Indicator variable adjustment for missing X-data in regression models

stats methodologists meeting

April 2016

Method

Intercept	X1 (complete)	X2
1	7.4	.
1	4.0	.
1	1.2	.
1	0.6	.
1	10.0	5.7
1	1.0	6.8
1	9.0	2.9
1	0.2	4.8
1	1.8	9.4
1	0.6	5.1
1	5.0	6.8

Intercept	X1 (complete)	X2 (replace mvs)	I2 (indicator)
1	7.4	0	1
1	4.0	0	1
1	1.2	0	1
1	0.6	0	1
1	10.0	5.7	0
1	1.0	6.8	0
1	9.0	2.9	0
1	0.2	4.8	0
1	1.8	9.4	0
1	0.6	5.1	0
1	5.0	6.8	0

Method (extension)

Intercept	X1 (complete)	X2 (replace mvs)	I2 (indicator)	I2.X1 interaction
1	7.4	0	1	7.4
1	4.0	0	1	4.0
1	1.2	0	1	1.2
1	0.6	0	1	0.6
1	10.0	5.7	0	0
1	1.0	6.8	0	0
1	9.0	2.9	0	0
1	0.2	4.8	0	0
1	1.8	9.4	0	0
1	0.6	5.1	0	0
1	5.0	6.8	0	0

regression of Y on X1 (ignoring X2)

regression of Y on X1 and X2

Method (standard)

Intercept	X1 (complete)	X2 (replace mvs)	I2 (indicator)
1	7.4	0	1
1	4.0	0	1
1	1.2	0	1
1	0.6	0	1
1	10.0	5.7	0
1	1.0	6.8	0
1	9.0	2.9	0
1	0.2	4.8	0
1	1.8	9.4	0
1	0.6	5.1	0
1	5.0	6.8	0

coefficient for X1 is a mixture of regression of Y on X1 ignoring X2, and Y on X1 eliminating X2

RSS (error estimate) comes from a mix of ‘fully adjusted’ and ‘partially adjusted’ observations

indicator method(s) go way back

- Cohen, J., & Cohen, P. (1975) *Applied multiple regression/correlation analysis for the behavioral sciences*. New York: John Wiley (~p274)

back in 2002…

- genetic association study between SNPs (at the Cathepsin K locus) and bone mineral density (BMD) values in a cohort of 3000 perimenopausal Scottish women

- also had data on other factors affecting BMD: time post-menopausal, HRT, BMI…

- competing methods:
  - approaches based on imputation
  - complete case
  - adjustment by indicator variables

is the indicator-variable method ‘ok’?

- Jones M P (1996) Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression. JASA 91 no. 443, pp222-230

- “…missing-indicator methods show unacceptably large biases in practical situations and are not advisable in general.”

- why ‘biased’?

Method (standard)

