

# Measurement in the Social Sciences: Models for Binary Data

Dave Armstrong

University of Oxford  
Department of Politics and International Relations  
Center for Research Methods in the Social Sciences

t: 01865 285956  
e: david.armstrong@politics.ox.ac.uk  
w: <http://users.ox.ac.uk/~polf0104>

Office: 1101 Manor Road Building (Politics Dept)  
Office Hours: TWTh 1400-1500

May 15, 2008

1 / 34

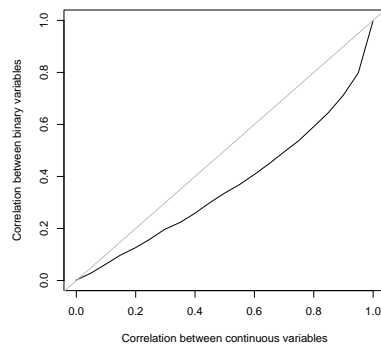
## Questions We Will Answer Today

- When might it be appropriate to use a traditional factor model on Binary data?
  - When it is appropriate, how do we do it?
- When it is not appropriate to use factor analysis on binary data, what is the alternative?
- How to evaluate when these models are appropriate

2 / 34

## Correlation for Binary Variables

- It is not *necessarily, always* wrong to use correlation when talking about binary variables. There are a couple of different reasons for this.
- If you can think of your binary variable as the “dichotomization” of some underlying continuous trait, then it might make sense.
- The graph to the right plots the pattern of correlation between 2 continuous variables ( $\mu = 0$ ,  $\sigma = 1$ ) on the  $x$ -axis and the correlation between the dummy variables constructed by the following rule:  $d = 1$  if  $x \geq 0$ , else  $d = 0$ , on the  $y$ -axis.



3 / 34

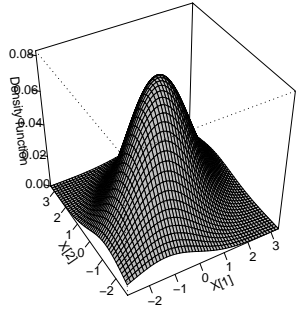
## Correlations for Binary Variables

- The relationship on the previous slide only holds when the latent continuous variables are cut symmetrically at the mean of the bivariate density.
- The next few slides will show what happens to this relationship when they are cut at other places than their means.
- The results are both not necessarily what you would expect, and further, they will call into question the utility of this approach unless there is a strong theoretical reason to assume otherwise.

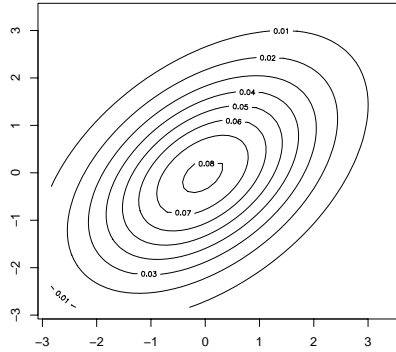
4 / 34

### Bivariate Density

These are two different representations of a bivariate density. The first is a 3-D surface plot and the second is a contour plot.



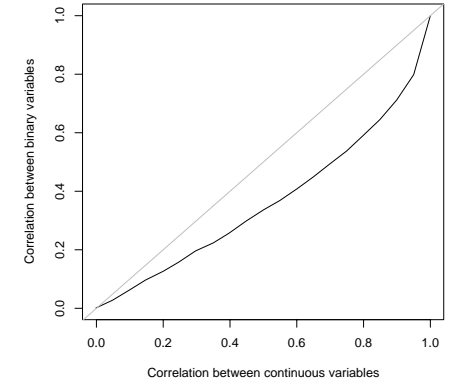
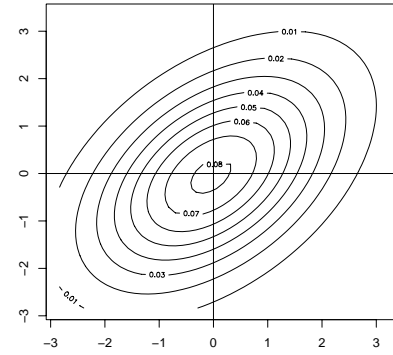
(a) 3-D Surface



