Social relations models for binary, ranked and ordinal relations

567 Statistical analysis of social networks

Peter Hoff

Statistics, University of Washington
Adapting the normal SRM

Relational data are often valued, non-binary.

The standard SRM is appropriate for **normal** valued relational data.

**However,** relational data is often neither binary nor normal:
- Links can be **absent** or **continuous** within the same network.
  - meaning $Y_{i,j}$ is either 0 or some arbitrary real number;
  - communication networks (time spent communicating);
- Links can be **absent** or **ordinal** within the same network.
  - conflict networks (negative, positive or zero relation);
  - ranked nominations (friends are ranked, non-friends are not).

The SRM can be adapted via **ordinal probit models** to handle such data.
An **ordinal variable** has a meaningful ordering to the possible outcomes. This is in contrast to a **categorical** (non-ordered) variable.

**Ordinal variable**
- continuous: (all real numbers)
- discrete: (counts, ranks, etc.)

**Categorical variable**
- non-orderable categories (religion, ethnicity, etc.)
The simplest ordinal variable is a binary random variable $Y \in \{0, 1\}$. Let

\[ \Pr(Y = 1) = \theta \]
\[ \Pr(Y = 0) = 1 - \theta \]

This model for $Y$ has the following **latent variable representation**:

\[ \mu = \Phi^{-1}(\theta) \]
\[ Z \sim N(\mu, 1) \]
\[ Y = 1 \times (Z > 0) \]

Here, $\Phi^{-1}(\theta)$ is the $\theta$-quantile of the standard normal distribution.
Probit representation

To confirm the representation, recall

- If $Z \sim N(\mu, 1)$ then $Z - \mu \sim N(0, 1)$;
- If $Z \sim N(\mu, 1)$ then $Z = \mu + \epsilon$, where $\epsilon \sim N(0, 1)$;
- If $\epsilon \sim N(0, 1)$ then $-\epsilon \sim N(0, 1)$.

Now do the calculation:

$$
\Pr(Y = 1) = \Pr(Z > 0) \\
= \Pr(-Z < 0) \\
= \Pr([-Z + \mu] < \mu) \\
= \Pr(\epsilon < \mu) = \Phi(\mu) = \theta
$$
Probit representation

\[ \theta = .15, \mu = \Phi^{-1}(\theta) = -1.04 \]
Probit regression

Now suppose we have

\textbf{binary data} $Y_1, \ldots, Y_n$

we want to relate to

\textbf{explanatory variables} $x_1, \ldots, x_n$

\textbf{Latent variable model:}

\[ \epsilon_1, \ldots, \epsilon_n \sim \text{i.i.d. } N(0, 1) \]

\[ Z_i = \beta^T x_i + \epsilon_i \]

\[ Y_i = 1 \times (Z_i > 0) \]

Under this latent variable model, the $Y_i$’s are independent and

\[ \Pr(Y_i = 1) = \Pr(Z_i > 0) \]

\[ = \Pr(\beta^T x_i + \epsilon_i > 0) \]

\[ = \Pr(-\epsilon_i < \beta^T x_i) \]

\[ = \Pr(\epsilon_i < \beta^T x_i) = \Phi(\beta^T x_i) \]
This latent variable model is exactly the same as **probit regression**:

\[
Pr(Y_1 = y_1, \ldots, Y_n = y_n | x_1, \ldots, x_n) = \prod_{i=1}^{n} \Phi(\beta^T x_i)^{y_i} [1 - \Phi(\beta^T x_i)]^{1-y_i}
\]

Compare to **logistic regression**:

\[
Pr(Y_1 = y_1, \ldots, Y_n = y_n | x_1, \ldots, x_n) = \prod_{i=1}^{n} \left(\frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right)^{y_i} \left(\frac{1}{1+e^{\beta^T x_i}}\right)^{1-y_i}
\]

In fact, logistic regression has a latent variable representation also:

\[
\epsilon_1, \ldots, \epsilon_n \sim \text{i.i.d. } L(0, 1)
\]

\[
Z_i = \beta^T \mathbf{x}_i + \epsilon_i
\]

\[
Y_i = 1 \times (Z_i > 0)
\]
Let’s examine the effect of age difference on dominance.
Sheep dominance data

