Chapter 14 Dummy Variables

Def. A dummy (or indicator) variable is an artificial variable used to represent a categorical predictor in a regression model.

Dichotomous categorical predictor variable

(1) Gender: Male (M) or Female (F)

(2) Treatment: Placebo or Aspirin

(3) Treatment: Traditional treatment or new treatment

Code dummy variable D to have 2 values corresponding to the two levels of the predictor variable.

— usually: baseline (placebo) $D=0$
treatment (aspirin) $D=1$.

— alternatively: $\left.\begin{array}{l}\text{baseline } D=-1 \\ \text{treatment } D=1.\end{array}\right\}$ or any other coding.

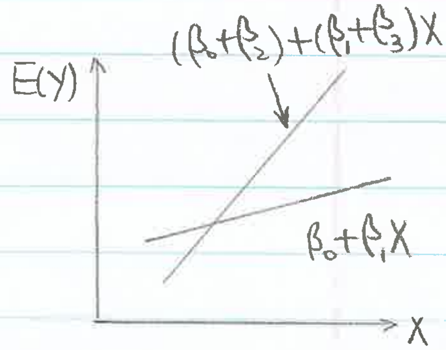
Note: Interpretation of results will depend upon how you code D in the regression model.

Incorporating D into the model

Response: Y

Predictors: X, D (dummy variable for categorical variable assuming = 0, 1)

Possible Models

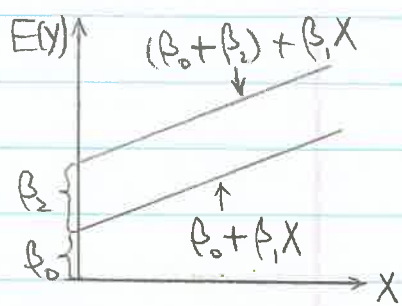


(1) $E(Y) = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 DX$

$$= \begin{cases} \beta_0 + \beta_1 X & \text{when } D = 0 \\ (\beta_0+\beta_2) + (\beta_1+\beta_3) X & \text{when } D = 1. \end{cases}$$

— different slops and intercepts.

Most general



(2) $E(Y) = \beta_0 + \beta_1 X + \beta_2 D$

$$= \begin{cases} \beta_0 + \beta_1 X & \text{when } D = 0 \\ (\beta_0+\beta_2) + \beta_1 X & \text{when } D = 1. \end{cases}$$

— Same slope, different intercepts in 2 categories.

Parallel lines

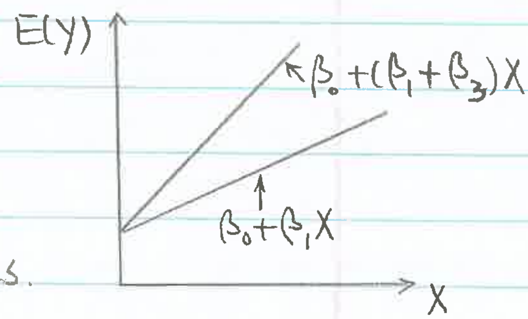
"The treatment effect is additive"

Difference in intercept is β_2.

(3) $E(Y) = \beta_0 + \beta_1 X + \beta_3 DX$

$$= \begin{cases} \beta_0 + \beta_1 X & \text{when } D=0 \\ \beta_0 + (\beta_1+\beta_3)X & \text{when } D=1. \end{cases}$$

— Same intercept, different slopes.



Difference in slope is β_3.

Concurrent

(4) $E(Y) = \beta_0 + \beta_1 X$

— response for 2 categories is the same

Coincident

Important hypothesis tests

(1) Are the two lines parallel?

Test $H_0: \beta_3 = 0$ vs. $H_a: \beta_3 \neq 0$

— Use t test.

(2) Do the two lines have the same intercept?

Test $H_0: \beta_2 = 0$ vs. $H_a: \beta_2 \neq 0$.

— use t test.

(3) Is response for 2 categories the same?
Test $H_0: \beta_2 = \beta_3 = 0$ vs. H_a: at least one of β_2, β_3 is not 0
— use F test.

Hierarchical Models

$(1) > (2)\ (\beta_3 = 0) > (4)\ (\beta_2 = 0).$

$(1) > (3)\ (\beta_2 = 0) > (4)\ (\beta_3 = 0).$

Note: Model (3) is not contained in (2), $(2) \not> (3)$, and vice versa.

Polynomial Models with Dummy Variable D

General form of second-order model:

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \alpha_0 D + \alpha_1 X D + \alpha_{11} X^2 D + \varepsilon$$

Note: (1) Don't include any higher-order terms of D since
$D^k = D$ for any k.
(2) Don't count D for order.

