Environmental Health Perspectives
Vol. 87, pp. 69–75, 1990

On the Consequences of Model Misspecification in Logistic Regression

by Melissa Dowd Begg* and Stephen Lagakos†

Logistic regression models are commonly used to study the association between a binary response variable and an exposure variable. Besides the exposure of interest, other covariates are frequently included in the fitted model in order to control for their effects on outcome. Unfortunately, misspecification of the main exposure variable and the other covariates is not uncommon, and this can adversely affect tests of the association between the exposure and response. We allow the term "misspecification" to cover a broad range of modeling errors including measurement errors, discretizing continuous explanatory variables, and completely excluding covariates from the model. This paper reviews some recent results on the consequences of model misspecification on the large sample properties of likelihood score tests of association between exposure and response.

Introduction

Data analysts are often interested in assessing the association between a response variable and an explanatory variable. In addition to collecting data on the response variable and explanatory variable of interest, they may also collect data on other covariates in order to control for the covariates' effects. Unfortunately, misspecification of the main explanatory variable and the other covariates is not uncommon, and this can affect tests of the association between the explanatory variable and the response. This paper reviews some recent results on the consequences of model misspecification on the validity and power of tests of association between an explanatory variable and a binary response variable.

Suppose that x denotes the explanatory variable of interest, and $\mathbf{z}$ denotes a vector of other explanatory variables. For simplicity of presentation, we shall hereafter refer to x and $\mathbf{z}$ as the exposure variable and covariates, respectively. When the outcome of interest is binary, logistic regression models are commonly used to study the association between exposure and response. If Y denotes the binary response, then the relationship between exposure and response is modeled as:

$$logit\ Pr(Y = 1|x,\mathbf{z}) = \theta + \alpha x + \beta'\mathbf{z} \qquad (1)$$

where θ, α, and β are unknown parameters. The null hypothesis of no association between exposure and outcome can be expressed as:

$$H_0\colon \alpha = 0$$

Given n independent and identically distributed observations of the form $(Y_i, x_i, \mathbf{z}_i)$, $i = 1,2,\ldots,n$, this hypothesis can be assessed using the likelihood score test of $\alpha = 0$, say $Q(x,\mathbf{z})$, which is asymptotically equivalent to tests of $\alpha = 0$ based on the maximum likelihood estimator of α and on the likelihood ratio statistic (*1*).

Suppose that x_i^* denotes a misspecified version of x_i, and $\mathbf{z}_i^*$ denotes a misspecified version of $\mathbf{z}_i$. We consider the test statistic, say $Q(x^*,\mathbf{z}^*)$, having the same functional form as $Q(x,\mathbf{z})$, but with x_i and $\mathbf{z}_i$ replaced by x_i^* and $\mathbf{z}_i^*$, respectively. We want to know the properties of this new statistic for different types of misspecification. We allow the term "misspecification" to cover a broad range of modeling errors for x and $\mathbf{z}$. It can include measurement error, mismodeling the functional form of x or $\mathbf{z}$ (e.g., using x^* = weight instead of x = weight^2), discretizing a continuous x or $\mathbf{z}$, or completely excluding covariates from the model. This misspecification can be arbitrary, but we require throughout this discussion that the distribution of Y conditional on x, x^*, $\mathbf{z}$, and $\mathbf{z}^*$ be equal to the distribution of Y conditional on x and $\mathbf{z}$ alone. In words, this means that once we have x and $\mathbf{z}$, x^* and $\mathbf{z}^*$ provide no additional information about Y. This paper investigates the consequences of using $Q(x^*,\mathbf{z}^*)$ rather than $Q(x,\mathbf{z})$ as a test of $\alpha = 0$. The issue of estimation, albeit interesting, will only be discussed briefly for the problem of omitted covariates.

There are various ways of assessing the ramifications of using $Q(x^*,\mathbf{z}^*)$ instead of $Q(x,\mathbf{z})$. We will focus on the asymptotic distributions of $Q(x,\mathbf{z})$ and $Q(x^*,\mathbf{z}^*)$ because their exact distributions are, in general, intractable.

*Division of Biostatistics, Columbia University, School of Public Health, 600 W. 168 Street, New York, NY 10032.
†Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115.
Address reprint requests to M. D. Begg, Division of Biostatistics, Columbia University, School of Public Health, 600 W. 168 Street, New York, NY 10032.

