The Stata Journal

Users of any of the software, ideas, data, or other materials published in the Stata Journal or the supporting files understand that such use is made without warranty of any kind, by either the Stata Journal, the author, or Stata Corporation. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the Stata Journal is to promote free communication among Stata users.

The *Stata Technical Journal* (ISSN 1536-867X) is a publication of Stata Press, and Stata is a registered trademark of Stata Corporation.

The Stata Journal publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those teaching, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

For more information on the Stata Journal, including information for authors, see the web page

http://www.stata-journal.com

Subscriptions are available from Stata Corporation, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates:

Subscriptions mailed to US and Canadian addresses:

3-year subscription (includes printed and electronic copy)	$153
2-year subscription (includes printed and electronic copy)	$110
1-year subscription (includes printed and electronic copy)	$ 59

Subscriptions mailed to other countries:

3-year subscription (includes printed and electronic copy)	$225
2-year subscription (includes printed and electronic copy)	$158
1-year subscription (includes printed and electronic copy)	$ 83
3-year subscription (electronic only)	$153

Back issues of the Stata Journal may be ordered online at

http://www.stata.com/bookstore/sj.html

The Stata Journal is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to The Stata Journal, Stata Corporation, 4905 Lakeway Drive, College Station TX 77845, USA, or email sj@stata.com.

The Stata Journal (2003)
3, Number 1, pp. 32–46

Intra-class correlation in random-effects models for binary data

Germán Rodríguez
Princeton University
grodri@princeton.edu

Irma Elo
The University of Pennsylvania
popelo@pop.upenn.edu

Abstract. We review the concept of intra-class correlation in random-effects models for binary outcomes as estimated by Stata's `xtprobit`, `xtlogit`, and `xtclog`. We consider the usual measures of correlation based on a latent variable formulation of these models and note corrections to the last two procedures. We also discuss alternative measures of association based on manifest variables or actual outcomes and introduce a new command `xtrho` for computing these measures for all three types of models.

Keywords: st0031, intra-class correlation, random-effects, probit, logit, complementary log-log, Pearson's r, Yule's Q

1 Introduction

Random-effects models are used in the analysis of clustered or longitudinal data, where the usual assumption of independence of the responses is not appropriate. The measurement of the extent to which the observations in a cluster or within an individual are correlated is often of interest. In this note, we discuss measures of intra-class correlation in random-effects models for binary outcomes.

We start with the classical definition of intra-class correlation for continuous data (Longford 1993, Chapter 2). We then consider the usual extensions to binary outcomes based on a latent-variable formulation of generalized linear models with binomial errors and link probit, logit, or complementary log-log (Fahrmeir and Tutz 1994). In this process, we note a couple of errors in Stata's `xtlogit` and `xtclog` as documented in version 7. (Both corrected in Stata 8.)

We also consider alternative measures of intra-class correlation based on manifest rather than latent variables. The possible outcomes for two observations on the same group or individual may be viewed as a two by two contingency table, and we consider measures of association based on Pearson's correlation coefficient and measures based on the odds-ratio such as Yule's Q coefficient. We describe the calculation of these measures for probit, logit, and complementary log-log models, using numerical integration procedures for the last two. Finally, we introduce a new command, `xtrho`, that can be used to compute these measures.

2 Linear models

Let Y_{ij} represent a continuous outcome for the jth observation in the ith group. The usual linear mixed-effects model estimated by `xtreg` assumes that

$$Y_{ij} = \mu + u_i + e_{ij} \tag{1}$$

where μ is a constant and $u_i \sim N(0, \sigma_u^2)$ independently of $e_{ij} \sim N(0, \sigma_e^2)$. In this model, $E(Y_{ij}) = \mu$, $\text{var}(Y_{ij}) = \sigma_u^2 + \sigma_e^2$, and $\text{cov}(Y_{ij}, Y_{ik}) = \sigma_u^2$ for two observations in the same group (and 0 otherwise). We could introduce covariates simply by writing $x'_{ij}\beta$ instead of μ, but we will keep things simple to focus on the covariance structure. The correlation between any two observations in the same group is, from the standard definition of Pearson's correlation coefficient (as the ratio of the covariance to the square root of the product of the variances),

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \tag{2}$$

Although ρ is defined in the conventional way, it turns out to represent the ratio of the variance of the random effect u_i to the total variance and thus can be interpreted as the proportion of variance explained by clustering. (Note that we have not squared ρ. If we did, we would obtain the proportion of variance of one response explained by another response in the same group, not the proportion explained by clustering.)

