A GENERALIZED MODEL FOR ANALYSIS OF NONINDEPENDENT OBSERVATIONS

by

Grace E. Kissling

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1357

Spetember 1981

A GENERALIZED MODEL FOR ANALYSIS OF NONINDEPENDENT OBSERVATIONS

by

Grace E. Kissling

A Dissertation submitted to the Faculty of
The University of North Carolina at Chapel Hill
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the
Department of Biostatistics.

Chapel Hill

1981

Approved by:

Laurence L. Kupper
Adviser

Adviser

Allen James Wilcox
Reader

ABSTRACT

GRACE E. KISSLING. A Generalized Model for Analysis of Nonindependent Observations. (Under the direction of LAWRENCE L. KUPPER and MICHAEL D. HOGAN.)

Occupational pregnancy history studies are presently being done to identify potential occupational hazards to human reproduction. Because the entire pregnancy history of each couple is collected in these studies, the problem of possibly nonindependent pregnancy outcomes within the same family arises. Usual statistical methods which treat the pregnancy as the unit of analysis assume that all observations are mutually independent events. It has been found in several studies that within a family, for example, the occurrence of a fetal loss is associated with greater chances of subsequent fetal losses. Thus, the statistical assumption that each pregnancy within the same family is independent of all other pregnancies within the family may not be appropriate.

Statistical methods for dealing with nonindependent pregnancies have been investigated in the analysis of animal litter studies. Generalized models, such as the beta-binomial model, have been found useful for analyzing these animal litter studies. In the beta-binomial model, is is assumed that the probability of fetal loss, p, has a binomial distribution within each litter and that p varies from litter to litter according to a beta distribution. Within an animal litter, all fetuses experience the same risk factors at the same time. However, within a human family, pregnancies occur over time, so that factors

affecting the outcome may change from pregnancy to pregnancy.

A uniform-logistic model analogous to the beta binomial model has been developed. In this model, a logistic regression model is used to incorporate potential risk factors associated with each pregnancy. For a dichotomous outcome, Y=0 or Y=1, and a set of risk factors, $\underset{\sim}{Z}$, a logistic model has the form,

$$\Pr(Y=1 \mid \beta_0^*, \underset{\sim}{\beta}, \underset{\sim}{Z}) = 1/\{1 + \exp[-(\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z})]\}.$$

In the uniform-logistic model, the constant term, β_0^*, is assumed to have a continuous uniform distribution over the sample of families. In using this model, only conditional independence of pregnancies within the same family is assumed. The events are not assumed to be unconditionally independent.

The uniform-logistic model has been applied to a subset of data from an occupational pregnancy history study. In these data, the estimated model parameters are nearly the same as those of a usual logistic model.

ACKNOWLEDGMENTS

I am extremely thankful for the guidance and support that my co-advisers, Drs. Lawrence L. Kupper and Michael D. Hogan, have given me. I would also like to thank the other members of my committee, Drs. C. M. Suchindran, M. J. Symons, and A. J. Wilcox for their suggestions and ideas during the course of this dissertation.

I especially wish to thank my parents for their encouragement, love, and support that they have constantly provided me. The graphs in Chapter II were drawn by my brother, Steve.

I would also like to thank Violette Kasica and Joan Kempthorne for their constant encouragement. Violette also took care of many final details after I left Chapel Hill, for which I am grateful.

Finally, I would like to acknowledge the financial support provided by the National Institute of Environmental Health Sciences Grant No. 5-T32-ES07018-01 and 02 during my course work and provided by my parents during my dissertation. The excellent typing was done by Mrs. Gay Hinnant. I also acknowledge the role that the National Institute for Occupational Safety and Health has had in this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

REVIEW OF THE LITERATURE

1.1 Introduction

An error often made in analyzing occupational pregnancy history studies lies in the selection of a sampling unit which is incompatible with assumptions necessary to perform the chosen statistical analysis method. There has been an intensive debate about the correct sampling unit to be used in experiments performed on litters of animals (Kalter [1974]; Staples and Haseman [1974]; Becker [1974]; Weil [1974]; and Palmer [1974]). However, to this date, little discussion has appeared in the literature concerning identification of the proper sampling unit in studies of human pregnancy histories. Standard statistical procedures generally assume that all sampling units are mutually independent. If assumptions of independence of sampling units are violated, then results of any statistical tests may be invalid (Haseman and Hogan [1975]; Kupper and Haseman [1978]). Most statistical procedures currently used on human pregnancy history data treat each individual conception as the sampling unit, thus, each conception is assumed to be an independent event. These statistical procedures in current use include chi-square tests, Fisher's exact test, and comparisons of rates and proportions.

If all conceptions within a family have one or both parents in common, it may not be reasonable to assume that outcomes of these conceptions are independent events. The statistical implications of

nonindependence of events in human reproduction have rarely been discussed in the literature. Cohen (1976) and Dobbins, et al. (1978) briefly mention the possibility of nonindependence in analyzing human pregnancy histories. Selevan (1977) has further recognized the problem and the lack of statistical techniques to deal with it. With the exception of some work with stochastic models, statistical methods appropriate for the analysis of pregnancy histories have not been developed. In view of the scarcity of statistical methods to deal with nonindependence problems in human studies, the proposed development of an appropriate statistical methodology will draw heavily upon an analogous situation encountered in animal reproduction experiments and its associated statistical methodology. This analogy will be discussed later.

1.2 Nonindependence of Events Within the Same Family

In animal studies, statistical analyses which treat the fetus as the sampling unit, when outcomes within a litter are not independent, artificially enlarge the sample size (Weil [1970]; Haseman and Hogan [1975]). Treating the fetus as the sampling unit amounts to counting the same thing more than once. When the sample size is artificially large, statistical tests may indicate significant differences exist among exposure groups when, in fact, the differences are not significant. An analysis ignoring maternal factors common to several pregnancies will tend to underestimate the true standard errors of the differences (Williams [1975]).

It has been recognized in human studies that occurrence of a fetal loss is associated with increased chances of subsequent fetal loss within the same family (Stein et al. [1975]). This example of a

pregnancy outcome being associated with one or more subsequent pregnancy outcomes is evidence that pregnancies within a family cannot be assumed to be independent events. Other factors such as parity and increasing maternal age may also influence pregnancy outcomes within a family in such a manner that the outcomes cannot be assumed independent. Dependence among outcomes is apparent when the variation within groups, such as families, is smaller than the variation among groups. This difference in variation is called heterogeneity. Before applying usual statistical techniques, it is usually recommended that a preliminary test of homogeneity be carried out (Haseman and Kupper [1979]).

One such test of homogeneity was derived by Potthoff and Whittinghill (1966). Suppose that there are k families in which the response variables is the occurrence of fetal loss. Potthoff's test is of the null hypothesis that the probability of fetal loss, p, is the same value within each of the k families. Under the assumption of homogeneity (i.e., under the null hypothesis), the responses in all k families are from a binomial distribution with a common probability of fetal loss, p. Potthoff's test was constructed to have maximal statistical power against the alternative hypothesis that the response has greater variability than expected under the null hypothesis. Potthoff claims that the "spreading out [excess variability] could be the result of lack of independence (and, more specifically, positive correlation) among the n_i elements of an individual sample [family]; [or] it could also result from different values of p in the different samples [families]". Thus, if the null hypothesis of homogeneity is rejected, it would not be correct to proceed with statistical analyses which assume independence or common probabilities of fetal loss among families.

1.3 Chi-square Testing of Contingency Tables

Suppose that the association between exposure and fetal loss is being investigated. A 2x2 contingency table, such as Table 1, can be constructed.

TABLE 1.1

2x2 Contingency Table

	Exposed	Unexposed	Total
Fetal Loss	a	b	n_1
No Fetal Loss	c	d	n_2
Total	m_1	m_2	N

The (uncorrected) chi-square statistic for this table is

$$X^2 = \frac{(ad-bc)^2 N}{n_1 n_2 m_1 m_2} .$$

Under the null hypothesis that there is no association between exposure and fetal loss, X^2 has a chi-square distribution with 1 degree of freedom. Assumptions implicit in testing the null hypothesis are:

(1) each of the N observations are mutually independent, and

(2) each of the m_1 observations in the exposed group has the same probability, p_1, of a fetal loss; similarly, each of the m_2 observations in the unexposed group has the same probability, p_2, of a fetal loss (Fleiss [1973]).

The null hypothesis tests that p_1 equals p_2 against the alternative that p_1 does not equal p_2.

When sample sizes are small or an event is very rare, Fisher's exact test is often used instead of the chi-square test. The same

assumptions are required for proper application of Fisher's test as are required for the chi-square test, and the same violations of assumptions will result in the same consequences discussed for the chi-square test. Fisher's exact test also assumes that the row and column totals of the 2x2 table are fixed.

In occupational pregnancy outcome studies, both assumptions (1) and (2) are violated. In several studies, it has been shown that within a family, the outcome of one pregnancy tends to influence the outcomes of subsequent pregnancies. For example, Stein et al. (1975) found that a "history of spontaneous abortion is associated with a higher frequency of subsequent abortions." They suggest that there may be a familial tendency toward repeated spontaneous abortions. Thus, it can not be assumed a priori that all pregnancy outcomes within a family are mutually independent events. Neither can it be assumed that all exposed women, for example, have the same probability of spontaneously aborting, as assumption (2) would imply.

It seems more reasonable to consider a distribution of risk within the exposed group and within the unexposed group. Such a distribution of risk would mean that individuals within each exposure group have different probabilities of displaying the response under study. The distributions of these probabilities are compared between the exposed and unexposed groups when testing for the effect of exposure in the response of interest. The concept of a distribution of risks within exposure categories has long been recognized in the animal litter experiment setting (Luning [1966]; Weil [1970]; Jensh [1970]; Haseman and Soares [1976]; Aeschbacher [1977]; Gaylor [1978]; and Haseman and Kupper [1979]). This concept, however, has been neglected in human

pregnancy outcome studies, and has not yet been incorporated into statistical analyses of these studies.

1.4 Inflation of Type I Error

The level of significance or the Type I error of a hypothesis test is the probability of rejecting the null hypothesis, when in fact, the null hypothesis is true. When the assumptions of mutual independence of observation and equal probabilities of fetal loss within exposure populations are not valid, the Type I error of chi-square tests and Fisher's exact test is increased (Palmer [1974]; Haseman and Soares [1976]; Vuataz and Sotek [1978]; Haseman and Kupper [1979]). The consequence of Type I error inflation is stating that population risks are statistically significantly different, when in fact, there may be no significant difference. Thus, p-values derived from chi-square or Fisher's statistics are misleading in that the null hypothesis of no association will be rejected with greater probability than stated by the p-value. Cohen (1976) also suggests that when sampling units are not mutually independent, usual chi-square statistics may not have a chi-square distribution, thus, significance tests which assume independence are invalid.

Simulations have been performed to examine the magnitude of Type I error inflation present in several sets of animal litter data. The degree of inflation appears to depend on the strength of the relationships between littermates. Haseman and Soares (1976) conducted simulations on CF1S mouse data and found that the Type I error inflated from the nominal 5% level to 13%. Haseman and Hogan (1975) also found inflation of Type I error in their simulations. In simulation studies,

Vuataz and Sotek (1978) detected an inflation from the nominal 5% level to a 30% Type I error. Fears, et al. (1977) investigated Fisher's exact test and derived mathematical expressions for the upper bound of possible Type I error.

Several investigators have attempted to "adjust" chi-square statistics to correct the Type I error. Cohen (1976) suggested that X^2 should be divided by the size of the largest family. However, it is not guaranteed that this adjustment will correct the chi-square statistic; it will only be an improvement over the unadjusted chi-square. Other authors have suggested forming ratios of X^2 to chi-square statistics derived from homogeneity tests (Haseman and Soares [1976]). Kruger (1970) suggested that the significance level of the chi-square test should be more stringent. However, the extent of Type I error inflation is unknown in any given set of data; thus, appropriate adjustments cannot be determined. It would be better to use other statistical procedures for which assumptions are more nearly satisfied.

1.5 Survivorship Analysis Methods

1.5.1 Life Table Models

Recently, the relationship between pregnancy outcome and factors of interest has been examined with survivorship analysis techniques. These techniques compare groups with respect to their distributions of time intervals between specified events. Survival time may be defined as the length of time of progression from one parity to the next as Hogue [1971], Erickson and Bjerkedal [1978], and Trost and Lurie [1980] have done. It could also be defined as the length of time the fetus survives from conception, as Shapiro et al. [1962] and Mellin [1962] have done.

Shapiro et al. [1962] and Mellin [1962] have adapted life table techniques to study the length of gestation. These fetal life tables were based on samples of pregnant women who were followed for the duration of one pregnancy. Shapiro et al. examined the effects of mother's age, gravidity, year, and outcome of the last pregnancy on fetal loss through comparisons of rates. However, these factors were not accounted for in the construction of fetal life tables. With a large number of observations, it would be possible to control for variables such as maternal age, gravidity, outcome of previous pregnancies, pregnancy spacing, and other factors in the life tables through stratification. An additional disadvantage of using fetal life tables is that gestational age is not reliably reported in most cases.

Other investigators have constructed life tables of the length of time to progress from a specified parity to the next. Erickson and Bjerkedal [1978] examined the time intervals between the first and second live births and between the second and third live births in a large group of Norwegian mothers. They suspected that very long or very short intervals between pregnancies are associated with adverse outcomes. However, pregnancy spacing may be affected by changes in parental habits, socioeconomic status, age, gravidity, etc., as well as hazardous exposures in the workplace. Thus, it may be difficult to determine which factors are most responsible for pregnancy spacing. Dobbins et al. [1978] examined progression from one parity to the next in an occupational setting. They constructed parity life tables, that is, life tables for the length of time spent at parities 0 to 1, 1 to 2, and 2 to 3. They then compared survival curves for exposed and unexposed women. Namboodiri et al. (1980) have also presented an

occupational parity life table. Their life table involves stochastic transitions among four states: (contracepting, working), (contracepting, not working), (noncontracepting, working), and (noncontracepting, not working). Large sample sizes are necessary for construction of this kind of life tables. It would be difficult to retain large sample sizes in occupational studies while examining transitions from one parity to the next. In an example, Namboodiri et al. (1980) examined the transition from parity 0 to parity 1 of all currently married white women in the United States in 1965. Sample sizes of occupational studies are too small to permit a reliable life table construction similar to the work fertility life table.

In life table construction, as in chi-square testing, it is assumed that all events are independent. When considering one pregnancy per family, it is valid to assume independence of events. However, life table methods would not be appropriate to apply to complete pregnancy histories in which several pregnancies within a family may be dependent events unless the life tables were parity specific.

1.5.2 Cox Regression

Recently, there has been interest in using Cox regression to model child spacing. Cox regression models produce a nonparametric estimate of hazard functions and can incorporate covariables. (A hazard function is a function over time of the probability of an event occurring during a time interval given that it has not occurred before the beginning of the interval.) After obtaining an estimate of the hazard function, survival curves can be estimated. Cox regression requires several assumptions. First, the observations are assumed to be mutually

independent. It is also assumed that hazard functions evaluated at different values of the covariables are proportional, i.e., constant multiples of the background hazard. This implies that hazard functions as functions of time can not intersect. Such intersection could indicate that there is an interaction between one or more of the independent variables and time. In addition, since Cox regression is based on maximum likelihood estimation of the coefficients of the independent variables, the sample sizes should be large enough for asymptotic maximum likelihood properties to hold before performing tests of hypothesis (Cox [1972]). Desirable asymptotic properties include approximate normality of maximum likelihood estimates. Hypothesis testing is based on the assumption of approximate normality of the estimated coefficients. Trost and Lurie (1980) have been the first to investigate Cox regression models for child spacing. They considered the time from marriage to birth of the first child as the "survival time" to model. They also stated that times between first and second births, second and third births, etc., could be modeled in the same way. By considering the first birth only, the assumption of independence of events is justifiable. However, it is not clear that hazards for different values of the covariables will be constant multiples of each other at each point in time. Also, Trost and Lurie completely ignored fetal loss. Several other investigators (e.g., see Hogue [1971]) claim that the occurrence of a fetal loss increases the risk of further loss. Also, it is thought that fetal loss may decrease the time between live births because of the desire to "replace the loss." Thus, the hazard function for time between live births is probably affected by the occurrence of fetal loss. Since occupational populations are often small, there may not be enough

pregnancies for maximum likelihood properties, such as approximate normality of the estimated coefficients, to hold for hypothesis testing, especially if times between fourth and fifth pregnancies, for example, are being modeled.

Life tables methods and Cox regression can use the information conveyed by censored observations. Censored observations would be, for example, observations for which the family withdrew from the study before completion of childbearing. Families which had not completed childbearing at the time of the study would also be considered censored observations. Proper treatment of censored observations is necessary when it is known that pregnancy histories are incomplete for some families.

1.5.3 Summary of Survivorship Analysis Methods

Survivorship analysis methods consider the individual pregnancy rather than the entire pregnancy history. These methods could be applied to pregnancy history data, but some modifications in the usual procedures may be needed. Fetal life tables should not be used unless reliable data about length of gestation are obtained. Parity life tables are usually limited to the low parities because of sample size considerations. In life table construction and in survival time modeling, such as Cox regression, fetal loss should not be ignored. Even after proper modifications are made to survivorship methods, the effect of exposure on length of time between events does not have as clear an interpretation as the effect of exposure on the occurrence of these events. Therefore, it seems more profitable to first investigate statistical methods for analyzing frequencies of events for which a clear interpretation of any exposure related differences exists.

1.6 Stochastic Models

Sheps and Menken (1973) have thoroughly discussed the use of stochastic models for modeling events related to pregnancy. Stochastic models are usually used for estimating probabilities of changing from a state of being to another state and the time spent in any of the possible states before a transition occurs. For example, a stochastic model could be used to estimate transition probabilities and expected waiting times among the states: susceptible to conception, contracepting, pregnant, normal delivery, spontaneous abortion, induced abortion, not susceptible to conception. The estimated transition probabilities can be arranged in matrix form which is referred to as the transition matrix. An advantage of using a stochastic model is that nonindependence of events within families can be built into the model. Shachtman and Hogue (1976) have presented a model in which several transition matrices are estimated. For example, a separate transition matrix is estimated for the following situations: the first pregnancy is a live birth, the first pregnancy is an abortion, and pregnancies after the first live birth. The transition from one situation to another takes nonindependence into account.

When stochastic models are used, it is assumed that the population is homogeneous; that is, that all members have identical characteristics. This may not be a reasonable assumption to make when dealing with data from occupational surveys. Spilerman (1972) has suggested a regression method to adjust for differing characteristics among members of the population. However, the regression method would not be valid in pregnancy history studies in which several observations may be dependent.

Another assumption which is usually made when using stochastic models is that the Markov property holds (Sheps and Menken [1973]). The Markov property states that an individual's probability of transition depends only on his or her present state and not on any previous states in which the individual may have been. It may be necessary to redefine states to satisfy the Markov property.

Estimation of transition probabilities requires a large number of observations. Occupational studies may not always supply enough observations to construct a useful stochastic model. In addition, the populations are usually heterogeneous, in violation of the important homogeneous population assumption. Thus, stochastic models may be used for large studies, however, great care must be taken to satisfy the Markov property and to deal with heterogeneity.

1.7 Comparisons of Observed and Expected Numbers of Events

In 1914, Greenwood and Yule reported a method to test for the effects of birth order and maternal age on the occurrence of some defect of interest. The Greenwood-Yule method compares observed frequencies of defects with expected frequencies of defects. Expected frequencies are derived from estimating the probability that an individual has a defect, assuming that every individual is equally likely to have the defect. This estimation is based on the data available in the study rather than national or regional expected frequencies.

Norton (1952) examined the association between neurosis and maternal age and birth order with the Greenwood-Yule method. McKeown and Record (1955) compared the Greenwood-Yule method to a chi-square rate comparison method. They noted that the Greenwood-Yule method may

be in error if each pregnancy history is not complete. For example, if several mothers have not reached the end of their reproductive periods at the time of the study, the estimated probability of a defect may be in error since births occurring after the study are not observed.

After data from occupational histories are stratified by birth order and/or maternal age group, stratum sizes may not be large enough to detect differences if they exist. It would be possible to stratify by exposure level and other independent variables of interest, however, the stratum-specific sample sizes may be so greatly reduced that any test for differences would be meaningless. Occupational studies may include pregnancy histories which are not complete, introducing error into the analysis with the Greenwood-Yule method. It is also not clear how to treat pregnancy histories for which exposure levels change.

Levine et al. (1980) have developed a method for monitoring fertility impairment in industry. Observed numbers of births within the working population are compared to predicted numbers of births. Predicted numbers of births are calculated from tables of national birth probabilities specific for maternal birth cohort, age, parity, and race. This method was applied to a group of thirty-six male factory employees and a decrease in fertility was attributed to exposure to chemicals in the workplace (Levine et al. [1981]).

1.8 Nonparametric Methods

Nonparametric statistical methods make fewer assumptions about the underlying distribution of the data than usual parametric methods. The Cox regression model previously discussed is one such nonparametric method. Cox regression relies on large sample sizes for approximate normality of the regression coefficients and valid hypothesis testing.

Two nonparametric tests commonly used to analyze animal litter data are the Mann-Whitney U-test and Jonckheere's test. The Mann-Whitney U-test is a nonparametric method which uses the mother as the sampling unit. As an example, suppose that the proportion of affected fetuses is calculated for each litter. The Mann-Whitney U-test tests that the distributions of these proportions are the same in the exposed and control groups (e.g. see Conover [1977]). Independent variables could be taken into account by stratification before testing. Since all fetuses within an animal litter are born at the same time, littermates experience the same conditions. In human pregnancies, however, conditions such as parity and maternal age change from pregnancy to pregnancy. It is not clear how to adjust for differing independent variables within a family when using the Mann-Whitney U-test.

Jonckheere's test is another nonparametric method which is often used in dose-response investigations on animal litters. Jonckheere's test is useful in detecting monotonic trends in response. However, here too, it is not clear how to deal with the independent variables which may change from pregnancy to pregnancy within a human family.

Nonparametric methods are generally less powerful statistically than parametric methods. Less power means that greater sample sizes would be needed to detect a specified difference with nonparametric methods than with parametric methods (Haseman and Soares [1976]). In addition, there is no simple way to adjust for continuous independent variables. However, nonparametric methods have the advantage of making only mild assumptions about the underlying probability distribution.

1.9 Generalized Models

Mathematical modeling is useful in examining the relationship between exposure and response while controlling for possibly confounding variables. Some examples of mathematical modeling include multiple linear regression and multiple logistic regression. In several animal teratology studies, a special form of modeling has been found useful. Models of this special form are known as generalized models.