Under a specific sequence of models, we show that as the sample size goes to infinity,

$$Q(x,\mathbf{z}) \xrightarrow{L} N(\mu,1)$$

and

$$Q(x^*,\mathbf{z}^*) \xrightarrow{L} (\mu^*,1)$$

where $\xrightarrow{L}$ denotes convergence in distribution, and $\mu = 0$ when $\alpha = 0$. We can assess the asymptotic validity of $Q(x^*,\mathbf{z}^*)$ by studying μ^* when $\alpha = 0$. By comparing the magnitudes of μ and μ^* when $\alpha \neq 0$, we can assess asymptotic relative efficiency. This general theoretical result is then simplified analytically and evaluated numerically to examine particular types of misspecification.

In the first section of the paper we describe the general formulation of the problem. In the following three sections we look at a variety of situations in which mismodeling occurs. We first consider the situation in which the exposure is mismodeled in the absence of covariates. Next come cases in which the exposure is misspecified, but the covariates are modeled properly. Finally, we study cases in which both exposure and covariates are misspecified. All of these scenarios are followed by examples relating theoretical results to their consequences in practice.

General Formulation

Let Y_i denote the binary outcome, x_i the exposure, and $\mathbf{z}_i$ the vector of covariates for the ith of n independent observations. Then the likelihood score test of $\alpha = 0$ based on model in Eq. (1) takes the form:

$$Q(x,\mathbf{z}) = \sum_{i=1}^{n} x_i\left(Y_i - \frac{e^{\hat{\theta}+\hat{\beta}'\mathbf{z}_i}}{1 + e^{\hat{\theta}+\hat{\beta}'\mathbf{z}_i}}\right)/\sqrt{w}$$

where θ and $\hat{\beta}$ are the restricted maximum likelihood estimators (MLEs) of parameters θ and β when $\alpha = 0$, and w arises from the sample information matrix (*1*). Although the exact distribution of $Q(x,\mathbf{z})$ is quite complex, it can be shown to be asymptotically $N(0, 1)$ when $\alpha = 0$. This is used in practice to compute significance levels.

To obtain an approximation to the distribution of $Q(x,\mathbf{z})$ when $\alpha \neq 0$, one can use its asymptotic distribution for a sequence of contiguous alternative models to Eq. (1). This leads to the result (*1,2*) that

$$Q(x,\mathbf{z}) \xrightarrow{L} N(\mu,1),$$

where μ depends on θ, α, β and the joint distribution of x and $\mathbf{z}$. The magnitude of μ reflects the asymptotic power of $Q(x,\mathbf{z})$; the larger $|\mu|$ is, the larger the asymptotic power.

Now consider the asymptotic distribution of $Q(x^*,\mathbf{z}^*)$. To derive this limiting distribution, we again specify a particular sequence of contiguous alternative models in which the fitted model approaches the true model Eq. (1) as n goes to infinity. In this way, it can be shown (*2*) that $Q(x^*,\mathbf{z}^*)$ is also asymptotically normal:

$$Q(x^*,\mathbf{z}^*) \xrightarrow{L} N(\mu^*,1)$$

where μ^* depends on θ, α_0, β, the joint distribution of x, x^*, $\mathbf{z}$, and $\mathbf{z}^*$. Thus, $Q(x^*,\mathbf{z}^*)$ is asymptotically valid if $\mu^* = 0$ when $\alpha = 0$. The asymptotic relative efficiency (ARE) of $Q(x^*,\mathbf{z}^*)$ to $Q(x,\mathbf{z})$, when the former is valid, is given by $(\mu^*/\mu)^2$. The ARE can be interpreted loosely as the ratio of sample sizes needed to achieve the same power. For example, if the ARE is 0.9, then the correct test attains the same power as the mismodeled test with about 90% as many observations. In the following sections, we will evaluate this result when the model has no covariates, when the model contains correctly specified covariates, and when the model contains misspecified forms for both exposure and covariates. In each setting, we consider conditions for $Q(x^*,\mathbf{z}^*)$ to be valid and then examine its efficiency relative to the correct test.