(b) Contour

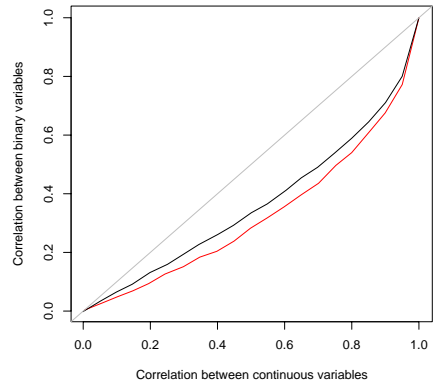
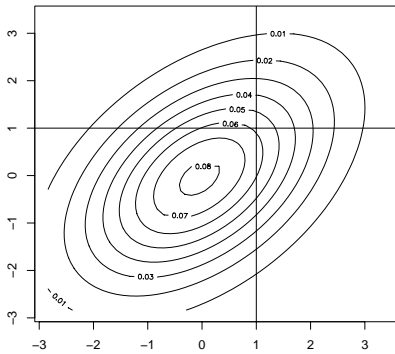
### Ideal Situation

Here, we will just show the situation shown before, cut-points at the mean of the bivariate density.



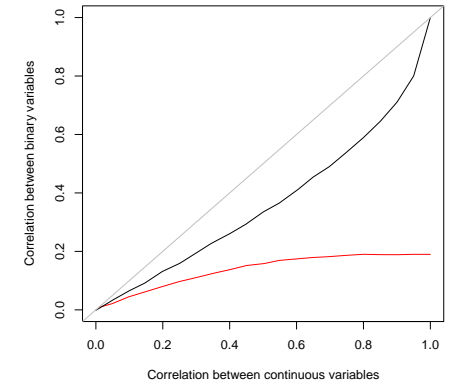
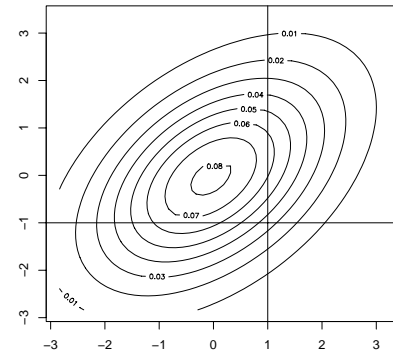
### Symmetric - both high

Here, the cut-points are both at 1. The black-line represents the best-case scenario and the red-line represents the relationship for this case.



### Symmetric - one high, one low

Here, one cut-points is high and the other is low. The black-line represents the best-case scenario and the red-line represents the relationship for this case.



## Redux

- When the cut-points slicing up the underlying latent density are symmetric at the mean of the bivariate density, then correlations are a bit of an underestimate of the true relationship, but maybe not that much.
- As these cut-points depart from the mean of the bivariate density, the results get marginally worse, so long as they remain symmetric.
- As the cut-points depart from symmetry, all bets are off.

9/34

## Correlation and $\phi$

- Another reason that the correlation is not *necessarily* the wrong to use is that there is a measure of association between categorical variables called  $\phi$  (phi) that essentially reduces to the Pearson correlation coefficient  $r$ .

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

- This is appropriate if the data are “truly” categorical - that is it is inappropriate to think of the variables as just being a grouping of some “perfectly-precise, normally-distributed random variable”.
- You could easily factor analyze a matrix of  $\phi$ 's. It does, however, become more difficult to think about what the communality estimates might be, though if not 1.

