Paper 270-28

Estimation of prevalence ratios when PROC GENMOD does not converge

James A. Deddens, NIOSH & University of Cincinnati, Cincinnati, OH
Martin R. Petersen, NIOSH, Cincinnati, OH
Xiudong Lei, University of California, Berkeley, CA

ABSTRACT

When studying a prevalent outcome, it is often of interest to estimate the prevalence ratio instead of the odds ratio. In SAS one can use PROC GENMOD with the binomial distribution and the log link function. Unlike the logistic model, the log-binomial model places restrictions on the parameter space, and the maximum likelihood estimate (MLE) might occur on the boundary of the parameter space, in which case PROC GENMOD will not converge to the correct estimate. We propose a method that uses PROC GENMOD to correctly estimate the MLE. The method consists of expanding the original data set to include a large number of copies of the original data set together with one copy of the original data set with cases and controls reversed. The estimated standard error of the prevalence ratio on the expanded data set is then "adjusted" to obtain the correct estimate of the standard error of the prevalence ratio. We provide a SAS MACRO to implement our new method. In addition we present an exact method for the one independent variable setting. We also provide a SAS MACRO to implement this exact method. The new approximation method yielded estimates which were close to the exact maximum likelihood estimates and to the true parameters. By comparison, the Cox proportional hazard approach did not perform nearly as well as the new method. The exact method can be used easily with single independent variable models, while the approximation method can be used with either single or multiple independent variable models.

INTRODUCTION

Recently there has been much discussion and interest in the literature concerning the appropriateness of estimating prevalence ratios versus odds ratios in cross sectional studies (Axelson et al. (1994), Lee (1994,1995), Lee & Chia (1993,1995), Ma & Wong (1999), McNutt et al. (1999), Skov et al. (1998), Stromberg (1994,1995), Thompson et al. (1998), Yu & Zhang (1999), Zhang & Yu (1998)). When the mean duration of the disease is known, the odds ratio may be preferable because it can be used to estimate the incidence rate ratio (Stromberg (1994)). However, the prevalence ratio is often more interpretable than the odds ratio (Axelson et al. (1994)). The odds ratio can be used to approximate the prevalence ratio, but the approximation is good only if the prevalence is low. Lee (1994) and Lee & Chia (1993) recommended using the Cox proportional hazard model to estimate the prevalence ratio. Skov et al.,(1998) recommended using the log-binomial model, which directly models the prevalence ratio, and showed that for dichotomous variables this method was better than the Cox proportional hazard method.

We define the prevalence ratio (PR) as: $P(Y_i=1|X_j+1)/P(Y_i=1|X_j)$, where Y_i is a 0/1 variable with $Y_i = 1$ indicating the outcome of interest, and X_j is a covariate of interest. (If there are other independent variables in the model, then the PR is adjusted for those variables.) For example, $Y_i = 1$ might indicate that the person has back pain, and X_j might be a measure of work stress.

The PR for X_j can be estimated (partial likelihood estimate) from the Cox proportional hazard model as $\exp(\hat{\beta}_1)$, adjusted for the other terms in the model, by using a constant time and treating $Y_i = 0$ as censored. The model may be written as

$$h(t|\mathbf{X}_i) = h_0(t)\exp(\beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}).$$

The beta's ($\beta_1, \beta_2, \ldots, \beta_k$) are unknown parameters, $h(t|\mathbf{X}_i)$ is a hazard function for the i^{th} vector of independent variables ($\mathbf{X}_i$), $h_0(t)$ is the baseline hazard, X_{1i} is the covariate of interest evaluated in the i^{th} observation, and the other X_{ji}'s are other covariates evaluated in the i^{th} observation. The proportional hazard model can be fit with the following code:

```
PROC PHREG;
MODEL TIME*Y(0)=X/TIES=BRESLOW;,
```

where TIME = 2 - Y.

