1

Regression Models with categorical predictor variables

common examples:

- sex (male / female)
- SES (low / middle / high)
- intervention (treatment / control)

example: y = weight

(turkey data) age is a continuous (numerical) input.

(Georgia, Virginia, Wisconsin) state of origin is a categorical input (g, v, w)

data:

group 1	group 2	$\cdots$	group c
(X_{11}, Y_{11})	(X_{21}, Y_{21})		(X_{c1}, Y_{c1})
$\vdots$	$\vdots$	$\cdots$	$\vdots$
(X_{1n_1}, Y_{1n_1})	(X_{2n_2}, Y_{2n_2})		(X_{cn_c}, Y_{cn_c})

We will model the categorical input using indicator variables.

Let $$I_1 = 1\{g\} = \begin{cases} 1, & \text{if state = 'georgia'} \\ 0, & \text{otherwise} \end{cases}$$

$$I_2 = 1\{v\} = \begin{cases} 1, & \text{if state = 'virginia'} \\ 0, & \text{otherwise} \end{cases}$$

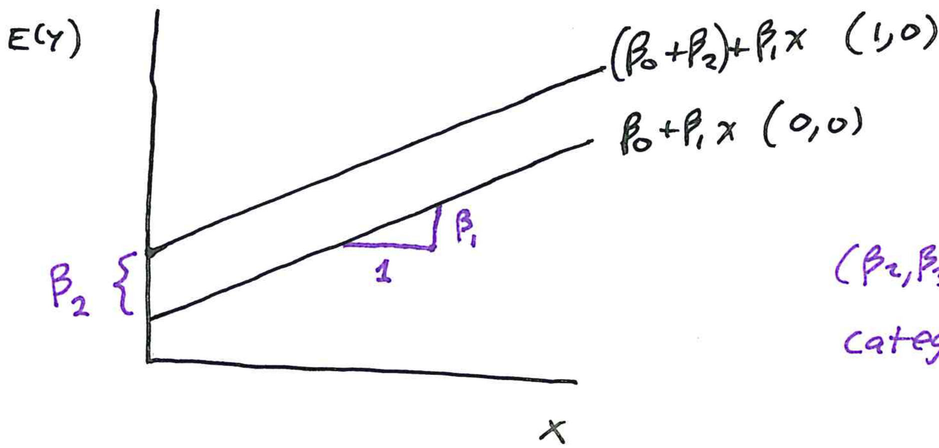
$(I_1, I_2) = (1,0)$ if g, $(0,1)$ if v, $(0,0)$ if w

(2)

<u>model :</u> $E(Y) = \beta_0 + \beta_1 X + \beta_2 I_1 + \beta_3 I_2$

$$\Longleftrightarrow \quad E(Y) = \begin{cases} (\beta_0 + \beta_2) + \beta_1 X, & \text{if } g \\ (\beta_0 + \beta_3) + \beta_1 X, & \text{if } v \\ \beta_0 + \beta_1 X, & \text{if } w \end{cases}$$

(Note that we use one less indicator variable than number of categories)



β_1 is the effect of continuous input X on response

(β_2, β_3) is the effect of the categorical input on response

<u>parameter interpretation :</u>

$$\beta_1 = \frac{\partial E(Y)}{\partial X_1}$$

(effect of the continuous input)

β_1 is the difference in mean weight from a 1 year increase in age, with origin held constant.

$$\beta_2 = E(Y \mid g, X) - E(Y \mid w, X)$$

(effect of categorical input is defined by multiple parameters)

β_2 is the difference in mean weight between georgia and wisconsin, with age held constant.

3

$\beta_3 = E(Y \mid v, X) - E(Y \mid w, X)$ (diff. in mean weight between v and w.)

$\beta_2 - \beta_3 = E(Y \mid g, X) - E(Y \mid v, X)$ (diff. in mean weight between g and v)

Test for categorical input effects: $(H_0: \beta_2 = \beta_3 = 0)$

(F): $E(Y) = \beta_0 + \beta_1 X + \beta_2 I_1 + \beta_3 I_2$ [mA = lm (Y ~ X + group) OR mA = lm(y ~ X + i1 + i2)]

(R): $E(Y) = \beta_0 + \beta_1 X$ [ML = lm (Y ~ X)]

anova(mL, mA), $F^* = 68.81$, $p = .000$

[We accept the model which includes origin as a categorical predictor for turkey weight. (test for partial effect)]

Estimating categorical predictor effects: confint (mA)

g - w : CI for $\beta_2 = [-2.375, -1.462]$, $t_2 = -9.506$

data supports the hypothesis that Georgia turkeys weigh less than Wisconsin turkeys

v - w : CI for $\beta_3 = [-2.67, -1.71]$, $t_3 = -10.367$

data supports the hypothesis that Virginia turkeys weigh less than Wisconsin turkeys

$g-v$: CI for $\beta_2 - \beta_3 = [-0.22, 0.77]$, $t_{23} = 1.25$
($p_{23} = .242$)

The data is compatible with a model that combines g, v into a single group.

Some quick distribution theory : $\underline{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix}$, $\underline{a}' = [0\ 0\ 1\ -1]$

$\underline{b} \sim N\left(\underline{\beta}, \sigma^2 (X'X)^{-1}\right)$, $E(\underline{a}'\underline{b}) = \underline{a}'\underline{\beta}$

$Var(\underline{a}'\underline{b}) = \sigma^2 \cdot \underline{a}'(X'X)^{-1}\underline{a}$

So,

$SE(\underline{a}'\underline{\beta}) = \sqrt{\underline{a}' \widehat{Cov}(\underline{b}) \underline{a}}$, CI for $\underline{a}'\underline{\beta} = \underline{a}'\underline{b} \pm t_{\frac{\alpha}{2}}^{(n-p)} SE(\underline{a}'\underline{b})$

Estimating continuous predictor effects

CI for $\beta_1 = [0.43, 0.545]$, $t = 18.908$

We accept the model which includes age as a continuous predictor for turkey weight. (test for partial effect)

5

Interaction Model :

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 I_1 + \beta_3 I_2 + \beta_4 X \cdot I_1 + \beta_5 X \cdot I_2$$

table:

group	intercept	slope
1: (1,0)	$\beta_0 + \beta_2$	$\beta_1 + \beta_4$
2: (0,1)	$\beta_0 + \beta_3$	$\beta_1 + \beta_5$
3: (0,0)	β_0	β_1

[The interaction model implies that the regression functions will have unequal slopes.]

(effect of continuous input depends on the level of the categorical input)

test for interaction effect : $(H_0 : \beta_4 = \beta_5 = 0)$.

(F): interaction model mI = lm(y~x+i1+i2+I(x*i1)+I(x*i2))

(R): additive model mA = lm(y~x+i1+i2)

turkey data: $F^* = 0.5204$, $p = .6156$ anova(mA, mI)

[~~The additive model~~
The observed data is compatible with the additive model ~~observed data~~.]

see scatterplot