*Environmental Health Perspectives*
*Vol. 79, pp. 223-227, 1989*

# Statistical Analysis of Epidemiologic Data of Pregnancy Outcomes

by William J. Butler* and Lorrie A. Kalasinski†

**In this paper, a generalized logistic regression model for correlated observations is used to analyze epidemiologic data on the frequency of spontaneous abortion among a group of women office workers. The results are compared to those obtained from the use of the standard logistic regression model that assumes statistical independence among all the pregnancies contributed by one woman. In this example, the correlation among pregnancies from the same woman is fairly small and did not have a substantial impact on the magnitude of estimates of parameters of the model. This is due at least partly to the small average number of pregnancies contributed by each woman.**

## Introduction

Epidemiologic studies of pregnancy outcome may focus on one or several adverse health effects, including spontaneous abortion, stillbirth, lowered birth weight, malformations, and developmental abnormalities. Some of these outcomes are dichotomous (i.e., malformation: present or absent); others can be measured on a continuous scale (i.e., birth weight). Sometimes the epidemiologist will transform an outcome variable measured on a continuous scale to be a dichotomous outcome variable (i.e., birth weight < 2500 g: yes or no).

The epidemiologist is usually interested in examining the association of the dichotomous outcome variable with a number of potential risk factors, while controlling for the effects of known risk factors. The need to examine simultaneously the effects of many risk factors makes it necessary to conduct a multivariate statistical analysis of the data (*1*).

In analyzing pregnancy outcomes, attention is restricted to those women who have experienced at least one pregnancy. Of course, this analysis strategy is not designed to examine the effects of exposure on the fertility of the study population, which could be addressed using other methods (*2*).

Data are usually collected for each of the pregnancies experienced by each of the women included in the study. Depending on the group being studied, the number of pregnancies experienced by each woman could range from 1 to some number such as 9 or 10. However, most women who have experienced at least 1 pregnancy will have experienced only between 1 and 4 pregnancies.

Some of the risk factors of interest to the researcher are a characteristic of the woman and will remain constant for all of her pregnancies (i.e., race, age at menarche, etc.). Other risk factors may change during the reproductive lifetime of the woman but may be recorded in a manner such that they remain constant across all the pregnancies for the woman (i.e., education, religion, etc.). Both types of variables will be referred to as woman-level predictors since they characterize the woman. Other risk factors characterize a particular pregnancy and may vary from pregnancy to pregnancy (i.e., smoking, alcohol consumption, prenatal care, woman's age, etc.). These types of variables will be referred to as pregnancy-level predictors. Note that though pregnancy-level predictors may vary across pregnancies, it is possible that they will remain constant for some women. For example, a woman might choose to smoke cigarettes during all of her pregnancies so that the value of the smoking variable would be the same for all of her pregnancies.

The logistic regression model is the most commonly selected statistical model for the multivariate analysis of epidemiologic data with a dichotomous outcome variable (*1*). However, the correlation among pregnancies contributed by the same woman violates the assumptions of the usual methods of estimation and testing for this model. The question then arises as to whether an appropriate statistical analysis can be carried out by adjusting either the data that are collected or the statistical model used in the analysis.

Some researchers have restricted their analysis to only one pregnancy from each woman; the most recent

*Failure Analysis Associates, 2225 East Bayshore Road, Palo Alto, CA 94303.
†Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109.

pregnancy is usually selected. Though the assumptions of the statistical model are met by this restricted data set, there is obviously a substantial loss of information and, therefore, statistical power. Other researchers have chosen to ignore the clustering and treat the observations as independent. These researchers might argue that the effect of the correlation is relatively small since (a) the magnitude of the correlation among pregnancies from the same woman is small and (b) the average number of pregnancies contributed by each woman is relatively low. Alternatively, researchers have used all the pregnancies experienced by all the women but have included the outcomes of a woman's previous pregnancies as a predictor variable for the outcome of her subsequent pregnancies. For example, a predictor variable for whether a woman's second pregnancy resulted in a spontaneous abortion would be whether or not her first pregnancy had resulted in a spontaneous abortion. This approach is similar to that often taken in time-series analysis and is designed to control for the effects of women whose pregnancies have a high probability of aborting.