Generalized models have been used in animal teratology studies--the dominant-lethal study, in particular--to model pregnancy outcomes within litters. Littermates tend to be more alike than fetuses from different families. Therefore, the variability in response is greater among families than within families. Generalized models can be used to assume a probability distribution for the outcomes within a family, then to assume another probability distribution for the variability of response among families. This seems to be a realistic picture of the actual situation in animal litters. There also seems to be some intuitive justification for this approach in human pregnancy histories.

The most commonly used generalized model in animal litter data has been the beta-binomial model (e.g., see Williams [1975]; VanRyzin [1975]; Aeschbacher et al. [1977]; Altham [1978]). Suppose, for example, that the outcome of interest is the occurrence of fetal death. Under the beta-binomial model, it is assumed that fetal death is binomially distributed within each litter with parameter p, where p is the probability of fetal death. Also, p is assumed to have a beta distribution among litters. Depending on the values of the two parameters of a beta distribution, the distribution can take on a wide range of shapes, such as J-shaped, U-shaped, and reverse J-shaped.

Thus, within a litter, each fetus has the same probability of fetal death, but the probability of fetal death varies among litters. Conditional upon the value of p, each fetus is treated as an independent observation. However, unconditionally, the observations may be dependent. Independence does not have to be assumed, however, Kupper and Haseman (1978) have developed a correlated binomial model in which the correlation between fetuses within a litter can be estimated. Vuataz (1978) has found that in animal teratology studies, the beta-binomial model produces more reliable results than chi-square tests. Kupper and Haseman found that their correlated binomial model fit data from teratology studies as well as the beta-binomial model.

Other forms of generalized models have been suggested. McCaughran and Arnold (1976) proposed a negative binomial model in which the number of fetal deaths has a Poisson distribution within each litter, and the Poisson parameter has a gamma distribution among litters. However, this model does not take litter size into account, and there is no limit on the number of fetal deaths theoretically possible within each litter. Luning et al. (1966) have proposed a model similar to the beta-binomial model in which the number of fetal deaths has a binomial distribution within each litter, but the probability of fetal death has a normal distribution among litters. This model does not restrict the probability of fetal death to being between 0 and 1 as a probability should be restricted.

Generalized models allow for variation in response within families, as well as among families. This seems to be a more realistic situation than having to assume constant proportions within families and within exposure groups as the chi-square test requires. Thus, a

statistical analysis with an appropriate generalized model should be more valid than the usual statistical procedures which have been discussed. In addition, the true Type I error should be closer to the stated level of the test.

1.10 Summary and Statement of the Problem

Several statistical methods have been used or proposed to analyze pregnancy history data. These methods include chi-square testing of contingency tables, survivorship analysis methods, stochastic models, and comparisons of birth rates. When the individual pregnancy is considered as the unit of analysis, most of these methods assume that the units are mutually independent. Thus, in the application of these methods to entire pregnancy histories, this assumption of independence is violated. Ignoring dependence when it exists leads to test statistics which reject the null hypothesis of no effect with greater probability than the stated nominal Type I level of the test.

The animal litter data literature suggests two methods for analyzing human pregnancy histories--nonparametric methods and generalized models. In an animal litter, all fetuses are born at the same time, so all experience nearly the same conditions. In human families, however, births occur one after another, with time separating them. Thus, a family's children may not have the same sets of covariables. It would be difficult to incorporate covariables into nonparametric analyses. Mathematical modeling seems to be a reasonable approach to analyzing human pregnancy histories. Generalized models which have been applied to animal litter data include the beta-binomial model and the negative binomial model. In these models, each conception is treated

as an observation, however, using a compound distribution approach, maternal heterogeneity is also modeled.

CHAPTER II

THEORETICAL DEVELOPMENT OF THE UNIFORM-LOGISTIC MODEL

2.1 Introduction

A generalized model for analyzing pregnancy history data is developed in this chapter. This model is based on a logistic regression model with the modification that the constant term has a continuous uniform distribution. Hence, the model will be referred to as a uniform-logistic model.

Each mother may have several pregnancies, and with each pregnancy, is a set of factors which may be associated with the outcome of the pregnancy. This model incorporates the sets of factors for each pregnancy and maternal heterogeneity to model the probability of a pregnancy displaying some event of interest. The events considered will generally be adverse. Thus, the response is dichotomized into presence or absence of an adverse event during the pregnancy.

As with the beta-binomial model for animal litter studies, it is assumed that the underlying probability of the event of interest varies from mother to mother. For a particular pregnancy, the probability of the adverse outcome of interest depends on the mother's underlying probability of the outcome, in addition to covariables associated with the pregnancy. These covariables include factors which are thought to influence the occurrence of adverse events during pregnancy, such as exposures in the workplace, mother's age at pregnancy, gravidity, race,

and smoking and drinking habits. Conditional upon the underlying probability of an adverse outcome, pregnancies within a family are assumed to be independent events. However, the events are not necessarily unconditionally independent.

In Section 2.2, the unconditional covariance between random variables is expressed in terms of conditional and expected covariances. In Section 2.3, the uniform-logistic model is formulated. In Section 2.4, the background risk of an adverse event is found; Section 2.5 deals with the unconditional probability of an adverse outcome. Section 2.6 applies the results of Section 2.2 to derive an expression for the unconditional covariance. The likelihood function is derived in Section 2.7.

2.2 Unconditional Covariance

To show that conditional independence does not necessarily lead to unconditional independence, consider the following. Let Y_i and Y_j denote two random variables with parameter θ. The covariance between Y_i and Y_j, not conditional upon θ can be written in terms of covariances conditional upon θ. The covariance between Y_i and Y_j, conditional upon θ is

$$\text{cov}[(Y_i,Y_j)|\theta] = E_{Y_iY_j}[Y_iY_j|\theta] - E_{Y_i}[Y_i|\theta]\,E[Y_j|\theta]. \qquad (2.2.1)$$

The expectation with respect to θ is

$$E_\theta\{\text{cov}[(Y_i,Y_j)|\theta]\} = E_\theta\{E_{Y_iY_j}[Y_iY_j|\theta]\} - E_\theta\{E_{Y_i}[Y_i|\theta]\,E_{Y_j}[Y_j|\theta]\} \qquad (2.2.2)$$

which reduces to

$$E_\theta\{\text{cov}[(Y_i,Y_j)|\theta]\} = E_{Y_iY_j}[Y_iY_j] - E_\theta\{E_{Y_i}[Y_i|\theta]\,E_{Y_j}[Y_j|\theta]\}. \qquad (2.2.3)$$

Now, consider the covariance of the expected values of Y_i and Y_j, conditional upon θ.

$$\text{cov}\{E_{Y_i}[Y_i|\theta],\ E_{Y_j}[Y_j|\theta]\} = E_\theta\{E_{Y_i}[Y_i|\theta]\ E_{Y_j}[Y_j|\theta]\} - E_\theta\{E_{Y_i}[Y_i|\theta]\}\ E_\theta\{E_{Y_j}[Y_j|\theta]\} \tag{2.2.4}$$

This simplifies to

$$\text{cov}\{E_{Y_i}[Y_i|\theta], E_{Y_j}[Y_j|\theta]\} = E_\theta\{E_{Y_i}[Y_i|\theta]E_{Y_j}[Y_j|\theta]\} - E_{Y_i}[Y_k]E_{Y_j}[Y_j]. \tag{2.2.5}$$

Adding equation (2.2.3) and 2.2.5),

$$\begin{aligned} &E_\theta\{\text{cov}[(Y_i,Y_j)|\theta]\} + \text{cov}\{E_{Y_i}[Y_i|\theta], E_{Y_j}[Y_j|\theta]\} \\ &\quad = E_{Y_iY_j}[Y_iY_j] - E_\theta\{E_{Y_i}[Y_i|\theta]E_{Y_j}[Y_j|\theta]\} \\ &\qquad + E_\theta\{E_{Y_i}[Y_i|\theta]E_{Y_j}[Y_j|\theta]\} - E_{Y_i}[Y_i]E_{Y_j}[Y_j] \\ &\quad = E_{Y_iY_j}[Y_iY_j] - E_{Y_i}[Y_i]E_{Y_j}[Y_j] \\ &\quad = \text{cov}(Y_i,Y_j). \end{aligned} \tag{2.2.6}$$

Thus, the unconditional covariance between Y_i and Y_j can be written as the sum of the expected covariance between Y_i and Y_j with respect to θ and the covariance between the expected values of Y_i and Y_j, conditional upon θ.

Suppose that Y_i and Y_j are conditionally independent. Then

$$E_\theta\{\text{cov}[(Y_i,Y_j)|\theta]\} = 0,$$

and the unconditional covariance is

$$\text{cov}(Y_i,Y_j) = \text{cov}\{E_{Y_i}[Y_i|\theta],E_{Y_j}[Y_j|\theta]\}. \qquad (2.2.7)$$

This unconditional covariance may be different from zero. Thus, Y_i and Y_j may be conditionally independent but not unconditionally independent.

2.3 The Uniform-logistic Model

Using the same compound distribution approach as the beta-binomial model, the model proposed for analyzing human reproductive histories is a uniform-logistic model. In this uniform-logistic model, the response of interest is the occurrence or not of an adverse event during pregnancy.

To define notation, let Y_j be a dichotomous random variable such that

$$Y_j = \begin{cases} 1 & \text{if an adverse outcome occurs at pregnancy } j \\ 0 & \text{if no adverse outcome occurs at pregnancy } j. \end{cases}$$

Let $\underset{\sim}{X}_j = (X_{1j},X_{2j},\ldots,X_{pj})'$ denote the pxl vector of covariables associated with outcome Y_j and let $\underset{\sim}{\beta}' = (\beta_1,\beta_2,\ldots,\beta_p)$ denote the vector of regression coefficients associated with $\underset{\sim}{X}_j$. A logistic model for the probability of an adverse outcome for pregnancy j has the form:

$$\begin{aligned} \Pr(Y_j = 1|\beta_0,\underset{\sim}{\beta},\underset{\sim}{X}_j) &= \frac{1}{1 + \exp[-(\beta_0 + \beta_1X_{1j} + \beta_2X_{2j} + \ldots + \beta_pX_{pj})]} \\ &= \frac{1}{1 + \exp[-(\beta_0 + \sum_{i=1}^{p} \beta_iX_{ij})]} \\ &= \frac{1}{1 + \exp[-(\beta_0 + \underset{\sim}{\beta}'\underset{\sim}{X}_j)]} \qquad (2.3.1) \end{aligned}$$

Letting X_i^* denote a 'low risk' value of the i^{th} covariable and $\underset{\sim}{X}^* = (X_1^*, X_2^*, \ldots, X_p^*)'$, the above expression can be rewritten as

$$\Pr(Y_j = 1 | \beta_0, \underset{\sim}{\beta}, \underset{\sim j}{X}) = \frac{1}{1 + \exp[-(\beta_0 + \sum_{i=1}^{p} \beta_i X_i^* + \sum_{i=1}^{p} \beta_i (X_{ij} - X_i^*))]} \tag{2.3.2}$$

Let $\beta_0^* = \beta_0 + \sum_{i=1}^{p} \beta_i X_i^*$ and $\underset{\sim j}{Z} = \underset{\sim j}{X} - \underset{\sim}{X}^*$. Then

$$\Pr(Y_j = 1 | \beta_0^*, \underset{\sim}{\beta}, \underset{\sim j}{Z}) = \frac{1}{1 + \exp[-(\beta_0^* + \sum_{i=1}^{p} \beta_i Z_{ij})]}$$

$$= \frac{1}{1 + \exp[-(\beta_0^* + \underset{\sim}{\beta}' \underset{\sim j}{Z})]} . \tag{2.3.3}$$

When $\underset{\sim j}{Z} = \underset{\sim}{0}$, i.e., when $\underset{\sim j}{X} = \underset{\sim}{X}^*$,

$$\Pr(Y_j = 1 | \beta_0^*, \underset{\sim}{\beta}, \underset{\sim}{X}^*) = \frac{1}{1 + \exp[-(\beta_0^*)]} . \tag{2.3.4}$$

This is the 'low risk' or 'background' probability of the adverse outcome.

A uniform probability distribution can be imposed on β_0^*. For example,

$$f_{\beta_0^*}(\beta_0^*) = \begin{cases} \frac{1}{b-a} & \text{for } a < \beta_0^* < b \\ 0 & \text{otherwise.} \end{cases} \tag{2.3.5}$$

Thus, the uniform-logistic model is a compounding of a logistic model conditional upon its constant term, β_0^*, and a uniform distribution imposed on β_0^*.

2.4 Distribution of the 'Background Risk'

Let

$$R = \Pr(Y = 1 \mid \beta_0^*, \underset{\sim}{\beta}, \underset{\sim}{X}^*)$$

$$= \frac{1}{1 + \exp[-(\beta_0^*)]}$$

$$= \frac{\exp[\beta_0^*]}{1 + \exp[\beta_0^*]} \quad . \tag{2.4.1}$$

R denotes the 'background risk' of the adverse outcome of interest. Assuming that β_0^* has a uniform distribution as in (2.3.5), the distribution of R can easily be found.

$$R = \frac{\exp[\beta_0^*]}{1 + \exp[\beta_0^*]} \tag{2.4.3a}$$

$$R + R \exp[\beta_0^*] = \exp[\beta_0^*] \tag{2.4.3b}$$

$$R = (1-R) \exp[\beta_0^*] \tag{2.4.3c}$$

$$\frac{R}{1-R} = \exp[\beta_0^*] \tag{2.4.3d}$$

$$\log \frac{R}{1-R} = \beta_0^* \tag{2.4.3e}$$

$$\frac{d\beta_0^*}{dR} = \frac{1-R}{R} \frac{(1-R)-(-R)}{(1-R)^2} \tag{2.4.4a}$$

$$\frac{d\beta_0^*}{dR} = \frac{1}{R(1-R)} \tag{2.4.4b}$$

Thus

$$f_R(R) = f_{\beta_0^*}(R) \frac{d\beta_0^*}{dR}$$

$$= \begin{cases} \dfrac{1}{(b-a)R(1-R)} & \text{if } 0 < \dfrac{e^a}{1+e^a} < R < \dfrac{e^b}{1+e^b} < 1 \\ 0 & \text{otherwise} \end{cases} .$$

This is a portion of a concave function between 0 and 1 which is symmetric about 1/2. (See Figure 2.1) The portion of the function belonging to $f_R(R)$ depends on the choice of a and b.

2.5 Unconditional Probability of an Adverse Event

2.5.1 The Unconditional Probability for One Pregnancy

The probability of an adverse outcome on the j^{th} pregnancy, unconditional on β_0^*, is found by integrating the conditional probability with respect to $f_{\beta_0^*}(\beta_0^*)$.

$$\begin{aligned} \Pr(Y_j = 1) &= \int_{-\infty}^{\infty} \Pr(Y_j = 1 | \beta_0^*) f_{\beta_0^*}(\beta_0^*) d\beta_0^* \\ &= \int_a^b \frac{1}{1 + \exp[-(\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_j)]} \frac{1}{b-a} d\beta_0^* \\ &= \frac{1}{b-a} \int_a^b \frac{\exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]}{1 + \exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]} d\beta_0^* \end{aligned} \qquad (2.5.1)$$

Noting that the integral has the form $\int \frac{du}{u}$,

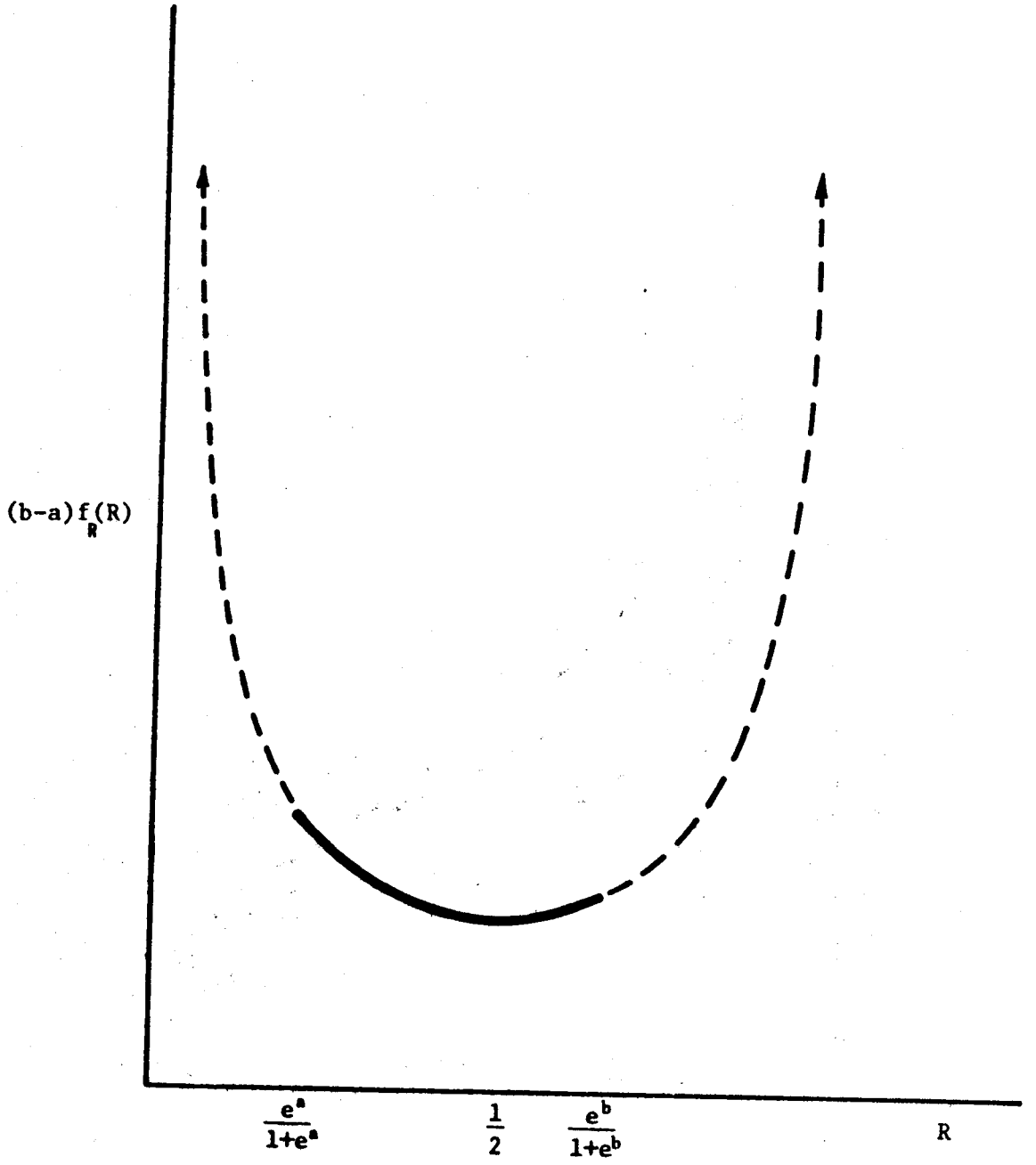
$$\Pr(Y_j=1) = \frac{1}{b-a} \log(1 + \exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]) \Big|_{\beta_0^*=a}^{b}$$

$$= \frac{1}{b-a} \log \frac{1 + \exp[b + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]}{1 + \exp[a + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]} . \qquad (2.5.2)$$

This function has a shape similar to the logistic function. (See Figure 2.2.)

Figure 2.1

Graph of $(b-a)f_R(R)$ as a function of R



where $f_R(R) = \begin{cases} \frac{1}{(b-a)R(1-R)}, 0 < \frac{e^a}{1+e^a} < R < \frac{e^b}{1+e^b} < 1 \\ 0 \quad , \text{otherwise} \end{cases}$

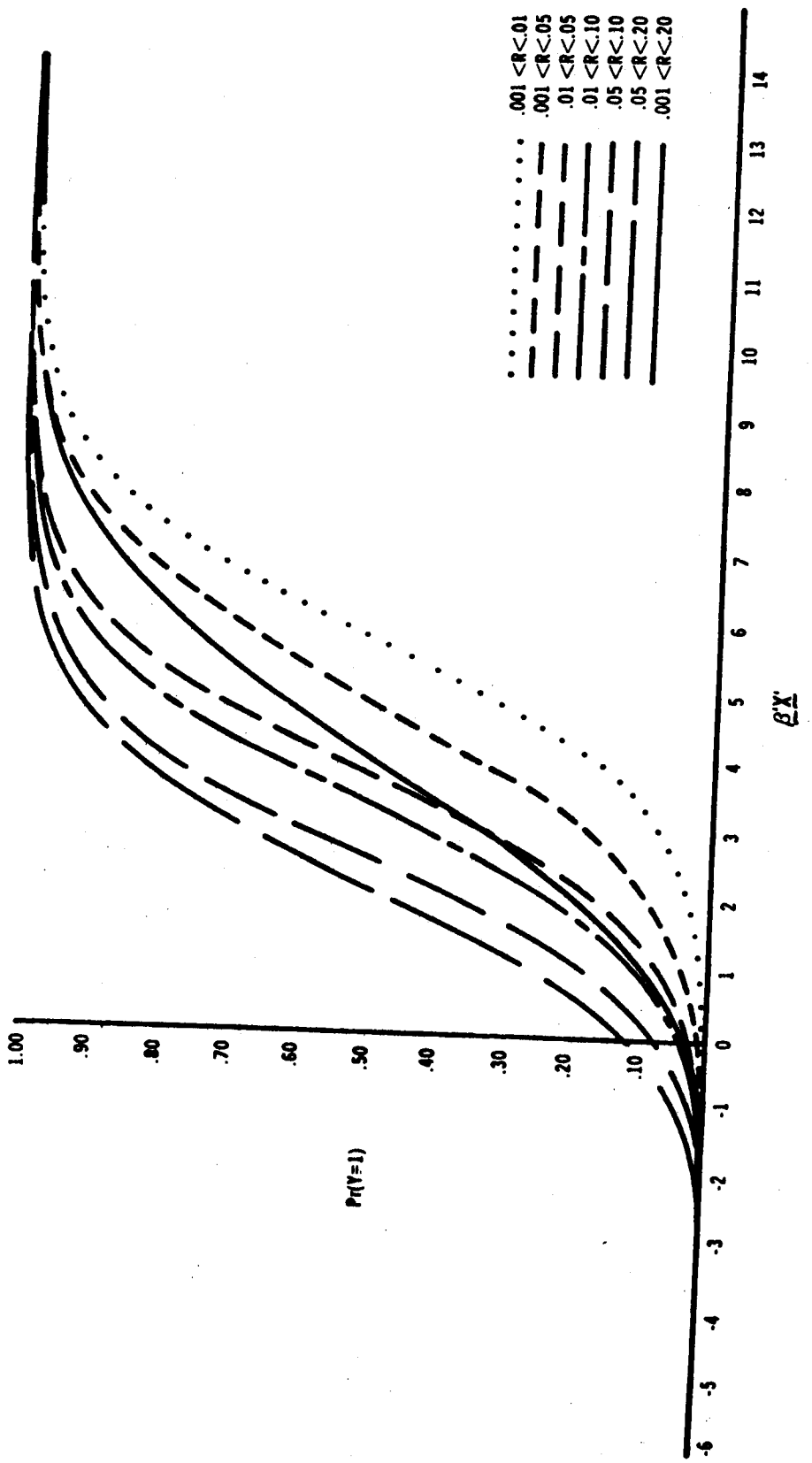


Figure 2.2

The uniform-logistic probability, Pr(Y=1), as a Function of $\beta'X$.