3 Probit models

In random-effects probit models as estimated by `xtprobit`, we assume that conditional on unobserved random effects u_i, the outcomes are realizations of independent Bernoulli random variables Y_{ij} with probabilities depending on u_i. Specifically, we assume that the conditional probability of a positive outcome given the random effect u_i is

$$\pi_{ij} = \Pr(Y_{ij} = 1 | u_i) = \Phi(\eta + u_i)$$

where Φ is the standard normal c.d.f. and η is a constant. (In the more general case with covariates, we would write $x'_{ij}\beta$ instead of η.) The inverse transformation Φ^{-1} is, of course, the probit leading to the model

$$\Phi^{-1}(\pi_{ij}) = \eta + u_i \tag{3}$$

Estimation of this model requires integrating out u_i to obtain the unconditional distribution of the outcomes Y_{ij}, which are of course correlated within groups.

3.1 Latent correlation in probit models

The probit model can be reformulated in terms of a continuous *latent variable* Y_{ij}^* such that the outcome is positive if, and only if, Y_{ij}^* is above a threshold. The latent variable follows the usual linear mixed-effects model in (1), but with mean η. The threshold can be taken to be zero with no loss of generality, as any other value can be absorbed into the constant. As is often the case with latent variables, it turns out that the scale of Y_{ij}^* is not identified either. To see this point, note that the conditional probability of a positive outcome is

$$\Pr(Y_{ij}^* > 0|u_i) = \Pr(e_{ij} > -\eta - u_i) = \Phi\{(\eta + u_i)/\sigma_e\}$$

and depends only on the ratio of η and u_i to σ_e. To maintain consistency with (3), we take $\sigma_e = 1$. This means that all coefficients are scaled in terms of the within-group standard deviation.

A nice feature of the latent variable formulation is that it allows us to compute the intra-class correlation using the same formula given in (2) for continuous outcomes, except that $\sigma_e^2 = 1$, so we now have

$$\rho_{\text{probit}} = \frac{\sigma_u^2}{\sigma_u^2 + 1} \tag{4}$$

This is the formula used in `xtprobit` and is correct as long as you realize that it refers to correlations in the *latent* scale.

The Stata 7 Reference Manual Volume 4 illustrates the use of several `xt` commands using data from a subsample of the National Longitudinal Survey of Youth (NLSY). The `union.dta` subset has union membership information from 1970–88 for 4,434 women aged 14–26 in 1968. An analysis using `xtprobit` yields an estimated intra-class correlation of 0.6367, which is reproduced in Table 1 for later reference. We interpret this result as indicating that unobserved individual characteristics (the u_i) account for 64% of a woman's propensity to belong to a union in different years (the latent variable Y_{ij}^*).

Table 1: Intra-class correlations for union membership data

Model	Manual	Revised
Probit	0.6367	0.6367
Logit	0.8417	0.6175
C-log-log	0.7611	0.6595

3.2 Manifest association in probit models

An alternative way to look at intra-class correlation is to focus on the actual dichotomous outcomes. We view the joint distribution of two observations Y_{ij} and Y_{ik} in the same group as a two-by-two contingency table. The cell probabilities can be computed in terms of two types of quantities: the marginal probability of a positive outcome, which for a probit model is

$$\pi_{\cdot 1} = \Pr(Y_{ij} = 1) = \Phi(\eta/\sqrt{1 + \sigma_u^2})$$

and the joint probability of two positive outcomes in the same group, which is

$$\pi_{11} = \Pr(Y_{ij} = 1, Y_{ik} = 1) = \Phi_2(\eta/\sqrt{1+\sigma_u^2}, \eta/\sqrt{1+\sigma_u^2}, \rho)$$

where Φ_2 denotes the standard bivariate normal distribution as computed by Stata's `binorm`. Note that these quantities depend on both σ_u^2 and η. For simplicity, we continue to take η as fixed, but we could easily introduce covariates by writing $\eta_{ij} = x'_{ij}\beta$ instead.

We are now in a position to compute any standard measure of association for dichotomous variables. For example, Pearson's correlation coefficient between two outcomes in the same group (and with the same marginal probability) is

$$r = \frac{\pi_{11} - \pi_{.1}^2}{\pi_{.1}(1-\pi_{.1})} \tag{5}$$

This formula follows directly from the general definition of Pearson's coefficient. (Note that we are using r and ρ to distinguish manifest and latent measures rather than sample and population quantities.) The dashed line on the left panel of Figure 1 plots latent ρ as a function of σ_u. The solid lines plot the values of manifest r corresponding to values of η chosen to produce conditional probabilities of 0.5, 0.2, 0.1, 0.05, 0.025, and 0.01 (from top to bottom) when the random effect is zero. Because the correlation is a symmetric function of η, these curves also represent conditional probabilities increasing from 0.5 to 0.99. The graph shows that the manifest correlation is always less than the latent correlation when both are measured using Pearson's coefficient, although both measures approach one as $\sigma_u \to \infty$ for any η. The difference increases with the magnitude of η for any fixed value of σ_u.

