A Correlation Metric for Cross-Sample Comparisons Using Logit and Probit

July 1, 2011

Bamberg (German Stata User Group Meeting)

KRISTIAN BERNT KARLSON w/ Richard Breen and Anders Holm

SFI – The Danish National Centre of Social Research

Department of Education, Aarhus University

CONTENTS

ISSUE: INTERACTION TERMS

Interaction effects in logit/probit models not identified

Allison (1999): Differences in true effects conflated by differences in conditional error variance (i.e., heteroskedasticity)

ISSUE: INTERACTION TERMS

Assume: binary y, manifestation of latent y*.

$$y^* = \alpha + \beta x + s\omega$$

Following standard econometrics, a logit coefficient identifies:

$$b = \frac{\beta}{s}$$

Beta = effect from underlying linear reg. model of y* on x

s = (function of) latent error standard deviation, sd(y*|x)

ISSUE: INTERACTION TERMS

Allison noted problem when comparing effects across groups:

$$d = b_2 - b_1 = \frac{\beta_2}{s_2} - \frac{\beta_1}{s_1}$$

We cannot identify difference of interest:

$$d^* = \beta_2 - \beta_1$$

SOLUTION: A REINTERPRETATION OF THE LOGIT COEFFICIENT

Interaction terms = identification issue not easily resolved!

We suggest a new strategy.

Shift of focus from differences in effects (not identified) to differences in correlations (identified).

= possible solution to problem identified by Allison (1999)
in some situations met in real applications

SOLUTION: A REINTERPRETATION OF THE LOGIT COEFFICIENT

We show how to derive, from a logit/probit model, the correlation between an observed predictor, x, and the latent variable, y*, assumed to underlie the binary variable, y:

$$r_{y^*x} = \frac{b \times sd(x)}{\sqrt{b^2 \operatorname{var}(x) + \operatorname{var}(\omega)}} = \frac{\operatorname{cov}(x, y^*)}{sd(x)sd(y^*)}$$

where b is a logit/probit coefficient and var(ω) the variance of a standard logistic/normal variable ($\pi^2/3$ for logit, 1 for probit).

SOLUTION: A REINTERPRETATION OF THE LOGIT COEFFICIENT

It follows that:

$$b = \frac{r_{y^*x}}{\sqrt{1 - r_{y^*x}^2}} \frac{sd(\omega)}{sd(x)}$$

Thus:

$$d = \frac{r_{y^*x,2}}{\sqrt{1 - r_{y^*x,2}^2}} \frac{sd(\omega)}{sd(x_2)} - \frac{r_{y^*x,1}}{\sqrt{1 - r_{y^*x,1}^2}} \frac{sd(\omega)}{sd(x_1)}$$

SOLUTION: A REINTERPRETATION OF THE LOGIT COEFFICIENT

Uses of the correlation metric for comparisons:

\+ interest in the relative positions of individuals (or other units of analysis) within a group, e.g., countries, regions, cohorts.

\- interest in the absolute positions of individuals within groups

\- interest in group-differences in effects, but not the within-group relative positions (e.g., gender, ethnicity).

EXAMPLE: TRENDS IN IEO IN THE US

Thanks to Uli Kohler, -nlcorr- implements the new metric.

EXAMPLE: Did IEO decline across cohorts born in 20th century?

GSS DATA

* Five 10-year birth cohorts, 1920 to 1969.
* Outcome: high school graduation (y=0/1, y* = educ. propensity)
* Predictor: Parental SES (papres80)

Corrrelation of interest = corr(SES, y*), over cohorts!

EXAMPLE: TRENDS IN IEO IN THE US

Previous research, argument for using logit coefficients:

‘differences in [social] background effects … cannot result from changing marginal distributions of either independent or dependent variables because such changes do not affect [the parameter estimates]’ (Mare 1981: 74, parentheses added).

But given our reexpression of the logit coefficent, differences in logit effects across groups (cohorts) will also reflect differences in sd(x).

EXAMPLE: TRENDS IN IEO IN THE US

Trends with logit coefficients

```
. esttab m1 m2 m3 m4 m5
```

	1920-1929	1930-1939	1940-1949	1950-1959	1960-1969
	(1)	(2)	(3)	(4)	(5)
	hs	hs	hs	hs	hs
hs					
papres80	0.0510***	0.0495***	0.0488***	0.0567***	0.0515***
	(8.77)	(9.10)	(9.03)	(11.86)	(9.83)
_cons	-1.197***	-0.600**	0.102	0.0228	0.164
	(-5.18)	(-2.81)	(0.48)	(0.12)	(0.79)
N	2016	2457	3894	5302	4870

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

THE DANISH NATIONAL CENTRE FOR SOCIAL RESEARCH

EXAMPLE: TRENDS IN IEO IN THE US

Trends with correlations

```
. nlcorr logit hs papres80 [pw=wtssall], over(coh6cat)
```

Covariate and coh6cat	NL_Corr	Fisher	Std. Err.	z	sig.
papres80					
1920-1929	.2760257	.2833748	.0314611	9.007151	1.93e-18
1930-1939	.2865121	.2947623	.0314897	9.36059	7.51e-20
1940-1949	.3040799	.314009	.0336668	9.326957	1.03e-19
1950-1959	.3711105	.3897103	.0312976	12.45175	1.71e-34
1960-1969	.3518855	.3675941	.0358131	10.26424	1.06e-23

EXAMPLE: TRENDS IN IEO IN THE US

Trends with correlations, decomposed

```
. nlcorr logit hs papres80 [pw=wtssall], over(coh6cat) altout
```

Covariate and coh6cat	NL_Corr	Fisher	Std. Err.	Ratio	Std. Dev.|X
papres80					
1920-1929	.2760257	.2833748	.0314611	.2871826	10.21205
1930-1939	.2865121	.2947623	.0314897	.2990492	10.96442
1940-1949	.3040799	.314009	.0336668	.3191948	11.87381
1950-1959	.3711105	.3897103	.0312976	.39965	12.78491
1960-1969	.3518855	.3675941	.0358131	.3759288	13.24407

EXAMPLE: TRENDS IN IEO IN THE US

Trends with correlations, contrasts, statistical tests

```
. nlcorr logit hs papres80 [pw=wtssall], over(coh6cat) base(1)
(1 missing value generated)
```

Covariate and coh6cat	Corr. Diff.	Fisher Diff.	z	sig.
papres80				
1920-1929	0	0	0	.
1930-1939	.0104864	.0113875	.3787369	.7426636
1940-1949	.0280542	.0306343	1.115983	.4280562
1950-1959	.0950848	.1063356	4.062163	.0002083
1960-1969	.0758599	.0842193	3.178677	.0051037

CONCLUSION

Correlation metric to be preferred in some situations
-- a solution to the issue identified by Allison (1999)

Example: Evidence on trends in IEO different when correlation metric used (compared to logit coefficients).

WP: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1857431

A Reinterpretation of Coefficients from Logit, Probit, and Other Non-Linear Probability Models: Consequences for Comparative Sociological Research