2.5.2 The Logistic Probability as a Limiting Case

As b approaches a, the unconditional probability from the uniform-logistic model approaches the logistic probability. This is seen by applying L'hospital's Rule to the uniform-logistic unconditional probability.

$$\lim_{b\to a} \Pr(Y_j=1) = \lim_{b\to a} \frac{1}{b-a} \log \frac{1 + \exp[b + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]}{1 + \exp[a + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]}$$

$$= \lim_{b\to a} \frac{1 + \exp[a + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]}{1 + \exp[b + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]} \frac{\exp[b + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]}{1 + \exp[a + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]}$$

$$= \lim_{b\to a} \frac{\exp[b + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]}{1 + \exp[b + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]}$$

$$= \frac{\exp[a + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]}{1 + \exp[a + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]} \qquad (2.5.3)$$

$$= \Pr(Y_j = 1) \text{ under a logistic model.}$$

This result is intuitively reasonable because as b approaches a, a smaller and smaller interval is being put around β_0^* of the logistic model. Ultimately, the interval is the point, β_0^*.

2.6 Unconditional Covariance for the Uniform-logistic Model

As derived in Section 2.2, the unconditional covariance can be expressed as

$$\operatorname{cov}(Y_i,Y_j) = E_\theta\{\operatorname{cov}[Y_i,Y_j]|\theta\} + \operatorname{cov}_\theta\{E_{Y_i}[Y_i|\theta],E_{Y_j}[Y_j|\theta]\} \qquad (2.6.1)$$

In the particular case of the uniform-logistic model, θ is replaced by β_0^*. Thus, the unconditional covariance between two pregnancies associated with the same mother is found as follows:

$$\operatorname{cov}(Y_i,Y_j) = E_{\beta_0^*}\{\operatorname{cov}[(Y_i,Y_j)|\beta_0^*]\} + \operatorname{cov}_{\beta_0^*}\{E_{Y_i}[Y_i|\beta_0^*],E_{Y_j}[Y_j|\beta_0^*]\} \qquad (2.6.2)$$

Assuming that Y_i and Y_j are conditionally independent,

$$E_{\beta_0^*}\{\mathrm{cov}[(Y_i,Y_j)|\beta_0^*] = 0$$

and equation (2.6.2) reduces to

$$\mathrm{cov}(Y_i,Y_j) = \mathrm{cov}_{\beta_0^*}\{E_{Y_i}[Y_i|\beta_0^*], E_{Y_j}[Y_j|\beta_0^*]\}$$

$$= E_{\beta_0^*}\{E_{Y_i}[Y_i|\beta_0^*]E_{Y_j}[Y_j|\beta_0^*]\} - E_{\beta_0^*}\{E_{Y_i}[Y_i|\beta_0^*]\}E_{\beta_0^*}\{E_{Y_j}[Y_j|\beta_0^*]\}$$

$$= E_{\beta_0^*}\left\{\frac{1}{1+\exp[-(\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_i)]}\ \frac{}{1+\exp[-(\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_j)]}\right\}$$

$$- E_{\beta_0^*}\left\{\frac{1}{1+\exp[-(\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_i)]}\right\} E_{\beta_0^*}\left\{\frac{1}{1+\exp[-(\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_j)]}\right\} \qquad (2.6.3)$$

Note that when Y_i is a 0-1 indicator variable,

$$E_{Y_i}[Y_i|\beta_0^*] = \Pr[Y_i = 1|\beta_0^*].$$

Thus,

$$E_{\beta_0^*}\{E_{Y_i}[Y_i|\beta_0^*]\} = E_{\beta_0^*}\{\Pr(Y_i = 1|\beta_0^*)\}$$

which was found in equation (2.5.2) to be

$$E_{\beta_0^*}\{E_{Y_i}[Y_i|\beta_0^*]\} = \frac{1}{b-a}\log\frac{1+\exp[b + \underset{\sim}{\beta}'\underset{\sim}{Z}_i]}{1+\exp[a + \underset{\sim}{\beta}'\underset{\sim}{Z}_i]}, \qquad (2.6.4a)$$

and similarly for Y_j,

$$E_{\beta_0^*}\{E_{Y_j}[Y_j|\beta_0^*]\} = \frac{1}{b-a}\log\frac{1+\exp[b + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]}{1+\exp[a + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]}. \qquad (2.6.4b)$$

Now, only the integral,

$$\int_a^b \frac{1}{1+\exp[-(\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_i)]} \frac{1}{1+\exp[-(\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_j)]} \frac{1}{b-a} d\beta_0^*$$

$$= \int_a^b \frac{\exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_i]}{1+\exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_i]} \frac{\exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]}{1+\exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]} \frac{1}{b-a} d\beta_0^*$$

remains to be evaluated. Let

$$t = \exp[\beta_0^*], \quad c_i = \exp[\underset{\sim}{\beta}'\underset{\sim}{Z}_i], \quad \text{and} \quad c_j = \exp[\underset{\sim}{\beta}'\underset{\sim}{Z}_j] .$$

Then

$$\beta_0^* = \log t \text{ and } d\beta_0^* = \frac{1}{t} dt,$$

and

$$\int_a^b \frac{\exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_i]}{1+\exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_i]} \frac{\exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]}{1+\exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_j]} \frac{1}{b-a} d\beta_0^*$$

$$= \frac{1}{b-a} \int_{e^a}^{e^b} \frac{c_i t}{1+c_i t} \frac{c_j t}{1+c_j t} \frac{1}{t} dt$$

$$= \frac{c_i c_j}{b-a} \int_{e^a}^{e^b} \frac{t}{(1+c_i t)(1+c_j t)} dt. \qquad (2.6.5)$$

A partial fraction expansion of the integrand is

$$\frac{t}{(1+c_i t)(1+c_j t)} = \frac{A}{1+c_i t} + \frac{B}{1+c_j t}$$

$$= \frac{A + Ac_j t + B + Bc_i t}{(1+c_i t)(1+c_j t)} \qquad (2.6.6)$$

Therefore, $A + Ac_j t + B + Bc_i t = t$ where

$$\left.\begin{aligned} A + B &= 0 \\ Ac_j + Bc_i &= 1 \,. \end{aligned}\right\} \tag{2.6.7}$$

Solving this system of equations for A and B results in

$$\left.\begin{aligned} A &= \frac{-1}{c_i - c_j} \\ B &= \frac{1}{c_i - c_j} \end{aligned}\right\} \tag{2.6.8}$$

So,

$$\frac{c_i c_j}{b-a}\int_a^b \frac{t}{(1+c_i t)(1+c_j t)}\,dt$$

$$= \frac{c_i c_j}{b-a}\int_a^b \left(\frac{-1}{c_i-c_j}\,\frac{1}{1+c_i t} + \frac{1}{c_i-c_j}\,\frac{1}{1+c_j t}\right) dt$$

$$= \frac{c_i c_j}{b-a}\left[\frac{-1}{c_i-c_j}\,\frac{1}{c_i}\log(1+c_i t) + \frac{1}{c_i-c_j}\,\frac{1}{c_j}\log(1+c_j t)\right]\Bigg|_{t=a}^{b}$$

$$= \frac{1}{b-a}\left[\frac{-c_j}{c_i-c_j}\log\frac{1+c_i e^b}{1+c_i e^a} + \frac{c_i}{c_i-c_j}\log\frac{1+c_j e^b}{1+c_j e^a}\right]$$

$$= \frac{1}{b-a}\left[\frac{-\exp[\beta' Z_j]}{\exp[\beta' Z_i] - \exp[\beta' Z_j]}\log\frac{1+\exp[b+\beta' Z_i]}{1+\exp[a+\beta' Z_i]}\right.$$

$$\left. + \frac{\exp[\beta' Z_i]}{\exp[\beta' Z_i] - \exp[\beta' Z_j]}\log\frac{1+\exp[b+\beta' Z_j]}{1+\exp[a+\beta' Z_j]}\right]$$

$$= \frac{1}{b-a}\left[\frac{1}{1-\exp[-\beta'(Z_j - Z_i)]}\log\frac{1+\exp[b+\beta' Z_i]}{1+\exp[a+\beta' Z_i]}\right.$$

$$\left. + \frac{1}{1-\exp[\beta'(Z_j - Z_i)]}\log\frac{1+\exp[b+\beta' Z_j]}{1+\exp[a+\beta' Z_j]}\right] \tag{2.6.9}$$

Substituting this evaluation of the integral into the covariance expression, equation (2.6.3),

$$\text{cov}(Y_i,Y_j) = \frac{1}{b-a}\left\{\frac{-\exp[\beta'(Z_j - Z_i)]}{1 - \exp[\beta'(Z_j - Z_i)]} \log \frac{1 + \exp(b + \beta'Z_i]}{1 + \exp[a + \beta'Z_i]}\right.$$

$$\left. + \frac{1}{1 - \exp[\beta'(Z_j - Z_i)]} \log \frac{1 + \exp[b + \beta'Z_j]}{1 + \exp[a + \beta'Z_j]}\right\}$$

$$- \left(\frac{1}{b-a}\right)^2 \log \frac{1 + \exp[b + \beta'Z_i]}{1 + \exp[a + \beta'Z_i]} \log \frac{1 + \exp[b + \beta'Z_j]}{1 + \exp[a + \beta'Z_j]}$$

$$= \frac{-\exp[\beta'(Z_j - Z_i)]}{1 - \exp[\beta'(Z_j - Z_i)]} \Pr(Y_i=1) + \frac{1}{1 - \exp[\beta'(Z_j - Z_i)]}\Pr(Y_j=1)$$

$$- \Pr(Y_i=1)\Pr(Y_j=1) \qquad (2.6.10)$$

2.7 The Likelihood Function

2.7.1 Mathematical Background

Suppose that there are n_ℓ pregnancies within the ℓ^{th} family and that k_ℓ of these result in adverse outcomes, where $0 \leq k_\ell \leq n_\ell$ for $\ell = 1,2,\ldots,N$. The observations are ordered such that $Y_1,\ldots,Y_{k_\ell}$ had been adversely affected and $Y_{k_\ell+1},\ldots,Y_{n_\ell}$ were not affected. Then, assuming conditional independence, the likelihood function for the ℓ^{th} family is

$$L_\ell = \Pr(Y_1=1,Y_2=1,\ldots,Y_{k_\ell}=1,Y_{k_\ell+1}=0,\ldots,Y_{n_\ell}=0)$$

$$= \int_a^b \Pr(Y_1=1|\beta_0^*)\Pr(Y_2=1|\beta_0^*)\ldots\Pr(Y_{k_\ell}=1|\beta_0^*)\Pr(Y_{k_\ell+1}=0|\beta_0^*)\ldots\Pr(Y_{n_\ell}=0|\beta_0^*)$$

$$\times f_{\beta_0^*}(\beta_0^*)d\beta_0^*$$

(2.7.1)

The overall likelihood for the study population is the product of all family likelihoods, L_ℓ, $\ell = 1,2,\ldots,N$.

$$L = \prod_{\ell=1}^{N} L_\ell$$

For

$$\Pr(Y_j = 1 | \beta_0^*) = \frac{1}{1 + \exp[-(\beta_0^* + \underset{\sim}{\beta}'(\underset{\sim}{X}_j - \underset{\sim}{X}^*))]}$$

and β_0^* having a uniform distribution, U(a,b), the likelihood for family ℓ, not conditional upon β_0^*, is

$$L_\ell = \int_a^b \frac{\exp(\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_1]}{1 + \exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_i]} \cdots \frac{\exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_{k_\ell}]}{1 + \exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_{k_\ell}]} \frac{1}{1 + \exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_{k_\ell+1}]}$$

$$\cdots \frac{1}{1 + \exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_{n_\ell}]} \frac{1}{b-a} d\beta_0^*$$

$$= \frac{1}{b-a} \int_a^b \frac{\prod_{i=1}^{k_\ell} \exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_i]}{\prod_{j=1}^{n_\ell} (1 + \exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_j])} d\beta_0^*. \qquad (2.7.2)$$

This is easier to integrate if it can be written as a partial fraction expansion. The partial fraction expansion is based on the following.

Consider the rational function, $g(t) = \frac{\phi(t)}{f(t)}$ where $\phi(t)$ is a polynomial of degree less than the degree of $f(t)$. If $f(t)$ can be factored as

$$f(t) = (t-a)(t-b) \cdots (t-m)$$

where $a,b,\ldots,m$ are distinct, real roots of $f(t)$, then $g(t)$ can be written as

$$g(t) = \frac{\phi(t)}{f(t)} = \frac{A}{t-a} + \frac{B}{t-b} + \cdots + \frac{M}{t-m} \qquad (2.7.3)$$

where $A = \frac{\phi(a)}{f'(a)}$, $B = \frac{\phi(b)}{f'(b)}$, ... , $M = \frac{\phi(m)}{f'(m)}$

(Gradshteyn and Ryzik [1980])

Making the substitutions:

$$c_j = \exp[\beta' Z_j]$$

$$t = \exp[\beta_0^*]$$

$$\beta_0^* = \log t$$

$$d\beta_0^* = \frac{1}{t}\, dt.$$

Then

$$\frac{1}{b-a}\int_a^b \frac{\prod_{i=1}^{k_\ell} \exp[\beta_0^* + \beta' Z_i]}{\prod_{j=1}^{n_\ell} (1 + \exp[\beta_0^* + \beta' Z_j])}\, d\beta_0^* = \frac{1}{b-a}\int_{e^a}^{e^b} \frac{\prod_{i=1}^{k_\ell}(c_i t)}{\prod_{j=1}^{n_\ell}(1 + c_j t)} \frac{1}{t}\, dt \qquad (2.7.4)$$

2.7.2 Case 1. $k_\ell = 0$

$$\frac{1}{b-a}\int_{e^a}^{e^b} \frac{\prod_{i=1}^{k_\ell}(c_i t)}{\prod_{j=1}^{n_\ell}(1 + c_j t)t}\, dt = \frac{1}{b-a}\int_{e^a}^{e^b} \frac{1}{t\prod_{j=1}^{n_\ell}(1 + c_j t)}\, dt$$

$$= \frac{1}{b-a}\int_{e^a}^{e^b} \frac{1}{t\prod_{j=1}^{n_\ell} c_j (t-(-\frac{1}{c_j}))}\, dt \qquad (2.7.5)$$

The integrand $\frac{1}{t\prod_{j=1}^{n_\ell} c_j (t-(-\frac{1}{c_j}))}$ has the form $\frac{\phi(t)}{f(t)}$ where $\phi(t) = \frac{1}{\prod_{j=1}^{n_\ell} c_j}$

and $f(t) = t\prod_{j=1}^{n_\ell} (t-(-\frac{1}{c_j}))$.

The roots of $f(t)$ are $0, \frac{-1}{c_1}, \frac{-1}{c_2}, \ldots, \frac{-1}{c_{n_\ell}}$ and

$$f'(t) = 1 \cdot \prod_{j=1}^{n_\ell} \left(t-\left(-\frac{1}{c_j}\right)\right) + \sum_{h=1}^{n_\ell} t \prod_{\substack{j=1 \\ j\neq h}}^{n_\ell} \left(t-\left(-\frac{1}{c_j}\right)\right).$$

Thus

$$\frac{1}{t\prod_{j=1}^{n_\ell} c_j\left(t-\left(-\frac{1}{c_j}\right)\right)} = \frac{1/\prod_{j=1}^{n_\ell} c_j}{\prod_{j=1}^{n_\ell}\left(\frac{1}{c_j}\right) + 0}\,\frac{1}{t}$$

$$+ \sum_{m=1}^{n_\ell} \frac{1/\prod_{j=1}^{n_\ell} c_j}{\prod_{j=1}^{n_\ell}\left(\left(-\frac{1}{c_m}\right)-\left(\frac{-1}{c_j}\right)\right) + \sum_{h=1}^{n_\ell}\left(\frac{-1}{c_m}\right)\prod_{\substack{j=1 \\ j\neq h}}^{n_\ell}\left(\left(\frac{-1}{c_m}\right)-\left(\frac{-1}{c_j}\right)\right)} \frac{1}{t-\left(\frac{-1}{c_m}\right)}$$

$$= \frac{1}{t} + \sum_{m=1}^{n_\ell} \frac{1/\prod_{j=1}^{n_\ell} c_j}{\prod_{j=1}^{n_\ell}\left(\left(\frac{-1}{c_m}\right)-\left(\frac{-1}{c_j}\right)\right) + \sum_{h=1}^{n_\ell}\left(\frac{-1}{c_m}\right)\prod_{\substack{j=1 \\ j\neq h}}^{n_\ell}\left(\left(\frac{-1}{c_m}\right)-\left(\frac{-1}{c_j}\right)\right)} \frac{1}{t-\left(\frac{-1}{c_m}\right)}$$

(2.7.6)

Note that $\prod_{j=1}^{n_\ell} \frac{c_m-c_j}{c_m c_j} = 0$ for all m and that $\prod_{\substack{j=1 \\ j\neq h}}^{n_\ell} \frac{c_m-c_j}{c_m c_j} = 0$ unless $h = m$.

Therefore, equation (2.7.6) reduces to

$$\frac{1}{t} + \sum_{m=1}^{n_\ell} \frac{-1/\prod_{j=1}^{n_\ell} c_j}{\frac{1}{c_m}\prod_{\substack{j=1 \\ j\neq m}}^{n_\ell}\frac{1}{c_j}\prod_{\substack{j=1 \\ j\neq m}}^{n_\ell}\frac{c_m-c_j}{c_m}} \frac{c_m}{1+c_m t}$$

$$= \frac{1}{t} - \sum_{m=1}^{n_\ell} \frac{c_m^{n_\ell}}{\prod_{\substack{j=1 \\ j\neq h}}^{n_\ell}(c_m - c_j)} \frac{1}{1+c_m t}$$

Thus, for $k_\ell = 0$, the likelihood factor from family ℓ is

$$L_\ell = \frac{1}{b-a}\int_a^b \frac{1}{\prod_{j=1}^{n_\ell}(1+\exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_j])}\, d\beta_0^*$$

$$= \frac{1}{b-a}\int_{e^a}^{e^b}\left\{\frac{1}{t} - \sum_{m=1}^{n_\ell}\frac{c_m^{n_\ell}}{\prod_{\substack{j=1\\ j\neq m}}^{n_\ell}(c_m-c_j)}\,\frac{1}{1+c_m t}\right\} dt$$

$$= \frac{1}{b-a}\left\{\log t - \sum_{m=1}^{n_\ell}\frac{c_m^{n_\ell}}{\prod_{\substack{j=1\\ j\neq m}}^{n_\ell}(c_m-c_j)}\,\frac{1}{c_m}\log(1+c_m t)\right\}\Bigg|_{t=e^a}^{e^b}$$

$$= \frac{1}{b-a}\left\{(b-a) - \sum_{m=1}^{n_\ell}\frac{c_m^{n_\ell-1}}{\prod_{\substack{j=1\\ j\neq m}}^{n_\ell}(c_m-c_j)}\log\frac{1+e^b c_m}{1+e^a c_m}\right\}$$

$$= 1 - \frac{1}{b-a}\sum_{m=1}^{n_\ell}\frac{(\exp[\beta'\underset{\sim}{Z}_m])^{n_\ell-1}}{\prod_{\substack{j=1\\ j\neq m}}^{n_\ell}\{\exp[\beta'\underset{\sim}{Z}_m]-\exp[\beta'\underset{\sim}{Z}_j]\}}\log\frac{1+\exp[b+\beta'\underset{\sim}{Z}_m]}{1+\exp[a+\beta'\underset{\sim}{Z}_m]} \qquad (2.7.8)$$

2.7.3 Case 2. $1 \le k_\ell \le n_\ell$

$$\frac{1}{b-a}\int_a^b \frac{\prod_{i=1}^{k_\ell}\exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_i]}{\prod_{j=1}^{n_\ell}(1+\exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_j])}\, d\beta_0^* = \frac{1}{b-a}\int_{e^a}^{e^b}\frac{\prod_{i=1}^{k_\ell}(c_i t)}{\prod_{j=1}^{n_\ell}(1+c_j t)}\,\frac{dt}{t}$$

$$= \frac{1}{b-a}\int_{e^a}^{e^b}\frac{t^{k_\ell-1}\prod_{i=1}^{k_\ell}c_i}{\prod_{j=1}^{n_\ell}(1+c_j t)}\, dt \qquad (2.7.9)$$

The integrand, $\dfrac{t^{k_\ell-1}\prod_{i=1}^{k_\ell} C_i}{\prod_{j=1}^{n_\ell}(1+c_j t)}$, is of the form $\dfrac{\phi(t)}{f(t)}$ where

$$\phi(t) = \frac{t^{k_\ell-1}\prod_{i=1}^{k_\ell} c_i}{\prod_{j=1}^{n_\ell}(1+c_j t)} \text{ and } f(t) = \prod_{j=1}^{n_\ell}(t-(\tfrac{-1}{c_j})).$$

The distinct, real roots of f(t) are $\frac{-1}{c_1}, \frac{-1}{c_2}, \ldots, \frac{-1}{c_{n_\ell}}$ and by equation (2.7.3),

$$\frac{t^{k_\ell-1}\prod_{i=1}^{k_\ell} C_i}{\prod_{j=1}^{n_\ell}(1+c_j t)} = \sum_{m=1}^{n_\ell} \frac{\phi(\frac{-1}{c_m})}{f'(\frac{-1}{c_m})} \frac{1}{t-(\frac{-1}{c_m})}$$

$$= \sum_{m=1}^{n_\ell} \frac{(\frac{-1}{c_m})^{k_\ell-1}(\prod_{i=1}^{k_\ell} c_i)/(\prod_{j=1}^{n_\ell} c_j)}{\sum_{h=1}^{n_\ell}\prod_{\substack{j=1\\ j\neq h}}^{n_\ell}((\frac{-1}{c_m})-(\frac{-1}{c_j}))} \frac{1}{t-(\frac{-1}{c_m})}$$

$$= \sum_{m=1}^{n_\ell} \frac{(-1)^{k_\ell-1}(\frac{1}{c_m})^{k_\ell-1}\prod_{i=1}^{k_\ell} c_i}{(\prod_{j=1}^{n_\ell} c_j)\sum_{h=1}^{n_\ell}\prod_{\substack{j=1\\ j\neq h}}^{n_\ell}(\frac{c_m-c_j}{c_m c_j})} \frac{c_m}{1+c_m t}. \qquad (2.7.10)$$