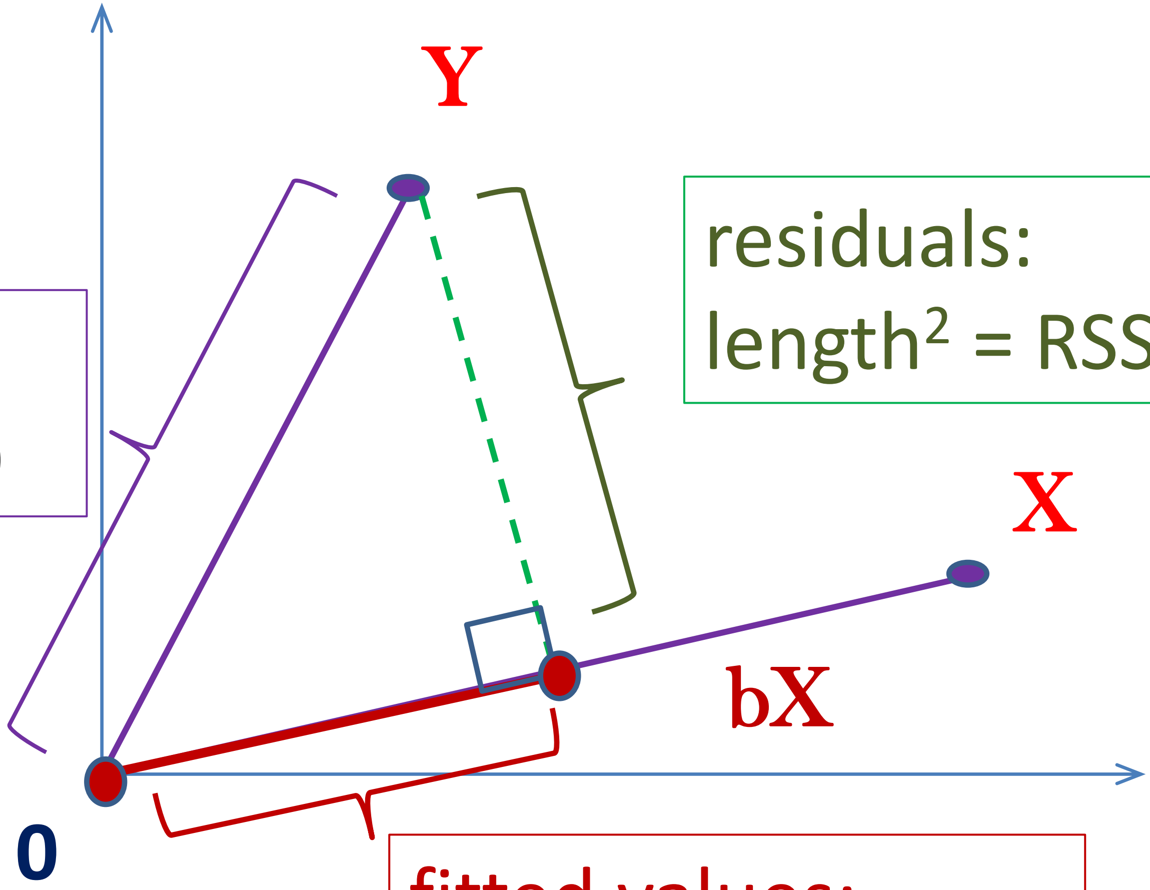
Intercept	X1 (complete)	X2 (replace mvs)	I2 (indicator)
1	7.4	0	1
1	4.0	0	1
1	1.2	0	1
1	0.6	0	1
1	10.0	5.7	0
1	1.0	6.8	0
1	9.0	2.9	0
1	0.2	4.8	0
1	1.8	9.4	0
1	0.6	5.1	0
1	5.0	6.8	0

coefficient for X1 is a mixture of regression of Y on X1 *ignoring* X2, and Y on X1 *eliminating* X2