Let

- \( \text{dom}[i,j] = \) indicator that \( i \) has dominated \( j \) at least once;
- \( \text{aged}[i,j] = \) age\( i \) - age\( j \).

\[
\text{mean(\text{aged}[\text{dom}\!=\!1],na.rm=TRUE )}
\]

## [1] 2.184

\[
\text{mean(\text{aged}[\text{dom}\!=\!0],na.rm=TRUE )}
\]

## [1] -1.079051

\[
\text{summary( \text{glm}(\text{dom}\sim\text{aged},\text{family}\!=\!\text{binomial}) )$coef}
\]

##
### Estimate Std. Error z value Pr(>|z|)
### (Intercept) -0.8357694 0.08729001 -9.574628 1.022245e-21
### aged 0.2260270 0.02317097 9.754752 1.760303e-22

\[
\text{summary( \text{glm}(\text{dom}\sim\text{aged},\text{family}\!=\!\text{binomial(link}\!=\!\text{probit}) )$coef}
\]

##
### Estimate Std. Error z value Pr(>|z|)
### (Intercept) -0.5138332 0.05111585 -10.05232 8.972511e-24
### aged 0.1385751 0.01337034 10.36437 3.601223e-25
Comparing logistic and probit regression output:

- $\hat{\beta}$’s are on different scales;
- inference (z-scores) are typically similar.

Both models are **binary regression models** of the form:

$$
Pr(Y_1 = y_1, \ldots, Y_n = y_n | x_1, \ldots, x_n) = \prod_{i=1}^{n} g(\beta^T x_i)^{y_i} [1 - g(\beta^T x_i)]^{1-y_i},
$$

where $g^{-1}$ is called the **inverse-link function**.

**Link functions:**

- **logistic regression**: $g^{-1}(\mu) = \exp(\mu) /[1 + \exp(\mu)];$
- **probit regression**: $g^{-1}(\mu) = \Phi(\mu);$  
- **other binary regression**: $g^{-1}(\mu)$ is a strictly increasing function.
SRM for binary relational data

Recall the latent variable representation of probit regression:

\[ Z_{i,j} = \beta^T x_{i,j} + \epsilon_{i,j} \]
\[ Y_{i,j} = 1 \times (Z_{i,j} > 0) \]

We could estimate \( \beta \) with a probit regression analysis, but what about network dependence in the data?

Could there be

- across-row heterogeneity/within row correlation?
- across-column heterogeneity/within column correlation?
- within dyad correlation?
SRM for binary relational data

Recall the latent variable representation of probit regression:

\[ Z_{i,j} = \beta^T x_{i,j} + \epsilon_{i,j} \]
\[ Y_{i,j} = 1 \times (Z_{i,j} > 0) \]

Dependence on the \( Y \)-scale can be induced by dependence on the \( Z \)-scale:

\[ \epsilon_{i,j} = a_i + b_j + e_{i,j} \]
\[ \{(a_1, b_1), \ldots, (a_n, b_n)\} \sim \text{i.i.d. } \mathcal{N}(0, \Sigma_{ab}) \]
\[ \{(e_{i,j}, e_{j,i}) : i \neq j\} \sim \text{i.i.d. } \mathcal{N}(0, \Sigma_e) \]

\[ \Sigma_{ab} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \quad \Sigma_e = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \]

Note that the variance of \( e_{i,j} \) is fixed at 1.

The scales of \( e_{i,j} \) and \( \beta \) are not separately identifiable.
SRM probit model:

\[ Y_{i,j} = 1 \times (Z_{i,j} > 0) \]
\[ Z_{i,j} = \beta^T x_{i,j} + \epsilon_{i,j} \]
\[ \epsilon_{i,j} = a_i + b_j + e_{i,j}, \]

and \( \{a_i, b_j, e_{i,j}\} \) are random effects as described previously.