10/34

## Tetrachoric correlation

- If the two variables are “binned” realizations of some underlying continuous variable, then you could use a tetrachoric correlation.
- Stata uses an approximation for this as discussed below, where  $\tilde{r}$  is the tetrachoric correlation. Assume we have the following cross-tabulation:

	0	1
0	a	b
1	c	d

then:

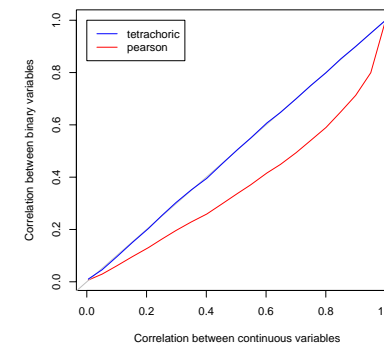
$$\alpha = \frac{a \times d}{b \times c}$$

$$\tilde{r} = \frac{\alpha^{\frac{\pi}{4}} - 1}{\alpha^{\frac{\pi}{4}} + 1}$$

11/34

## Tetrachoric vs. Pearson Correlations(1)

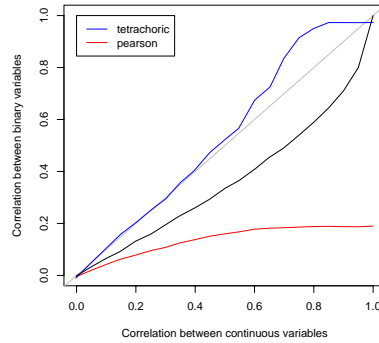
- As you can see by the graph at the right, the tetrachoric correlation is almost a perfect reproduction of the underlying correlation of the bivariate normal data.



12/34

## Tetrachoric vs. Pearson Correlations(2)

- This breaks down when the cut-points are markedly asymmetric, this is the situation where one cut-point = 1 and the other = -1. When the underlying correlation is high, the tetrachoric correlation overstates the true correlation.



13/34

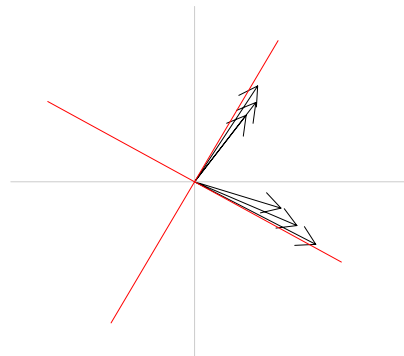
## Underlying Continuous Correlation Matrix

	[, 1]	[, 2]	[, 3]	[, 4]	[, 5]	[, 6]
[1, ]	1.000	0.354	0.537	0.091	0.049	0.096
[2, ]	0.354	1.000	0.431	0.071	0.127	0.088
[3, ]	0.537	0.431	1.000	0.092	0.035	0.045
[4, ]	0.091	0.071	0.092	1.000	0.425	0.316
[5, ]	0.049	0.127	0.035	0.425	1.000	0.350
[6, ]	0.096	0.088	0.045	0.316	0.350	1.000

14/34

## Factor Analysis of underlying correlation matrix

var	$a_{j1}$	$a_{j2}$	$U_j$
v1	0.659	0.065	0.560
v2	0.524	0.111	0.712
v3	0.814	0.019	0.336
v4	0.073	0.604	0.629
v5	0.032	0.697	0.512
v6	0.063	0.506	0.739
$\lambda_m^2$	1.384	1.126	



15/34

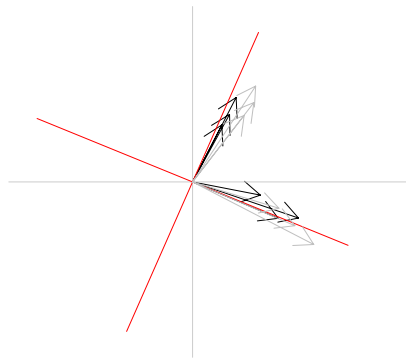
## Generating Categorical Measures of Association

1. Draw continuous variables with the correlation matrix above and all with mean 0.
2. Then, we generate dummy variables such that  $d = 1$  if  $x_j \geq 0$  and 0 otherwise.
3. Then, we'll generate a correlation matrix, which we'll call a  $\phi$  matrix and a tetrachoric correlation matrix.
4. Finally, we'll perform exactly the same procedure to see what we get.

