ 1 & 27 & 0 & 1 \\ 1 & 28 & 0 & 1 \\ 1 & 26 & 0 & 1 \\ \hline 1 & 21 & 0 & 0 \\ 1 & 27 & 0 & 0 \\ 1 & 29 & 0 & 0 \\ 1 & 23 & 0 & 0 \\ 1 & 25 & 0 & 0 \end{bmatrix} \end{array}; \quad Y = \begin{bmatrix} 13.3 \\ 8.9 \\ 15.1 \\ 10.4 \\ 13.1 \\ 12.4 \\ 13.2 \\ 11.8 \\ 11.5 \\ 14.2 \\ 15.4 \\ 13.1 \\ 13.8 \end{bmatrix}$$

$$Y = X\beta + \epsilon, \quad \hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} 1.43 \\ 0.48 \\ -1.92 \\ -2.19 \end{pmatrix}.$$

Fitted equation is

$$\hat{Y} = 1.43 + 0.48X - 1.92Z_1 - 2.19Z_2.$$

Model for	(Z_1, Z_2)	Fitted Model
G	(1,0)	$\hat{Y} = 1.43 + 0.48X - 1.92 = -0.49 + 0.4868X$
V	(0,1)	$\hat{Y} = 1.43 + 0.48X - 2.19 = -0.76 + 0.4868X$
W	(0,0)	$\hat{Y} = 1.43 + 0.48X.$

3 parallel fitted linear lines with different intercepts.

Three sets of data, Straight line models:

$$Y = X_0(\beta_0+\beta_1 X) + Z_1(\gamma_0+\gamma_1 X) + Z_2(\delta_0+\delta_1 X) + \epsilon$$
$$= \beta_0 + \beta_1 X + \gamma_0 Z_1 + \gamma_1 Z_1 X + \delta_0 Z_2 + \delta_1 Z_2 X + \epsilon$$

Coefficient Matrix:
$$\underset{\sim}{X} = \begin{array}{c} \begin{array}{cccccc} X_0 & X & Z_1 & Z_2 & Z_1X & Z_2X \end{array} \\ \begin{bmatrix} 1 & 28 & 1 & 0 & 28 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 25 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \gamma_0 \\ \gamma_1 \\ \delta_0 \\ \delta_1 \end{pmatrix}$$

$$Y = \underset{\sim}{X}\beta + \epsilon \xrightarrow{OLS} \hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{Y} = 2.475 + 0.445x - 3.454 Z_1 + 0.061(Z_1 X) - 2.775 Z_2 + 0.025(Z_2 X)$$

Three sets of straight lines are:

$\hat{Y} = -0.979 + 0.5060X$ Setting $Z_1 = 1, Z_2 = 0$

$\hat{Y} = -0.300 + 0.4700X$ Setting $Z_1 = 0, Z_2 = 1$

$\hat{Y} = 2.475 + 0.445X$ Setting $Z_1 = 0, Z_2 = 0$

<u>Note that</u> these are exactly what one would find if one fits each subset of data separately.