Note that $\prod_{\substack{j=1\\ j\neq h}}^{n_\ell} \frac{c_m-c_j}{c_m c_j} = 0$ unless $h = m$. Therefore, equation (2.7.10) reduces to

$$\sum_{m=1}^{n_\ell} \frac{(-1)^{k_\ell-1}(\frac{1}{c_m})^{k_\ell-1}\prod_{i=1}^{k_\ell} c_i}{(\prod_{j=1}^{n_\ell} c_j)(\prod_{\substack{j=1\\ j\neq m}}^{n_\ell}(c_m-c_j))(\prod_{\substack{j=1\\ j\neq m}}^{n_\ell}\frac{1}{c_j})(\frac{1}{c_m})^{n_\ell-1}} \frac{c_m}{1+c_m t}$$

$$= \sum_{m=1}^{n_\ell} \frac{(-1)^{k_\ell-1}\left(\prod_{i=1}^{k_\ell} c_i\right)c_m^{n_\ell-k_\ell-1}}{c_m \prod_{\substack{j=1\\ j\neq m}}^{n_\ell}(c_m-c_j)} \frac{1}{1+c_m t} \qquad (2.7.11)$$

Thus, for $1 \le k_\ell \le n_\ell$, the likelihood factor from family ℓ is

$$L_\ell = \frac{1}{b-a}\int_a^b \frac{\prod_{i=1}^{k_\ell}(\exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_i])}{\prod_{j=1}^{n_\ell}(1+\exp[\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z}_j])}\, d\beta_0^*$$

$$= \frac{1}{b-a}\int_{e^a}^{e^b} \sum_{m=1}^{n_\ell} \frac{(-1)^{k_\ell-1}\left(\prod_{i=1}^{k_\ell} c_i\right)c_m^{n_\ell-k_\ell}}{\prod_{\substack{j=1\\ j\neq m}}^{n_\ell}(c_m-c_j)} \frac{1}{1+c_m t}\, dt$$

$$= \frac{1}{b-a}\sum_{m=1}^{n_\ell} \frac{(-1)^{k_\ell-1}\left(\prod_{i=1}^{k_\ell} c_i\right)c_m^{n_\ell-k_\ell}}{\prod_{\substack{j=1\\ j\neq m}}^{n_\ell}(c_m-c_j)} \frac{1}{c_m}\log\frac{1+e^b c_m}{1+e^a c_m}$$

$$= \frac{1}{b-a}\sum_{m=1}^{n_\ell} \frac{(-1)^{k_\ell-1}\left(\prod_{i=1}^{k_\ell}\exp[\underset{\sim}{\beta}'\underset{\sim}{Z}_i]\right)(\exp[\underset{\sim}{\beta}'\underset{\sim}{Z}_m])^{n_\ell-k_\ell-1}}{\prod_{\substack{j=1\\ j\neq m}}^{n_\ell}\{\exp[\underset{\sim}{\beta}'\underset{\sim}{Z}_m] - \exp[\underset{\sim}{\beta}'\underset{\sim}{Z}_j]\}}\log\frac{1+\exp[b+\underset{\sim}{\beta}'\underset{\sim}{Z}_m]}{1+\exp[a+\underset{\sim}{\beta}'\underset{\sim}{Z}_m]}$$

$$= \frac{(-1)^{k_\ell-1}\prod_{i=1}^{k_\ell}\exp[\underset{\sim}{\beta}'\underset{\sim}{Z}_i]}{b-a}\sum_{m=1}^{n_\ell}\frac{(\exp[\underset{\sim}{\beta}'\underset{\sim}{Z}_m])^{n_\ell-k_\ell-1}}{\prod_{\substack{j=1\\ j\neq m}}^{n_\ell}\{\exp[\underset{\sim}{\beta}'\underset{\sim}{Z}_m]-\exp[\underset{\sim}{\beta}'\underset{\sim}{Z}_j]\}}\log\frac{1+\exp[b+\underset{\sim}{\beta}'\underset{\sim}{Z}_m]}{1+\exp[a+\underset{\sim}{\beta}'\underset{\sim}{Z}_m]}\,. \qquad (2.7.12)$$

2.7.4 The Combined Form of the Likelihood Function

The results of subsections 2.7.2 and 2.7.3 can be combined into one expression for the likelihood function. Let

$$\Delta_{k_\ell} = \begin{cases} 1 & \text{if } k_\ell = 0 \\ 0 & \text{if } 1 \le k_\ell \le n_\ell. \end{cases}$$

Then the likelihood for family ℓ is:

$$L_\ell = \Delta_{k_\ell} + \frac{(-1)^{k_\ell - 1} \prod_{i=1}^{k_\ell} \exp[\beta' Z_i]}{b-a} \sum_{m=1}^{n_\ell} \frac{(\exp[\beta' Z_m])^{n_\ell - k_\ell - 1}}{\prod_{\substack{j=1 \\ j \neq m}}^{n_\ell} \{\exp[\beta' Z_m] - \exp[\beta' Z_j]\}}$$

$$\log \frac{1 + \exp[b + \beta' Z_m]}{1 + \exp[a + \beta' Z_m]} \qquad (2.7.13)$$

The likelihood for the study population is

$$L = \prod_{\ell=1}^{N} L_\ell$$

The coefficients, β, can now be estimated using maximum likelihood estimation techniques.

2.8 Invariance of β to the Choice of the 'Low Risk' Vector, X^*

Let $Z_j = X_j - X^*$ for some 'low risk' vector, X^*. The likelihood for the ℓ^{th} family in terms of the Z_j's, β, a, and b is

$$L_\ell = L_\ell(Z, \beta, a, b) = \int_a^b \frac{\prod_{i=1}^{k} \exp[\beta_0^* + \beta' Z_i]}{\prod_{j=1}^{n} (1 + \exp[\beta_0^* + \beta' Z_j])} \frac{1}{b-a} d\beta_0^*.$$

Let $W_j = Z_j - c = X_j - (X^* + c)$ for arbitrary vector, c. Then $\beta_0^{(c)} = \beta_0 + \beta'(X^* + c) = (\beta_0 + \beta' X^*) + \beta' c = \beta_0^* + \beta' c$ where β_0 is the

constant term from a logistic model. Therefore, if $\beta_0^* \sim U(a,b)$ then $\beta_0^{(c)} \sim U(a+\underset{\sim}{\beta}'\underset{\sim}{c},\ b+\underset{\sim}{\beta}'\underset{\sim}{c}) = U(a^{(c)},b^{(c)})$. The likelihood for the ℓ^{th} family in terms of the $\underset{\sim}{W}_j$'s, $\underset{\sim}{\beta}$, $a^{(c)}$, and $b^{(c)}$ is

$$L_\ell = L_\ell(\underset{\sim}{W},\underset{\sim}{\beta},a^{(c)},b^{(c)}) = \int_{a^{(c)}}^{b^{(c)}} \frac{\prod_{i=1}^{k} \exp[\beta_0^{(c)}+\underset{\sim}{\beta}'\underset{\sim}{W}_i]}{\prod_{j=1}^{n} (1+\exp[\beta_0^{(c)}+\underset{\sim}{\beta}'\underset{\sim}{W}_j])} \frac{1}{b^{(c)}-a^{(c)}}\, d\beta_0^{(c)}.$$

It will be shown that $\underset{\sim}{\beta}$ maximizes $L_\ell(\underset{\sim}{Z},\underset{\sim}{\beta},a,b)$ and $L_\ell(\underset{\sim}{W},\underset{\sim}{\beta},a^{(c)},b^{(c)})$.

Since $\underset{\sim}{W}_j = \underset{\sim}{Z}_j - \underset{\sim}{c}$, $\underset{\sim}{Z}_j = \underset{\sim}{W}_j + \underset{\sim}{c}$.

$$L_\ell(\underset{\sim}{Z},\underset{\sim}{\beta},a,b) = \int_a^b \frac{\prod_{i=1}^{k} \exp[\beta_0^*+\underset{\sim}{\beta}'(\underset{\sim}{W}_i+\underset{\sim}{c})]}{\prod_{j=1}^{n} (1+\exp[\beta_0^*+\underset{\sim}{\beta}'(\underset{\sim}{W}_j+\underset{\sim}{c})])} \frac{1}{b-a}\, d\beta_0^*$$

$$= \Delta_k + \frac{(-1)^{k-1} \prod_{i=1}^{k} \exp[\underset{\sim}{\beta}'(\underset{\sim}{W}_i+\underset{\sim}{c})]}{b-a} \times \sum_{m=1}^{n} \frac{\{\exp[\underset{\sim}{\beta}'(\underset{\sim}{W}_m+\underset{\sim}{c})]\}^{n-k-1}}{\prod_{\substack{j=1 \\ j\neq m}}^{n} \{\exp[\underset{\sim}{\beta}'(\underset{\sim}{W}_m+\underset{\sim}{c})]-\exp[\underset{\sim}{\beta}'(\underset{\sim}{W}_j+\underset{\sim}{c})]\}} \times \log \frac{1+\exp[b+\underset{\sim}{\beta}'(\underset{\sim}{W}_m+\underset{\sim}{c})]}{1+\exp[a+\underset{\sim}{\beta}'(\underset{\sim}{W}_m+\underset{\sim}{c})]}$$

$$= \Delta_k + \frac{(-1)^{k-1} \exp[k\underset{\sim}{\beta}'\underset{\sim}{c}] \prod_{i=1}^{k} \exp[\underset{\sim}{\beta}'\underset{\sim}{W}_i]}{b-a} \times \sum_{m=1}^{n} \frac{\exp[\underset{\sim}{\beta}'\underset{\sim}{c}(n-k-1)]\{\exp[\underset{\sim}{\beta}'\underset{\sim}{W}_m]\}^{n-k-1}}{\prod_{\substack{j=1 \\ j\neq m}}^{n} \exp[\underset{\sim}{\beta}'\underset{\sim}{c}]\{\exp[\underset{\sim}{\beta}'\underset{\sim}{W}_m]-\exp[\underset{\sim}{\beta}'\underset{\sim}{W}_j]\}}$$

$$\times \log \frac{1 + \exp[b+\beta'c+\beta'W_m]}{1 + \exp[a+\beta'c+\beta'W_m]}$$

$$= \Delta_k + \frac{(-1)^{k-1} \prod_{i=1}^{k} \exp[\beta'W_i] \exp[k\beta'c]\exp[\beta'c(n-k-1)]}{b-a} \frac{}{\exp[\beta'c(n-1)]}$$

$$\times \sum_{m=1}^{n} \frac{\{\exp[\beta'W_m]\}^{n-k-1}}{\prod_{\substack{j=1 \\ j\neq m}}^{n} \{\exp[\beta'W_m]-\exp[\beta'W_j]\}} \log \frac{1 + \exp[b+\beta'c+\beta'W_m]}{1 + \exp[a+\beta'c+\beta'W_m]}$$

$$= \Delta_k + \frac{(-1)^{k-1} \prod_{i=1}^{k} \exp[\beta'W_i]}{(b+\beta'c) - (a+\beta'c)} \sum_{m=1}^{n} \frac{\{\exp[\beta'W_m]\}^{n-k-1}}{\prod_{\substack{j=1 \\ j\neq m}}^{n} \{\exp[\beta'W_m] - \exp[\beta'W_j]\}}$$

$$\times \log \frac{1 + \exp[(b+\beta'c)+\beta'W_m]}{1 + \exp[(a+\beta'c)+\beta'W_m]}$$

$$= \int_{a+\beta'c}^{b+\beta'c} \frac{\prod_{i=1}^{k} \exp[\beta_0^{(c)}+\beta'W_i]}{\prod_{j=1}^{n} (1+\exp[\beta_0^{(c)}+\beta'W_j])} \frac{1}{(b+\beta'c) - (a+\beta'c)} d\beta_0^{(c)}$$

$$= \int_{a^{(c)}}^{b^{(c)}} \frac{\prod_{i=1}^{k} \exp[\beta_0^{(c)}+\beta'W_i]}{\prod_{j=1}^{n} (1 + \exp[\beta_0^{(c)}+\beta'W_j])} \frac{1}{b^{(c)}-a^{(c)}} d\beta_0^{(c)}$$

where $\beta_0^{(c)} \sim U(a^{(c)},b^{(c)})$

$$= L_\ell(W,\beta,a^{(c)},b^{(c)}).$$

$L_\ell(Z,\beta,a,b) = L_\ell(W,\beta,a^{(c)},b^{(c)})$, so if β is a maximum likelihood estimator of $L_\ell(Z,\beta,a,b)$, it is also a maximum likelihood estimator of $L_\ell(W,\beta,a^{(c)},b^{(c)})$; β is invariant to the choice of X^*.

CHAPTER III

AN APPLICATION OF THE UNIFORM-LOGISTIC MODEL TO REAL DATA

3.1 Introduction

In Chapter II, a uniform-logistic model was developed. In this chapter, the model will be applied to data from an occupational pregnancy history study. The estimates of the uniform-logistic model will be compared to those of the logistic model. The uniform-logistic model provides for covariance between events within the same family, whereas the logistic model assumes mutual independence of all events. For this reason, it is expected that the uniform-logistic model should be better suited for pregnancy history data than the usual logistic model.

In Section 3.2, the analysis strategy is briefly discussed. The population to be analyzed is described in Section 3.3. Section 3.4 discusses the results of a stratified analysis on the data. Section 3.5 presents the modeling results for both the uniform-logistic model and the logistic model. This modeling was performed treating the exposure as a continuous variable. The goodness of fit of the model is also evaluated. Modeling was repeated in Section 3.6, dichotomizing the exposure variable as no and any exposure. An interpretation of the final model is presented in Section 3.7. In Section 3.8, qualifications on the results of the uniform-logistic model are discussed.

3.2 Brief Description of the Analysis Strategy

The ensuing analysis follows the strategy recommended by Kleinbaum, Kupper and Chambless (1981). They advocate the use of stratified analysis in conjunction with mathematical modeling. The purpose of the stratified analysis is to identify potential confounders and effect modifiers. The mathematical modeling discussed in particular by Kleinbaum et al. (1981) was logistic regression. Estimated odds ratios can be found from the parameter estimates of logistic regression models. The uniform-logistic model is similar to the logistic model, so it as assumed that the same modeling strategy and odds ratio estimation will also apply to the uniform-logistic model.

The variable selection strategy used in the modeling stage is described in detail by Kleinbaum et al. (1981). This strategy is summarized as follows. First, a set of risk factors potentially associated with the response is identified. Next, a model containing these factors (or functions of them) as main effects and as many first and second order interactions of the main effects with exposure as possible without resulting in multicollinearity is fit to the data. Then, in a series of forward and/or backward steps, interaction terms are added or removed from the model. After there are only significant interaction terms remaining in the model, the main effects not involved in the interactions are removed sequentially to increase the precision of the estimated parameters. The process is stopped when removal of any more main effects would change the coefficients of the variables remaining in the model, in particular, the coefficients of the exposure and interaction terms. In a logistic model, such a change in the coefficients of the exposure and interaction terms would result in a

change in the estimated odds ratios. Thus, reduction of the model is terminated just before the estimated odds ratios become unstable and no longer valid. This strategy aims for a valid estimate of the odds ratio which is as precise (i.e., with the smallest variance) as possible.

3.3 Description of the Population Analyzed

Part of a population of workers exposed to substance X in the workplace is analyzed here. It was suspected that this substance was associated with adverse outcomes in pregnancies occurring to the couples. Consequently, a survey of reproductive outcomes of these couples was carried out. One response variable of interest was fetal loss. This outcome was dichotomized into whether the pregnancy terminated as a live birth or as a fetal loss. Here, a fetal loss is defined as any pregnancy in which the fetus did not survive until live birth. Thus, the term, "fetal loss," includes what are commonly referred to as miscarriages and stillbirths. The independent variables used in this analysis are exposure, gravidity, mother's age at pregnancy, race, occurrence of prior fetal loss, and mother's smoking and drinking habits during pregnancy.

The population analyzed consisted of 248 white pregnancies from 146 families. Of these 248 pregnancies, there were 37 fetal losses and 211 live births. Proportions of fetal losses occurring at different levels of the independent variables are presented in Figures 3.1-3.6.