At this meeting, a number of adjustments in the standard statistical models for the analysis of correlated data from animal teratogenicity studies have been discussed by Chinchilli and Clark (*3*) and might be considered for use with epidemiologic data. However, in the standard teratogenicity experiment there are no pregnancy-level (or fetus-level) predictors, only woman-level (or dam-level) predictors. It is also usually the case in epidemiologic studies that the pregnancy-level predictors are of particular interest. Thus, extensions of these statistical models are necessary to analyze epidemiologic data.

The question addressed in this paper is the magnitude of the effect of ignoring this correlation among pregnancies from the same woman. The results of a multivariate statistical analysis using the standard logistic regression model, which assumes independence of all the observations, is compared to the results obtained from a model recently proposed by Rosner (*4*) for the analysis of correlated dichotomous outcome data. The empirical comparison presented here uses data recently collected on the outcomes of pregnancies of women employed in clerical and administrative support jobs by the state of Michigan (*5*). Computer software to perform the estimation and hypothesis testing is available from Dr. Rosner.

## Logistic Regression Model

### Independent Data

For simplicity, the logistic regression model is first presented for the special case in which there are only two pregnancy-level predictors and a single woman-level predictor. For purposes of explanation, smoking and alcohol consumption will be the two pregnancy-level predictors and race will be the single woman-level predictor. All three variables will be dichotomous. The outcome variable is whether or not the pregnancy resulted in a spontaneous abortion. The extension of the model to include a greater number of woman- and pregnancy-level predictors is straightforward and is used in the data analysis example.

The logistic regression model is written

$$P(D_{ij} = 1 \mid U_{1i}, X_{1ij}, X_{2ij}) = 1/[1 + \exp(-(B_0 + B_1U_{1i} + G_1X_{1ij} + G_2X_{2ij}))]$$

where $i = 1, \ldots, I$ indexes women;
$j = 1, \ldots, n_i$ indexes the pregnancies contributed by woman $i$;
$D_{ij}$ = 1 if pregnancy $j$ from woman $i$ resulted in a spontaneous abortion and = 0 otherwise;
$U_{1i}$ = race of the woman, the single woman-level predictor; 0 = white; 1 = nonwhite;
$X_{1ij}$ = smoking during pregnancy, a pregnancy-level predictor; 0 = no, 1 = yes;
$X_{2ij}$ = alcohol consumption during pregnancy, a pregnancy-level predictor; 0 = no, 1 = yes

Note that the woman-level predictor $U_{1i}$ is not subscripted by the letter $j$ since the woman-level predictor remains constant for all the pregnancies experienced by woman $i$.

The logistic regression model provides a convenient structure for describing the relative increase in the risk of a spontaneous abortion for each of the three risk factors as well as for describing the risk of a spontaneous abortion for any combination of the risk factors. The odds ratio, OR, is frequently used to describe the relative increase in risk and is calculated by taking the antilog of the coefficients for each of the risk factors (*1*). For example, the odds ratio for smoking equals

$$\text{OR(Smoking during pregnancy)} = \exp(G_1).$$

In addition, the risk of a pregnancy resulting in a spontaneous abortion for a white woman who smoked but did not consume alcohol during her pregnancy is given by

$$P(D_{ij} = 1 \mid U_{1i}=1, X_{1ij}=1, X_{2ij}=0) = 1/[1 + \exp(-(B_0 + B_1 + G_1))].$$

Importantly, the risk of a spontaneous abortion for a woman whose predictor variables are all equal to 0 is given by

$$P(D_{ij} = 1 \mid U_{1i}=0, X_{1ij}=0, X_{2ij}=0) = 1/[1 + \exp(-(B_0))].$$

Using this model, the statistical analysis of the data reduces to obtaining estimates and standard errors for the four parameters of the model: $B_0$, $B_1$, $G_1$, $G_2$. This is done using maximum likelihood methods. Of course, the adequacy of the model should be assessed before drawing firm conclusions (*6*).