The log-binomial model can be written as:

$$P(Y_i=1|\mathbf{X}_i) = \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}),$$

where for each vector $\mathbf{X}_i$, Y_i has a binomial distribution. The prevalence ratio (PR) for X_1 is then estimated (maximum likelihood estimate) as $\exp(\hat{\beta}_1)$, adjusted for the other terms in the model. The log-binomial model can be fit with the following code:

```
PROC GENMOD;
MODEL Y=X=BIN LINK=LOG INTERCEPT=-4;
```

(The start value of -4 for the intercept has worked well in practice.) It turns out, however, that for many situations with quantitative covariates, the maximum likelihood estimate (MLE) for the log-binomial model is on the boundary of the parameter space, which means that the predicted prevalence

$$\hat{P}(Y_i=1|\mathbf{X}_i) = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki})$$

will be equal to 1.

PROC GENMOD, which maximizes the likelihood by finding the point at which the derivative is equal to zero, is unable to find an estimate in such a situation because the iterative procedure does not converge. With one independent variable, the problem usually occurs when $Y_i = 1$ for every observation for which X is at its maximum or when $Y_i = 1$ for every observation for which X is at its minimum. When more than one independent variable is involved, the problem can also occur when $Y_i = 1$ for every observation involving certain combinations of the levels of independent variables.

In this article we give an example illustrating the problem. We also propose a new method that is easily implemented using PROC GENMOD, and which will yield an approximate maximum likelihood solution. Finally we present an exact method that is also easily implemented using PROC GENMOD, although it is more difficult to implement in the multiple independent variable setting.

ILLUSTRATIVE EXAMPLE

Consider the following simple example with a quantitative covariate:

X	1	2	3	4	5	6	7	8	9	10
Y	0	0	0	0	1	0	1	1	1	1

For the log-binomial model, the parameter space is:

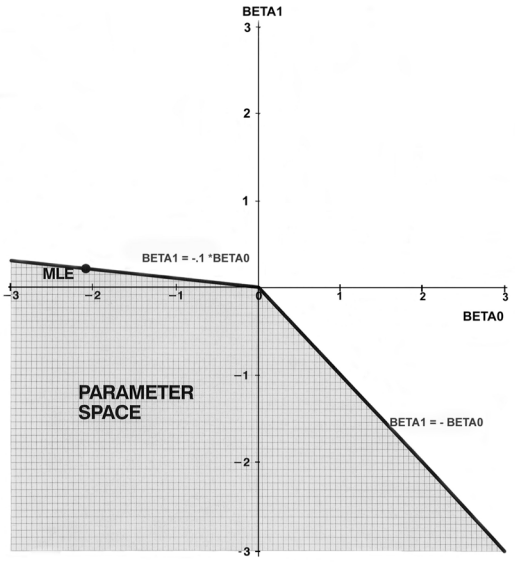
$\{ (\beta_0 \ \beta_1)$: $0 \leq \exp(\beta_0 + \beta_1 X) \leq 1$ for X = 1,2,...,10$\}$

or

$\{ (\beta_0 \ \beta_1)$: $\beta_0 + \beta_1 X \leq 0$ for X = 1,2,...,10 $\}$

Figure 1 illustrates this region. Note that this region has a boundary consisting of 2 straight-line segments.

FIGURE 1
PARAMATER SPACE FOR ILLUSTRATIVE EXAMPLE



In order to find the maximum of the likelihood function, one needs to check for interior points where the derivative is 0 and then also check all values on the boundary. In this example the derivative is never 0 inside the region and hence the maximum (if it exists) occurs somewhere on the boundary. In fact the maximum occurs at $\hat{\beta}_0 = -2.094$, $\hat{\beta}_1 = 0.2094$. Standard software packages will be unable to find this maximum likelihood estimate. For example, PROC GENMOD (using the default start values) will report $\hat{\beta}_0 = -0.827$, $\hat{\beta}_1 = 0.0827$ (which is a point on the boundary), but SAS version 8.1 will also give a warning that the procedure did not converge.

METHODS

We now propose a new method (COPY method) that involves modifying the data so that the MLE in the modified data approximates the MLE from the original data. The MLE in the modified data set is always inside the interior of the parameter space, and so this approximate MLE can be found using PROC GENMOD. More specifically, we propose that one create a new data set that contains c-1 copies of the original data and 1 copy of the original data with the Y_i values interchanged (1's changed to 0's and 0's changed to 1's). For example, one could use c = 1000 and have 999 copies of the original data and 1 copy of the original data with all Y-values interchanged. In order to account for the increased sample size, one then multiplies the estimated standard error, $\hat{SE}$, of $\hat{\beta}_1$ by the square root of c (for this example, the square root of 1,000).

