EPRS8550
Multiple Regression

1. Regression Equation
2. Regression Analysis
3. Testing the Regression Coefficients
4. Comparing Regression Models
5. Multicollinearity
6. Standardized Regression Coefficient

1. Regression Equation

Simple regression

$$\hat{Y} = b_0 + b_1 X$$

Multiple regression

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 \quad \text{(Two quant. X's)}$$

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k \quad \text{(k quant. X's)}$$

Raw Data

Y (Achieve)	X1(Hours)	X2(Motivation)	$\hat{Y}$
16	2	15	16.67
14	3	12	16.29
20	4	17	19.18
19	4	14	17.95
23	5	18	20.44
20	6	20	22.10
22	6	16	20.46
25	8	24	25.43
26	8	23	25.02
24	9	22	25.46
mean 20.90	5.50	18.10	
var 14.99	5.39	16.30	
SD 3.872	2.321	4.04	

Correlation Matrix

	Y	X1	X2
Y	1	.8840	.8745
X1		1	.8826
X2			1

(a) Slopes

Recall $b = r\frac{s_y}{s_x}$ (for a simple reg.)

Now,

$$b_1 = \frac{r_{y1} - r_{y2}\cdot r_{12}}{1 - r_{12}^2}\cdot\frac{s_y}{s_{x_1}} = \frac{.8840 - (.8745)(.8826)}{1 - (.8826)^2}\cdot\frac{3.872}{2.321} = .846$$

$$b_2 = \frac{r_{y2} - r_{y1}\cdot r_{12}}{1 - r_{12}^2}\cdot\frac{s_y}{s_{x_2}} = \frac{.8745 - (.8840)(.8826)}{1 - (.8826)^2}\cdot\frac{3.872}{4.040} = .409$$

(b) Intercept

Recall $b_0 = \bar{Y} - b_1\bar{X}$

Here, $b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 = 20.9 - (.846)(5.5) - (.409)(18.1) \approx 8.85$

If you use just x1 or just X2 (i.e., simple reg)

$\hat{Y} = 12.8 + 1.47X_1$ or

$\hat{Y} = 5.79 + .84X_2$

Now, we have

$$\hat{Y} = 8.85 + .846X_1 + .409X_2$$

Q: How would you interpret .846?
Q: How about 8.85?

(Note) we assumed no interaction.

$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2$ is the model with an interaction.

2. Regression Analysis

Once you get $\hat{Y}$, then the rest is the same as the simple regression.

$$SS_{tot} = SS_{reg} + SS_{res}$$
$$\sum (Y - \bar{Y})^2 = \sum (Y' - \bar{Y})^2 + \sum (Y - Y')^2$$
$$134.9 = 110.832 + 24.068$$

(Note) SS_{res} with X1 alone is 29.49, X2 alone is 31.74.

ANOVA table

source	SS	df	MS	F
Reg	110.832	k=2	55.42	16.12
Res	24.068	N-k-1=7	3.43	
Tot	134.900	N-1=9		

$F^*_{\alpha,k,N-k-1} = F^*_{.05,2,7} = 4.74$

$H_0: \beta_1 = \beta_2 = 0$ (or equivalent to $H_0: R^2 = 0$)

$$R^2 = \frac{110.832}{134.9} = .822$$ (R^2=.78 with X1 alone, .76 with X2 alone.)

Recall

$$F = \frac{R^2/k}{(1-R^2)/(N-k-1)} = \frac{.822/2}{(1-.822)/7} = 16.12$$

Conclude that together X1 and X2 explain a statistically sig. amt of variation in Y.

3. Testing the Regression Coefficients

Recall for the simple regression:

$H_0: \beta = 0$

$t = \frac{b}{s_b}$ where $s_b = \frac{\sqrt{MSres}}{s_x \cdot \sqrt{N-1}}$

Here,

$H_0: \beta_1 = 0$

$t = \frac{b_1}{s_{b_1}} = \frac{.846}{.5669} = 1.49$ where

$$s_{b_1} = \frac{\sqrt{MSres}}{s_{x1} \cdot \sqrt{N-1} \cdot \sqrt{1-r_{12}^2}} = \frac{1.852}{2.32 \cdot \sqrt{10-1} \cdot \sqrt{1-.7797}} = .5669$$

$H_0: \beta_2 = 0$

$t = \frac{b_2}{s_{b_2}} = \frac{.409}{.325} = 1.234$ where

$$s_{b_2} = \frac{\sqrt{MSres}}{s_{x2} \cdot \sqrt{N-1} \cdot \sqrt{1-r_{12}^2}} = \frac{1.852}{4.04 \cdot \sqrt{10-1} \cdot \sqrt{1-.7797}} = .325$$

$t_{crit} = t_{\alpha/2, N-k-1} = t_{.05,7} = 2.365$

Why non-significant?

Confidence interval for slope

$b_1 \pm t_{\alpha/2, N-k-1} \cdot (s_{b_1})$ $\qquad$ $b_2 \pm t_{\alpha/2, N-k-1} \cdot (s_{b_2})$

$.846 \pm 2.365 \cdot (.5669)$ $\qquad$ $.409 \pm 2.365 \cdot (.32556)$

$.846 \pm 1.34$ $\qquad$ $.409 \pm .7669$

$(-.483, 2.197)$ $\qquad$ $(-.3689, 1.1709)$

4. Comparing Regression Models

$H_0: \beta_1 = \beta_2 = 0$ or $H_0: R^2 = 0$ an omnibus test

Suppose the null hypothesis is rejected.