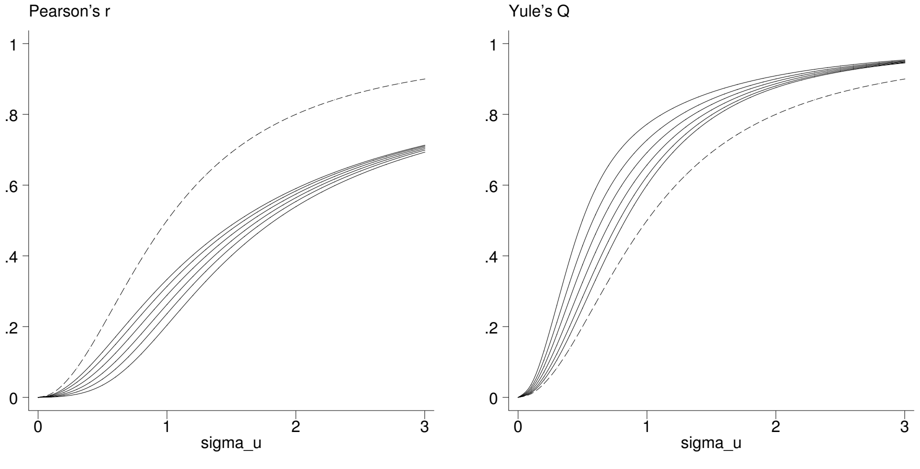

Figure 1: Latent and manifest intra-class correlation for probit models

As noted in the classic text by Bishop et al. (1975, Chapter 11), all the standard measures of association for a two-by-two table are essentially functions of Pearson's correlation coefficient or functions of the odds ratio, which in the present context can be written as

$$\alpha = \frac{\pi_{11}(1 - 2\pi_{.1} + \pi_{11})}{(\pi_{.1} - \pi_{11})^2}$$

Perhaps the best-known measure depending on the odds ratio is Yule's Q, defined as

$$Q = \frac{\alpha - 1}{\alpha + 1} \tag{6}$$

For two-by-two tables, Q coincides with Goodman and Kruskal (1959)'s γ and thus has a nice interpretation as the difference between the probabilities of like and unlike orderings of the responses of two randomly chosen pairs from different groups. The right panel of Figure 1 shows latent ρ using a dashed line and values of manifest Q using solid lines for the same values of η as the left panel. Yule's Q always exceeds the latent correlation, and the difference increases with $|\eta|$ for fixed σ_u.

Yule (1912) was troubled by the dependence of Q on the margins. He proposed standardizing the two-by-two table so that it would have 50:50 margins while preserving the odds ratio, and then taking the difference between the diagonal and off-diagonal probabilities as a measure of 'colligation', Y. This measure can be computed as

$$Y = \frac{\sqrt{\alpha} - 1}{\sqrt{\alpha} + 1} \tag{7}$$

and turns out to be equivalent to Pearson's r for the standardized table. In our context, Yule's Y always lies between Q and r and is thus closer to latent ρ, but still depends on η for fixed σ_u. Latent ρ itself, by the way, coincides with the tetrachoric correlation coefficient proposed by Pearson (1900).

The choice between latent and manifest measures is not obvious. The latent correlation has the advantage of not depending on the marginal distribution, while the manifest association has the advantage of referring more directly to observable quantities.

4 Logit models

A similar development applies to random-effects logit models as estimated by `xtlogit`. We assume that conditional on random effects u_i the observations have independent Bernoulli distributions with probabilities

$$\pi_{ij} = \Pr(Y_{ij} = 1 | u_i) = F(\eta + u_i)$$

where F is the standard logistic distribution with c.d.f. $F(\eta) = e^\eta/(1 + e^\eta)$. The inverse transformation F^{-1} is the logit, leading to the model

$$\text{logit}(\pi_{ij}) = \log \frac{\pi_{ij}}{1 - \pi_{ij}} = \eta + u_i \tag{8}$$

4.1 Latent correlation in logit models

This model also admits a latent variable formulation, except that this time the individual error terms e_{ij} are assumed to follow standard logistic rather than standard normal distributions. To be precise, we assume that $Y_{ij} = 1$ if, and only if, $Y^*_{ij} > 0$, just as before. We further assume that Y^*_{ij} follows the linear mixed model in (1) with mean η and $u_i \sim N(0, \sigma^2_u)$, but e_{ij} now has a logistic distribution with mean 0 and variance σ^2_e. Just as before, we note that the scale of the latent variable is not identified. For convenience, we take e_{ij} to have a *standard* logistic distribution, which happens to have variance $\pi^2/3$ or approximately 3.29. (We could, of course, set the variance to one for comparability with probit models, but then we would lose some convenience in calculation, not to mention the fact that coefficients in standard logit models have a nice interpretation in terms of log-odds ratios.)