By introducing this augmented data, one hopes to obtain a new maximum likelihood estimate which is approximately equal to the correct estimate, but which is not on the boundary. The goal is to choose a number of copies (for example, 1,000) which is large enough so that the new estimate is sufficiently close to the MLE, but not so many copies that the new estimate is also so close to the boundary that the software cannot estimate it. We did not experience this latter problem with PROC GENMOD in SAS version 8.1. However, occasionally PROC GENMOD did not converge when the number of copies was too small.

In theory, an exact MLE can be found by noting that if the log-binomial model fails to converge because the observed prevalence = 1 and the maximum likelihood predicted prevalence = 1 for some observation $\mathbf{X}_m$, then the search for maximum likelihood estimators can be restricted to estimators which yield a predicted prevalence of 1 when $\mathbf{X}_i = \mathbf{X}_m$. The likelihood can be rewritten and PROC GENMOD can be used to obtain the maximum likelihood solutions. The formal theorem and proof can be obtained from the authors.

The problem with using the theorem is that it only applies when the maximum likelihood estimated prevalence is 1 for some observed vector $\mathbf{X}_m$, but the maximum likelihood estimates are unknown (before the theorem is applied). The fact that the model fails to converge is a good indication that the solution is on the boundary (and thus the estimated prevalence is 1), although bad start values can also be the problem. In order to find the levels of the independent variables for which this occurs, grouping and plotting the data should be done. This works best for one dimensional graphs. For multidimensional data, the COPY method may be essential. For log-binomial regression on one independent variable, if the estimated prevalence is 1, it will occur at the largest or smallest X_i. Thus the exact method can be used when PROC GENMOD fails to converge, if one can determine which maximum likelihood estimated prevalence is equal to 1. In fact, one can repeat the rewriting of the likelihood and obtain a solution when more than one estimated prevalence is 1, such as with a U-shaped quadratic relationship between $\log[P(Y_i = 1|X)]$ and X. When one is fitting many models, however, it is easier to use the COPY method, which also worked well for simple models in our simulations.

We will describe several simulation studies that show the appropriateness of the COPY method. We will study the effects of prevalence, slope (β_1), and sample size (n). In addition, we will examine choices for the number of copies (c).

SIMULATIONS

Simulations were performed for the situation of one continuous covariate. Data were generated according to the log-binomial model, with X uniformly distributed from 0 to 10. The prevalence at X = 5 varied among 0.1, 0.3, 0.5, 0.7, and 0.9. Three values were chosen for β_1, namely 0, medium, and large, where medium and large depended on the prevalence. The intercept, β_0, was then determined from the prevalence at X = 5 and the slope, β_1. Thus, there were 15 basic simulations, and the sample size was set at n = 100. Selected simulations were repeated with n = 1,000. In the first set of simulations, we used two versions of the COPY method, one with 100 copies (COPY100) and one with 1,000 copies (COPY1000). We also used PROC GENMOD to compute the results of the log-binomial model and SAS PROC PHREG to compute the results of the Cox proportional hazard model for the unmodified data. Each simulation involved 1,000 replications of the data (same X's, different Y's). All hypothesis

tests were considered significant if the Wald p-value was less than or equal to 0.05.

We also did some limited simulations (namely 6) for the 2 independent variable case: {1. n = 100, prevalence = 0.9, $\beta_1 = \beta_2 = 0.01$}, {2. n = 100, prevalence = 0.5, $\beta_1 = \beta_2 = 0$}, {3. n = 900, prevalence = 0.9, $\beta_1 = \beta_2 = 0.01$}, {4. n = 900, prevalence = 0.9, $\beta_1 = \beta_2 = 0$}, {5. n = 900, prevalence = 0.5, $\beta_1 = \beta_2 = 0.05$}, and {6. n = 900, prevalence = 0.5, $\beta_1 = \beta_2 = 0$}. The independent variables took on integer values between 1 and the square root of n, inclusive..