Possible

1. One of the variables explains most of variation and the second variable explains very little.
2. Each variable explains a significant amount of variation.

$$F = \frac{(R^2_{FM} - R^2_{RM})/(k_{FM} - k_{RM})}{(1 - R^2_{FM})/(N - k_{FM} - 1)}$$

Suppose you would like to test if you should have X2 in the model when the model already has X1.

$H_0: \beta_2 = 0$

FM: $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$

RM: $\hat{Y} = b_0 + b_1 X_1$

$$F = \frac{(.822 - .78)/(2-1)}{(1-.822)/10-2-1} = 1.578 \quad \text{NS}$$

$$F^*_{\alpha, k_{FM} - k_{RM}, N - k_{FM} - 1} = F^*_{.05,1,7} = 5.59$$

Conclusion: X2 does not appear to explain a

significant amount of information in test scores beyond what is already explained by x1.

Suppose, this time, you would like to test if you should have X1 in the model when the model already has X2.

$H_0: \beta_1 = 0$

FM: $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$

RM: $\hat{Y} = b_0 + b_2 X_2$

$$F = \frac{(.822 - .76)/(2-1)}{(1-.822)/10-2-1} = 2.232 \quad \text{NS}$$

$$F^*_{\alpha, k_{FM}-k_{RM}, N-k_{FM}-1} = F^*_{.05,1,7} = 5.59$$

Conclusion: X1 does not appear to explain a significant amount of information in test scores beyond what is already explained by x2.

3 and 4 above produce the same results ($t^2 = F$).

4 is more flexible! Why?

Example:

FM: $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5$

RM: $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$

Note on Multiple Correlation

$$\sqrt{R^2} = R$$

$$R^2_{y.12} = \frac{r^2_{y1} + r^2_{y2} - 2 \cdot r_{y1} \cdot r_{y2} \cdot r_{12}}{1 - r^2_{12}}$$

$$R_{y.12} = \sqrt{\frac{r^2_{y1} + r^2_{y2} - 2 \cdot r_{y1} \cdot r_{y2} \cdot r_{12}}{1 - r^2_{12}}}$$

5. Multicollinearity

Problems

1. Limits the size of R.

Example: Multiple R = .46

	X1 (Reading)	X2 (Writing)	Y (Grade)
X1	1	.58	.33
X2		1	**.45**
Y			1

Reading add only .01 to the prediction of German Grade above and beyond writing.

2. Makes determining the importance of a given predictor difficult because the effects of the predictors are confounded due to the correlation among them.

3. Increases the variances of the regression coefficients. The greater the variances, the more unstable the prediction equation will be.

Diagnosis

1. Examine correlation matrix. (Not always indicates multicollinearity, however)
2. Examine b and Sb.
    - Large changes in b when a variable is added or deleted
    - Non-significant b for important IVs
    - Unexpected sign of b
    - Wide confidence interval
3. VIF (Variance Inflation Factors)

    $VIF = 1/(1 - R_j^2)$

    R_j^2 is the squared multiple correlation for predicting jth predictor from all other predictors. (X_j is regressed on the k - 1 other X variables.)

    If VIF > 10, then reason for concern, Myers (1990)

4. Tolerance = $(1 - R_j^2)$ Tolerance < .10 --- problem

What to do

1. Combine predictors that are highly correlated (add them to form a single measure).
2. Conduct a principal components analysis to reduce the number of predictors.
3. Use “ridge regression”. See Myers (1990).

6. b vs. β (beta = Standardized Regression Coefficient)

Raw scores of Y and X -- b
Z scores of Y and X -- β

Recall $Y' = 8.85 + .846\ X_1 + .409\ X_2$

Now $Z_{Y'} = .507Z_{X1} + .427Z_{X2}$

$$\beta_1 = b_1 \frac{s_{X_1}}{s_Y} = .846 \cdot \frac{2.321}{3.872} = .507 \qquad \beta_2 = b_2 \frac{s_{X_2}}{s_Y} = .409 \cdot \frac{4.040}{3.872} = .427$$

Recall

$$b_1 = \frac{r_{y1} - r_{y2} \cdot r_{12}}{1 - r_{12}^2} \cdot \frac{s_Y}{s_{X_1}} \qquad b_2 = \frac{r_{y2} - r_{y1} \cdot r_{12}}{1 - r_{12}^2} \cdot \frac{s_Y}{s_{X_2}}$$

Therefore

$$\beta_1 = \frac{r_{y1} - r_{y2} \cdot r_{12}}{1 - r_{12}^2} = \frac{.8840 - .8745 \cdot .8826}{1 - .8826^2} = .507$$

$$\beta_2 = \frac{r_{y2} - r_{y1} \cdot r_{12}}{1 - r_{12}^2} = \ldots = .427$$

What happens when $r_{12} = 0$?

$$\beta_1 = r_{y1} \qquad \beta_2 = r_{y2}$$

Should we use b or beta?

	b	beta
Advantages	Easy to interpret	Compare across multiple betas
Disadvantages	Cannot directly compare b's	Affected by s_x*

* The variation of X directly affects the size of beta

Reservations regarding the use of beta were expressed in various textbooks

"... should rarely if ever used" (Darlington, 1990)
"...seldom find beta to be useful" (Judd & McClell, 1989)