FIGURE 3.1. PROPORTION OF FETAL LOSSES BY EXPOSURE

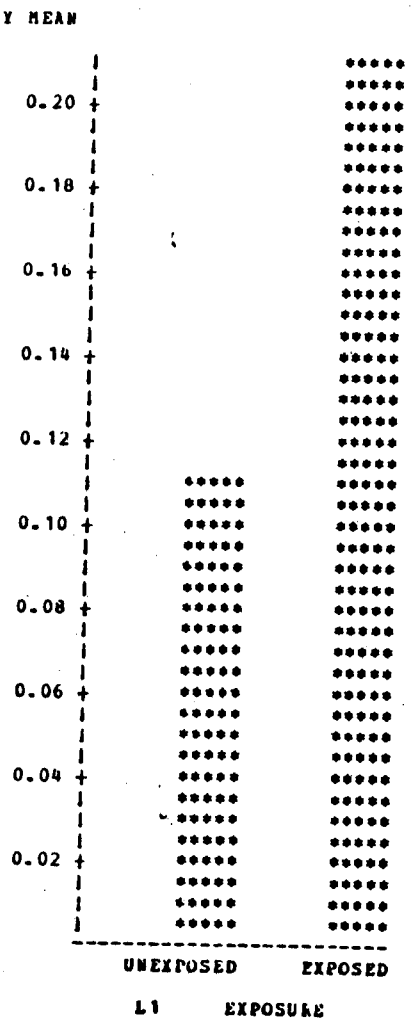


FIGURE 3.2. PROPORTION OF FETAL LOSSES BY GRAVIDITY WITHIN EXPOSURE LEVEL

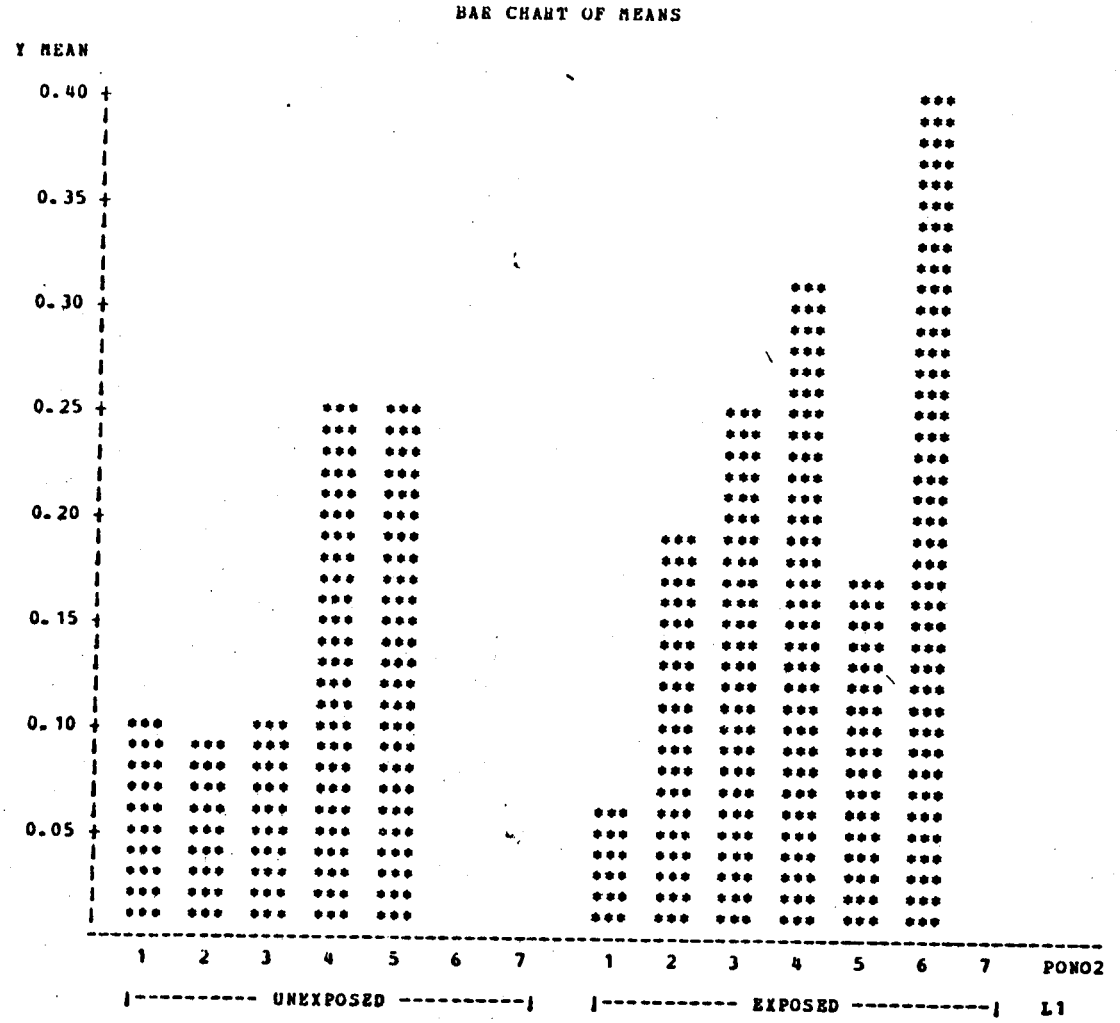


FIGURE 3.3. PROPORTION OF FETAL LOSSES BY MOTHER'S AGE AT PREGNANCY WITHIN EXPOSURE LEVEL

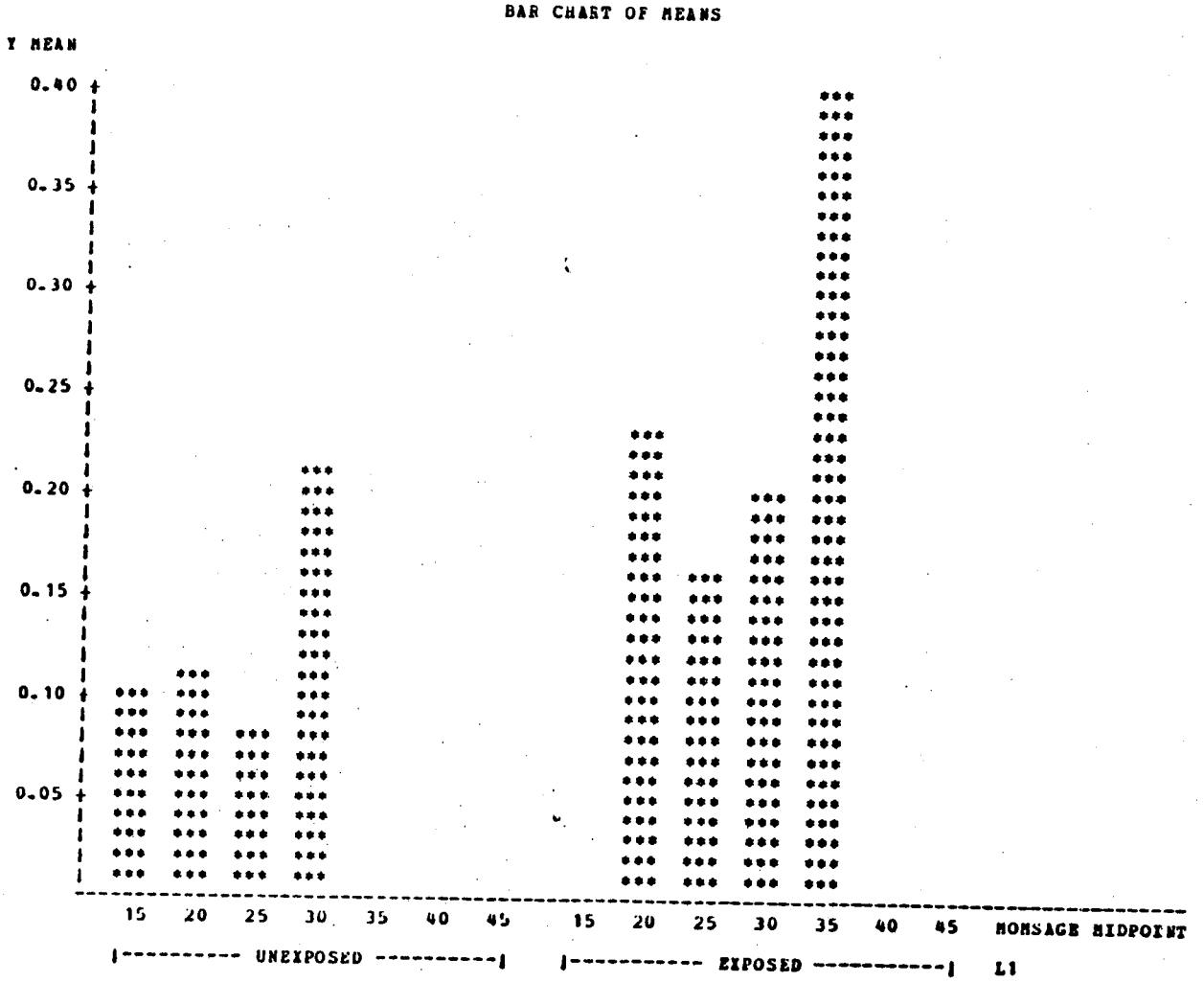


FIGURE 3.4. PROPORTION OF FETAL LOSSES BY PRIOR LOSS WITHIN EXPOSURE LEVEL

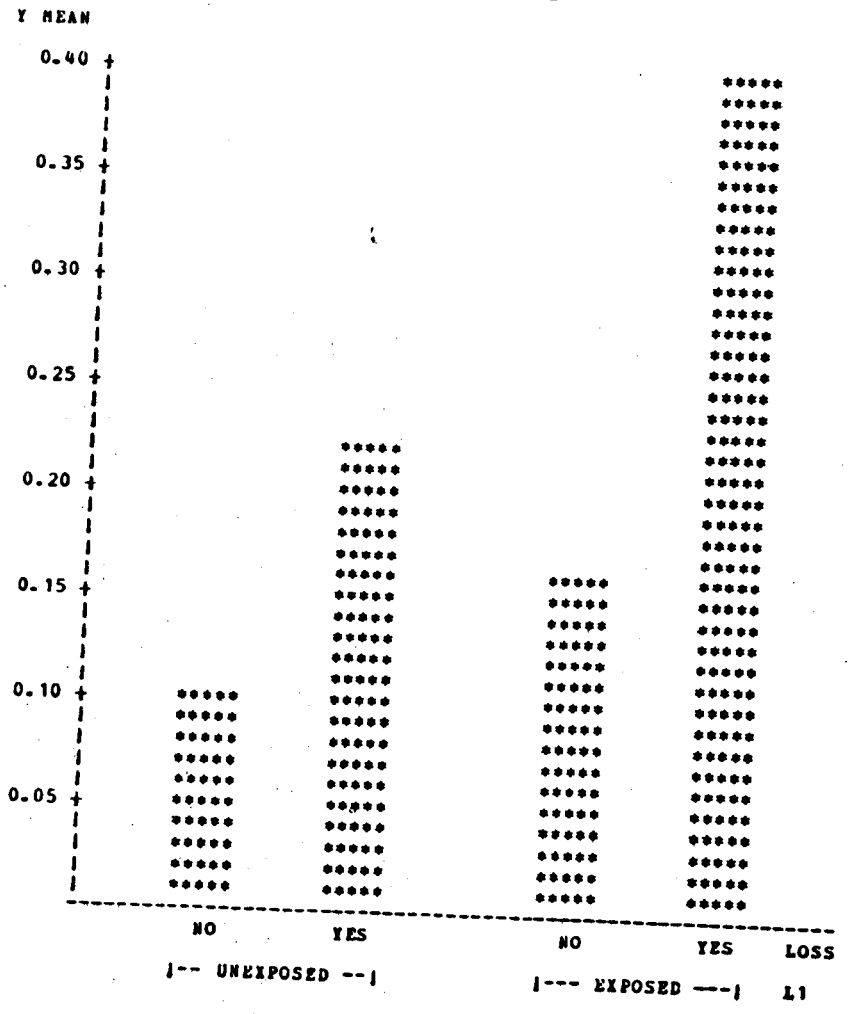


FIGURE 3.5. PROPORTION OF FETAL LOSSES BY SMOKING HABITS WITHIN EXPOSURE LEVEL

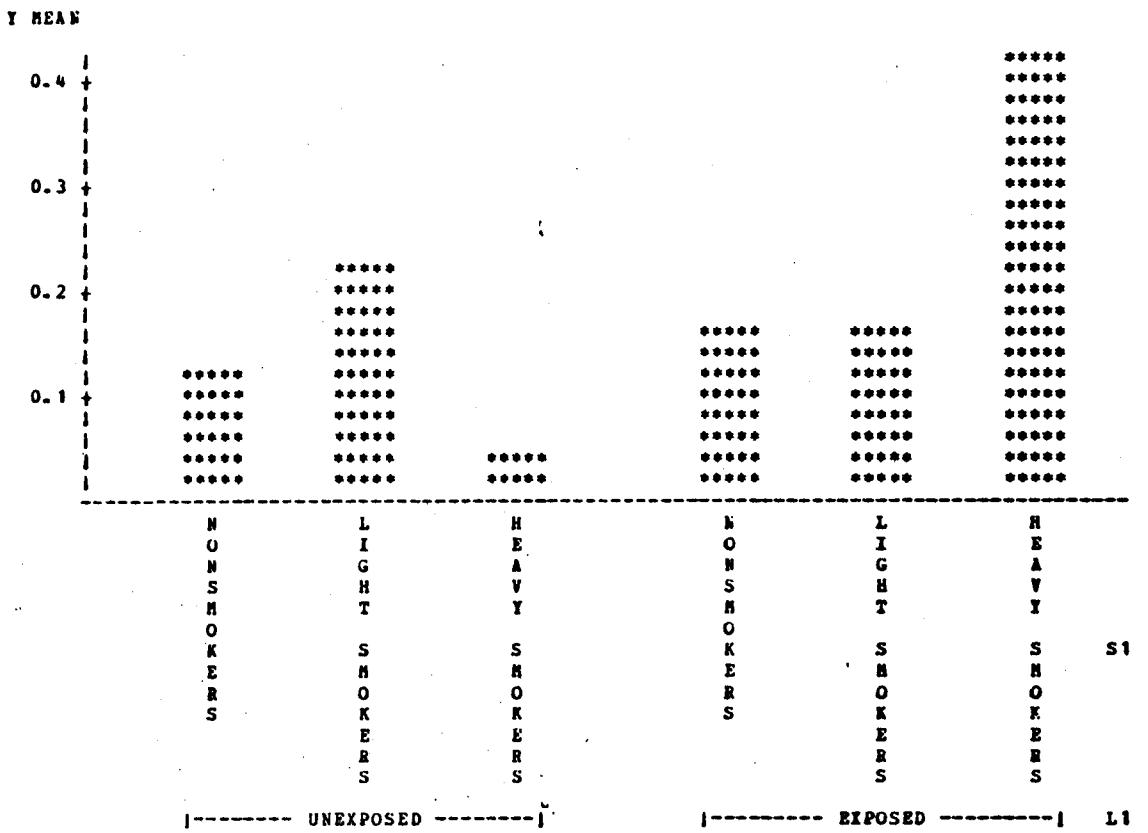
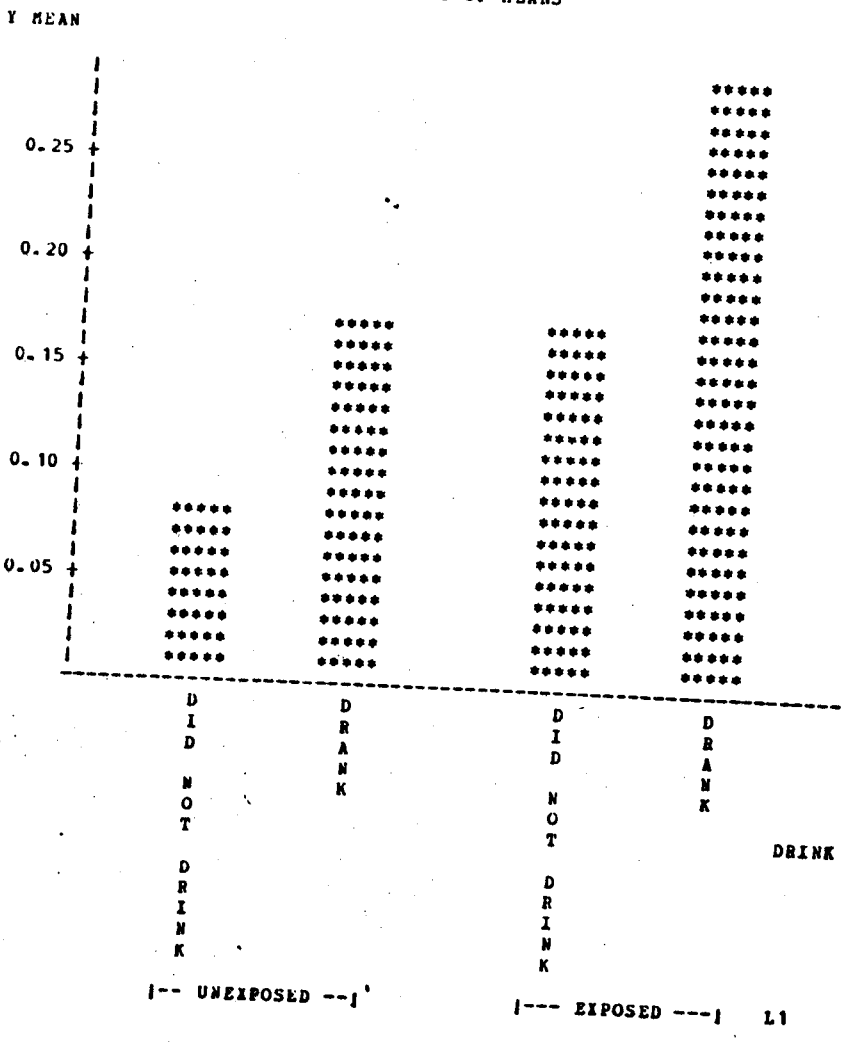


FIGURE 3.6. PROPORTION OF FETAL LOSSES BY DRINKING HABITS WITHIN EXPOSURE LEVEL



3.4 Results of a Stratified Analysis

An analysis of the association between exposure and pregnancy outcome, stratified by levels of the independent variables of interest is presented in Tables 3.1-3.6. In Table 3.1, the crude table, it appears that exposure approximately doubles the odds of fetal loss over the odds at no exposure, without controlling for the other factors which may affect pregnancy outcome. The estimated variances of the odds ratios are based on a Taylor Series approximation. These may not be good approximations for tables with small cell sizes. However, the stratified analysis was performed as a crude guide to later mathematical modeling strategies. Thus, the above 95% confidence intervals of the estimated odds ratios are not of great interest at this point.

It appears that the odds ratios are fairly uniform across strata of Gravidity, Mother's Age at Pregnancy, Prior Loss, and Mother's Drinking Habits during pregnancy. However, in Table 3.5, the odds ratios across strata of Mother's Smoking Habits during pregnancy were quite different. For smokers of one or more packs of cigarettes per day, the odds ratio was much larger than the odds ratios for nonsmokers and light smokers (less than one pack per day). The nonuniformity of odds ratios across strata of Smoking suggests that Smoking may be an effect modifier.

3.5 Results of the Modeling, Treating Exposure as a Continuous Variable

3.5.1 Definitions of the Variables

For this analysis, the variables were defined as follows:

$$Y_j = \begin{cases} 1 & \text{, if a fetal loss occurs at pregnancy } j \\ 0 & \text{, if no fetal loss occurs at pregnancy } j \end{cases}$$

TABLE 3.1

EXPOSURE BY FETAL LOSS

Overall Table

	Exposed	Not Exposed	Total
Fetal Loss	20	17	37
No Fetal Loss	75	136	211
Total	95	153	248

$\chi^2(1) = 4.5450$ p = .0165

OR = 2.1333 (1.054, 4.319)*

*95% confidence interval

TABLE 3.2

EXPOSURE BY FETAL LOSS, STRATIFIED BY GRAVIDITY

Stratified by:

Gravidity

1

	Exposed	Not Exposed	Total
Fetal Loss	1	7	8
No Fetal Loss	16	63	79
Total	17	70	87

OR = .562 (.064,4.907)*

$\chi^2(1)$ = .2745 p = .3001

2

	Exposed	Not Exposed	Total
Fetal Loss	4	4	8
No Fetal Loss	17	39	56
Total	21	43	64

OR = 2.294 (.513,10.265)*

$\chi^2(1)$ = 1.2060 p = .1361

3

	Exposed	Not Exposed	Total
Fetal Loss	7	2	9
No Fetal Loss	21	18	39
Total	28	20	48

OR = 3.0000 (.552,16.306)*

$\chi^2(1)$ = 1.6872 p = .0970

4+

	Exposed	Not Exposed	Total
Fetal Loss	8	4	12
No Fetal Loss	21	16	37
Total	29	20	49

OR = 1.5238 (.389,5.968)*

$\chi^2(1)$ = .36086 p = .2740

OR (adj.) = 1.7415 (.779,3.894)*

OR (crude) = 2.1333 (1.054,4.319)*

M-H $\chi^2(1)$ = 1.7674 p = .0919

*95% confidence interval (based on a Taylor Series approximation)

TABLE 3.3

EXPOSURE BY FETAL LOSS, STRATIFIED
BY MOTHER'S AGE AT PREGNANCY

Stratified by:

Mother's Age at Pregnancy

< 24 years old

	Exposed	Not Exposed	Total
Fetal Loss	9	11	20
No Fetal Loss	35	99	134
Total	44	110	154

OR = 2.3143
(.885,6.054)*

$\chi^2(1) = 3.0200$
p = .0411

≥ 24 years old

	Exposed	Not Exposed	Total
Fetal Loss	11	6	17
No Fetal Loss	40	37	77
Total	51	43	94

OR = 1.6958
(.570,5.047)*

$\chi^2(1) = .90352$
p = .17092

OR (adj.) = 2.0200 (.984,4.155)*

OR (crude) = 2.1333 (1.054,4.319)*

M-H $\chi^2(1)$ = 3.6256 p = .0284

*95% confidence interval (based on a Taylor Series approximation)

TABLE 3.4

EXPOSURE BY FETAL LOSS, STRATIFIED
BY OCCURRENCE OF PRIOR LOSS

Stratified by:

Prior Loss

No Prior Loss

	Exposed	Not Exposed	Total
Fetal Loss	12	13	25
No Fetal Loss	63	122	185
Total	75	135	210

OR = 1.7876
(.771,4.147)*

$\chi^2(1)$ = 1.8568
p = .0865

Prior Loss

	Exposed	Not Exposed	Total
Fetal Loss	8	4	12
No Fetal Loss	12	14	26
Total	20	18	38

OR = 2.3333
(.560,9.717)*

$\chi^2(1)$ = 1.3493
p = .1227

OR (adj.) = 1.9148 (.928,3.953)*

OR (crude) = 2.1333 (1.054,4.319)*

M-H $\chi^2(1)$ = 3.1486 p = .0380

*95% confidence interval (based on a Taylor Series approximation)

TABLE 3.5

EXPOSURE BY FETAL LOSS, STRATIFIED BY MOTHER'S SMOKING HABITS DURING PREGNANCY

Stratified by:

Smoking

Nonsmoking

	Exposed	Not Exposed	Total
Fetal Loss	10	12	22
No Fetal Loss	53	97	150
Total	63	109	172

OR = 1.5252 (.618,3.765)*

$\chi^2(1) = .8418$

p = .1794

Smoking
< 1 pack/day

	Exposed	Not Exposed	Total
Fetal Loss	2	4	6
No Fetal Loss	11	14	25
Total	13	18	31

OR = .6364 (.098,4.138)*

$\chi^2(1) = .2188$

p = .3200

Smoking
≥ 1 pack/day

	Exposed	Not Exposed	Total
Fetal Loss	8	1	9
No Fetal Loss	11	25	36
Total	19	26	45

OR = 18.182 (2.022,163.521)*

$\chi^2(1) = 9.8198$

p = .0009

OR (adj.) = 1.7789 (.829,3.815)*

OR (crude) = 2.1333 (1.054,4.319)*

M-H $\chi^2(1)$ = 4.2236 p = .0199

*95% confidence interval (based on a Taylor Series approximation)

TABLE 3.6

EXPOSURE BY FETAL LOSS, STRATIFIED BY MOTHER'S DRINKING HABITS DURING PREGNANCY

Stratified by:

Drinking

Did Not Drink

	Exposed	Not Exposed	Total
Fetal Loss	11	9	20
No Fetal Loss	53	97	150
Total	64	106	170

OR = 2.2369 (.872,5.740)*

$\chi^2(1) = 2.8906$ p = .0445

Did Drink

	Exposed	Not Exposed	Total
Fetal Loss	9	8	17
No Fetal Loss	22	39	61
Total	31	47	78

OR = 1.9943 (.673,5.910)*

$\chi^2(1) = 1.5607$ p = .1058

OR (adj.) = 2.1293 (1.045,4.339)*

OR (crude) = 2.1333 (1.054,4.319)*

M-H $\chi^2(1)$ = 4.4171 p = .0178

*95% confidence interval (based on a Taylor Series approximation)

$$E_j = \begin{cases} \text{the estimated exposure level, if exposed} \\ \text{the unexposed value, if not exposed} \end{cases}$$

$GRAV_j$ = Mother's gravidity* at pregnancy j

MA_j = Mother's age at the time of pregnancy j

$$LOSS_j = \begin{cases} 1 & \text{, if any fetal loss had occurred to the mother prior to pregnancy j} \\ 0 & \text{, if no fetal loss had occurred to the mother prior to pregnancy j} \end{cases}$$

$$Smoking_j = \begin{cases} \text{Nonsmokers at the time of pregnancy j} \\ \text{Light smokers, less than one pack per day} \\ \text{Heavy smokers, one of more packs per day} \end{cases}$$

Smoking was coded as two indicator variables.

$$SMK1_j = \begin{cases} 1 & \text{, if a light smoker at pregnancy j} \\ 0 & \text{, otherwise} \end{cases}$$

$$SMK2_j = \begin{cases} 1 & \text{, if a heavy smoker at pregnancy j} \\ 0 & \text{, otherwise} \end{cases}$$

$$DRNK_j = \begin{cases} 1 & \text{, if the mother was in the habit of drinking at the time of pregnancy j} \\ 0 & \text{, if the mother was not in the habit of drinking at the time of pregnancy j} \end{cases}$$

---

*Gravidity is defined as the number of pregnancies a woman has had, regardless of their outcomes.

In the development of the uniform-logistic model (Section 2.3), a "low-risk" vector, $\underset{\sim}{X}^*$, was subtracted from the observed vector of independent variables, $\underset{\sim}{X}_j$. This was done to give a theoretical interpretation to the uniform distribution imposed β_o^*. It is shown in Section 2.8 that $\underset{\sim}{\beta}$ is invariant to the choice of the "low risk" vector, $\underset{\sim}{X}^*$. In other words, the estimated effects of the independent variables, $\underset{\sim}{\beta}$, are not affected by the vector of "low risk" values of the independent variables, $\underset{\sim}{X}^*$, which are subtracted from the observed vector of independent variables, $\underset{\sim}{X}_j$.

For this analysis, the "low risk" vector subtracted from the vector of independent variables was

$$\underset{\sim}{X}^* = \begin{pmatrix} E^* \\ GRAV^* \\ MA^* \\ LOSS^* \\ SMK1^* \\ SMK2^* \\ DRNK^* \end{pmatrix} = \begin{pmatrix} \text{'unexposed value'} \\ 2.0 \\ 24.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix}$$

From the epidemiological literature, the lowest risk age at pregnancy seems to be about 24 years of age. Similarly, women of gravidity 2 seem to be at low risk. The absence of prior fetal loss, smoking, and drinking are also presumed to be associated with a low risk of fetal loss.

Under the uniform-logistic model (see Section 2.4), the distribution of the probability of fetal loss, R, in a population having the characteristics, $\underset{\sim}{X}^*$, has the form,

$$f_R(R) = \begin{cases} \frac{1}{b-a}\,\frac{1}{R(1-R)} & , \quad 0 < \frac{e^a}{1+e^a} < R < \frac{e^b}{1+e^b} < 1 \\ 0 & , \quad \text{otherwise.} \end{cases}$$

The parameters of the uniform distribution, a and b, depend on $\underset{\sim}{X}^*$, so another choice of a "low risk" vector, $\underset{\sim}{X}^*$, may lead to a different distribution of background risk of fetal loss associated with the new $\underset{\sim}{X}^*$. (It still has the form of $f_R(R)$.) However, the effects of the independent variables, $\underset{\sim}{\beta}$, remain the same regardless of the choice of $\underset{\sim}{X}^*$.

