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

• An issue!
• A solution?
• An example: Trends in IEO in the US
• A conclusion
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 $x$  
$s = (\text{function of})$ latent error standard deviation, $\text{sd}(y^*|x)$
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, $y$:

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

where $b$ is a logit/probit coefficient and $\var(\omega)$ 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
* Outcome: high school graduation (y=0/1, y* = educ. propensity)
* Predictor: Parental SES (papres80)

Correlation 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 coefficient, 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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) hs</td>
<td>(2) hs</td>
<td>(3) hs</td>
<td>(4) hs</td>
<td>(5) hs</td>
</tr>
<tr>
<td>hs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>papres80</td>
<td>0.0510***</td>
<td>0.0495***</td>
<td>0.0488***</td>
<td>0.0567***</td>
<td>0.0515***</td>
</tr>
<tr>
<td></td>
<td>(8.77)</td>
<td>(9.10)</td>
<td>(9.03)</td>
<td>(11.86)</td>
<td>(9.83)</td>
</tr>
<tr>
<td>_cons</td>
<td>-1.197***</td>
<td>-0.600**</td>
<td>0.102</td>
<td>0.0228</td>
<td>0.164</td>
</tr>
<tr>
<td></td>
<td>(-5.18)</td>
<td>(-2.81)</td>
<td>(0.48)</td>
<td>(0.12)</td>
<td>(0.79)</td>
</tr>
<tr>
<td>N</td>
<td>2016</td>
<td>2457</td>
<td>3894</td>
<td>5302</td>
<td>4870</td>
</tr>
</tbody>
</table>
```
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
EXAMPLE: TRENDS IN IEO IN THE US

Trends with correlations

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

<table>
<thead>
<tr>
<th>Covariate and coh6cat</th>
<th>NL_Corr</th>
<th>Fisher</th>
<th>Std. Err.</th>
<th>z</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>papres80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1920–1929</td>
<td>.2760257</td>
<td>.2833748</td>
<td>.0314611</td>
<td>9.007151</td>
<td>1.93e–18</td>
</tr>
<tr>
<td>1940–1949</td>
<td>.3040799</td>
<td>.314009</td>
<td>.0336668</td>
<td>9.326957</td>
<td>1.03e–19</td>
</tr>
<tr>
<td>1950–1959</td>
<td>.3711105</td>
<td>.3897103</td>
<td>.0312976</td>
<td>12.45175</td>
<td>1.71e–34</td>
</tr>
</tbody>
</table>
EXAMPLE: TRENDS IN IEO IN THE US

Trends with correlations, decomposed

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>papres80 1920-1929</td>
<td>.2760257</td>
<td>.2833748</td>
<td>.0314611</td>
<td>.2871826</td>
<td>10.21205</td>
<td></td>
</tr>
<tr>
<td>1930-1939</td>
<td>.2865121</td>
<td>.2947623</td>
<td>.0314897</td>
<td>.2990492</td>
<td>10.96442</td>
<td></td>
</tr>
<tr>
<td>1940-1949</td>
<td>.3040799</td>
<td>.314009</td>
<td>.0336668</td>
<td>.3191948</td>
<td>11.87381</td>
<td></td>
</tr>
<tr>
<td>1950-1959</td>
<td>.3711105</td>
<td>.3897103</td>
<td>.0312976</td>
<td>.39965</td>
<td>12.78491</td>
<td></td>
</tr>
<tr>
<td>1960-1969</td>
<td>.3518855</td>
<td>.3675941</td>
<td>.0358131</td>
<td>.3759288</td>
<td>13.24407</td>
<td></td>
</tr>
</tbody>
</table>
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)
```

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>papres80 1920–1929</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1930–1939</td>
<td>.0104864</td>
<td>.0113875</td>
<td>.3787369</td>
<td>.7426636</td>
</tr>
<tr>
<td>1940–1949</td>
<td>.0280542</td>
<td>.0306343</td>
<td>1.115983</td>
<td>.4280562</td>
</tr>
<tr>
<td>1950–1959</td>
<td>.0950848</td>
<td>.1063356</td>
<td>4.062163</td>
<td>.0002083</td>
</tr>
</tbody>
</table>
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).

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