least-squares fitting

$$|Y|^2 = \sum_{i=1}^{n} Y_i^2$$

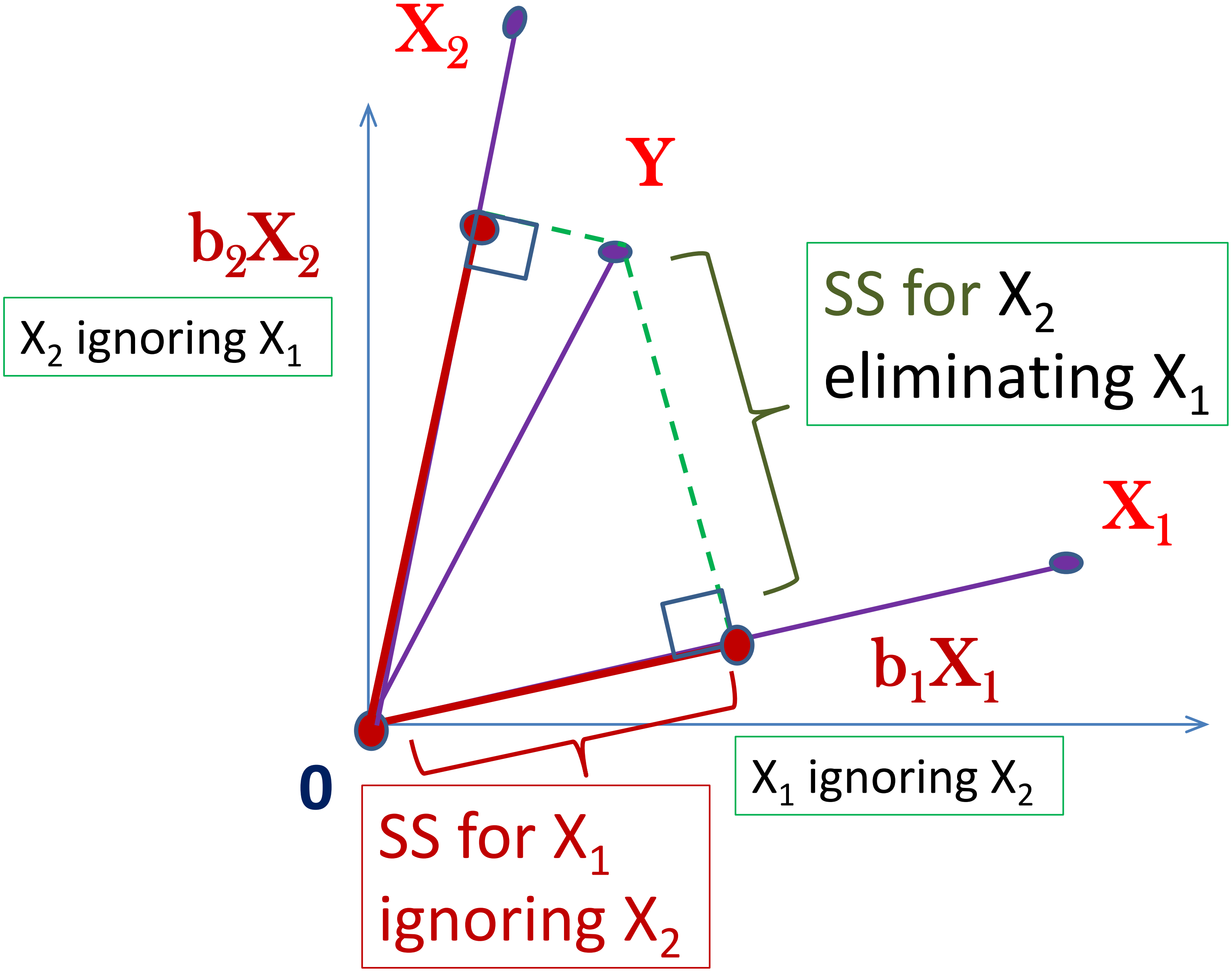
length² = total SS (for Y)