The parameters of the uniform-logistic model were estimated using maximum likelihood techniques. The program, MAXLIK (Kaplan and Elston [1972]), provided the likelihood maximization algorithms. Trial runs with the uniform-logistic model had indicated that estimates of the parameters from a logistic model were acceptable initial estimates of the parameters of the uniform-logistic model in the sense that they lead to a solution that maximized the likelihood function. Thus, initial estimates for the parameters of the uniform-logistic model were supplied by the maximum likelihood estimates of the parameters of the related logistic model.

3.5.2 Model Fitting

Following the modeling strategy of Kleinbaum, et al. (1981), a model containing main effects (Exposure, Gravidity, Prior Loss, Mother's Age at Pregnancy, Smoking, and Drinking habits) and first and second order interactions of the main effects with Exposure was first considered. This resulted in a singular information matrix due to an exact singularity. Next, a model containing main effects and first order interactions with Exposure was fit. No additional second order

interactions with Exposure were significant enough ($p < .20$) to enter this model. The estimated parameters and their standard errors for this model are presented in Table 3.7 (MODEL 1A). In this table, the columns of z-values are the parameter estimates divided by their standard errors. By the asymptotic normality properties of maximum likelihood estimates, each z asymptotically has a normal distribution. Z^2 is the usual statistic, $[\hat{\beta}_j^2/\hat{var}(\hat{\beta}_j)]$, which has a chi-square distribution with 1 degree of freedom, asymptotically. The probability of observing the estimated parameter or one more extreme in the same direction under the null hypothesis that the true parameter is 0 is listed in the column under "1-sided p." The first order interactions of all main effects with Exposure except Smoking x Exposure were insignificant, so they were dropped from the model, leading to MODEL 2A (Table 3.8). A likelihood ratio test of the contribution of these interactions supported their removal from the model ($\chi^2(4) = .90$, n.s.). The Smoking x Exposure interaction remained significant ($\chi^2(2) = 6.32$, $p = .04$). Next, the insignificant main effects were candidates for removal from the model. Since the Smoking x Exposure interaction was significant, the lower order terms, Smoking and Exposure were required to remain in the model. In addition, Mother's Age at Pregnancy was retained in all models. This was done to ensure that the vectors of independent variables would be distinct within each family. The requirement of distinct vectors of independent variables arose from the development of the uniform-logistic likelihood function in Section 2.7.1. The first main effect to be removed from the model was Gravidity ($p = .295$). MODEL 3A contains main effects except Gravidity and the Smoking x Exposure interaction. This did not change

TABLE 3.7

MODEL 1A: MAIN EFFECTS AND FIRST ORDER INTERACTIONS WITH EXPOSURE

	LOGISTIC				UNIFORM-LOGISTIC			
	MLE	SE	z	1-sided p	MLE	SE	z	1-sided p
$\hat{\beta}_0$	-2.5261	.4272	-5.913	< .0001				
					$\hat{a}$ -3.4818	0.0		
					$\hat{b}$ -2.2159	0.0		
E	.0248	.0353	.702	.241	.0241	0.0		
GRAV	-.0087	.2560	-.034	.486	-.0278	0.0		
LOSS	1.1485	.7380	1.556	.060	1.3049	0.0		
MA	.0178	.0759	.234	.406	.0209	0.0		
SMK1	.8246	.6655	1.239	.108	.8311	0.0		
SMK2	-1.2396	.9680	-1.281	.100	-1.3599	0.0		
DRNK	.8241	.5966	1.381	.084	.8286	0.0		
GRAV x E	-.0130	.0215	-.605	.272	-.0128	0.0		
LOSS x E	.0142	.0591	.240	.405	.0135	0.0		
MA x E	.0038	.0052	.731	.232	.0035	0.0		
SMK1 x E	-.0555	.0659	-.842	.200	-.0565	0.0		
SMK2 x Z	.1465	.0673	2.177	.015	.1454	0.0		
DRNK x E	-.0226	.0477	-.474	.318	-.0197	0.0		
$-2 \log \hat{L}$	184.69				184.70 (nearly singular information matrix)			

TABLE 3.8

MODEL 2A: MAIN EFFECTS AND SMOKING x EXPOSURE INTERACTION

	LOGISTIC				UNIFORM-LOGISTIC			
	MLE	SE	z	1-sided p	MLE	SE	z	1-sided p
$\hat{\beta}_0$	-2.4227	.3540	-6.844	< .0001				
$\hat{a}$					-2.9559	1.1972	-2.469	.007
$\hat{b}$					-1.9574	.5767	-3.394	.0003
E	.0185	.0267	.693	.245	.0185	.0270	.685	.246
GRAV	-.1165	.1841	-.633	.263	-.1074	.1997	-.538	.295
LOSS	1.1824	.5220	2.265	.012	1.1431	.7699	1.485	.069
MA	.0605	.0493	1.227	.110	.0596	.0496	1.202	.115
SMK1	.8674	.6551	1.324	.093	.8880	.6582	1.349	.089
SMK2	-1.1338	.9552	-1.187	.118	-1.1300	.9636	-1.173	.120
DRNK	.6302	.4212	1.496	.067	.6310	.4276	1.476	.070
SMK1 x E	-.0688	.0617	-1.115	.133	-.0693	.0618	-1.121	.131
SMK2 x E	.1280	.0632	2.025	.021	.1278	.0630	2.029	.021
$-2 \log \hat{L}$		185.57				185.60		

Likelihood Ratio Test for removing all interactions except Smoking x Exposure

$$-2 \log \frac{\hat{L}(\text{MODEL 2A})}{\hat{L}(\text{MODEL 1A})}$$

LOGISTIC: $= 185.57 - 184.69 = .88 \sim \chi^2(4)$ (n.s.)

UNIFORM-LOGISTIC: $= 185.60 - 184.70 = .90 \sim \chi^2(4)$ (n.s.)

TABLE 3.9

MODEL 3A: EXPOSURE, PRIOR LOSS, MOTHER'S AGE AT PREGNANCY, SMOKING, DRINKING, AND SMOKING x EXPOSURE

	LOGISTIC				UNIFORM-LOGISTIC			
	MLE	SE	z	1-sided p	MLE	SE	z	1-sided p
$\hat{\beta}_0$	-2.4339	.3549	-6.858	< .0001	$\hat{a}$ -2.7014	1.0391	-2.600	.005
					$\hat{b}$ -1.8368	.5586	-3.288	.0006
E	.0176	.0268	.657	.255	.0154	.0273	.564	.286
LOSS	1.0164	.4496	2.261	.012	.8944	.5939	1.506	.066
MA	.0445	.0428	1.040	.149	.0488	.0452	1.080	.140
SMK1	.9075	.6507	1.395	.082	.9192	.6644	1.384	.083
SMK2	-1.1298	.9498	-1.189	.117	-1.1039	.9688	-1.139	.127
DRNK	.6181	.4193	1.474	.070	.5896	.4368	1.350	.088
SMK1 x E	-.0714	.0616	-1.159	.123	-.0695	.0625	-1.112	.133
SMK2 x E	.1237	.0626	1.976	.024	.1296	.0651	1.991	.023
$-2 \log \hat{L}$	185.98				185.88			

Likelihood Ratio Test for removing Gravidity from MODEL 2A

$$-2 \log \frac{\hat{L}(\text{MODEL 3A})}{\hat{L}(\text{MODEL 2A})}$$

$= 185.98 - 185.57$
$= .41 \sim \chi^2(1)$ (n.s.)

$= 185.88 - 185.60$
$= .28 \sim \chi^2(1)$ (n.s.)

Range of the underlying risk of fetal loss, R.
$.0629 < R < .1374$

TABLE 3.10

MODEL 4A: EXPOSURE, PRIOR LOSS, MOTHER'S AGE AT PREGNANCY, SMOKING, SMOKING x EXPOSURE

	LOGISTIC				UNIFORM-LOGISTIC				
	MLE	SE	z	1-sided p		MLE	SE	z	1-sided p
$\hat{\beta}_0$	-2.2523	.3248	-6.934	< .0001	$\hat{a}$	-2.7474	1.2289	-2.236	.013
					$\hat{b}$	-1.8279	.6316	-2.894	.002
E	.0132	.0269	.491	.311		.0187	.0280	.668	.252
LOSS	1.0083	.4469	2.256	.012		.9586	.7157	1.339	.090
MA	.0526	.0422	1.246	.106		.0535	.0466	1.148	.125
SMK1	1.0434	.6438	1.621	.053		1.1763	.6872	1.712	.043
SMK2	-.8345	.9058	-.921	.178		-.7679	.9505	-.808	.209
SMK1 x E	-.0591	.0608	.972	.165		-.0656	.0632	-1.038	.150
SMK2 x E	.1235	.0614	2.011	.022		.1258	.0794	1.584	.056
$-2 \log \hat{L}$	188.10					187.82			

Likelihood Ratio Test for removing Drinking from MODEL 3A

$$-2 \log \frac{\hat{L}(\text{MODEL 4A})}{\hat{L}(\text{MODEL 3A})}$$

$= 188.10 - 185.98$
$= 2.12 \sim \chi^2(1)$
$p = .1454$

$= 187.82 - 185.88$
$= 1.94 \sim \chi^2(1)$
$p = .1637$

TABLE 3.11

ESTIMATES FROM THE LOGISTIC MODELS

	MODEL 1A	MODEL 2A	MODEL 3A	MODEL 4A
Estimated Coefficient of:				
E	.0248	.0185	.0176	.0132
SMK1 x E	-.0555	-.0688	-.0714	-.0591
SMK2 x E	.1465	.1280	.1237	.1235
Standard Error for Coefficient of:				
E	.0353	.0267	.0268	.0269
SMK1 x E	.0659	.0617	.0616	.0608
SMK2 x E	.0673	.0632	.0626	.0614
Estimated Odds Ratio at $(E_1-E_0)=15$				
Nonsmokers		1.320	1.302	1.219
Light Smokers		.470	.446	.502
Heavy Smokers		9.002	8.327	7.772
95% C.I. for the $\hat{OR}$ at $(E_1-E_0)=15$				
Nonsmokers		(.602,2.895)	(.593,2.860)	(.553,2.685)
Light Smokers		(.066,3.329)	(.086,2.313)	(.098,2.564)
Heavy Smokers		(1.221,66.352)	(1.589,43.632)	(1.542,39.181)

MODEL 1A: Main Effects + 1st Order Interactions with Exposure

MODEL 2A: Main Effects + Smoking x Exposure Interaction

MODEL 3A: Exposure, Prior Loss, Mother's Age at Pregnancy, Smoking, Drinking, Smoking x Exposure

MODEL 4A: Exposure, Prior Loss, Mother's Age at Pregnancy, Smoking, Smoking x Exposure

TABLE 3.12

ESTIMATES FROM THE UNIFORM-LOGISTIC MODELS

	MODEL 1A	MODEL 2A	MODEL 3A	MODEL 4A
Estimated Coefficient of:				
E	.0241	.0185	.0154	.0187
SMK1 x E	-.0565	-.0693	-.0695	-.0656
SMK2 x E	.1454	.1278	.1296	.1258
Standard Error for Coefficient of:				
E	0.0	.0270	.0273	.0281
SMK1 x E	0.0	.0618	.0625	.0632
SMK2 x E	0.0	.0630	.0651	.0794
Estimated Odds Ratio at $(E_1-E_0)=15$				
Nonsmokers		1.320	1.260	1.324
Light Smokers		.467	.444	.495
Heavy Smokers		8.976	8.802	8.736
95% C.I. for the $\hat{OR}$ at $(E_1-E_0)=15$				
Nonsmokers		(.597,2.919)	(.564,2.814)	(.580,3.019)
Light Smokers		(.066,3.323)	(.084,2.359)	(.092,2.669)
Heavy Smokers		(1.221,65.969)	(1.540,50.326)	(.929,82.189)

MODEL 1A: Main Effects + 1st Order Interactions with Exposure

MODEL 2A: Main Effects + Smoking x Exposure Interaction

MODEL 3A: Exposure, Prior Loss, Mother's Age at Pregnancy, Smoking, Drinking, Smoking x Exposure

MODEL 4A: Exposure, Prior Loss, Mother's Age at Pregnancy, Smoking, Smoking x Exposure

the estimated odds ratios by much (see Tables 3.11, 3.12), so validity of the odds ratios was preserved after dropping Gravidity from MODEL 2A. The next least significant variable was Drinking ($p = .088$). Removing Drinking from MODEL 3A led to MODEL 4A. This did not increase the precision of the parameter estimates. Since Drinking was on the borderline of significance and since it has been suspected of being associated with fetal loss, Drinking was left in the model. Therefore, the final model chosen was MODEL 3A. MODEL 3A contains Exposure, Prior Loss, Mother's Age at Pregnancy, Smoking, Drinking, and Smoking x Exposure.

3.5.3 Estimation of the Odds Ratio

Next, estimates of the odds ratios based on the logistic and the uniform-logistic models, MODEL 3A, were found. Since there was a significant Smoking x Exposure interaction in the model, the odds ratios across strata of Smoking are expected to be nonuniform. Therefore, odds ratios will be estimated for each level of Smoking. Interactions of all other independent variables with Exposure were not significant in the model, so the estimated odds ratios based on the model will be identical across strata of these independent variables.

Using the notation in Kleinbaum et al. (1981), the logistic model is

$$P(X) = \Pr(D = 1 \mid E, V_1, \ldots, V_{p_1}, W_1, \ldots, W_{p_2})$$

$$= \frac{1}{1 + \exp\left[-\left(\alpha + \beta E + \sum_{i=1}^{p_1} \gamma_i V_i + E \sum_{j=1}^{p_2} \delta_j W_j\right)\right]},$$

where

E = exposure

$$D = \begin{cases} 1 & \text{, if a fetal loss occurs} \\ 0 & \text{, if no fetal loss occurs} \end{cases}$$

$V_1, V_2, \ldots V_{p_1}$ are the functions of the potential confounders (main effects)

and $W_1, W_2, \ldots, W_{p_2}$ are the potential effect modifiers (interactions).

Given a continuous exposure variable, E, the odds ratio, OR, comparing the odds of fetal loss at exposure E_1 with the odds of fetal loss at exposure E_0 is

$$OR = \frac{p_1}{1-p_1} \Big/ \frac{p_0}{1-p_0} = \frac{p_1(1-p_0)}{p_0(1-p_1)} .$$

Under the logistic model,

$$p_0 = \frac{\exp[\alpha + \beta E_0 + \sum_i \gamma_i V_i + E_0 \sum_j \delta_j W_j]}{1 + \exp[\alpha + \beta E_0 + \sum_i \gamma_i V_i + E_0 \sum_j \delta_j W_j]}$$

and

$$p_1 = \frac{\exp[\alpha + \beta E_1 + \sum_i \gamma_i V_i + E_1 \sum_j \delta_j W_j]}{1 + \exp[\alpha + \beta E_1 + \sum_i \gamma_i V_i + E_1 \sum_j \delta_j W_j]} .$$

Thus,

$$OR = \frac{\dfrac{\exp[\alpha+\beta E_1+\sum_i \gamma_i V_i+E_1 \sum_j \delta_j W_j]}{1 + \exp[\alpha+\beta E_1+\sum_i \gamma_i V_i+E_1 \sum_j \delta_j W_j]} \cdot \dfrac{1}{1 + \exp[\alpha+\beta E_0+\sum_i \gamma_i V_i+E_0 \sum_j \delta_j W_j]}}{\dfrac{1}{1 + \exp[\alpha+\beta E_1+\sum_i \gamma_i V_i+E_1 \sum_j \delta_j W_j]} \cdot \dfrac{\exp[\alpha+\beta E_0+\sum_i \gamma_i V_i+E_0 \sum_j \delta_j W_j]}{1 + \exp[\alpha+\beta E_0+\sum_i \gamma_i V_i+E_0 \sum_j \delta_j W_j]}}$$

$$= \frac{\exp[\alpha+\beta E_1+\sum_i \gamma_i V_i+E_1 \sum_j \delta_j W_j]}{\exp[\alpha+\beta E_0+\sum_i \gamma_i V_i+E_0 \sum_j \delta_j W_j]}$$

$$= \exp[\beta(E_1-E_0) + (E_1-E_0) \sum_j \delta_j W_j].$$

The estimated odds ratio, $\hat{OR}$, is given by

$$\hat{OR} = \exp[\hat{\beta}(E_1-E_0) + (E_1-E_0) \sum_j \hat{\delta}_j W_j].$$

$$\hat{var}(\log(\hat{OR})) = \hat{var}[\hat{\beta}(E_1-E_0) + (E_1-E_0) \sum_j \hat{\delta}_j W_j]$$

$$= (E_1-E_0)^2 \{\hat{var}(\hat{\beta}) + \sum_j W_j^2 \hat{var}(\hat{\delta}_j)$$

$$+ 2 \sum_j W_j \hat{cov}(\hat{\beta},\hat{\delta}_j) + 2 \sum_j \sum_{i>j} W_i W_j \hat{cov}(\hat{\delta}_i,\hat{\delta}_j)\}$$

and a $100(1-\alpha)\%$ confidence interval for $\hat{OR}$ is given by

$$\exp[\log(\hat{OR}) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{var}(\log(\hat{OR}))}].$$

Based on the estimates from the <u>logistic</u> model, MODEL 3A,

$$\hat{OR} = \exp[\hat{\beta}(E_1-E_0) + (E_1-E_0)\sum_j \hat{\delta}_j W_j]$$

$$= \exp[(.0176)(E_1-E_0) + (E_1-E_0)(SMK1)(-.0714) + (E_1-E_0)(SMK2)(.1237)]$$

and $\hat{var}(\log(\hat{OR})) = (E_1-E_0)^2 \{(.0007166) + (SMK1)^2 (.0037936)$

$$+ (SMK2)^2(.0039146) + 2(SMK1)(-.0006885)$$

$$+ 2(SMK2)(-.0007287) + 2(SMK1)(SMK2)(.0006671)\}.$$

The estimated odds ratios, variances of log $(\hat{OR})$, and 95% confidence intervals are tabulated in Tables 3.13 - 3.15.

TABLE 3.13

ESTIMATED ODDS RATIOS UNDER THE LOGISTIC MODEL

	$\hat{OR}$		
(E_1-E_0)	Nonsmoking	Smoking < 1 pack/day	Smoking ≥ 1 pack/day
5	1.092	.764	2.027
10	1.192	.584	4.108
15	1.302	.446	8.327
20	1.422	.341	16.878
25	1.553	.260	34.209

TABLE 3.14

ESTIMATED VARIANCE OF LOG $(\hat{OR})$ UNDER THE LOGISTIC MODEL

	$\hat{var}(\log(\hat{OR}))$		
(E_1-E_0)	Nonsmoking	Smoking < 1 pack/day	Smoking ≥ 1 pack/day
5	.0179	.0783	.0793
10	.0717	.3133	.3174
15	.1612	.7050	.7141
20	.2866	1.2533	1.2695
25	.4479	1.9583	1.9836

TABLE 3.15

ESTIMATED 95% CONFIDENCE INTERVALS FOR $\hat{OR}$
UNDER THE LOGISTIC MODEL

$$\exp[\log(\hat{OR}) \pm 1.960 \sqrt{\hat{var}(\log(\hat{OR}))}]$$

(E_1-E_0)	Nonsmoking	Smoking < 1 pack/day	Smoking ≥ 1 pack/day
5	(.840,1.419)	(.442,1.322)	(1.167,3.520)
10	(.705,2.015)	(.195,1.749)	(1.362,12.394)
15	(.593,2.860)	(.086,2.313)	(1.589,43.632)
20	(.498,4.060)	(.038,3.060)	(1.854,153.603)
25	(.418,5.765)	(.017,4.046)	(2.164,540.757)

Based on the estimates from the uniform-logistic model, MODEL 3A,

$$\hat{OR} = \exp[(.0154)(E_1-E_0) + (-.0695)(SMK1)(E_1-E_0) + (.1296)(SMK2)(E_1-E_0)]$$

and

$$\hat{var}(\log(\hat{OR})) = (E_1-E_0)^2\{(.000747) + (SMK1)^2(.003906) + (SMK2)^2(.004234)$$
$$+ 2(SMK1)(-.000714) + 2(SMK2)(-.000732)$$
$$+ 2(SMK1)(SMK2)(.000667)\}.$$

The estimated odds ratios, variances of $\log(\hat{OR})$, and 95% confidence intervals are tabulated in Tables 3.16 - 3.18.

The estimated odds ratios from the logistic model and the uniform-logistic model are about the same. The uniform-logistic model produces slightly larger 95% confidence intervals.

TABLE 3.16

ESTIMATED ODDS RATIOS UNDER THE UNIFORM-LOGISTIC MODEL

	$\hat{OR}$		
(E_1-E_0)	Nonsmoking	Smoking < 1 pack/day	Smoking $\geq$ 1 pack/day
5	1.080	.763	2.065
10	1.166	.582	4.263
15	1.260	.444	8.802
20	1.361	.339	18.174
25	1.470	.259	37.525

TABLE 3.17

ESTIMATED VARIANCE OF LOG($\hat{OR}$) UNDER THE UNIFORM-LOGISTIC MODEL

	$\hat{var}(\log(\hat{OR}))$		
(E_1-E_0)	Nonsmoking	Smoking < 1 pack/day	Smoking $\geq$ 1 pack/day
5	.0187	.0806	.0879
10	.0747	.3225	.3517
15	.1681	.7256	.7913
20	.2988	1.2900	1.4068
25	.4669	2.0156	2.1981

TABLE 3.18

ESTIMATED 95% CONFIDENCE INTERVALS FOR $\hat{OR}$ UNDER THE UNIFORM-LOGISTIC MODEL

$$\exp[\log(\hat{OR}) \pm 1.960 \sqrt{\hat{var}(\log(\hat{OR}))}]$$

(E_1-E_0)	Nonsmoking	Smoking < 1 pack/day	Smoking ≥ 1 pack/day
5	(.826,1.412)	(.437,1.331)	(1.155,3.692)
10	(.683,1.993)	(.191,1.772)	(1.333,13.631)
15	(.564,2.814)	(.084,2.359)	(1.540,50.326)
20	(.466,3.972)	(.037,3.149)	(1.778,185.811)
25	(.385,5.608)	(.016,4.179)	(2.053,686.011)

3.5.4 Goodness of Fit of the Logistic and Uniform-logistic Models

The goodness of fit of the logistic model and the uniform-logistic model, MODEL 3A, was tested using a contingency table based test proposed by Hosmer and Lemeshow (1980). For the logistic model, let

$$\pi(\underset{\sim}{x}) = Pr(Y=1|\underset{\sim}{X}=\underset{\sim}{x}) = \frac{1}{1 + \exp[-(\beta_0 + \underset{\sim}{\beta}'\underset{\sim}{X})]}$$

where no assumptions are made about the distribution of $\underset{\sim}{X}$. The maximum likelihood estimate of $\pi(\underset{\sim}{x})$ is

$$\hat{\pi}(\underset{\sim}{X}) = \frac{1}{1 + \exp[-(\hat{\beta}_0 + \hat{\underset{\sim}{\beta}}'\underset{\sim}{X})]} .$$

TABLE 3.19

ESTIMATED ODDS RATIOS AND 95% CONFIDENCE INTERVALS UNDER THE LOGISTIC AND UNIFORM-LOGISTIC MODELS

$\hat{OR}$
(95% C.I.)