## Correlated Data

Rosner's (*4*) model is written in a very similar way:

$$P(D_{ij} = 1 \mid U_{1i}, X_{1ij}, X_{2ij}) = 1 / [1 + \exp(-(B_{0i} + B_1 U_{1i} + G_1 X_{1ij} + G_2 X_{2ij}))]$$

except now each woman has her own intercept, $B_{0i}$. For the special case of all three predictors taking on the value of 0, this model reduces to

$$P(D_{ij} = 1 \mid U_{1i}=0, X_{1ij}=0, X_{2ij}=0) = 1 / [1 + \exp(-(B_{0i}))] = p_i,$$

where $p_i$ is the probability that each of the $n_i$ pregnancies of woman $i$ will result in the adverse outcome. This probability has a subscript $i$, indicating that it is specific to woman $i$. Insufficient information precludes estimating $p_i$ (or $B_{0i}$) for each woman, but it is possible to estimate the distribution of the $p_i$ across women. Some women have higher probabilities of aborting and their $p_i$ is higher (closer to 1), whereas other women have lower probabilities of aborting and their $p_i$ is smaller (closer to 0). Controlling for the effect of the women then reduces to parameterizing the distribution of $p_i$. Rosner's (*4*) model assumes that the woman's probabilities of aborting a pregnancy follow a beta distribution. The beta distribution is described by two parameters, denoted by $s$ and $t$, which specify the expected value and variance of the distribution:

$$E(p_i) = s/(s+t) = P$$

$$\mathrm{Var}(p_i) = P(1-P)/(s+t+1) = P(1-P)\,r.$$

The expected value for a women's probability of aborting a pregnancy is $P$, and the variance among these probabilities can be high or low depending on the value of $r$.

The parameter $r$ serves the same role as the intralitter correlation coefficient in animal teratogenicity studies (*7*) and the same role as the intracluster correlation coefficient in population sampling (*8*). When $r$ is large, say, close to 1, then the variance is large. This means that the women differ greatly in their probabilities of aborting. On the other hand, when $r$ is small, say, close to 0, then the variance is small. This means that the women have similar probabilities of aborting.

The beta-binomial model is routinely used for the analysis for animal teratogenicity studies where the sampling unit is the dam and there are no fetus-level (that is, pregnancy-level) predictors (*9*). Rosner's model extends the beta-binomial model to accommodate the pregnancy-level predictors. The intralitter correlation, $r$, among dams in animal teratogenicity experiments is usually in the range of 0.05 to 0.10 (*7*), but there are no reports of its magnitude for human populations.

As with the standard logistic regression model, use of Rosner's approach reduces to obtaining estimates for the parameters of the model: $s$, $t$, $B_1$, $G_1$, $G_2$. As before (*4*), these estimates are obtained from the data using maximum likelihood methods.

## Example

In August 1985, approximately 5300 women employed by the state of Michigan in clerical and administrative support jobs were asked to complete health questionnaires as part of a study on office work and the health of women (*5*). The identification of risk factors for spontaneous abortions and stillbirths was one of the primary purposes of the study.

All women who indicated on the health questionnaire that they had experienced at least one pregnancy since January 1, 1980, while working for the state of Michigan were contacted for a personal interview. Attention is restricted here to pregnancies that occurred during the years 1980 to 1985 that resulted in a singleton birth, stillbirth, or spontaneous abortion. Stillbirths and spontaneous abortions are combined to comprise the adverse outcome category.

Most women contributed only one or two pregnancies in this analysis; the average number of pregnancies contributed by each woman is 1.3 (Table 1). One could analyze a single pregnancy from each woman, thus using only 628 (77%) of the 815 pregnancies.

**Table 1. Number of pregnancies contributed by each woman during 1980–1985, while employed by the state of Michigan.**

| Number of pregnancies contributed | Number of women | Total number of pregnancies |
|---|---|---|
| 1 | 472 (75%) | 472 (58%) |
| 2 | 132 (21%) | 264 (32%) |
| 3 | 20 (3%) | 60 (7%) |
| 4+ | 4 (1%) | 19 (2%) |
| Total | 628 (100%) | 815 (100%) |
| Average number of pregnancies contributed by each woman | | 815/628 = 1.3 |

This would result in a loss of statistical power; the magnitude of the loss in power would be greater if there is only a small correlation among the pregnancies from the same woman.