16/34

### Factor Analysis of $\phi$ matrix

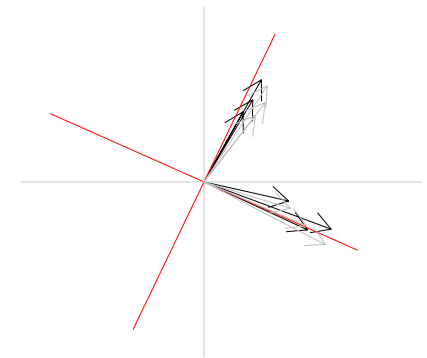
var	$a_{j1}$	$a_{j2}$	$U_j$
v1	0.554	-0.008	0.692
v2	0.401	0.082	0.832
v3	0.664	0.040	0.557
v4	0.039	0.468	0.779
v5	0.034	0.579	0.663
v6	0.022	0.390	0.846
$\lambda_m^2$	0.913	0.715	



17/34

### Factor Analysis of $\tilde{r}$ matrix

var	$a_{j1}$	$a_{j2}$	$U_j$
v1	0.680	-0.010	0.537
v2	0.504	0.103	0.735
v3	0.805	0.051	0.349
v4	0.050	0.579	0.661
v5	0.044	0.712	0.490
v6	0.028	0.488	0.760
$\lambda_m^2$	1.370	1.096	



18/34

### Should you do FA on either $\phi$ or $\tilde{r}$ ?

- If the cross-tabs for your observed variables are roughly symmetric and heavily populated on the diagonal.
  - If this is not the case, we'll see in the second part of class what models are generally used for these types of relationships.
- Just remember, if you use  $\phi$ , you're likely underestimating the relationship and with cross-tabs that are marked departures from symmetry, these discrepancies could be quite large.
- Remember though  $\tilde{r}$  performs better than  $\phi$  in the case of asymmetry, it is still not perfect.
- In sum, it's not necessarily a bad idea, you just need to be careful to check your results to make sure that asymmetry is not causing big problems.

19/34

### A short note on the theoretical model

- The graphical model presented last class still characterizes our understanding about how the latent concept relates to the observed variables.
- There is, however, another bit of added complexity. We also have some cut-points  $\tau_j$  such that if  $x_j \geq \tau_j$  then  $d_j = 1$  and  $d_j = 0$  otherwise.
- The traditional factor model assumes that these  $\tau$ 's are arrayed randomly around the center of the continuous variable.



20/34

## Cumulative Scaling

- What happens if the observed variables are *not* symmetric?
- There is a broad set of techniques that might be jointly referred to as “cumulative scaling” techniques that deal with just this type of data.
- We will talk about two of these models that are estimable in Stata: Mokken scale and Rasch model.

21 / 34

## Graphical Representation

- The idea here is that the cut-points are not randomly arrayed around the center. They are systematically arrayed from left to right and their ordering tells us something about 1) the ordering of the items with respect to the latent dimension and 2) the ordering of individuals with respect to the latent dimension.

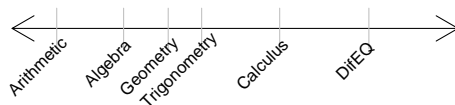


- The cut-points tell you “how much of the underlying continuous trait do you need to observe a 1 on the observed variable”.

22 / 34

## The Quintessential Example: Math Test

- These methods originate from and are most commonly used in testing.
- Imagine a math test where there are 6 questions, one each dealing with - arithmetic, algebra, geometry, trigonometry, calculus, and differential equations.
- If we think about these in terms of traditional correlation, we would expect that in general, people would miss questions at random - that is there would be no particular pattern in which questions people miss.
- In a cumulative scale, we would expect people to stop answering questions correctly based on their mathematical ability. We would expect something like this:



23 / 34

## More Math Example

- We would expect someone who has just finished high school to be able to answer maybe up to the geometry or trigonometry question, but it would be unlikely they would be able to answer the differential equation question.
- Similarly, if someone could answer the differential equation question, we would expect that they would also be able to answer all the other questions.
- Let's take this understanding and start to think of how we might operationalize this model.