In view of this development, we can compute the latent intra-class correlation using the same general formula as in linear models except that $\sigma^2_e = \pi^2/3$, so we now have

$$\rho_{\text{logit}} = \frac{\sigma^2_u}{\sigma^2_u + \pi^2/3} \tag{9}$$

The Stata 7 manual takes the individual variance as one, but the code has since been patched to use (9) instead. Note that using one instead of the logistic variance would lead to overestimation of ρ.

As shown in Table 1, the logit estimate of the intra-class correlation for the union data based on (9) is 0.6175 (rather than 0.8417 using a variance of one) and is very similar to the probit estimate.

4.2 Manifest association in logit models

We now consider the marginal and joint probabilities required to compute measures of manifest association. The marginal probability of a positive outcome given η and σ_u is

$$\pi_{.1} = \Pr\{Y_{ij} = 1\} = \int_{-\infty}^{+\infty} \frac{\exp(\eta + \sigma_u z)}{1 + \exp(\eta + \sigma_u z)} \phi(z) dz$$

where $\phi(z)$ denotes the standard normal density. The joint probability of two positive outcomes in the same group and with the same linear predictor η is

$$\pi_{11} = \Pr\{Y_{ij} = 1, Y_{ik} = 1\} = \int_{-\infty}^{+\infty} \left\{ \frac{\exp(\eta + \sigma_u z)}{1 + \exp(\eta + \sigma_u z)} \right\}^2 \phi(z) dz$$

Unfortunately, these integrals do not have a closed form and must be computed by numerical quadrature as explained further below. The dashed line in Figure 2 shows the intra-class correlation in the *latent* scale as a function of σ_u, computed using (9). It also shows the intra-class association in the *manifest* scale as a function of σ_u and η. We show Pearson's r on the left panel and Yule's Q on the right, computed using

(5) and (6), respectively. The horizontal scale on the graph was chosen to represent the same range of latent correlation as in the probit graph and can be made comparable dividing by $\pi/\sqrt{3}$, the same transformation used to make logit and probit coefficients comparable. The values of η were chosen to produce conditional probabilities of 0.5, 0.2, 0.1, 0.05, 0.025, and 0.01 when the random effect is zero, just as in Figure 1. We can see that, after appropriate scaling, the latent and marginal correlations for logit models behave very much like their probit counterparts.

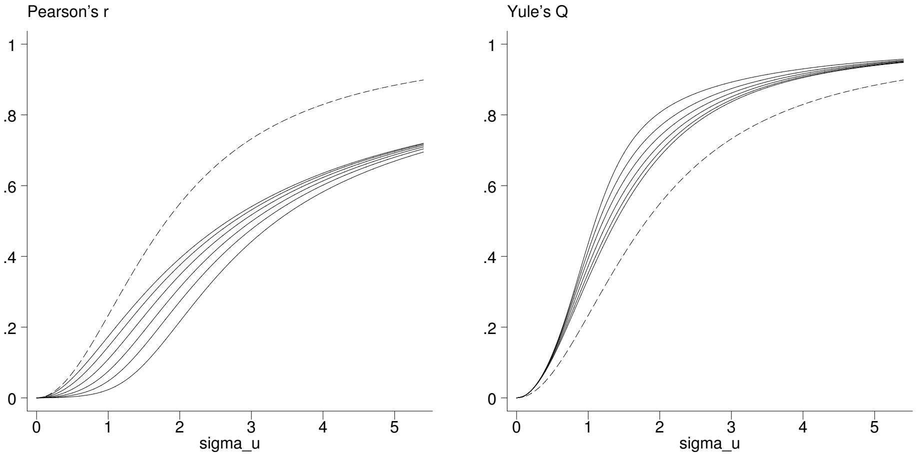

Figure 2: Latent and manifest intra-class correlation for logit models

The marginal and joint probabilities can be computed using Gaussian quadrature, which is the method used by Stata to fit the model. We found that accurate estimation of these probabilities over the entire range of interest shown in Figure 2 required a large number of quadrature points. We also tried adaptive Gaussian quadrature as described by Liu and Pierce (1994). This procedure rescales the evaluation points so that the integrand is sampled in a more appropriate region and produced accurate estimates over our range of interest using fewer quadrature points. The method we finally used, however, is based on Crouch and Spiegelman (1990) and relies on a trapezoid rule that combines simplicity with remarkable accuracy. The rule approximates