Misspecified Exposure in the Absence of Covariates

The special case in which there is a misspecified exposure in the absence of other covariates has received considerable attention, and the reader is directed to the papers of Lagakos (*3*) and Tosteson and Tsiatis (*4*), as well as the references contained therein. Let us denote the correctly specified score test by $Q(x)$, and the misspecified version by $Q(x^*)$. The limiting distribution of $Q(x^*)$ can be derived for a sequence of contiguous alternative models to the true model. Such a sequence is described by Eq. (2):

$$\textit{logit}\ Pr\,(Y = 1|x) = \theta + \frac{\alpha_0}{\sqrt{n}}\,x \qquad (2)$$

where θ and α_0 are unknown parameters. It follows that $Q(x^*)$ is asymptotically normal, with mean μ^* and variance 1. It can be shown that $Q(x^*)$ is asymptotically valid (*2–4*); hence, the misspecification of x does not distort the asymptotic size of the score test of $\alpha = 0$. However, misspecification does affect the score test's power to detect an association between the exposure and response variables. With the above results, it can be shown (*3,4*) that the asymptotic relative efficiency (ARE) of $Q(x^*)$ versus $Q(x)$ is given by:

$$\text{ARE}[x^*:x] = [\textit{correlation}(x^*,x)]^2 \qquad (3)$$

Thus, the consequences of misspecifying exposure are reflected by the squared correlation between the fitted and correct measures of exposure. Recall that the ARE can be thought of as the ratio of sample sizes required by two tests in order to achieve the same power. This result says that correlation squared provides a way to make this comparison. Scale and location changes to x or x^* will not alter the ARE, since correlation enjoys the property of location/scale invariance. Furthermore, because correlation is a symmetric quantity, the ARE

of $Q(x^*)$ to $Q(x)$ when x is the appropriate exposure variable is equal to the ARE of $Q(x)$ to $Q(x^*)$ when x^* is the appropriate exposure variable.

The equality of the ARE for model misspecification to the square of the correlation arises not only in logistic models, but in a broad range of other settings. These include MLE tests based on linear models for measured response, MLE tests from logistic models for binary response, and likelihood ratio tests from logistic models for dichotomous response (*3*,*4*).

In order to get a sense for the effects of mismodeling, let us consider the consequences of a particular kind of misspecification, discretizing a continuous exposure variable. Other examples, including mismodeling a continuous exposure, misspecifying the dose metameter in a test for trend, misclassifying a categorical exposure, and errors in measurement, have been discussed elsewhere (*3*).

Discretizing a Continuous Exposure. Data on a measured exposure variable are often grouped into k categories prior to statistical analysis. Examples include classifying systolic blood pressure measurements as high or low, dividing age into 10-year categories, or grouping measured exposure levels of a potential carcinogen into categories of low, middle, or high. Discretization of a continuous exposure may occur because that is the only available information on that variable, or perhaps because the appropriate functional form relating the exposure to the outcome variable is unknown. The general result assures us that discretization does not distort the size of the test; however, it can cause a loss in power. We want to know how large this loss is, and whether there are rules for picking categories which will minimize this loss.

If we want to group a continuous exposure into several categories, a few choices must be made. First we must decide on k, the number of categories. Once k is selected, we must choose the $(k-1)$ cutpoints that form the boundaries for k intervals. Finally we must decide upon the value, say x_j^*, of x^* when x falls into interval j. For a given k and cutpoints, it is easily shown (*5*) that the optimal choice for x_j^* is $x_j^* = \theta_j$, where θ_j is the mean of x within interval j. The corresponding ARE, obtained by simplifying Eq. (3), is given by:

$$\text{ARE}[x^*:x] = \sum_{j=1}^{k} \pi_j(\theta_j - \theta)^2/var(x) \qquad (4)$$

where π_j is the probability that x falls into the jth interval, and $\theta = E(x)$. Connor (*5*) derives this same result as an optimization criterion for categorizing a continuous exposure that is linearly related to a dichotomous outcome variable. He provides an iterative algorithm for finding the optimal cutpoints for k intervals. In general, the optimal intervals are not equiprobable.

To illustrate the numerical results that arise from Eq. (4), let us consider instances in which x is distributed uniformly, normally, and exponentially. Results for these examples are displayed in Table 1. Even when x is split into as few as three categories, the optimization solution with nonequiprobable intervals maintains reasonably good relative efficiency. Serious loss in efficiency can occur, however, if equiprobable intervals are used (*3*). For example, consider an exposure x that follows an exponential distribution, but has been divided into four discrete categories. We see from Table 1 that the ARE$[x^*:x]$ is about 89% when the optimal intervals are used; however, this ARE reduces to 73% when equiprobable intervals are used. More generally, the table reveals the following interesting results. If x follows a uniform or normal distribution, the cost of using equiprobable intervals is not too great. But if x is exponential, the consequences of using equiprobable intervals are much more severe. This result gives rise to some simple guidelines for discretizing a measured exposure. If one feels fairly sure that the distribution of x is nearly symmetric, the choice of equiprobable intervals is reasonably safe. If x's distribution is highly skewed, one should strictly adhere to the optimization criterion for choosing intervals.