24 / 34

## Cumulative Model and Cross-Tabs

- Let's think about just two of these questions - let's say arithmetic and algebra. What would we expect the cross-tab to look like?

		Algebra	
		Wrong	Right
Arithmetic	Wrong	$a$	$b$
	Right	$c$	$d$

- So, what sorts of responses make sense?  $a$  does - it is possible that someone with no math ability would get both questions wrong,  $d$  does because we would expect someone with higher math ability to get both right,  $c$  does as it would be reasonable that someone with a bit of math ability could answer an arithmetic question, but not an algebra question.  $b$  is the only cell that should not have many observations.

25 / 34

## Cumulative Model and Asymmetry

- The cumulative scaling model is one that allows for (and expects) asymmetry in the cross-tabulation of any 2 variables.
- It would be possible to see symmetric tables if:
  - Your sample is a bimodal distribution of geniuses and idiots, or
  - you have two items that have nearly identical cut-points.
- If observations are providing "incorrect" answers in systematic and substantively interesting ways, the cumulative scale can model that.

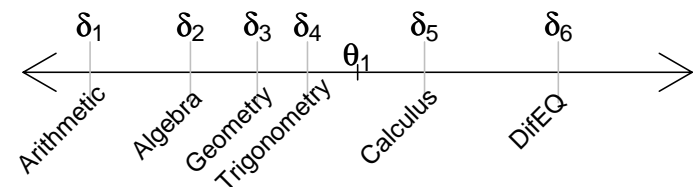
26 / 34

## Formal statement of the model

- We have two different sets of objects we're trying to model - the cut-points (we'll call  $\delta_j$  where  $j$  indexes observed variables and the individual abilities or ideal points, we'll call these  $\theta_i$  where  $i$  indexes individuals.
- For  $j = 1, 2, \dots, k$ , if  $\theta_i < \delta_j$ , then  $\theta_i < \delta_{j+1} < \dots < \delta_k$ .
- Conversely, if  $\theta_i > \delta_j$ , then  $\theta_i > \delta_{j-1} > \dots > \delta_1$ .
- As a result of the above, we would expect that  $\sum_{i=1}^n x_{i,j} > \sum_{i=1}^n x_{i,j+1}$ . Fewer people total will answer a harder question correctly than an easier one.
- Let's go back to the math example...

27 / 34

## Back to the Math Test



- The model says that since  $\delta_4 < \theta_1 < \delta_5$ , that this person should get the trigonometry question right and the calculus question wrong.

28 / 34

## Model Assumptions

- This model assumes unidimensionality, though there are 2-dimensional IRT models as well. They're more difficult to implement, so we're not going to talk about them much here, but if you're interested, there are lots of references.
- The only other model assumption is called "double monotonicity" which states that two conditions must hold simultaneously:
  - For any fixed  $\theta_i$ ,  $p(x_1 = 1) \geq p(x_2 = 1) \geq \dots \geq p(x_k = 1)$  - the probabilities of a positive response should be monotonically decreasing.
  - For any fixed  $\delta_j$ ,  $p(v_{1j} = 1) \leq p(v_{2j} = 1) \leq \dots \leq p(v_{nj} = 1)$  - the probability that any observation gives a positive response should be monotonically related to the ordering of the ideal points.