RESULTS

ILLUSTRATIVE EXAMPLE REVISITED

Consider again the illustrative example with a quantitative covariate: Using the exact method, it can be easily shown that for the log-binomial model, the maximum likelihood estimates are $\hat{\beta}_0 = -2.0936$ and $\hat{\beta}_1 = 0.2094$ with estimated standard errors 1.0208 and 0.1021, respectively. PROC GENMOD fails to converge for this example, although it does provide an answer, which varies depending on the start values used. The results of using the Cox proportional hazards method and our COPY method (with 1,000 copies) are displayed in table 1. Note that the Cox proportional hazard method does not estimate an intercept. Our COPY method approximates the exact maximum likelihood estimates to within 2 or 3 decimal places, while the Cox proportional hazard method estimate is more than 50 percent too large. The $\hat{SE}$ of $\hat{\beta}_1$ using the Cox proportional hazard method is almost twice as large as the estimated standard error from the exact method, while the COPY method is correct to 3 decimal places. Finally (not shown), the COPY method results in the correct conclusion of a relationship (p = 0.0403, Wald test), while the Cox proportional hazard method does not (p = 0.0978). (For the exact maximum likelihood, p=0.0403.)

SIMULATION RESULTS

Table 2 shows the percent of times PROC GENMOD (on the unmodified data) fails to converge for the 15 simulations with n = 100. Convergence was determined automatically by SAS, Version 8.1. This version is more reliable than previous versions in notifying the user that the estimates did not converge. PROC GENMOD tended to fail when either the prevalence or the slope was high. PROC GENMOD failed about half the time at the combination of both extremes (prevalence = 0.9 and high slope = 0.02).

Although one could always use one of the COPYc methods (COPY method with a total of c copies), it is better to use PROC GENMOD when it converges and one of the COPYc methods when PROC GENMOD does not converge. Using this approach, the average slope and standard error, over all simulations, was calculated and is shown in table 3. When PROC GENMOD converges, GEN+COPY100 and GEN+COPY1000 are obviously the same. For all simulations, the average slopes were close to the true slopes (from table 2). Interestingly, the COPY100 method usually yielded estimates that were slightly closer to the true values than the COPY1000 method, although this may be due to chance or the simulations chosen. The proportional hazards approach is about as good as this method with respect to estimating the slope. However the standard errors using the proportional hazards approach are too large especially for large prevalence.

Table 4 shows the size (α) and power for GEN+COPY100, GEN+COPY1000, and PHREG for the 1,000 replications with n = 100. The size is defined as the proportion of times the procedure rejected H_0: β_1 = true slope, and the power is similarly defined for H_0: $\beta_1 = 0$. (The size and power are the same when the true slope = 0.) For the six simulations for which PROC GENMOD always converged (table 2), the size and power for GEN+COPY100 and GEN+COPY1000 were naturally the same. For other simulations, GEN+COPY100 generally had the same or smaller power than GEN+COPY1000.

It can also be seen in table 4 that the proportional hazards approach (PROC PHREG) has smaller size and power than any of the other methods. For prevalence = 0.7 and prevalence = 0.9 the results for PROC PHREG were very bad.

There were three simulated data sets for which COPY100 did not converge. Although the largest predicted prevalence was never 1, all were close to 1. For COPY1000, they were even closer to 1, but the increased sample size (for the modified data) allowed COPY1000 to converge. For one of the 3 data sets, COPY10000 was run, and it also converged even though its largest predicted prevalence was even closer to 1.

Simulations were also done with a sample size of 1,000 for 3 situations: prevalence = 0.1 and slope = 0, prevalence = 0.9 and slope = 0, and prevalence = 0.9 and slope = 0.02. The results were similar to before. However, when the slope = 0.02 and prevalence = 0.9, PROC GENMOD failed to converge for 40.8 percent of the simulations. Again using PROC GENMOD when it converged and the COPY method when PROC GENMOD failed to converge, the powers increased, relative to n = 100, and the sizes decreased to 0.024 and 0.033 for COPY100 and COPY1000, respectively.