(E_1-E_0)	Nonsmoking		Smoking < 1 pack/day		Smoking ≥ 1 pack/day	
	Logistic	Uniform-logistic	Logistic	Uniform-logistic	Logistic	Uniform-logistic
5	1.092 (.840,1.419)	1.080 (.826,1.412)	.764 (.442,1.322)	.763 (.437,1.331)	2.027 (1.167,3.520)	2.065 (1.155,3.692)
10	1.192 (.705,2.015)	1.166 (.683,1.993)	.584 (.195,1.749)	.582 (.191,1.772)	4.108 (1.362,12.394)	4.263 (1.333,13.631)
15	1.302 (.593,2.860)	1.260 (.564,2.814)	.446 (.086,2.313)	.444 (.084,2.359)	8.327 (1.589,43.632)	8.802 (1.540,50.320)
20	1.422 (.498,4.060)	1.361 (.466,3.972)	.341 (.038,3.060)	.339 (.037,3.140)	16.878 (1.854,153.603)	18.174 (1.778,185.811)
25	1.553 (.418,5.765)	1.470 (.385,5.608)	.260 (.017,4.046)	.259 (.016,4.179)	34.209 (2.164,540.757)	37.525 (2.053,686.011)

In this test, the frequencies of predicted probabilities of observing Y=1, $\hat{\pi}(x)$, falling within certain intervals over the range, (0,1), are cross-tabulated by the frequencies of the observed outcome, Y=0 or 1. Under the null hypothesis that the logistic model fits the data, the expected frequencies of observing Y=1 and Y=0 are found. These expected frequencies depend on how the range of predicted probabilities, (0,1), is divided into intervals. For this application of the test, the predicted probabilities were divided into g quantiles of equal size (g=5, quintiles and g=10, deciles). The expected number of observations for which Y=1 in the j^{th} quantile is the sum of the $\hat{\pi}(x)$'s in the j^{th} quantile. The expected number of observations for which Y=0 in the j^{th} quantile is the marginal total number of $\hat{\pi}(x)$'s in the j^{th} quantile minus the expected number of observations for which Y=1. A chi-square statistic of the form,

$$X^2 = \sum_j \frac{(O_j - E_j)^2}{E_j}$$

is used to test the fit of the model. (O_j is the observed frequency in category j and E_j is the expected frequency in category j.)

Mathematically, the test is formulated as follows. Suppose for $j = 1,2,\ldots,g$, there are $\hat{n}_{\cdot j}$ of the predicted probabilities, $\hat{\pi}_i(x)$ such that $c^*_{j-1} \leq \hat{\pi}_i(x) < c^*_j \cdot c^*_j$, $j = 1,\ldots,g$ are chosen such that the $\hat{n}_{\cdot j}$'s are all equal; $c^*_0 = 0$, $c^*_g = 1$. Let $J_j = \{i : c^*_{j-1} \leq \hat{\pi}_i(x) < c^*_j\}$ denote the set of indices of $\hat{\pi}_i(x)$ such that $c^*_{j-1} \leq \hat{\pi}_i(x) < c^*_j$. Then the $\hat{\pi}(x)$'s can be tabulated in a 2 x g contingency table.

Quantiles of $\hat{\pi}(x)$

Observed Outcome Y		c^*_{j-1}		c^*_j		
	0	$\hat{n}^*_{01}$	. . .	$\hat{n}^*_{0j}$	. . .	$\hat{n}^*_{0g}$
	1	$\hat{n}^*_{11}$	. . .	$\hat{n}^*_{1j}$	. . .	$\hat{n}^*_{1g}$
		$\hat{n}^*_{\cdot 1}$		$\hat{n}^*_{\cdot j}$		$\hat{n}^*_{\cdot g}$	n

$\sum_{r\epsilon J_i} \hat{\pi}_r$ is the sum of the predicted probabilities of observing Y=1 in the interval, (c^*_{j-1}, c^*_j). Thus, $\sum_{r\epsilon J_i} \hat{\pi}_r$ is the expected value of the number of outcomes for which Y=1, $\hat{n}^*_{1j}$, in the j^{th} interval under the null hypothesis that the logistic model fits adequately. The chi-square logistic model is

$$\hat{C}^*_g = \sum_{j=1}^{g} \left[\frac{(\hat{n}^*_{1j} - \sum_{r\epsilon J_j} \hat{\pi}_r)^2}{\sum_{r\epsilon J_j} \hat{\pi}_r} + \frac{(\hat{n}^*_{0j} - (\frac{n}{g} - \sum_{r\epsilon J_j} \hat{\pi}_r))^2}{(\frac{n}{g} - \sum_{r\epsilon J_j} \hat{\pi}_r)} \right]$$

$$= \sum_{j=1}^{g} \frac{(\hat{n}^*_{1j} - \sum_{r\epsilon J_j} \hat{\pi}_r)^2}{(\sum_{r\epsilon J_j} \hat{\pi}_r)(1 - g \sum_{r\epsilon J_j} \frac{\hat{\pi}_r}{n})} .$$

$\hat{C}^*_g$ approximately has a chi-square distribution with (g-2) degrees of freedom under the null hypothesis that the logistic model fits.

This goodness of fit test was also applied to the uniform-logistic

model. As an example, suppose that g=5. Then the 2x5 table of frequencies of estimated probabilities falling within each quintile by the observed outcome for the uniform-logistic model, MODEL 3A, is presented in Table 3.20.

TABLE 3.20

EXAMPLE OF THE GOODNESS OF FIT TEST FOR THE UNIFORM-LOGISTIC MODEL, MODEL 3A

n^*_{ij}		Quintiles of $\hat{\pi}$ 1	2	3	4	5	
Y	0	48	46	42	43	32	211
	1	1	4	8	7	17	37
		49	50	50	50	49	248
	j	1	2	3	4	5	
	$\sum_{r\epsilon J_j} \hat{\pi}$	2.86464	4.54992	6.40644	9.25688	18.20226	

The test statistic for this example is:

$$\hat{C}^*_5 = \sum_{j=1}^{5} \frac{(\hat{n}^*_{1j} - \sum_{r\epsilon J_i} \hat{\pi}_r)^2}{(\sum_{r\epsilon J_j} \hat{\pi}_r)(1 - \frac{5}{248} \sum_{r\epsilon J_j} \hat{\pi}_r)}$$

$$= \frac{(1 - 2.86464)^2}{(2.86464)(1 - \frac{5}{248}(2.86464))} + \frac{(4 - 4.54992)^2}{(4.54992)(1 - \frac{5}{248}(4.54992))}$$
$$+ \frac{(8 - 6.40644)^2}{(6.40644)(1 - \frac{5}{248}(6.40644))} + \frac{(7 - 9.25688)^2}{(9.25688)(1 - \frac{5}{248}(9.25688))}$$
$$+ \frac{(17 - 18.20226)^2}{(18.20226)(1 - \frac{5}{248}(18.20226))}$$

$$= 2.6184 \sim \chi^2(5 - 2) = \chi^2(3); \quad p = .454.$$

Both the logistic model and the uniform-logistic model fit the data adequately in that the lack of fit was insignificant. However, in the following table, it can be seen that the logistic model fits better than the uniform-logistic model.

TABLE 3.21

GOODNESS OF FIT STATISTICS FOR MODEL 3A

		Number of quantiles, g	
		5	10
Model	Logistic	$\chi^2(3) = 1.859$ $p = .60$	$\chi^2(8) = 7.449$ $p = .49$
	Uniform-logistic	$\chi^2(3) = 2.618$ $p = .45$	$\chi^2(8) = 11.664$ $p = .17$

The difference in lack of fit between the logistic model and the uniform-logistic model is more striking than the differences among the estimates and standard errors of the two models. It may be that the "background risk" of fetal loss is so subtle that the uniform-logistic model is not detecting it in the data; thus, the parameter estimates for the two models are nearly the same. In a different set of data, the uniform-logistic model may fit better than the logistic model. The performance of the uniform-logistic model should not be judged on the results from this one set of data.

3.6 Results of Modeling, Exposure Treated as a Dichotomous Variable

3.6.1 Definitions of the Variables

The same modeling procedure was repeated with exposure treated as a dichotomous variable, i.e.,

$$E_j = \begin{cases} 1 & \text{, if exposed (low exposure) at pregnancy j} \\ 0 & \text{, if not exposed at pregnancy j} \end{cases}$$

All of the other variables were defined as in Section 3.5.1.

3.6.2 Model Fitting

Again, the modeling strategy advocated by Kleinbaum et al. (1981) was followed. The first model constructed contained main effects and first and second order interactions of the main effects with Exposure. The information matrix was singular, so no estimates could be obtained. Next, a model containing main effects and first order interactions with Exposure was fit. No second order interactions were significant enough ($p < .20$) to be added to this model. The estimated parameters, their standard errors, z-values, and 1-sided p-values for the main effects plus first order interactions model are presented in Table 3.22. (MODEL 1B). As before, the first order interactions of all main effects with Exposure except Smoking x Exposure were insignificant. The Smoking x Exposure interaction was significant with $\chi^2(2) = 6.021$, $p = .05$. A likelihood ratio test of the contribution of the insignificant interactions supported their removal from the model ($\chi^2(4) = 1.00$, n.s.). Next, the insignificant main effects were candidates for removal from the model. Since the Smoking x Exposure interaction was significant, the lower order terms, Smoking and Exposure were required to remain in the model. Mother's Age at Pregnancy, again, was retained in all models to ensure that the vectors of independent variables would be distinct within each family. The first main effect to be removed from the model was Gravidity ($p = .269$). MODEL 3B contains all main effects except Gravidity and Smoking x Exposure interactions. This did not

TABLE 3.22

MODEL 1B: MAIN EFFECTS AND FIRST ORDER INTERACTIONS WITH EXPOSURE

	LOGISTIC				UNIFORM-LOGISTIC			
	MLE	SE	z	1-sided p	MLE	SE	z	1-sided p
$\hat{\beta}_0$	-2.5436	.4443	-5.725	< .0001	$\hat{a}$ -2.8403	1.2328	-2.304	.010
					$\hat{b}$ -1.9807	.6524	-3.036	.001
E	.4158	.6382	.652	.257	.4137	.6475	.639	.261
GRAV	-.1247	.2772	-.450	.327	-.0935	.3000	-.312	.378
LOSS	1.2505	.7723	1.619	.053	1.0910	.9732	1.121	.131
MA	.0238	.0774	.307	.379	.0161	.0836	.193	.423
SMK1	.6570	.6832	.962	.168	.6771	.7011	.966	.167
SMK2	-1.5737	1.1184	-1.407	.080	-1.5419	1.1449	-1.347	.089
DRNK	.9705	.6154	1.577	.057	.9840	.6301	1.562	.059
GRAV x E	-.0227	.3802	-.060	.476	-.0332	.3884	-.085	.466
LOSS x E	.0099	1.0870	.009	.496	.0056	1.1034	.005	.498
MA x E	.0639	.1016	.629	.265	.0799	.1234	.647	.259
SMK1 x E	-.5790	1.1748	-.493	.311	-.5595	1.1994	-.466	.321
SMK2 x E	2.8897	1.3265	2.254	.012	3.1012	1.4130	2.195	.014
DRNK x E	-.7023	.9055	-.776	.219	-.7688	.9504	-.809	.209
$-2 \log \hat{L}$	184.06				184.05			

TABLE 3.23

MODEL 2B: MAIN EFFECTS AND SMOKING x EXPOSURE INTERACTION

	LOGISTIC				UNIFORM-LOGISTIC			
	MLE	SE	z	1-sided p	MLE	SE	z	1-sided p
$\hat{\beta}_0$	-2.3711	.3581	-6.621	< .0001				
$\hat{a}$					-3.1612	1.1777	-2.684	.004
$\hat{b}$					-2.0088	.5612	-3.579	.0002
E	.2519	.4975	.506	.306	.2512	.4967	.506	.306
GRAV	-.1116	.1838	-.607	.272	-.1171	.1902	-.616	.269
LOSS	1.1419	.5206	2.193	.014	1.2266	.7368	1.665	.048
MA	.0643	.0494	1.302	.097	.0617	.0491	1.257	.104
SMK1	.7775	.6685	1.163	.122	.7840	.6607	1.187	.118
SMK2	-1.4659	1.1062	-1.325	.093	-1.5092	1.1138	-1.355	.088
DRNK	.5815	.4239	1.372	.085	.6033	.4235	1.425	.077
SMK1 x E	-.9202	1.0958	-.840	.200	-.9229	1.0846	-.851	.197
SMK2 x E	2.6606	1.2598	2.1119	.017	2.6406	1.2456	2.120	.017
$-2 \log \hat{L}$		185.05				185.05		

Likelihood Ratio Test for removing all interactions except Smoking x Exposure

$$-2 \log \frac{\hat{L}(\text{MODEL 2B})}{\hat{L}(\text{MODEL 1B})}$$

$= 185.05 - 184.06$

$= .99 \sim \chi^2(4)$ (n.s.)

$= 185.05 - 184.05$

$= 1.00 \sim \chi^2(4)$ (n.s.)

Range of the underlying risk of fetal loss, R.
$.0407 < R < .1183$

TABLE 3.24

MODEL 3B: EXPOSURE, PRIOR LOSS, MOTHER'S AGE AT PREGNANCY, SMOKING, DRINKING, SMOKING x EXPOSURE

	LOGISTIC				UNIFORM-LOGISTIC			
	MLE	SE	z	1 sided p	MLE	SE	z	1-sided p
$\hat{\beta}_0$	-2.3748	.3589	-6.617	< .0001	$\hat{a}$ -2.7292	.6047	-4.513	< .0001
					$\hat{b}$ -1.7952	.5409	-3.319	.0004
E	.2181	.4953	.440	.315	.1967	.6405	.307	.379
LOSS	.9858	.4506	2.188	.014	.8715	.5629	1.548	.061
MA	.0491	.0431	1.139	.127	.0511	.0536	.953	.170
SMK1	.8088	.6648	1.217	.112	.8081	.8475	.954	.170
SMK2	-1.4844	1.1050	-1.343	.090	-1.4388	1.5105	-.952	.170
DRNK	.5694	.4223	1.348	.089	.5316	.5648	.941	.173
SMK1 x E	-.9456	1.0958	-.863	.194	-.9139	1.3967	-.654	.256
SMK2 x E	2.6326	1.2582	2.092	.018	2.6630	1.6873	1.578	.057
$-2 \log \hat{L}$		185.42				185.46		

Likelihood Ratio Test for removing Gravidity from MODEL 2B

$$-2 \log \frac{\hat{L}(\text{MODEL 3B})}{\hat{L}(\text{MODEL 2B})}$$

$= 185.42 - 185.05$

$= .37 \sim \chi^2(1)$ (n.s.)

$= 185.46 - 185.05$

$= .41 \sim \chi^2(1)$ (n.s.)

TABLE 3.25

MODEL 4B: EXPOSURE, PRIOR LOSS, MOTHER'S AGE AT PREGNANCY, SMOKING, SMOKING x EXPOSURE

	LOGISTIC				UNIFORM-LOGISTIC			
	MLE	SE	z	1-sided p	MLE	SE	z	1-sided p
$\hat{\beta}_0$	-2.1993	.3260	-6.746	< .0001				
$\hat{a}$					-2.2169	.9208	-2.408	.008
$\hat{b}$					-1.4785	.6005	-2.462	.007
E	.1209	.4918	.246	.403	.1400	.5204	.269	.394
LOSS	.9779	.4478	2.184	.014	.6700	.7740	.866	.193
MA	.0570	.0424	1.344	.089	.0651	.0495	1.315	.094
SMK1	.9269	.6587	1.407	.080	.9938	.7451	1.334	.091
SMK2	-1.2530	1.0830	-1.157	.129	-1.1420	1.1463	-.996	.160
SMK1 x E	-.7276	1.0793	-.674	.250	-.8002	1.1772	-.680	.248
SMK2 x E	2.6973	1.2474	2.162	.015	2.8916	1.4436	2.003	.022
$-2 \log \hat{L}$		187.20				186.81		

Likelihood Ratio Test for removing Drinking from MODEL 3B

$$-2 \log \frac{\hat{L}(\text{MODEL 4B})}{\hat{L}(\text{MODEL 3B})}$$

$= 187.20 - 185.42$

$= 1.78 \sim \chi^2(1)$

$p = .1822$

$= 186.81 - 185.46$

$= 1.35 \sim \chi^2(1)$

$p = .2453$

TABLE 3.26

ESTIMATES FROM THE LOGISTIC MODELS

	MODEL 1B	MODEL 2B	MODEL 3B	MODEL 4B
Estimated Coefficient of:				
E	.4158	.2519	.2181	.1209
SMK1 x E	-.5790	-.9202	-.9456	-.7276
SMK2 x E	2.9897	2.6606	2.6326	2.6973
Standard Error for Coefficient of:				
E	.6382	.4975	.4953	.4918
SMK1 x E	1.1748	1.0958	1.0958	1.0793
SMK2 x E	1.3265	1.2598	1.2482	1.2474
Estimated Odds Ratio				
Nonsmokers		1.286	1.234	1.128
Light Smokers		.513	.483	.545
Heavy Smokers		18.403	17.300	16.747
95% C.I. for the $\hat{OR}$				
Nonsmokers		(.485,3.411)	(.471,3.284)	(.430,2.959)
Light Smokers		(.073,3.583)	(.084,2.763)	(.080,3.724)
Heavy Smokers		(1.906,177.677)	(1.814,164.991)	(1.791,156.617)

MODEL 1B: Main Effects + 1st Order Interactions with Exposure

MODEL 2B: Main Effects + Smoking x Exposure

MODEL 3B: Exposure, Prior Loss, Mother's Age at Pregnancy, Smoking, Drinking, Smoking x Exposure

MODEL 4B: Exposure, Prior Loss, Mother's Age at Pregnancy, Smoking, Smoking x Exposure

TABLE 3.27

ESTIMATES FROM THE UNIFORM-LOGISTIC MODELS

	MODEL 1B	MODEL 2B	MODEL 3B	MODEL 4B
Estimated Coefficient of:				
E	.4137	.2512	.1967	.1400
SMK1 x E	-.5595	-.9229	-.9139	-.8002
SMK2 x E	3.1012	2.6406	2.6630	2.8916
Standard Error for Coefficient of:				
E	.6475	.4967	.6405	.5204
SMK1 x E	1.1994	1.0846	1.3967	1.1722
SMK2 x E	1.4130	1.2456	1.6873	1.4436
Estimated Odds Ratio				
Nonsmokers		1.286	1.217	1.150
Light Smokers		.511	.488	.517
Heavy Smokers		18.026	17.456	20.730
95% C.I. for the $\hat{OR}$				
Nonsmokers		(.486,3.403)	(.347,4.272)	(.415,3.190)
Light Smokers		(.075,3.489)	(.042,5.691)	(.063,4.255)
Heavy Smokers		(1.922,169.072)	(.833,365.63)	(1.486,289.160)

MODEL 1B: Main Effects + 1st Order Interactions with Exposure

MODEL 2B: Main Effects + Smoking x Exposure

MODEL 3B: Exposure, Prior Loss, Mother's Age at Pregnancy, Smoking, Drinking, Smoking x Exposure

MODEL 4B: Exposure, Prior Loss, Mother's Age at Pregnancy, Smoking, Smoking x Exposure

change the estimated odds ratio much (see Tables 3.26, 3.27), so validity of the odds ratio estimates was preserved. The next least significant variable was Drinking (p=.173). Removing Drinking from MODEL 3B produced MODEL 4B. The precision of the estimated odds ratios was not increased by removing Gravidity and Drinking from the model. In addition, the estimated odds ratios were somewhat different in MODEL 4B from those in MODEL 2B and MODEL 3B. This indicates a possible loss of validity of the estimated odds ratios. (See Tables 3.26, 3.27.) Thus, both Gravidity and Drinking were left in the model. The final model chosen was MODEL 2B. MODEL 2B contains Exposure, Gravidity, Prior Loss, Mother's Age at Pregnancy, Smoking, Drinking, and Smoking x Exposure.

3.6.3 Estimation of the Odds Ratios

Estimates of the odds ratios based on the logistic model and the uniform-logistic model, MODEL 2B, were found. Since there was a significant Smoking x Exposure interaction, the odds ratios across strata of Smoking are expected to be nonuniform. Therefore, odds ratios at each level of Smoking were estimated.

Using the notation for estimated odds ratios as given in Section 3.5.3 and the definition of E, the difference between exposed and unexposed, $(E_1-E_0) = 1$. Thus, $\hat{\beta}$ is the estimated increment in the log odds ratio associated with exposure. The term, $\hat{\delta}_j W_j$, is the estimated increment in the log odds ratio associated with the presence of exposure and risk factor W_j, for all j's. The estimated odds ratio, $\hat{OR}$, of the odds of fetal loss of exposed pregnancies versus the odds of fetal loss of unexposed pregnancies is

$$\hat{OR} = \exp[\hat{\beta} + \sum_j \hat{\delta}_j W_j].$$

Also,

$$\hat{var}(\log(\hat{OR})) = \hat{var}[\hat{\beta} + \sum_j \hat{\delta}_j W_j]$$

$$= \hat{var}(\hat{\beta}) + \sum_j W_j^2 \hat{var}(\hat{\delta}_j) + 2 \sum_j W_j \hat{cov}(\hat{\beta}, \hat{\delta}_j)$$

$$+ 2 \sum_i \sum_{j>i} W_i W_j \hat{cov}(\hat{\delta}_i, \hat{\delta}_j).$$

A $100(1-\alpha)\%$ confidence interval for the estimated odds ratio, $\hat{OR}$, is given by

$$\exp[\log(\hat{OR}) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{var}(\log(\hat{OR}))}\ ] .$$

Based on the estimates from the logistic model, MODEL 2B,

$$\hat{OR} = \exp[(.2519) + (-.9209)(SMK1) + (2.6606)(SMK2)]$$

and

$$\hat{var}(\log(\hat{OR})) = (.24754) + (SMK1)^2(1.20089) + (SMK2)^2(1.58697)$$
$$+ 2(SMK1)(-.23210) + 2(SMK2)(-.24808)$$
$$+ 2(SMK1)(SMK2)(.23353).$$

The estimated odds ratios, variances of $\log(\hat{OR})$ and 95% confidence intervals for $\hat{OR}$, based on the logistic model are presented in Tables 3.28-3.30.