Parameter estimation:

- The likelihood can’t be expressed in closed form;
- Bayesian parameter estimates can be obtained via MCMC. The latter is provided in the package \texttt{amen}. 
AME fit for binary relational data

Description:

An MCMC routine providing a fit to an additive and multiplicative effects (AME) regression model for binary relational data

Usage:

ame_bin(Y, X, rvar = TRUE, cvar = TRUE, dcor = TRUE, R = 0, seed = 1, nscan = 50000, burn = 500, odens = 25, plot = TRUE, print = TRUE)

Arguments:

Y: an n x n square relational matrix
X: an n x n x p array of covariates
rvar: logical: fit row random effects?
cvar: logical: fit column random effects?
dcor: logical: fit a dyadic correlation?
SRM probit in amen

```r
XD <- outer(X, X, "-")
fit_ame_bin <- ame(YB, XD, model = "bin")
```

### Graphs

- **SABR**
- **BETA**
- **sd.rowmean**
- **dyad.dep**
- **sd.colmean**
- **triad.dep**
The `summary` command provides a canned summary of the fitted model:

```
summary(fit_ame_bin)
```

##
## beta:
## pmean   psd  z-stat  p-val
## intercept  -0.654  0.149  -4.374   0
## .dyad       0.200  0.034   5.835   0

## Sigma_ab pmean:
## a  b
## a  0.603 -0.057
## b -0.057  0.172

## rho pmean:
## -0.359
Posterior analysis

Confidence intervals can be obtained via the `quantile` command:

```r
apply( fit_ame_bin$BETA , 2 , quantile, prob=c(.025,.5,.975))
## intercept .dyad
## 2.5% -0.9423416 0.1398794
## 50% -0.6554213 0.1985873
## 97.5% -0.4028070 0.2637526

apply( fit_ame_bin$SABR , 2 , quantile, prob=c(.025,.5,.975))
## va cab vb rho ve
## 2.5% 0.3111828 -0.26359446 0.07500074 -0.5818362 1
## 50% 0.5511227 -0.05253038 0.15423887 -0.3646930 1
## 97.5% 1.2468605 0.09875416 0.33123579 -0.1495458 1
```
Summary of results:

Strong evidence of an age effect:
- Older sheep dominate younger ones.

Row and column heterogeneity is not substantial:
- Compare $\hat{\sigma}_a^2 = 0.55$ and $\hat{\sigma}_b^2 = 0.15$ to the error variance $\sigma_e^2 = 1$.

There is evidence that dominance tends to go one-way in a dyad:
- The 95% CI for $\rho$ is $(-0.58, -0.15)$. 
Goodness of fit

The default amen plot provides four goodness of fit plots:

**Outdegree distribution:** Comparing outdegree distribution to simulated;

**Indegree distribution:** Comparing indegree distribution to simulated;

**Reciprocity:** Comparing fraction reciprocated ties to simulated fraction;

**Transitivity:** Comparing number of triangles to simulated number.

These statistics are computed with the commands

```
t_degree
t_recip
t_trans
```
Model comparison

Let’s use these statistics to evaluate the fit of three models:

*fit_bin_000*: the probit model, fix $\sigma_a^2 = \sigma_b^2 = \rho = 0$.

*fit_bin_001*: dyadic correlation $\rho$ estimated, $\sigma_a^2 = \sigma_b^2 = 0$.

*fit_bin_111*: full SRM covariance model.
fit_bin_000<-ame(YB,XD,model="bin",rvar=FALSE,cvar=FALSE,dcor=FALSE)
Model comparison

```r
fit_bin_001 <- ame(YB, XD, model = "bin", rvar = FALSE, cvar = FALSE, dcor = TRUE)
```
Model comparison

```r
fit_bin_111 <- ame(YB, XD, model = "bin")
```
The SRM model fit_bin_111 looks best based on these statistics, except possibly for the transitivity statistic (more on that soon).
Ordered probit models for ordinal data

The original sheep data consists of counts:

```
YV[1:8,1:8]
```

```
## V1 V2 V3 V4 V5 V6 V7 V8
## [1,] NA 0 0 0 0 0 0 1
## [2,] 0 NA 0 0 5 2 1 0
## [3,] 0 0 NA 0 7 4 0 0
## [4,] 0 0 8 NA 0 0 0 0
## [5,] 0 0 0 0 NA 1 0 0
## [6,] 0 0 7 0 NA 0 0
## [7,] 0 0 1 0 0 0 NA 0
## [8,] 0 1 1 4 5 3 0 NA
```

Ideally, this information should be taken into consideration.