The above simulations compare the COPY method to true population values. However, the COPY method estimates are really estimating the exact maximum likelihood estimates (which in turn estimate the true population values). In order to compare COPY100 with COPY1000, 1,000 data sets were simulated as above for a prevalence of 0.9, a slope of 0.02, and a maximum likelihood solution on the boundary. The results are given in table 5. The COPY1000 method is correct or only slightly off in the third decimal place on average, while COPY100 is off a bit more, especially for the intercept and its estimated standard error. According to 95% normality based prediction intervals on differences between the exact method and the COPY method (not shown), the intercept difference for COPY100 should be between -0.022 and -0.006, and the slope difference should be between -0.001 and 0.003 95 percent of the time. Similarly, the intercept difference for COPY1000 should be between -0.002 and -0.001, and the slope difference should be 0.000 (to 3 decimal places) 95 percent of the time.

The simulations with 2 independent variables were consistent with the results shown above for one independent variable.

DISCUSSION

When estimating prevalence ratios with the log-binomial model, Proc Genmod may not converge due to the estimate being on the boundary of the parameter space. This is most likely to happen when the prevalence is high and the prevalence ratio is large. When Proc Genmod fails to converge, the COPY method can provide excellent approximate estimates and standard errors. With 1000 copies, the estimates in our simulations were very close to the true MLEs for our sample of size 100. For datasets with more observations and more variables, 1000 copies can create a very large dataset. If the dataset is still large after removing unneeded variables, one can use as few as 100 copies and still get good estimates (table 5).

The iterative procedure in PROC GENMOD requires selecting starting values. Our experience indicates that often SAS selects starting values that are inappropriate for the log-binomial model. Therefore we strongly suggest that the user select initial values for the iterative procedure using the options available in SAS.

For example, we routinely use INTERCEPT = -4. This selects -4 as the starting value for β_0, and 0 as the starting values for the other β_i's. We also found that convergence was easier to obtain when the Y's were 0 or 1, rather than frequencies, although PROC GENMOD can do both.

When developing a model for the prevalence ratio the only assumptions are that the observations are independent and that the model is the correct form. With only one independent variable, the use of the prevalence ratio instead of the odds ratio requires that the relationship be "concave up", since $\exp(\beta_0+\beta_1X)$ is a concave up function of X, whereas the logistic function, $\exp(\beta_0+\beta_1X)/[1 + \exp(\beta_0+\beta_1X)]$, can be either concave up or concave down. Thus, with prevalence ratio modeling, if the relationship between the prevalence of Y and X is concave down, a transformation of X must be made, or a quadratic term must be included in the model.

CONCLUSIONS

In this article we have shown that PROC GENMOD may fail to converge when estimating the prevalence ratio using the log-binomial mode. In this situation we have shown that the proposed COPY method can be used to estimate the prevalence ratio when the MLE is on the boundary of the parameter space. It yielded a good approximation to the exact maximum likelihood estimates, as well as yielding good estimates of the true population parameters in our simulations. The exact maximum likelihood estimates can also be found if one has the time and the ability to find the $\mathbf{X}_m$ for which the estimated prevalence is 1. Without using these methods when the MLE is on the boundary of the parameter space, PROC GENMOD will not obtain the correct maximum likelihood estimates. Finally we have clearly shown that the use of the Cox proportional hazards model to estimate the prevalence ratio results in inflated standard errors, and, when the solution is on the boundary, incorrect estimates as well.

When estimating the prevalence ratio using the log-binomial model, we recommend first using PROC GENMOD, and if it fails to converge, then using the COPY1000 method.

The authors have written 2 SAS macros which will perform some of the methods suggested in this article. Readers can obtain copies from the NIOSH web page at http://www.cdc.gov/niosh/ext-supp-mat/pr-sasmac/ or from the authors at jdeddens@cdc.gov or mpetersen@cdc.gov.

REFERENCES

Axelson O, Fredriksson M, and Ekberg K.,(1994), Use of the prevalence ratio v the prevalence odds ratio as a measure of risk in cross sectional studies (Correspondence), *Occupational Environmental Med*icine,51:574

Lee J.,(1994), Odds ratio or relative risk for cross-sectional data? *International Journal of Epidemiology*,23:201-3.

