Factor + numeric covariate

one factor with J levels

one numeric covariate, x_i

model: $Y_i = b_0 + b_{j(i)} + c_1 x_i + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$$E(Y_i) = b_0 + b_{j(i)} + c_1 x_i$$

interpretation: c_1 = effect (slope) of x_i

$b_0 + b_{j(i)}$ = intercept = $E(Y_i)$ when $x_i = 0$

Example: y_i = 3pm temperature on one day in 1 of 3 cities

x_i = 7am temperature in same city on same day

$j(i)$ = 1, 2, or 3 (city)

$b_0 + b_1$ = expected 3pm temp when 7am temp is 0 in city 1

$b_0 + b_2$ = expected 3pm temp when 7am temp is 0 in city 2

$b_0 + b_3$ = expected 3pm temp when 7am temp is 0 in city 3

increasing 7am temp by 1 degree means that we expect 3pm temp to change by c_1 degrees

Note that the slope c_1 does not depend on city.

<u>Additive Model</u>: <u>effect</u> of one variable does not depend on the <u>value</u> of the other

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 & x_1 \\ 1 & 1 & 0 & 0 & x_2 \\ 1 & 0 & 1 & 0 & x_3 \\ 1 & 0 & 1 & 0 & x_4 \\ 1 & 0 & 0 & 1 & x_5 \\ 1 & 0 & 0 & 1 & x_6 \end{bmatrix} \qquad \text{rank}(X) = 4 \ \text{(probably)}$$

Drop intercept column (set $b_0 = 0$)

$b_0 = 0$ (no interpretation)

b_1 = intercept for city 1

b_2 = intercept for city 2

b_3 = intercept for city 3

Drop level 1 column (set $b_1 = 0$, R default)

b_0 = level 1 intercept $(b_0 + b_1)$

$b_1 = 0$

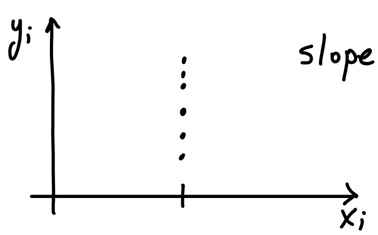
b_2 = level 2 intercept − level 1 intercept $(b_2 - b_1)$

b_3 = level 3 intercept − level 1 intercept $(b_3 - b_1)$

Digression on multicollinearity

In the multiple linear model, there is no assumption about the relationships between x_{i0}, x_{i1}, x_{i2}, etc. However, correlation in the design matrix can affect our ability to estimate things

simple linear model: $X = \begin{bmatrix} 1 & 9.2 \\ 1 & 9.2 \\ 1 & 9.2 \\ 1 & 9.2 \\ 1 & 9.2 \\ 1 & 9.2 \end{bmatrix}$



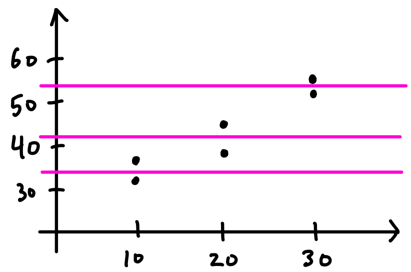
slope = ?

in that example $\text{rank}(X) = 1$, $(X^TX)^{-1}$ does not exist, $Var(\hat{\beta}_1) = \infty$

we need variation in the covariate to estimate a slope.

city temperature example:

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 & 10 \\ 1 & 1 & 0 & 0 & 10 \\ 1 & 0 & 1 & 0 & 20 \\ 1 & 0 & 1 & 0 & 20 \\ 1 & 0 & 0 & 1 & 30 \\ 1 & 0 & 0 & 1 & 30 \end{bmatrix} \qquad y = \begin{bmatrix} 32 \\ 37 \\ 45 \\ 39 \\ 52 \\ 55 \end{bmatrix}$$

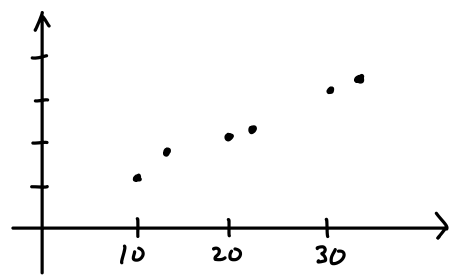


$\text{rank}(X) = 3$

Can't tell how much of the changes in y are due to the intercepts versus the slope

Need variation in 7am temp that's not collinear with city.