- In principle, throwing away information is not efficient.
- Effects of some covariates might distinguish high values of the relation.
Ordered probit models for ordinal data

Latent variable model:

\[ Z_{i,j} = \beta^T x_{i,j} + \epsilon_{i,j} \]
\[ \epsilon_{i,j} = a_i + b_j + e_{i,j} \]

Binary probit:

\[ Y_{i,j} = \begin{cases} 0 & \text{if } Z_{i,j} < 0 \\ 1 & \text{if } Z_{i,j} > 0 \end{cases} \]

What if \( Y_{i,j} \in \{y_1, \ldots, y_K\} = \mathcal{Y}\)?

Here, \( \mathcal{Y} \) is an ordered, countable set of possible values of \( Y_{i,j} \).
Ordered probit models for ordinal data

Latent variable model:

\[ Z_{i,j} = \beta^T x_{i,j} + \epsilon_{i,j} \]
\[ \epsilon_{i,j} = a_i + b_j + e_{i,j} \]

Ordered probit:

\[ Y_{i,j} = \begin{cases} 
  y_1 & \text{if } Z_{i,j} \in (-\infty, c_1) \\
  y_2 & \text{if } Z_{i,j} \in (c_1, c_2) \\
  \vdots & \\
  y_{K-1} & \text{if } Z_{i,j} \in (c_{K-2}, c_{K-1}) \\
  y_K & \text{if } Z_{i,j} \in (c_{K-1}, \infty) 
\end{cases} \]
Ordered probit models for ordinal data

```r
table(c(YB))
```
```
##
## 0 1
## 506 250
```

```r
table(c(YV))
```
```
##
## 0 1 2 3 4 5 6 7 8 9 10 11 12
## 506 100 59 34 20 13 5 7 5 1 3 1 2
```

```r
round( table(c(YV))/sum(table(c(YV))) ,3 )
```
```
##
## 0 1 2 3 4 5 6 7 8 9 10 11
## 0.669 0.132 0.078 0.045 0.026 0.017 0.007 0.007 0.007 0.001 0.004 0.001
## 12
## 0.003
```
Link function for ordered probit

\[ \begin{align*}
    p(z) & \quad \text{for} \quad z \in (-\infty, \infty) \\
    y & \quad \text{for} \quad y \in (0, 1)
\end{align*} \]
AME fit for ordinal relational data

Description:

An MCMC routine providing a fit to an additive and multiplicative effects (AME) regression model for ordinal relational data

Usage:

```r
ame_ord(Y, X, rvar = TRUE, cvar = TRUE, dcor = TRUE, R = 0, seed = 1, nscan = 50000, burn = 500, odens = 25, plot = TRUE, print = TRUE)
```

Arguments:

- **Y**: an n x n square relational matrix
- **X**: an n x n x p array of covariates
- **rvar**: logical: fit row random effects?
- **cvar**: logical: fit column random effects?
- **dcor**: logical: fit a dyadic correlation?
SRM probit in amen

```r
fit_ame_ord<-ame(YV,XD,model="ord")
```
Summary:

```
summary(fit_ame_ord)

##
## beta:
## pmean psd z-stat p-val
## .dyad 0.162 0.028 5.72 0
##
## Sigma_ab pmean:
##   a   b
## a 0.731 0.125
## b 0.125 0.224
##
## rho pmean:
## -0.388
```
Posterior analysis

Confidence intervals

```r
apply(fit_ame_ord$BETA, 2, quantile, prob=c(.025,.5,.975))
##             dyad
## 2.5% 0.1138031
## 50% 0.1615003
## 97.5% 0.2231967

apply(fit_ame_ord$SABR, 2, quantile, prob=c(.025,.5,.975))
## va  cab  vb   rho   ve
## 2.5% 0.3728079 -0.06367146 0.1060518 -0.5601718 1
## 50% 0.6811580  0.12026448 0.2057822 -0.3863772 1
## 97.5% 1.4126096  0.37453941 0.4239469 -0.2343322 1
```