Lee J.,(1995), Estimation of prevalence rate ratios from cross-sectional data: a reply, *International Journal of Epidemiology*,24:1066-7.

Lee J, and Chia KS.,(1993), Estimation of prevalence rate ratios for cross-sectional data: an example in occupational epidemiology, *British Journal of Industrial Medicine,*50:861-2.

Lee J, and Chia KS.,(1995), Prevalence odds ratio v prevalence ratio - a response, *Occupational Environmental Medicine*,52:781-2.

Ma S and Wong C.,(1999) Estimation of prevalence proportion rates (Letter), *International Journal of Epidemiol*ogy,28:175.

McNutt L, Hafner J, and Xue X.,(1999), Correcting the odds ratio in cohort studies of common outcomes (Letter), *Journal of the American Medical Association,* 282:529.

Skov T, Deddens J, Petersen M, and Endahl L.,(1998), Prevalence proportion ratios: estimation and hypothesis testing, *International Journal of Epidemiology,* 27:91-5.

Stromberg U.,(1994), Prevalence odds ratio v prevalence ratio, *Occupational Environmental Medicine*,51:143-4.

Stromberg U.,(1995), Prevalence odds ratio v prevalence ratio - some further comments (Correspondence), *Occupational Environmental Medicine,*52:143

Thompson M, Myers J, and Kriebel D.,(1998), Prevalence odds ratio or prevalence ratio in the analysis of cross sectional data: what is to be done?, *Occupational Environmental Medicine,*55:272-7.

Yu K and Zhang J.,(1999), Correcting the odds ratio in cohort studies of common outcomes (In reply), Journal of the American Medical Association,282:529.

Zhang J and Yu K. ,(1998), What's relative risk? A method of correcting the odds ratio in cohort studies of common outcomes, *Journal of the American Medical Accociation,*280:1690-1.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

James A. Deddens
NIOSH MS-R13
4676 Columbia Parkway
Cincinnati, OH 45226-1998
Work Phone: 1-513-556-4081
Fax: 1-513-841-4486
Email: james.deddens@math.uc.edu
or jdeddens@cdc.gov
Web: math.uc.edu/~deddens

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

Table 1. Comparison of Four Methods for the Illustrative Example

	$\hat{\beta}_0$	SE	$\hat{\beta}_1$	SE
EXACT MLE	-2.0936	(1.0208)	0.2094	(0.1021)
LOG-BINOMIAL (default start values)	-0.8273	(0.5892)	0.0827	(0.1272)
1,000 COPIES OF LOG-BINOMIAL	-2.0913	(1.0197)	0.2091	(0.1020)
COX PROPORTIONAL HAZARD			0.3251	(0.1964)

Table 2. Percentage of times PROC GENMOD failed to converge*

	Zero Slope			Medium Slope			High Slope		
Prevalence at X = 5	β_0	β_1	Percent	β_0	β_1	Percent	β_0	β_1	Percent
0.1	-2.3026	0.00	0.0%	-3.3026	0.20	0.0%	-4.3026	0.40	13.4%
0.3	-1.2040	0.00	0.0%	-1.7040	0.10	0.0%	-2.2040	0.20	13.1%
0.5	-0.6931	0.00	0.0%	-0.9431	0.05	0.1%	-1.1931	0.10	8.2%
0.7	-0.3567	0.00	0.0%	-0.5067	0.03	1.9%	-0.6567	0.06	28.2%
0.9	-0.1054	0.00	10.5%	-0.1554	0.01	19.4%	-0.2054	0.02	55.4%

* Based on 1,000 simulations of the log-binomial model with a sample size of 100 and a single X-variable, with uniform distribution [0, 10]

Table 3. Average Slope and Average Estimated SE for Combining GENMOD and the Copy Method*