$$\int_{-\infty}^{+\infty} F(\eta+\sigma z)^k \phi(z) dz \approx \frac{1}{h} \sum_{-\infty}^{+\infty} F\{\eta+\sigma(z_0+ih)\}^k \phi(z_0+ih)$$

where h is the step size, F is the logistic c.d.f., k is one for marginal and two for joint probabilities, and z_0 is an arbitrary value chosen for convenience. We compute the sum starting at z_0 given by the mode of the integrand and proceed towards the tails until the terms became negligible, repeatedly halving the step size until the desired accuracy is attained. The computations can be organized so that each step-halving reuses all the terms calculated previously. Crouch and Spiegelman (1990) show how to precompute the step size required to attain a given accuracy for logit models; our simplified approach works just as well and can also be applied to the next family of models.

5 Complementary log-log models

A third approach to modeling binary data, implemented in `xtclog`, uses the complementary log-log link

$$\log\{-\log(1-\pi_{ij})\} = \eta + u_i \tag{10}$$

where η is a constant and $u_i \sim N(0, \sigma_u^2)$ as before. The inverse of this transformation is the c.d.f. of the extreme value (or log-Weibull) distribution,

$$F(\eta + u_i) = 1 - \exp\{-\exp(\eta + u_i)\}$$

The complementary log-log link can also be obtained from the general latent variable formulation if we assume that the individual error terms e_{ij} have (reverse) extreme value distributions with c.d.f. $F(e_{ij}) = \exp\{-\exp(-e_{ij})\}$. This distribution is asymmetric, with a long tail to the right. It has mean equal to Euler's constant (0.577) and variance $\pi^2/6$ or about 1.645 (Johnson et al. 1995, Chapter 22). Under these assumptions, the latent variable Y_{ij}^* has mean $\eta + 0.577$ and variance $\sigma_u^2 + \pi^2/6$. It follows that for this model, we should calculate the latent intra-class correlation as

$$\rho_{\text{clog}} = \frac{\sigma_u^2}{\sigma_u^2 + \pi^2/6} \tag{11}$$

The formula reported in the Stata 7 manual and used by `xtclog` takes the individual variance to be one and therefore tends to overestimate the intra-class correlation, although not as much as would be the case for `xtlogit`. In Stata 8, all correlations have been corrected.

As shown in Table 1, the complementary log-log estimate of the intra-class correlation for the union data based on (11) is 0.6595 (compared with 0.7611 using a variance of one). Note that the revised estimate is much closer to the estimates obtained using the probit and logit links.

Measures of manifest association based on Pearson's r or Yule's Q can be computed along the same lines as for logit models. The marginal probability of a positive outcome is

$$\pi_{.1} = \Pr(Y_{ij} = 1) = \int_{-\infty}^{+\infty} 1 - \exp\big\{-\exp(\eta + \sigma_u z)\big\}\phi(z)dz$$

where $\phi(z)$ denotes the standard normal density. The joint probability of two positive outcomes in the same group and with the same linear predictor η is

$$\pi_{11} = \Pr(Y_{ij} = 1, Y_{ik} = 1) = \int_{-\infty}^{+\infty} \Big[1 - \exp\big\{-\exp(\eta + \sigma_u z)\big\}\Big]^2 \phi(z)dz$$

These integrals do not have a closed form solution but can be approximated easily and accurately using the trapezoid rule described in the previous section.

Figure 3 shows the latent and manifest measures for complementary log-log models. The horizontal scale covers the same range of latent correlations as the previous graphs and can be made comparable to the probit graph by dividing by $\pi/\sqrt{6}$, the same transformation used to make complementary log-log and probit coefficients comparable. The values of η were chosen to produce conditional probabilities of 0.01, 0.025, 0.05, 0.10, 0.20, 0.50, 0.80, 0.90, 0.95, 0.975 and 0.99 when the random effect is zero. Unlike the probit and logit links, the complementary log-log link is not a symmetric function of η, and this produces some interesting results.