Simple descriptive statistics for this group of pregnancies are presented in Table 2. The risk of a spontaneous abortion increases with age and birth order, is higher for those who smoke tobacco and drink alcohol, is greater for those with more education, and does not appear to depend on race. All of these associations, except for education, have previously appeared in the literature (*10*). Strictly speaking, standard chi-square tests are not appropriate for these data because of the correlation resulting from the pregnancies from the same women. For comparison purposes, it is useful to know that the standard chi-square tests are significant for all these positive associations.

The correlations (*11*) between outcomes of pregnancies from the same woman are all fairly small and comparable to those reported from the animal teratogenicity studies (Table 3). Explicitly incorporating the correlation of pregnancies from the same woman in the model is thus not expected to have a large effect on the estimation of the variance of the parameter estimates in the logistic regression model.

A comparison of the parameter estimates for the logistic regression models indicates that the correlation has very little effect on parameter estimation or hypothesis testing (Table 4). Indicator variables are used for each of the levels of the predictor variables. The point estimates for the coefficients are essentially the same for both models. The estimates for the standard errors of the coefficient estimates are also comparable except for the 4+ gravidity level.

There is also agreement in the intercept terms between the two models (Table 5). For the independence model, there is only one parameter, $B_0$, which is estimated to be −3.35. This corresponds to a baseline probability of 0.034. For the correlated model, the beta distribution parameters are estimated as $s = 0.323$ and $t = 8.004$. The baseline probability is thus estimated as 0.038, very close to the value obtained from the independence model. The intracluster correlation coefficient is $r = 0.11$, close to that observed for the crude analysis, which did not adjust for age, gravidity, alcohol, tobacco, or education. The hypothesis that the pregnancies from the same woman are statistically independent of each other ($H_0$: $r = 0.0$) is tested by comparing the difference in the $-2*\log(L)$ between the two models to a chi-square distribution with one degree of freedom (*4*). In this example, this hypothesis is rejected (difference = 7.5, $p < 0.001$), thus confirming,

**Table 2. Univariate analysis of outcomes of pregnancies that occurred during 1980–1985 while the woman was employed by the state of Michigan.**

| Variable | Number of spontaneous abortions and stillbirths/total number of pregnancies | Percent |
|---|---|---|
| Mother's age | | |
| ≤ 19 | 1/6 | 16.7 |
| 20–24 | 18/152 | 11.8 |
| 25–29 | 82/407 | 20.1 |
| 30–34 | 53/215 | 24.7 |
| 35–40 | 17/36 | 48.6 |
| Gravidity | | |
| 1 | 34/232 | 14.7 |
| 2 | 55/251 | 21.9 |
| 3 | 35/174 | 20.1 |
| 4+ | 47/158 | 29.7 |
| Alcohol | | |
| Yes | 24/61 | 39.3 |
| No | 147/754 | 19.5 |
| Smoking | | |
| Yes | 73/282 | 25.9 |
| No | 98/533 | 18.4 |
| Race | | |
| White | 107/520 | 20.6 |
| Black/other | 64/295 | 21.7 |
| Education | | |
| High school | 54/355 | 15.7 |
| College, 1–3 years | 94/373 | 25.2 |
| College, 4+ years | 23/107 | 23.7 |

**Table 3. Correlation of outcomes between pregnancies from the same woman, not controlling for known risk factors for spontaneous abortion and stillbirth.**

| | Pregnancy no. | | |
|---|---|---|---|
| Pregnancy no. | 1 | 2 | 3 |
| 1 | 1.00 | | |
| 2 | 0.10 | 1.00 | |
| 3 | 0.07 | 0.14 | 1.00 |