residuals: length² = RSS

fitted values: length² = fitted SS

least-squares fitting: X_1 and X_2



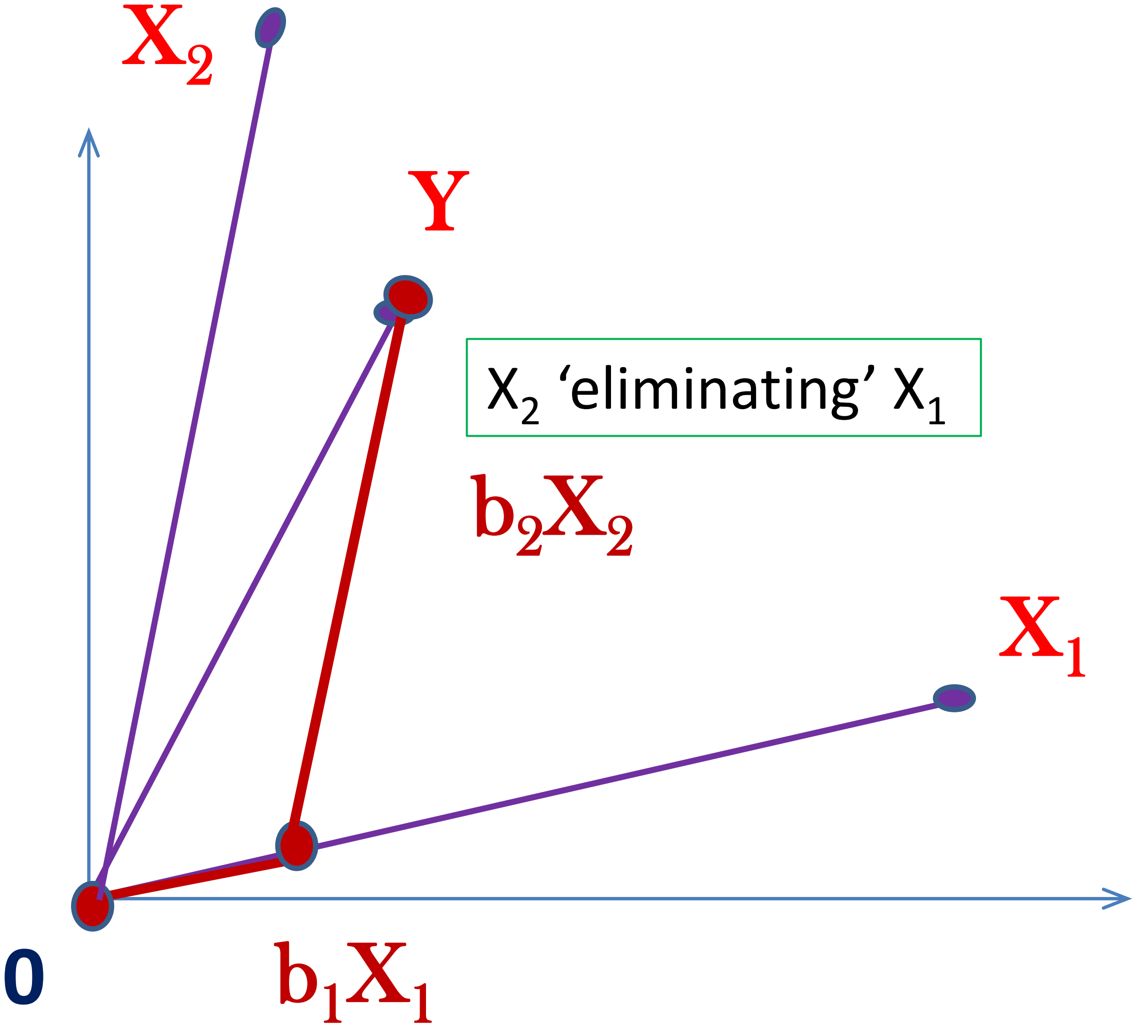
X_2 ignoring X_1

SS for X_2 eliminating X_1

X_1 ignoring X_2

SS for X_1 ignoring X_2

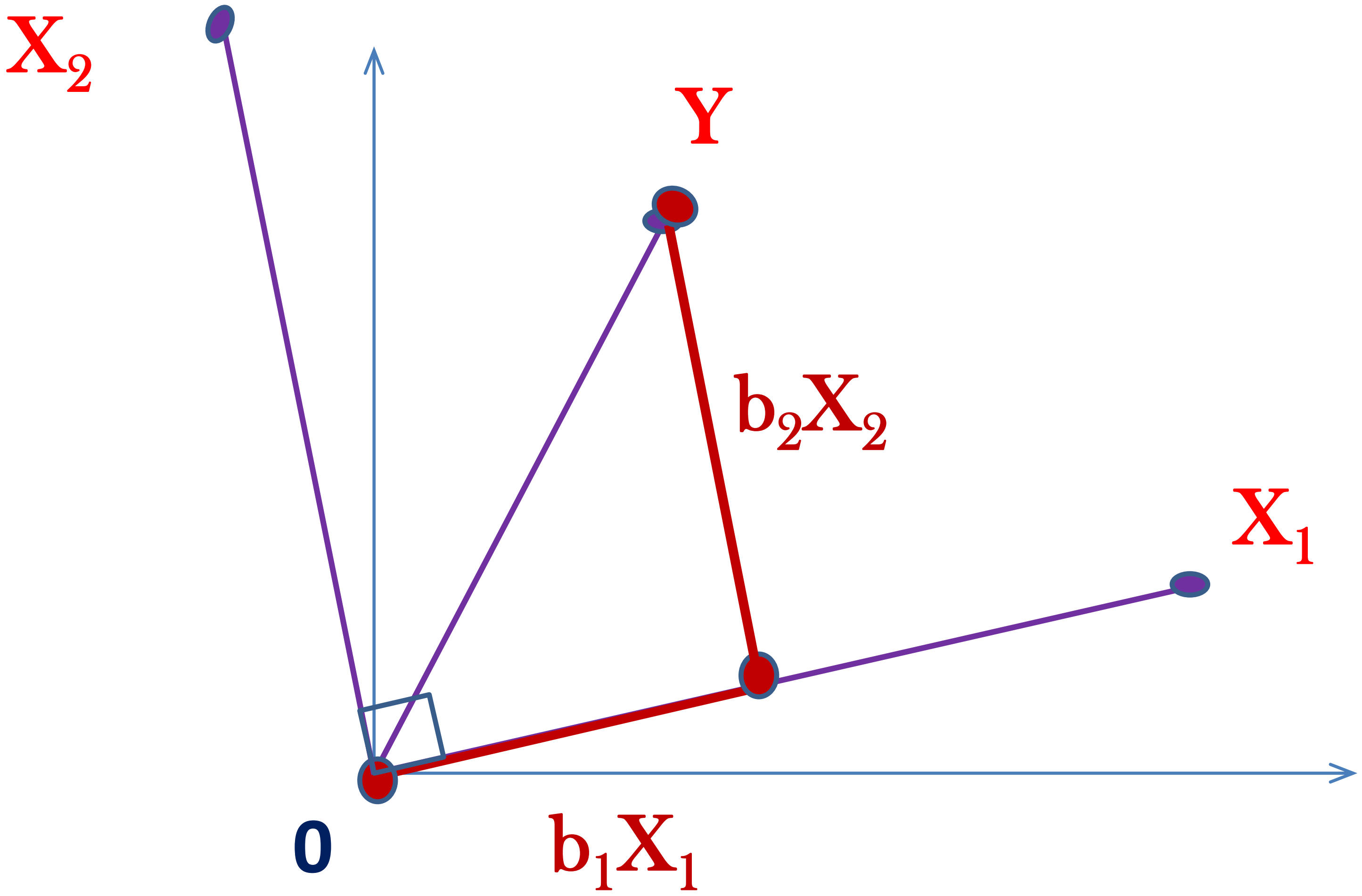
regression coeffs. for X_1 and X_2 together