29 / 34

## Assessing the Assumptions

- Consider the following table:

		$v_2$		
		0	1	
$v_1$	0	$f_{00}$	$f_{01}$	$f_{0+}$
	1	$f_{10}$	$f_{11}$	$f_{1+}$
	$\sum$	$f_{+0}$	$f_{+1}$	$n$

- Let's assume that  $\delta_1 < \delta_2$  that is to say,  $p(v_2) < p(v_1)$ . If this is the case, then we could expect to see  $f_{00}$ ,  $f_{10}$  and  $f_{11}$ , but *not*  $f_{01}$ . So, this is the cell that is in error.
- We want to see how much in error it is, that is, how different from our expectations it is. We'll denote our expected frequency as  $f_{01}^e$ .
- We can calculate  $f_{01}^e = \frac{f_{+1}f_{0+}}{n}$

30 / 34

## H-statistic

- We can calculate what Mokken called the  $H$ -statistic for any pair of items:

$$H_{jk} = 1 - \left( \frac{f_{01}}{f_{01}^e} \right)$$

where  $f_{01}$  is the error cell. (Note: the error cell could be  $f_{10}$  depending on how you set up your cross-tab, so be careful to make sure you're looking at the right cell).

- We can similarly create an  $H$ -statistic for any particular item by summing observed and expected frequencies across all of the pairs involving that particular variable:

$$H_j = 1 - \left( \frac{\sum^{pairs} f_{01}}{\sum^{pairs} f_{01}^e} \right)$$

31 / 34

## Evaluating $H$ -statistics

- Mokken suggested the following rules of thumb for evaluating these coefficients.
  - For  $H_j$ , it should be bigger than 0.3 and would be nice if it were above 0.5. Current thinking is that these limits might be a bit generous and that a more appropriate cut-off for acceptability is 0.4.
  - For  $H_{ij}$ , every one should be positive and statistically different from zero.
- $H$  is bound on the upper-end at 1, but has no particular lower bound. Negative values are suggestive that other models might be more appropriate for your data.
- We'll go through how Stata generates these numbers later in the worked example.

32 / 34

## Generating Scores

- As with all of these other methods, though we want to know the structure of the data, what we really want is an estimate for each individual of their position on the underlying dimension.
- There are a couple of possibilities here:
  1. By far the most common is to just sum up the number of ones:  $\sum_{j=1}^k x_{ij}$ . This should work because if you answered 4 questions right, the model assumes they're the first 4.
  2. Another possibility is to stick individuals past the cut-point of the hardest item to which they gave a positive response. This is a bit more difficult and if the model is right, there will be only marginal differences between these two methods.
- This is a bit unsatisfying as it has the same operationalization as the SRM, but remember the only reason we can aggregate data at all is because we can provide evidence that we have met the assumptions of some theoretical model. It is entirely possible that variables for which the SRM is deemed inappropriate could still be summed to create an estimate of the underlying dimension if they fit the Mokken model.

33 / 34

## Rasch Model

- The Mokken scaling model is also called a “non-parametric IRT (Item Response Theory)” model. It is non-parametric because it doesn't impose any particular functional form on the trace-line relating the latent variable to the probability of giving a positive answer to a particular variable.
- With the Rasch model, we are imposing a shape on the tracelines relating the latent variable to the probability of giving a positive response. Not surprisingly, it follows a logistic regression model:

$$\Pr(v_{ij} = 1) = \frac{1}{1 + e^{-(\beta_{0j} + \beta_1 \theta_i)}}$$

where  $\beta_1$  estimates the steepness of the logit curve and  $\beta_{0j}$  is a variable-specific intercept. Often times,  $\beta_1$  is set to 1.

34 / 34

## Rasch Model Likelihood Function

- The Rasch model is fit generally fit with conditional maximum likelihood and it is maximizing the following function:

$$\mathcal{L} = \prod_{i=1}^{n^*} \prod_{j=1}^k p(v_{ij} = 1)^{v_{ij}} (1 - p(v_{ij} = 1))^{1-v_{ij}}$$

where  $n^*$  is the total number of response patterns and  $k$  is the number of observed variables.

- We will look at the output from the Rasch model in the worked examples as well.

35 / 34

## Questions We Answered Today

- When might it appropriate to use a traditional factor model on Binary data?
  - When it is appropriate, how do we do it?
- When it is not appropriate to use factor analysis on binary data, what is the alternative?
- How to evaluate when these models are appropriate

36 / 34