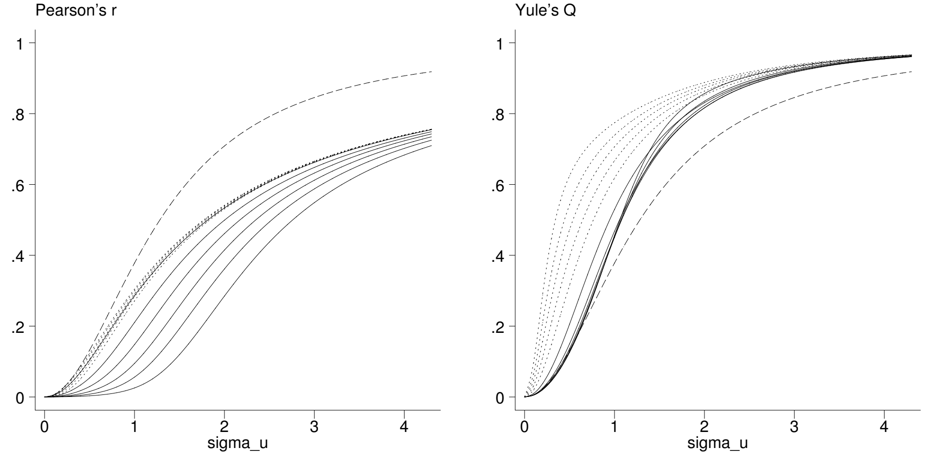

Figure 3: Latent and manifest intra-class correlation for c-log-log models

Figure 3 uses solid lines for values of $\eta \leq -0.3665$ corresponding to conditional probabilities below 0.5 and dotted lines for the rest. We see that Pearson's r varies substantially with η when the conditional probability is below one half (solid lines), with lower values for more negative η's, but doesn't vary much when the outcome is more likely (dotted lines). Yule's Q exhibits the oppositive behavior. This time, the solid lines corresponding to less frequent outcomes are tightly clustered, while the dotted lines corresponding to higher conditional probabilities fan out considerably, with Q increasing with η for fixed σ_u.

6 Inference about intra-class correlation

It may be useful to remind the reader that testing the hypothesis $H_0 : \sigma_u^2 = 0$, which is equivalent to the hypothesis $H_0 : \rho = 0$ of no intra-class correlation for any of the measures discussed here, requires special care because the postulated value lies on a boundary of the parameter space. In this case, the likelihood-ratio test does not have the usual χ^2 distribution with one degree of freedom but may be better approximated as a 50:50 mixture of χ^2s with zero and one degree of freedom, the approximation used by Stata; see Stram and Lee (1994) and Gutierrez et al. (2001). However, Pinheiro and Bates (2000, Section 2.4.1) note from simulations that this adjustment is not always successful. As an alternative, one can use the nominal one-degree-of-freedom test and treat the resulting p-value as a conservative approximation.

By the same token, the normal approximation to the distribution of the estimator of latent ρ may not be adequate, particularly for small amounts of clustering. Stata reports standard errors for σ_u and latent ρ but wisely computes confidence intervals working with $\log \sigma_u^2$, for which a normal approximation should be more reasonable. A similar approach can be used to compute confidence intervals for the measures of manifest association introduced here. Note, however, that by construction these intervals could never include the value $\rho = 0$ (or equivalently $\sigma_u^2 = 0$), so they should not be used in lieu of a test of significance.

In the union membership example, the 95% confidence intervals for $\log \sigma_u^2$ lead to 95% confidence intervals for latent ρ of (0.5974, 0.6372), using (9) for `xtlogit` and (0.6420, 0.6766) using (11) for `xtclog`.

7 Stata commands: xtrho and xtrhoi

We have written two commands that can be used to compute the marginal and joint probabilities of a positive outcome and three measures of intra-class manifest association: the odds ratio, Pearson's r and Yule's Q. You can obtain these commands while you are online by typing into Stata `net from http://opr.princeton.edu/stata`.

7.1 Syntax for xtrho

`xtrho` is a post-estimation command that can be used following a random effects `xtprobit`, `xtlogit`, or `xtclog`. The syntax is

```
xtrho [ , level(#) detail ]
```

By default, the calculations are done with η set to the median in the estimation sample and σ_u set to its maximum likelihood estimate. Confidence intervals are computed by setting σ_u to its lower and upper confidence bounds while keeping η at the median.

7.2 Options for xtrho

`level(#)` specifies the confidence level, in percent, for confidence intervals. The default is `level(95)` or as set by `set level`; see [R] **level**.

`detail` computes the measures with the linear predictor η set to percentiles 1, 25, 50, 75, and 99 in the estimation sample, and is useful to ascertain how the measures of association vary with observed characteristics.

7.3 Syntax for xtrhoi

`xtrhoi` is an immediate command that can be used to compute the measures for given values of η and σ_u for any of the three link functions. The syntax is `xtrhoi` η σ_u $\left[\ \textit{link}\ \right]$

where η and σ_u must be numbers and *link* must be one of `logit`, `probit`, or `clog`, with `logit` as the default.