TABLE 3.28

ESTIMATED ODDS RATIOS UNDER THE LOGISTIC MODEL

	$\hat{OR}$
Nonsmokers	1.286
Smokers, < 1 pack/day	.513
Smokers, ≥ 1 pack/day	18.403

TABLE 3.29

ESTIMATED VARIANCE OF LOG($\hat{OR}$) UNDER THE LOGISTIC MODEL

	$\hat{var}(\log(\hat{OR}))$
Nonsmokers	.2475
Smokers, < 1 pack/day	.9842
Smokers, $\geq$ 1 pack/day	1.3384

TABLE 3.30

ESTIMATED 95% CONFIDENCE INTERVALS FOR $\hat{OR}$ UNDER THE LOGISTIC MODEL

	95% Confidence Interval
Nonsmokers	(.485,3.411)
Smokers, < 1 pack/day	(.073,3.583)
Smokers, $\geq$ 1 pack/day)	(1.906,177.677)

Based on the estimates from the uniform-logistic model, MODEL 2B,

$$\hat{OR} = \exp[(.2512) + (-.9229)(SMK1) + (2.6406)(SMK2)$$

and

$$\hat{var}(\log(\hat{OR})) = (.24673) + (SMK1)^2(1.17642) + (SMK2)^2(1.55152) + 2(SMK1)(-.23112) + 2(SMK2)(-.24694) + 2(SMK1)(SMK2)(.23345).$$

The estimated odds ratios, $\hat{OR}$, variances of $\log(\hat{OR})$, and 95% confidence intervals for $\hat{OR}$ based on the uniform-logistic model, MODEL 2B, are presented in Tables 3.31-3.33. The estimated odds ratios and their 95% confidence intervals under stratified analysis, the logistic model, and the uniform-logistic model are presented in Table 3.34.

TABLE 3.31

ESTIMATED ODDS RATIOS UNDER THE UNIFORM-LOGISTIC MODEL

	$\hat{OR}$
Nonsmokers	1.286
Smokers, < 1 pack/day	.511
Smokers, $\geq$ 1 pack/day	18.026

TABLE 3.32

ESTIMATED VARIANCE OF LOG($\hat{OR}$) UNDER THE UNIFORM-LOGISTIC MODEL

	$\hat{var}(\log(\hat{OR}))$
Nonsmokers	.2467
Smokers, < 1 pack/day	.9609
Smokers, $\geq$ 1 pack/day	1.3044

TABLE 3.33

ESTIMATED 95% CONFIDENCE INTERVALS FOR $\hat{OR}$ UNDER THE UNIFORM-LOGISTIC MODEL

	95% Confidence Interval
Nonsmokers	(.485,3.403)
Smokers, < 1 pack/day	(.075,3.489)
Smokers, $\geq$ 1 pack/day	(1.922,169.067)

TABLE 3.34

ESTIMATED ODDS RATIOS AND 95% CONFIDENCE INTERVALS UNDER STRATIFIED ANALYSIS, THE LOGISTIC MODEL, AND THE UNIFORM-LOGISTIC MODEL

		$\hat{OR}$ (95% C.I.)	
	Stratified Analysis	Logistic Model	Uniform-logistic Model
Nonsmokers	1.525 (.618,3.765)	1.286 (.485,3.411)	1.286 (.485,3.403)
Smokers, < 1 pack/day	.636 (.098,4.138)	.513 (.073,3.583)	.511 (.075,3.489)
Smokers, ≥ 1 pack/day	18.182 (2.022,163.521)	18.403 (1.906,177.677)	18.026 (1.922,169.067)

It can be seen in Table 3.34 that the estimated odds ratios from the stratified analysis, the logistic model, and the uniform-logistic model are all about the same.

3.6.4 Goodness of Fit of the Logistic and Uniform-logistic models

The goodness of fit of the logistic and uniform-logistic models, MODEL 2B, was tested using the same method as described in Section 3.5.4. Again, both models fit the data adequately in that the lack of fit was insignificant. In Table 3.35, it can be seen that the logistic model fits better than the uniform-logistic model.

The fit of both models in which Exposure is treated as a dichotomous variable is better than the fit of the models in which Exposure is treated as a continuous variable. (See Table 3.21.)

TABLE 3.35

GOODNESS OF FIT STATISTICS FOR MODEL 2B

		Number of Quantiles, g	
		5	10
Model	Logistic	$\chi^2(3) = .840$ $p = .84$	$\chi^2(8) = 4.958$ $p = .76$
	Uniform-logistic	$\chi^2(3) = 1.563$ $p = .67$	$\chi^2(8) = 6.322$ $p = .61$

3.7 Interpretation of the Final Models

3.7.1 The Model Treating Exposure as a Continuous Variable, MODEL 3A

From MODEL 3A, it appears that Smoking is an effect modifier in that the effect of Exposure appears to be different at different levels of Smoking. Smoking has recently been suspected of being associated with higher risk of fetal loss. The positive sign of the estimated coefficient of Heavy Smoking x Exposure is consistent with this suspicion in that for heavy smokers, the higher the exposure is, the greater the estimated probability of fetal loss under both the logistic and uniform-logistic models. In addition to Smoking x Exposure, the main effects, Prior Loss, Mother's Age at Pregnancy, and Drinking were also significant factors in predicting the probability of fetal loss. Prior Loss, advanced Mother's Age at Pregnancy, and Drinking during pregnancy have also been suspected o. being risk factors for fetal loss. The positive signs of the coefficients of these main effects are consistent with this suspicion. In the presence of Drinking and/or Prior Loss and/or Advanced Mother's Age at Pregnancy, the estimated probability of fetal loss predicted by both models is higher than in the

absence of these factors. The coefficients of Exposure and Smoking are not very meaningful. As Kleinbaum, et al. (1981) have shown, only the coefficients of the highest order interactions and the main effects not involved in any interactions are unaffected by how the variables are coded. Depending on how the variables are coded, the coefficients of main effects involved in interaction terms could be negative or positive and of any magnitude. These main effect coefficients alone do not reflect the impact of the independent variables on fetal loss; interaction terms must also be considered. Thus, a test of these main effects would be meaningless.

3.7.2 The Model Treating Exposure as a Dichotomous Variable, MODEL 2B

The results and interpretation of the model treating exposure as a dichotomous variable, MODEL 2B, were essentially the same as the model treating exposure as a continuous variable, MODEL 3A. MODEL 2B was chosen as the final model for predicting the probability of fetal loss, given the set of independent variables for a particular pregnancy. In addition to the variables included in the continuous exposure variable model, Gravidity was retained in MODEL 2B. However, Gravidity did not significantly contribute to the prediction of fetal loss. The uniform-logistic model treating exposure as a dichotomous variable, MODEL 2B, fit the data better than the uniform-logistic model treating exposure as a continuous variable, MODEL 3A.

It is to be noted that the intercept of the logistic models, $\hat{\beta}_0$, falls within the interval, $(\hat{a},\hat{b})$, of the uniform-logistic model. In addition, the intervals, $(\hat{a},\hat{b})$, transform to reasonable ranges of

the underlying risk of fetal loss. In MODEL 3A, the range of underlying risk is (.063,.137). In MODEL 2B, the range is (.041,.118).

3.8 Qualifications on the Uniform-logistic Model

3.8.1 Disclaimer

The uniform-logistic model is a complex model which has been applied to one complex epidemiological data set. The behavior of the model has not yet been investigated. A study on simulated data with known parameters would be helpful in assessing the behavior of the model. Until more is known about the model, it would be wise to use the uniform-logistic model on real data with extreme caution, if at all. It may have been coincidental that the uniform-logistic model gave estimates very similar to the logistic model in this case. The estimates may be very different on some other data set.

3.8.2 Empirical Evidence in Support of the Uniform-logistic Model

Based on the data analyzed, the uniform-logistic model did not produce surprising results. First, it was possible to estimate the parameters of the uniform-logistic model using maximum likelihood techniques. Second, the estimated parameters seemed reasonable in the sense that the signs of the parameters were considered with the a priori expected effects of the variables. Third, the estimated parameters were slightly different from the parameters of the logistic regression model. Although the difference was slight, the standard errors of the parameters were generally more conservative (larger) in the uniform-logistic model than in the logistic model. This caused most of the p-values of these parameters to be slightly larger in the uniform-logistic model

than the logistic model. This would be expected, if, indeed, the covariance between observations is incorporated into the variance expressions for the parameters. Theoretically, under the uniform-logistic model, a nonzero covariance between pregnancies within the same family is allowed.

The slight differences observed between the uniform-logistic model and the logistic model may not be meaningful. It is possible that the differences "background risk" of fetal loss among families may be so subtle that they are almost undetectable with the uniform-logistic model in this set of data. This possibility is supported by a test of homogeneity as described by Potthoff and Whittinghill (1966). According to the results of this test on the data, there is no significant departure from homogeneity ($.10 < p < .90$). Thus, it appears that there may be no heterogeneity of risk of fetal loss among these families to be incorporated into the uniform-logistic model.

3.8.3 Heuristic Argument in Support of the Uniform-logistic Model

The rationale behind construction of the model provides heuristic support of the uniform-logistic model. In animal litter studies, the beta-binomial model is sometimes used to estimate the probability of fetal loss among litters. In the beta-binomial model, fetal losses within a litter are assumed to follow a binomial distribution with probability parameter, p. From mother to mother, the probability of fetal loss, p, is assumed to vary according to a beta distribution. Humans, however, do not have litters; their pregnancies occur over time. Factors associated with time change from pregnancy to pregnancy for a human. Therefore, the probability of fetal loss over a couple's

reproductive history should not be considered to have a simple binomial distribution. Thus, the beta-binomial model would probably not be appropriate for analyzing human pregnancy histories. The uniform-logistic model applies a logistic model to the probability of fetal loss in an attempt to adjust for factors which change from pregnancy to pregnancy. Rather than the probability of fetal loss having a beta distribution among mothers as in the beta-binomial model, the constant term in the logistic model is assumed to have a uniform distribution among couples. This is expressed mathematically as

$$\Pr(Y=1 \mid \beta_0^*, \underset{\sim}{\beta}, \underset{\sim}{Z}) = \frac{1}{1 + \exp[-(\beta_0^* + \underset{\sim}{\beta}'\underset{\sim}{Z})]}$$

where $\beta_0^* \sim U(a,b)$.

A given couple is assumed to have a β_0^* value between a and b, each value between a and b being equally likely.

There are two consequences of β_0^* having a uniform distribution. First, the underlying risk of fetal loss (i.e., the risk of fetal loss given a standard set of independent variables) is assumed to have a distribution in the population as in Figure 2.1. Thus, different mothers with the same set of independent variables may be at different risks of fetal loss. In many animal reproductive studies, the underlying risk of fetal loss has been adequately modeled by the beta distribution in the beta-binomial model. However, in humans, the underlying risk of fetal loss is a complex combination of many factors. The very slight differences between the uniform-logistic model and the logistic model observed with these data suggest that the underlying risk of fetal loss may not have been detected with the uniform-logistic model in this case. Second, a potentially nonzero covariance between

pregnancies within the same family is introduced. Thus, pregnancies are not treated as if they were unconditionally mutually independent events.

3.8.4 Goodness of Fit of the Uniform-logistic Model

The uniform-logistic model did not fit the data as well as the logistic model did, although the goodness of fit of the uniform-logistic model was adequate. Again, the uniform-logistic model is a complex model applied to a complex epidemiological data set. It is not clear for these data why the fit of the uniform-logistic model was not as good as the fit of the logistic model. However, for another data set, the model might fit better.

It can be shown that the logistic likelihood function is not a special case of the uniform-logistic likelihood function. Since the two likelihood functions are different, one would not expect the uniform-logistic model with its additional parameter to fit every set of data better than the logistic model.

3.8.5 An Interesting Observation

It was interesting to note that as the uniform-logistic model gave slightly higher p-values than the logistic model, for the most part, the p-values for Prior Loss was increased by the uniform-logistic model more than the p-value for any other variable. In MODEL 3A, the p-value for Prior Loss was .012 in the logistic model and was .066 in the uniform-logistic model. In MODEL 2B, the p-value for Prior Loss was .014 in the logistic model and was .048 in the uniform-logistic model. For the other independent variables, there was not much more

than a difference of .01 between the p-values from the logistic model and the uniform-logistic model. This may indicate that Prior Loss carries some information about the distribution of risk of fetal loss within the sample.

CHAPTER IV

SUGGESTIONS FOR FURTHER RESEARCH

4.1 Suggestions for Further Research

A uniform-logistic model has been developed and used on a set of real data. Although the modeling results seem reasonable and consistent with expectations, the performance of the uniform-logistic model can not be judged on this one example. Further work with this model is necessary to discover its strengths and weaknesses.

First, a simulation study could give insights to the behavior of the model. A population with known parameters and known correlations between observations within families could be generated. The sensitivity of the model to changes in the parameters and/or the correlations could then be investigated. Also, these models could be compared to the usual logistic models in which the assumption of mutual independence of observations is made. By making this comparison, the necessity of incorporating covariance into the model, as the uniform-logistic model does, could be assessed.

Second, other real data sets could be analyzed using a variety of methods, including the uniform-logistic model. This could give additional insights to the practicality of the uniform-logistic model.

Third, other distributions than the uniform distribution could be imposed on β_0^*. This may lead to better fitting models. It is also possible that other distributions may be more biologically meaningful.

However, at this time, there is not enough data to suggest any particular distribution.

Fourth, more theoretical work on the model is needed. In the construction of the likelihood function, it was assumed that within each family, the vectors of independent variables were distinct (see Section 2.7.1). If there are two or more equal vectors within a family, the likelihood function in its present form becomes undefined. It is very possible to construct the likelihood function to allow for non-distinct vectors of the independent variables within families. If this were done, it would be possible to analyze data entirely consisting of categorical variables without the possibility of an undefined likelihood function. To avoid this problem in the modeling done previously, a continuous variable, Mother's Age at Pregnancy, was included in all models.

To summarize, much more work should yet be done before the behavior of the uniform-logistic model can properly be assessed. The model has only been presented and demonstrated on one set of data in this dissertation.

REFERENCES

Aeschbacher, H.U.; Vuataz, L.; Sotek, J.; Stalder, R. (1977). "Use of the Beta-binomial Distribution in Dominant-Lethal Testing for 'Weak Mutagenic Activity' Part 1." Mutation Research 44, 369-390.

Altham, P.M.E. (1978). "Two Generalizations of the Binomial Distribution." Applied Statistics 27, No. 2, 162-167.

Becker, B.A. (1974). "The Statistics of Teratology." Teratology 9, 261-262.

Cohen, J.E. (1976). "The Distribution of the Chi-Square Statistic Under Clustered Sampling from Contingency Tables." Journal of the American Statistical Association 71, No. 355, 665-670.

Conover, W.J. (1971). Practical Nonparametric Statistics. John Wiley and Sons, Inc., New York.

Cox, D.R. (1972). "Regression Models and Life-Tables." Journal of the Royal Statistical Society, Series B 34, 187-220.

Dobbins, J.G.; Eifler, C.W.; Buffler, P.A. (1978). "The Use of Parity Survivorship Analysis in the Study of Reproductive Outcome." A paper prepared for the Symposium on Methodologic and Analytic Issues in Monitoring Human Populations for Reproductive Risks Associated with Environmental Exposures - Society for Epidemiologic Research Annual Meeting, 14-16 June 1978, Iowa City, Iowa.

Erickson, J.D., Bjerkedal, T. (1978). "Interpregnancy Interval: Association with Birthweight, Stillbirth, and Neonatal Death." Journal of Epidemiology and Community Health 32, No. 2, 124-130.

Fears, T.R.; Tarone, R.E.; Chu, K.C. (1977). "False-Positive and False-Negative Rates for Carcinogenicity Screens." Cancer Research 37, 1941-1945.

Fleiss, J.L. (1973). Statistical Methods for Rates and Proportions. John Wiley and Sons, Inc., New York.

Gaylor, D.W. (1978). "Methods and Concepts of Biometrics Applied to Teratology," in the Handbook of Teratology, Vol 4: Research Procedures and Data Analysis. J.G. Wilson and F.C. Fraser (eds.), Plenum Press, New York, 429-444.

Gradshteyn, I.S. and Ryzhik, I.M. (1980). Table of Integrals, Series and Products. Academic Press, New York.

Greenwood, M.; Yule, G.U. (1914). "On the Determination of Size of Family and of the Distribution of Characters in Order of Birth from Samples taken through Members of the Sibships." Journal of the Royal Statistical Society 77, 179-199.

Haseman, J.K.; Hogan, M.D. (1975). "Selection of the Experimental Unit in Teratology Studies." Teratology 12, 165-172.

Haseman, J.K.; Kupper, L.L. (1979). "Analysis of Dichotomous Response Data from Certain Toxicological Experiments." Biometrics 35, 281-293.

Haseman, J.K.; Soares, E.R. (1976). "The Distribution of Fetal Death in Control Mice and Its Implications on Statistical Tests for Dominant Lethal Effects." Mutation Research 41, 277-288.

Hogue, C. (1971). "Refilling the Empty Womb: Using Vital Statistics to Study Rapidity of Fetal Death Replacement." Master of Public Health Thesis from the University of North Carolina Department of Epidemiology.

Hosmer, D.W.; Lemeshow, S. (1980). "Goodness of Fit Tests for the Multiple Logistic Regression Model." Communications in Statistics A9(10), 1043-1069.

Jensh, R.P.; Brent, R.L.; Barr, Jr., M. (1970). "The Litter Effects as a Variable in Teratologic Studies of the Albino Rat." American Journal of Anatomy 128, 185-192.

Kalter, H. (1974). "Choice of the Number of Sampling Units in Teratology." Teratology 9, 257-258.

Kaplan, E.B.; Elston, R.C. (1972). "A Subroutine Package for Maximum Likelihood Estimation (MAXLIK)." Institute of Statistics Mimeo Series No. 823.

Kleinbaum, D.G.; Kupper, L.L.; Chambless, L.E. (1981). "Logistic Regression Analysis of Epidemiologic Data: Theory and Practice." To appear in Communications in Statistics.

Kruger, J. (1970). "Statistical Methods in Mutation Research," in Chemical Mutagenesis in Mammals and Man, F. Vogel and G. Rohrborn (eds.), Springer-Verlag, Heidelberg, 460-502.

Kupper, L.L.; Haseman, J.K. (1978). "The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments." Biometrics 34, 69-76.

Levine, P.J.; Symons, M.J.; Balogh, S.A.;et al. (1980). "A Method for Monitoring the Fertility of Workers. 1. Method and Pilot Studies." Journal of Occupational Medicine 22, No. 12, 781-791.

Levine, R.L.; Symons, M.J.; Balogh, S.A.; et al. (1981). "A Method for Monitoring the Fertility of Workers. 2. Validation of the Method Among Workers Exposed to Dibromochloropropane." Journal of Occupational Medicine 23, No. 3, 183-188.

Luning, K.G.; Sheridan, W.; Ytterborn, K.H.; Gullberg, U. (1966). "The Relationship Between the Number of Implantations and the Rate of Intra-Uterine Death in Mice." Mutation Research 3, 444-451.

McCaughran, D.A.; Arnold, D.W. (1976). "Statistical Models for Numbers of Implantation Sites and Embryonic Deaths in Mice." Toxicology and Applied Physiology 38, 325-333.

McKeown, T.; Record, R.G. (1955). "Maternal Age and Birth Order as Indices of Environmental Influence." (Unknown)

Mellin, G.W. (1962). "Fetal Life Tables: A Means of Establishing Perinatal Rates of Risk." Journal of the American Medical Association 180, No. 1, 11-14.

Namboodiri, N.K., Suchindran, C.M.; Wyman, K. (1980). "A Life Table Approach to the Study of Work-Fertility Relationships." A paper prepared for presentation at the Annual Meeting of the American Statistical Association, August, 1980.

Norton, A. (1952). "Incidence of Neurosis Related to Maternal Age and Birth Order." British Journal of Soc. Medicine 6, 253-258.

Palmer, A.K. (1974). "Statistical Analysis and Choice of Sample Units." Teratology 10, 301-302.

Potthoff, R.H.; Whittinghill, M. (1966). "Testing for Homogeneity: I. The Binomial and Multinomial Distributions." Biometrika 53, 167-182.

Selevan, S.G. (1977). "Reproductive History Studies - The Supporting Statement for Clearance for the Office of Management and Budget."

Shachtman, R.H.; Hogue, C.J. (1976). "Markov Chain Model for Events Following Induced Abortions." Operations Research 24, 916-932.

Shapiro, S.; Jones, E.W.; Densen, P.M. (1962). "A Life Table of Pregnancy Terminations and Correlates of Fetal Loss." Milbank Memorial Fund Quarterly XL, 8-45.

Sheps, M.C.; Menken, J.A. (1973). Mathematical Models of Conception and Birth. The University of Chicago Press, Chicago and London.

Spilerman, S. (1972). "The Analysis of Mobility Processes by the Introduction of Independent Variables into a Markov Chain." American Sociological Review 37, 277-294.

Staples, R.E.; Haseman, J.K. (1974). "Selection of Appropriate Experimental Units in Teratology." Teratology 9, 259-260.

Stein, Z.; Susser, M.; Warburton, D.; Wittes, J.; Kline, J. (1975). "Spontaneous Abortions as a Screening Device. The Effect of Fetal Survival on the Incidence of Birth Defects." American Journal of Epidemiology 102, No. 4, 275-290.

Trost, R.P.; Lurie, P. (1980). "Estimation of a Child Spacing Model with the Cox Regression Technique." A paper prepared for presentation at the Annual Meeting of the American Statistical Association, August, 1980.

VanRyzin, J. (1975). "Estimating the Mean of a Random Binomial Parameter with Trial Size Random." Sankhya 37, Series B, Part 1, 10-27.

Vuataz, L.; Sotek, J. (1978). "Use of the Beta-binomial Distribution in Dominant-Lethal Testing for 'Weak Mutagenic Activity', Part 2." Mutation Research 52, 211-230.

Williams, D.A. (1975). "The Analysis of Binary Responses from Toxicological Experiments Involving Reproduction and Teratogenicity." Biometrics 31, 949-952.

Weil, C.S. (1970). "Selection of the Valid Number of Sampling Units and a Consideration of Their Combination of Toxicological Studies Involving Reproduction, Teratogenesis, or Carcinogenesis." Food and Cosmetics Toxicology 8, 177-182.

Weil, C.S. (1974). "Choice of the Number of Sampling Units in Teratology." Teratology 10, 301.