As another example, consider the situation in which a continuous exposure is dichotomized into categories of none versus some. Under these special circumstances, it can be shown (*3*) that the ARE reduces to a simple function of the proportion unexposed (π) and the coefficient of variation (C) of the nonzero exposures:

$$\text{ARE}[x^*:x] = \pi/(\pi + C^2)$$

Lessening π causes the ARE to decrease slightly, but this loss is small. Increasing C, on the other hand, can lead to a great loss in power; the score test becomes highly inefficient when the coefficient of variation is large.

Misspecified Exposure and Correctly Specified Covariates

Now let us consider a more complex situation; suppose that x is misspecified in the presence of correctly specified covariates. The goal, then, is to study the behavior of $Q(x^*,\mathbf{z})$, the misspecified version of the score test. The asymptotic distribution of $Q(x^*,\mathbf{z})$ can be derived for a sequence of contiguous alternative models to Eq. (1); such a sequence is described by Eq. (1) with α replaced by $\alpha_0/\sqrt{n}$:

$$logit\ Pr(Y = 1 \mid x, \mathbf{z}) = \theta + \frac{\alpha_0}{\sqrt{n}} x + \beta'\mathbf{z}$$

Begg and Lagakos show (*2*) that the statistic $Q(x^*,\mathbf{z})$ is asymptotically valid, since $\mu^* = 0$ whenever $\alpha_0 = 0$. However, computation of the efficiency of $Q(x^*, \mathbf{z})$ relative to $Q(x,\mathbf{z})$ can be quite complex for the general case (*2*). But if we restrict attention to the special case in which scalar $\mathbf{z}$ is independent of x and x^*, we obtain a much simpler result. Therefore, suppose that the covariate $\mathbf{z}$ is independent of both x and x^*; that is, that the covariate is balanced across exposures, as in a ran-

Table 1. ARE[x^*:x] when discretizing a continuous exposure.[a]

k, number of intervals	Distribution of x	Optimal interval probabilities[b]	ARE[x^*:x] Using optimal intervals[b]	ARE[x^*:x] Using equiprobable intervals
2	Uniform	0.50,0.50	0.75	0.75
	Normal	0.50,0.50	0.65	0.65
	Exponential	0.80,0.20	0.65	0.48
3	Uniform	0.33,0.33,0.33	0.89	0.89
	Normal	0.27,0.46,0.27	0.81	0.79
	Exponential	0.64,0.29,0.07	0.82	0.64
4	Uniform	0.25,0.25,0.25,0.25	0.94	0.94
	Normal	0.16,0.34,0.34,0.16	0.88	0.86
	Exponential	0.53,0.30,0.14,0.04	0.89	0.73
5	Uniform	0.20,0.20,0.20,0.20,0.20	0.96	0.96
	Normal	0.11,0.24,0.31,0.24,0.11	0.92	0.90
	Exponential	0.45,0.29,0.17,0.07,0.02	0.93	0.77
6	Uniform	0.17,0.17,0.17,0.17,0.17,0.17	0.97	0.97
	Normal	0.07,0.18,0.25,0.25,0.18,0.07	0.94	0.93
	Exponential	0.39,0.27,0.18,0.10,0.04,0.01	0.95	0.82

[a] From Lagakos (*3*).
[b] From Connor (*5*).

domized clinical trial. Then the ARE of $Q(x^*,\mathbf{z})$ to $Q(x,\mathbf{z})$ can be approximated in the following way (*2*):

$$\text{ARE}[(x^*,\mathbf{z}):(x,\mathbf{z})] \approx [\textit{correlation}(x^*\mathbf{z},x\mathbf{z})]^2$$

Hence the square of the correlation of $x\mathbf{z}$ and $x^*\mathbf{z}$ approximates the ARE of $Q(x^*,\mathbf{z})$ to $Q(x,\mathbf{z})$. This result resembles the result obtained when there were no covariates in the model, except that now we must take $\mathbf{z}$ into account. By symmetry, the ARE of $Q(x,\mathbf{z})$ when $Q(x^*,\mathbf{z})$ is appropriate is also equal to correlation squared.