X_2 ‘eliminating’ X_1

X_1 ‘eliminating’ X_2

‘ignoring’=‘eliminating’ if $X_1 \perp X_2$



Jones M P (1996) Indicator and Stratification Methods for Missing Explanatory Variables…

- "…missing-indicator methods show unacceptably large biases in practical situations and are not advisable in general."

- bias (compared to complete case analysis) arises because of
  - correlation between X variables (as shown)
  - correlation between Y and incomplete X (because of effect on error estimate)
    - In Jones' examples this is taken as very high

increase in precision can outweigh bias

- (2002) it was reasonable to presume that the correlation between (CatK) genotype and the ‘other’ factors (age, BMI, HRT…) was low

- it was important to adjust for those factors where possible

- it was important to keep as many subjects as possible

- the main interest was in the power of the significance test of association between genotype and BMD

more recently…

- Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Stat*. 2007 February ; 61(1): 79–90

- Allison, Paul D. (2009) "Missing Data." Pp. 72-89 in The SAGE Handbook of Quantitative Methods in Psychology, edited by Roger E. Millsap and Alberto Maydeu-Olivares. Thousand Oaks, CA: Sage Publications Inc.

- Groenwold RHH, White IR, Donders ART, Carpenter JR, Altman DG, Moons KGM. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ : Canadian Medical Association Journal*. 2012;184(11):1265-1269.

more recently…

- Horton (2007): These ad-hoc approaches have the potential to induce bias and are not recommended (Jones 1996, Greenland & Finkle 1995).

- Allison (2009): Unfortunately, Jones (1996) proved that this method typically produces biased estimates of the regression coefficients, even if the data are MCAR… Although these methods probably produce reasonably accurate standard error estimates, the bias makes them unacceptable.

- Groenwold (2012): (‘Key Point’) In nonrandomized studies, the factor or test under study is often related to variables with missing values, in which case the missing indicator method typically results in biased estimates.

closing remarks

- notwithstanding the progress made in more sophisticated methods of handling missing data, simple methods can still have value

- ‘indicator variable’ methods will produce biased estimates of regression coefficients, and overestimate the residual variance (compared to ‘complete case’ analysis)

- but gains in precision can more than offset those losses (if you are careful)