These results are very similar to those obtained from the binary probit analysis using the dichotomized data.
The role of covariance

Regression modeling:

\[ Y_{i,j} = \beta^T x_{i,j} + \epsilon_{ij} \]
\[ Y = \langle X, \beta \rangle + E \]

OLS estimation:

\[ \hat{\beta} = (X^T X)^{-1} X^T y \]

Precision of OLS estimators:

Let \( C = \text{Cov}[E] \)

\[ \text{Cov}[\hat{\beta}] = (X^T X)^{-1} X^T C X (X^T X)^{-1} \]
\[ = (X^T X)^{-1} \sigma^2 \text{ if } C = \sigma^2 I \]

For networks and relational data, typically \( C \neq \sigma^2 I \).

Accurate standard errors can’t be obtained unless we know/estimate \( \text{Cov}[E] \).
Justifying the SRM

The social relations covariance model can by a symmetry principle.

**Exchangeability:**

1. randomly sample $n$ individuals from a population;
2. observe $\epsilon_{i,j} =$ directed relation between $i$th and $j$th person.

Consider a probability model $\Pr(E)$ for the possible outcomes of $E = \{\epsilon_{i,j}\}$

$$
\begin{bmatrix}
\text{NA} & -0.94 & 0.15 \\
-0.7 & \text{NA} & 0.63 \\
0.63 & -0.42 & \text{NA}
\end{bmatrix}
$$

$$
\begin{bmatrix}
\text{NA} & 0.63 & -0.42 \\
0.15 & \text{NA} & -0.94 \\
0.63 & -0.7 & \text{NA}
\end{bmatrix}
$$

Note the second matrix is the same as the first with (1,2,3) relabeled as (2,3,1).
Let $\pi$ be some permutation of $\{1, \ldots, n\}$.

Let $E = \{\epsilon_{i,j} : i \neq j\}$

$E_{\pi} = \{\epsilon_{\pi_i, \pi_j} : i \neq j\}$

**Exchangeability:** A probability distribution $\Pr(E)$ is exchangeable if

$$\Pr(E) = \Pr(E_{\pi})$$

for all $E$ and permutations $\pi$.

Exchangeability can be justified by

- random sampling of nodes from a population;
- symmetry in the uncertainty in the $\epsilon_{i,j}$'s.
Justifying the SRM

Interesting result:
Suppose

1. $\epsilon_{i,j}$'s are normal and mean zero;
2. our probability for $E = \{\epsilon_{i,j}\}$ is exchangeable.

Then

$$\epsilon_{i,j} = a_i + b_j + e_{i,j}$$

$$\{(a_1, b_1), \ldots, (a_n, b_n)\} \sim \text{i.i.d. } N(0, \Sigma_{ab})$$

$$\{(e_{i,j}, e_{j,i}) : i \neq j\} \sim \text{i.i.d. } N(0, \Sigma_e)$$

for some $\Sigma_{ab}$ and $\Sigma_e$ (Li and Loken, 2002).

Interpretation:

normality + exchangeability $\Rightarrow$ SRM
Justifying the SRM

There is a stronger result than this: If \( E \) is exchangeable, then

\[
\epsilon_{i,j} = a_i + b_j + e_{i,j}
\]

\[
\text{Cov}[(a_i, b_i)] = \Sigma_{ab}
\]

\[
\text{Cov}[(e_{i,j}, e_{j,i})] = \Sigma_e
\]

for some \( \Sigma_{ab}, \Sigma_e \).

**Interpretation:**

exchangeability \( \Rightarrow \) SRM covariance structure

**Implication:**

\[
\text{Cov}[\hat{\beta}] = (X^T X)^{-1} X^T C X (X^T X)^{-1},
\]

where \( C = \text{Cov}[(\epsilon_{i,j}, \epsilon_{j,i})] \).

Under exchangability, the SRM provides appropriate SEs and CIs for \( \beta \).