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 & 10 \\ 1 & 1 & 0 & 0 & 12 \\ 1 & 0 & 1 & 0 & 20 \\ 1 & 0 & 1 & 0 & 22 \\ 1 & 0 & 0 & 1 & 30 \\ 1 & 0 & 0 & 1 & 32 \end{bmatrix} \qquad y = \begin{bmatrix} 32 \\ 39 \\ 41 \\ 42 \\ 52 \\ 55 \end{bmatrix}$$



$\text{rank}(X) = 4$

A small positive slope will explain the data better than $c_1 = 0$

In fact, $\hat{c}_1 = \frac{1}{3}\left(\frac{y_2 - y_1}{2} + \frac{y_4 - y_3}{2} + \frac{y_6 - y_5}{2} \right)$

Adding another factor

factor 1 has J levels : (repub., dem, other)

factor 2 has K levels : (female, male)

y_i numeric response : (income)

model: $Y_i = b_0 + b_{j(i)} + c_{k(i)} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Example: 6 subjects.
$$X\underset{\sim}{b} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ c_1 \\ c_2 \end{bmatrix}$$

k

j	1 (F)	2 (M)
1 (R)	$E(Y_1)$	$E(Y_2)$
2 (D)	$E(Y_3)$	$E(Y_4)$
3 (I)	$E(Y_5)$	$E(Y_6)$

k

j	1	2
1	$b_0+b_1+c_1$	$b_0+b_1+c_2$
2	$b_0+b_2+c_1$	$b_0+b_2+c_2$
3	$b_0+b_3+c_1$	$b_0+b_3+c_2$

can interpret quantities that are expected values of linear combinations of observations.

$$E(Y_3) - E(Y_1) = (b_0+b_2+c_1) - (b_0+b_1+c_1) = b_2 - b_1$$
$$E(Y_4) - E(Y_2) = (b_0+b_2+c_2) - (b_0+b_1+c_2) = b_2 - b_1$$

$$E(Y_6) - E(Y_5) = (b_0+b_3+c_2) - (b_0+b_3+c_1) = c_2 - c_1$$
$$E(Y_2) - E(Y_1) = (b_0+b_1+c_2) - (b_0+b_1+c_1) = c_2 - c_1$$

Note that the effect of changing from $j=1$ to $j=2$ is $b_2 - b_1$, regardless of the level of factor 2.

Same for changing from $k=1$ to $k=2$ $(c_2 - c_1)$

<u>Effect</u> of changing one of the factor levels does not depend on the level of the other factor.

Another example of an additive model

rank$(X) = 4$, yet we have 6 columns and 6 parameters

need to drop 2 columns (e.g. set 2 parameters to 0)

R default: $b_1 = 0$, $c_1 = 0$

j \ k	1	2
1	b_0	$b_0 + c_2$
2	$b_0 + b_2$	$b_0 + b_2 + c_2$
3	$b_0 + b_3$	$b_0 + b_3 + c_2$

b_0 = EV for $j=1$, $k=1$

b_2 = EV for $(2,1) - (1,1)$ (or $(2,2) - (1,2)$)

b_3 = EV for $(3,1) - (1,1)$ (or $(3,2) - (1,2)$)

c_2 = EV for $(1,2) - (1,1)$ (or $(2,2) - (2,1)$)

(or $(3,2) - (3,1)$)

		k	
		1	2
j	1	*	*
	2	*	
	3	*	

Model dof = 4 < 6 = # of cells

consequence: can complete the table with incomplete information.

	1	2	3	4	5	6
1	*	*	*	*	*	*
2	*					
3	*					
4	*					

In general, model has

$J + K - 1$

model d.o.f.