7.4 Saved Results

The post-estimation and immediate commands are both r-class and save in `r()`:

Scalars

`r(mp)`	marginal probability	`r(r)`	Pearson's r
`r(jp)`	joint probability	`r(Q)`	Yule's Q
`r(or)`	odds ratio		

For `xtrho`, the saved quantities are the estimates at the median η.

8 Application to the union data

We illustrate these ideas by computing and interpreting measures of latent correlation and manifest association for the union data.

8.1 Latent propensity to unionize

We start by first fitting a random-effects logit model to the union data. This is the same model used in [R] **xtlogit** ([XT] for Stata 8).

```
. xtlogit union age grade not_smsa south southXt, i(id)
  (output omitted)
Random-effects logit                            Number of obs      =     26200
Group variable (i) : idcode                     Number of groups   =      4434

Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       5.9
                                                               max =        12

                                                Wald chi2(5)       =    221.95
Log likelihood  = -10556.294                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0092401   .0044368     2.08   0.037     .0005441    .0179361
       grade |   .0840066   .0181622     4.63   0.000     .0484094    .1196038
    not_smsa |  -.2574574   .0844771    -3.05   0.002    -.4230294   -.0918854
       south |  -1.152854   .1108294   -10.40   0.000    -1.370075   -.9356323
     southXt |   .0237933   .0078548     3.03   0.002     .0083982    .0391884
       _cons |   -3.25016   .2622898   -12.39   0.000    -3.764238   -2.736081
-------------+----------------------------------------------------------------
    /lnsig2u |   1.669888   .0430016                      1.585607     1.75417
-------------+----------------------------------------------------------------
     sigma_u |   2.304685   .0495526                      2.209582    2.403882
         rho |   .6175213   .0030872                      .5974278    .6372209
------------------------------------------------------------------------------
Likelihood ratio test of rho=0: chibar2(01) =  5978.89 Prob >= chibar2 = 0.000
```

The intra-class latent correlation ρ for this model is 0.6175, indicating a high correlation between a woman's propensity to be a union member in different years, after controlling for her education and residence. The estimate of σ_u can be interpreted as an ordinary logit coefficient by writing the random effect $u_{ij} \sim N(0, \sigma_u^2)$ as $\sigma_u z_{ij}$, where $z_{ij} \sim N(0,1)$. In this formulation, there is a parallel between the covariates x_{ij}, representing observed characteristics with coefficients β, and the standardized random effects z_{ij}, representing unobserved traits with coefficient σ_u. In the union data, the odds of belonging to a union in a given year for a woman who has unobserved propensity one standard deviation above the mean are about ten times the corresponding odds for a woman with average unobserved propensity and the same observed characteristics ($\exp(2.305) = 10.02$).

8.2 Manifest union membership

Next, we use `xtrho` to translate the results into quantities pertaining to observable outcomes.

```
. xtrho
Measures of intra-class manifest association in random-effects logit
Evaluated at median linear predictor
Measure          |   Estimate     [95% Conf.Interval]
-----------------+-------------------------------------
Marginal prob.   |     .22696      .22084     .233181
Joint prob.      |    .123255     .116043     .130688
Odds ratio       |    7.67092     7.12563     8.26475
Pearson's r      |    .408917     .390966     .426798
Yule's Q         |    .769344     .753865     .784128
```

For a woman whose observed propensity is at the sample median, the marginal probability of belonging to a union in any given year is 0.227. The joint probability of belonging to a union in two given years is 0.123. From these quantities, we can compute various measures of association pertaining to union membership in any two given years, say, 1975 and 1980 to fix ideas. Consider first the odds ratio of 7.67. This means that the odds of being a union member in 1980 for a woman who was a member in 1975 are nearly eight times the corresponding odds for a woman with the same observed characteristics who was not a member in 1975. Note the difference in interpretation between the two types of odds ratios discussed so far. The odds ratio of ten described earlier contrasts behavior in a given year for women with different unobserved propensities. The odds ratio of almost eight described here contrasts behavior in a given year for women with different observed behaviors in another year.

Pearson's correlation coefficient is 0.409, indicating much lower manifest than latent association. Squaring this coefficient, we see that union membership in a given year explains only about 17% of the variation in union behavior in another year. In contrast, persistent unobserved traits explain 62% of the latent propensity to belong to a union in a given year.