Important hypothesis tests

(1) $H_0: \alpha_0 = \alpha_1 = \alpha_{11} = 0$ vs. H_a: at least one of them not 0

(The models for two different levels are the same) — use F test.

(The models for two different levels are not the same)

(2) $H_0: \alpha_1 = \alpha_{11} = 0$ vs. H_a: at least one of α_1, α_{11} is not 0

(The treatment effect is additive) — use F test.

(The treatment effect is not additive)

Polytomous Categorical Predictors & Dummy Variables

Categorical variable takes on m distinct levels.

Treatment $\begin{cases} \text{Placebo} \\ \text{Aspirin} \\ \text{Tylenol} \end{cases}$ $m=3$.

In general, to represent the effects of a categorical predictor variable that takes on m possible levels, you need $m-1$ dummy variables: $D_1, D_2, \cdots, D_{m-1}$.

Usually each is coded by 0, 1.

	D_1	D_2	$\cdots$	D_{m-1}
Level 1	0	0	$\cdots$	0
Level 2	1	0	$\cdots$	0
Level 3	0	1	$\cdots$	0
$\vdots$				
Level m	0	0	$\cdots$	1

Ex. $m=3$

	D_1	D_2
Placebo	0	0
Aspirin	1	0
Tylenol	0	1

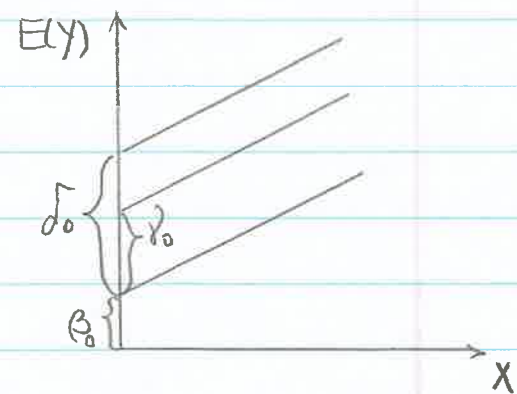
General model for $m=3$: $Y=\beta_0+\beta_1 X+\gamma_0 D_1+\gamma_1 D_1 X+\delta_0 D_2+\delta_1 D_2 X+\varepsilon$.

Important special cases

(i) $\gamma_1=\delta_1=0$.

$$E(Y)=\beta_0+\beta_1 X+\gamma_0 D_1+\delta_0 D_2$$

$$=\begin{cases}\beta_0+\beta_1 X & D_1=D_2=0\\ (\beta_0+\gamma_0)+\beta_1 X & D_1=1, D_2=0\\ (\beta_0+\delta_0)+\beta_1 X & D_1=0, D_2=1\end{cases}$$



— Three parallel lines.

β_0 gives baseline when $D_1=D_2=0$.
γ_0 — extra, additive effect of Aspirin.
δ_0 — extra, additive effect of Tylenol.

In general, choose a level as the baseline level to which all other levels will be compared.
For that level, $D_1=D_2=\cdots=D_{m-1}=0$.
Other levels are compared to that level by appropriate choice of codes.

(ii) $\gamma_0 = \gamma_1 = \delta_0 = \delta_1 = 0$

$E(Y) = \beta_0 + \beta_1 X$

— response for 3 categories is the same.

Coincident.

Important hypothesis tests

(1) $H_0: \gamma_1 = \delta_1 = 0$ vs. H_a: at least one of γ_1, δ_1 is not zero.

(Three lines are parallel or the treatment effect is additive) (Three lines are not parallel or the treatment effect is not additive)

— use F test.

(2) $H_0: \gamma_0 = \gamma_1 = \delta_0 = \delta_1 = 0$ vs. H_a: at least one of them not 0

(Three lines are identical or response for 3 categories is the same.) (Three lines are not identical or response for 3 categories is not the same.)

— use F test.

Note: (1) We can also include higher-order terms of X in the model.

— A polynomial model.

(2) Extension to the general case with p-1 predictors and m-1 dummy variables is similar.

Example 14.1: Bars of soap are scored for their appearance in a manufacturing operation. These scores are on a 1-10 scale, and the higher the score the better. The difference between operator performance and the speed of manufacturing line is believed to measurably affect the quality of the appearance. The following data were collected on this problem:

Operator	Line Speed	Appearance (Sum for 30 Bars)
A	150	255
A	175	246
A	200	249
B	150	260
B	175	223
B	200	231

(1) Using a dummy variable, fit a multiple regression model to these data and find the fitted line for each operator.

(2) Using $\alpha = 0.05$, determine whether operator differences are important in bar appearance.