		Zero Slope		Medium Slope		High Slope	
Prevalence at X=5	Method	$\hat{\beta}_1$	SE	$\hat{\beta}_1$	SE	$\hat{\beta}_1$	SE
0.1	GEN+COPY100	-0.002	0.105	0.218	0.111	0.416†	0.106†
	GEN+COPY1000	-0.002	0.105	0.218	0.111	0.422	0.105
	PHREG	-0.002	0.111	0.218	0.118	0.427	0.126
0.3	GEN+COPY100	-0.002	0.050	0.103	0.051	0.203†	0.048†
	GEN+COPY1000	-0.002	0.050	0.103	0.051	0.204	0.047
	PHREG	-0.002	0.060	0.103	0.061	0.205	0.063
0.5	GEN+COPY100	-0.000	0.032	0.050	0.032	0.102	0.031
	GEN+COPY1000	-0.000	0.032	0.050	0.032	0.102	0.031
	PHREG	-0.000	0.046	0.050	0.046	0.103	0.047
0.7	GEN+COPY100	-0.001	0.021	0.029	0.021	0.058	0.019
	GEN+COPY1000	-0.001	0.021	0.029	0.021	0.058	0.019
	PHREG	-0.001	0.039	0.029	0.039	0.059	0.039
0.9	GEN+COPY100	0.001	0.011	0.010	0.010	0.018†	0.009†
	GEN+COPY1000	0.001	0.011	0.010	0.010	0.018	0.009
	PHREG	0.001	0.034	0.010	0.034	0.020	0.034

*Based on 1,000 simulations of the log-binomial model with a sample size of 100 and a single X-variable, with uniform distribution [0, 10]. Same simulations as in table 2. If PROC GENMOD converged, the values of $\hat{\beta}_1$ and SE were used. If PROC GENMOD did not converge, the values of $\hat{\beta}_1$ and SE were taken from the COPY method (when it converged).

†When GENMOD failed to converge, COPY100 also failed to converge once. Values are used only when the model for COPY100 converged.

Table 4. Estimated Size and power for Combining GENMOD and the Copy Method*

Prevalence at X=5	Method	Zero Slope α	Zero Slope Power	Medium Slope α	Medium Slope Power	High Slope α	High Slope Power
0.1	GEN+COPY100	0.036	0.036	0.026	0.525	0.039†	0.998†
	GEN+COPY1000	0.036	0.036	0.026	0.525	0.059	0.998
	PHREG	0.027	0.027	0.017	0.466	0.025	0.996
0.3	GEN+COPY100	0.045	0.045	0.056	0.544	0.054†	0.993†
	GEN+COPY1000	0.045	0.045	0.056	0.544	0.066	0.993
	PHREG	0.018	0.018	0.022	0.367	0.014	0.976
0.5	GEN+COPY100	0.049	0.049	0.057	0.345	0.053	0.915
	GEN+COPY1000	0.049	0.049	0.057	0.345	0.058	0.915
	PHREG	0.005	0.005	0.007	0.094	0.008	0.639
0.7	GEN+COPY100	0.056	0.056	0.076	0.309	0.046	0.848
	GEN+COPY1000	0.056	0.056	0.078	0.309	0.052	0.848
	PHREG	0.000	0.000	0.001	0.013	0.000	0.196
0.9	GEN+COPY100	0.065	0.065	0.056	0.186	0.047†	0.566†
	GEN+COPY1000	0.103	0.103	0.059	0.229	0.061	0.617
	PHREG	0.000	0.000	0.000	0.000	0.000	0.000

*Same simulations as table 2 and table 3. If PROC GENMOD converged, its values for size and power were used. If PROC GENMOD did not converge, the values were taken from the COPY method (when it converged).

†When GENMOD failed to converge, COPY100 also failed to converge once. Values are used only when the model for COPY100 converged.

Table 5: Average Slope and average estimated SE for COPY100, COPY1000, and the Exact Method for a Prevalence of 0.9, a Slope of 0.02, and an Intercept of -0.2054 when PROC GENMOD fails to Converge*

Method	$\hat{\beta}_0$	$SE(\hat{\beta}_0)$	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$
COPY100	-0.222†	0.069†	0.022†	0.008†
COPY1000	-0.209	0.065	0.021	0.007
Exact	-0.208	0.064	0.021	0.006

*Based on 1,000 simulations of the log-binomial model with a sample size of 100 and a single X-variable, with uniform distribution [0, 10].

†COPY100 failed to converge once. Values are used only when the model for COPY 100 converged.