Finally, we come to Yule's Q, which is estimated as 0.769 when the linear predictor is set to the median. This means that if we picked at random two women with median observed characteristics as summarized by the linear predictor, the probability that their union memberships in two given years would be concordant exceeds the probability that they would be discordant by 77 percentage points. A pair is considered concordant if one of the women is a union member in both years and the other isn't. A pair is discordant if one woman becomes a union member while the other discontinues membership. Other combinations do not enter in the calculation.

The default output also shows a confidence interval for each measure. In our example, σ_u is estimated quite precisely, and consequently, all intervals are reasonably narrow. Recall, however, that these measures depend on observed characteristics via the linear predictor. We next use the `detail` option to explore how much these measures vary across the sample.

```
. xtrho, detail
Measures of intra-class manifest association in random-effects logit
Evaluated with linear predictor set at selected percentiles
Measure          |           p1         p25         p50         p75         p99
-----------------+-------------------------------------------------------------
Marginal prob.   |      .107702      .16166      .22696     .253184     .311292
Joint prob.      |      .045003      .07794     .123255     .142897     .189065
Odds ratio       |      9.49691     8.39124     7.67092     7.47801     7.16908
Pearson's r      |      .347578     .382257     .408917     .416721     .429884
Yule's Q         |      .809468     .787036     .769344     .764096     .755174
```

We see that the marginal probability of belonging to a union in a given year ranges from 0.108 to 0.311 as we move from the first to the 99th percentile in terms of observed propensity to belong to a union. As we would expect from the general results shown before, this variation affects Pearson's r and Yule's Q in opposite ways. Pearson's r is higher among women who are most likely to unionize, whereas Yule's Q, and the odds-ratio on which it is based, are higher among women least likely to belong to a union. Among women who have on average a thirty percent probability of belonging to a union, membership in a given year is associated with a seven-fold increase in the odds of being a union member in another year. But among women who have only a ten percent probability of unionizing, membership in one year is associated with almost a ten-fold increase in the odds of union membership in another year.

The immediate command `xtrhoi` can be used to explore these quantities for other values of σ_u or the linear predictor η.

8.3 A note on quadrature checks

We should note in closing that the model fitted here using 12 quadrature points (the default setting) does not quite pass the checks in `quadchk`, which refits the model using 8 and 16 points. Things look better if we use 30 points, the maximum that Stata allows. While some of the coefficients change, the general tenor of the results and conclusions remains the same, so we decided to keep the defaults to maintain consistency with the Stata manual.

9 Acknowledgment

The authors gratefully acknowledge support for this research from the National Institute of Child Health and Human Development grants R01 HD35277 and R01 HD36462.

10 References

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Crouch, E. A. C. and D. Spiegelman. 1990. The evaluation of integrals of the form $\int_{-\infty}^{+\infty} f(t)\exp(-t^2)dt$: application to logistic-normal models. *Journal of the American Statistical Association* 85: 464–469.

Fahrmeir, L. and G. Tutz. 1994. *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer.

Goodman, L. A. and W. H. Kruskal. 1959. Measures of association for cross-classifications. II: further discussion and references. *Journal of the American Statistical Association* 54: 123–163.

Gutierrez, R. G., S. Carter, and D. M. Drukker. 2001. sg160: On boundary-value likelihood-ratio tests. *Stata Technical Bulletin* 60: 15–18. In *Stata Technical Bulletin Reprints*, vol. 10, 269–273. College Station, TX: Stata Press.

Johnson, N. L., S. Kotz, and N. Balakrishnan. 1995. *Continuous Univariate Distributions*, vol. 2. 2d ed. New York: John Wiley & Sons.

Liu, Q. and D. A. Pierce. 1994. A note on gauss–hermite quadrature. *Biometrika* 81(3): 624–629.

Longford, N. 1993. *Random Coefficient Models*. London: Clarendon Press.

Pearson, K. 1900. Mathematical contribution to the theory of evolution. VII: on the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, Series A* 195: 1–7.

Pinheiro, J. C. and D. M. Bates. 2000. *Mixed-Effects Models in S and S-Plus*. New York: Springer.

Stram, D. O. and J. W. Lee. 1994. Variance components testing in the longitudinal mixed-effects model. *Biometrics* 50: 1171–1177.

Yule, G. U. 1912. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society* 75: 579–652.

About the Authors

Germán Rodríguez is Senior Research Demographer at Princeton University, where he teaches generalized linear models, does research on statistical models for demographic data with emphasis on multilevel models for discrete outcomes, and oversees demographic computing.

Irma Elo is Assistant Professor of Sociology at the University of Pennsylvania, where she works on racial and neighborhood disparities in infant health and other issues related to racial/ethnic and socioeconomic differences in mortality.