**Table 4. Multivariate logistic regression analysis of pregnancies that occurred during 1980–1985 while the woman was employed by the state of Michigan.**

| Variable | Independence coefficient (SE) | Correlated coefficient (SE) |
|---|---|---|
| Age | | |
| ≤ 19 | 1.06 (1.15) | 0.99 (1.00) |
| 20–24 | 0.00 (–) | 0.00 (–) |
| 25–29 | 0.32 (0.30) | 0.40 (0.30) |
| 30–34 | 0.40 (0.34) | 0.42 (0.33) |
| 35–40 | 1.02 (0.49) | 1.00 (0.50) |
| Gravidity | | |
| 1 | 0.00 (–) | 0.00 (–) |
| 2 | 0.32 (0.26) | 0.40 (0.27) |
| 3 | 0.12 (0.29) | 0.10 (0.30) |
| 4+ | 0.52 (0.25) | 0.48 (0.66) |
| Smoking | | |
| Yes vs. no | 0.47 (0.19) | 0.43 (0.19) |
| Alcohol | | |
| Yes vs. no | 0.99 (0.30) | 1.04 (0.30) |
| Race | | |
| White vs. other | −0.11 (0.20) | 0.03 (0.19) |
| Education | | |
| High school | 0.00 (–) | 0.00 (–) |
| College, 1–3 years | 0.60 (0.21) | 0.50 (0.20) |
| College, 4+ years | 0.48 (0.31) | 0.44 (0.29) |

**Table 5. Comparison of the intercepts of the logistic regression models.**

| | Independence | Correlated |
|---|---|---|
| Parameter estimates | $B_0 = -3.35$ | $s = 0.323$<br>$t = 8.004$ |
| Fitted probability of adverse outcome when all predictors = 0 | $1 / [1 + \exp(3.35)]$<br>$= 0.034$ | $0.323 / (8.004 + 0.323)$<br>$= 0.038$<br>$r = 1 / (8.004 + 0.323 + 1)$<br>$= 0.11$ |
| $-2*\log(L)$ | 461.87 | 454.39 |

as was suspected all along, that pregnancies from the same woman are positively correlated.

## Conclusions

Rosner's model with the random intercept term produces results that are almost identical to those obtained assuming independence. This is true for the estimates of the magnitude of association as well as for the statistical tests. The advantage of the Rosner model is the confidence in the interpretation of the results of the analysis achieved by specifically incorporating the random effects of the women.

The correlation among pregnancies from the same woman is low and is comparable to that reported in animal teratogenicity studies. The low correlation among pregnancies from the same woman and the low average number of pregnancies contributed by each woman are the reasons why the correlation does not have a large effect on the estimation and testing.

However, the correlation among pregnancies may have a greater effect if the average number of pregnancies contributed by each woman is large. This issue is currently being examined in the state of Michigan data set.

### REFERENCES


1. Kleinbaum, D. G., Kupper, L. L., and Morgenstern, H. Epidemiologic Research: Principles and Quantitative Methods. Lifetime Learning Press, Belmont, CA, 1982.
2. Levine, R. J., Symons, M. J., Balogh, S. A., Arndt, D. M., Kaswandik, N. T., and Gentile, J. W. A method for monitoring the fertility of workers. 1. Method and pilot studies. J. Occup. Med. 22: 781–791 (1980).
3. Chinchilli, V., and Clark, B. C. Trend tests for proportional responses in developmental toxicity experiments. Environ. Health Perspect. 79: 217–221 (1988).
4. Rosner, B. Multivariate methods in ophthalmology with application to other paired-data situations. Biometrics 40: 1025–1035 (1984).
5. Butler, W. J., and Brix, K. A. Reproductive outcomes of video display workers. Invited Paper, American Public Health Association, Las Vegas, NV, September 1986.
6. McCullagh, P., and Nelder, J. A. Generalized Linear Models. Chapman and Hall, London, 1983.
7. Haseman, J. K., and Kupper, L. L. Analysis of dichotomous response data from certain toxicological experiments. Biometrics 35: 281–293 (1979).
8. Cochran, W. G. Sampling Techniques. John Wiley and Sons, New York, 1977.
9. Williams, D. A. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics 37: 949–952 (1975).
10. Kline, J., and Stein, Z. Epidemiology of perinatal disorders: Spontaneous abortion (miscarriage). In: Perinatal Epidemiology (M. Bracken, Ed.), Oxford University Press, New York, 1984, pp. 23–51.
11. Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. Discrete Multivariate Analysis. MIT Press, Cambridge, MA, 1975.