As an example of this result, we will consider how choice of metameter can affect the performance of the trend test. We evaluate the ARE[$(x^*,\mathbf{z}):(x,\mathbf{z})$], allowing covariate $\mathbf{z}$ to follow different distributions.

Testing for Trend. Suppose x is an ordered categorical variable with k levels. These categories may represent dosage level in a rodent bioassay experiment or dose of medication in a clinical trial. We want to know whether or not response rates follow some trend in the levels of exposure x. The likelihood score test in this setting is equivalent to the well-known Cochran-Armitage test for trend. Use of this test requires selection of a metameter that quantifies the levels of exposure x. However, we usually do not know the correct metameter in advance. Thus, it is important to consider how using the wrong metameter for x affects the efficiency of the test. This problem has already been studied when there are no covariates (*3*); we now direct attention to the case in which a single, independent, correctly specified covariate $\mathbf{z}$ is present.

As an example let us consider an exposure with three levels. The chosen metameter can take on one of three basic shapes: linear, convex, or concave. We allow covariate $\mathbf{z}$ to follow the Bernoulli, normal, or exponential distribution. For a given distribution of $\mathbf{z}$, we can compute the ARE for a test based on one of the two non-optimal shapes relative to a test based on the optimal shape. (A subset of the values computed can be found in Tables 2 and 3.) Calculations show that the ARE's differ somewhat depending on the distribution of $\mathbf{z}$, but remain qualitatively the same. Briefly, numerical results show that in general convex (concave) metameters do quite well when the optimal weights are convex (concave). But choosing convex (concave) weights when the optimal weights are concave (convex) causes a great loss in efficiency. Linear weights, however, seem to enjoy fairly high relative efficiency, whether the optimal metameter is concave or convex. This simple scheme of results leads to rules of thumb for choosing a metameter for x. For example, if the dose metameter is believed almost certainly to be linear or convex, one should choose a mildly convex metameter. But if there is great uncertainty about the basic shape of the trend, linear weights are the safest bet. More generally, the similarity of these results with those in Lagakos (*3*) for the case of no covariates suggest that the effects of misspecifying x when covariates are correctly specified might be similar to those when there are no covariates.

Misspecified Exposure and Covariates

Let us now consider the situation in which both exposure and covariates are misspecified. Denote the test statistic in this case by $Q(x^*,\mathbf{z}^*)$. Again, one can derive the asymptotic distribution of the mismodeled test statistic by specifying a sequence of contiguous alternative models to Eq. (1) such that the fitted model approaches the true model under the null hypothesis as n goes to infinity. The limiting distribution of $Q(x^*,\mathbf{z}^*)$ has already been derived under very general conditions. This general approach specifies a sequence of alternative models to Eq. (1) in which α is replaced by $\alpha_0/\sqrt{n}$ and

Table 2. AREs for convex and concave metameters (x^*) when the true metameter (x) is linear (0,0.5,1) in tests for trend with k = 3 levels.[a]

		ARE[(x^*,z):(x,z)] when fitted metameter (x^*) is					
		Convex			Concave		
Distribution of covariate z	Relative group sizes	(0,0.2,1)	(0,0.1,1)	(0,0,1)	(0,0.8,1)	(0.0.9,1)	(0,1,1)
No covariate[b]	(0.33,0.33,0.33)	0.89	0.82	0.75	0.89	0.82	0.75
	(0.20,0.20,0.60)	0.93	0.89	0.84	0.91	0.84	0.77
	(0.20,0.30,0.50)	0.91	0.86	0.80	0.88	0.79	0.69
	(0.50,0.30,0.20)	0.87	0.79	0.69	0.91	0.86	0.80
	(0.60,0.20,0.20)	0.91	0.84	0.77	0.93	0.89	0.84
Normal (0, σ^2)	(0.33,0.33,0.33)	0.93	0.87	0.80	0.96	0.93	0.90
	(0.20,0.20,0.60)	0.97	0.95	0.92	0.98	0.96	0.94
	(0.20,0.30,0.50)	0.95	0.92	0.87	0.97	0.94	0.92
	(0.50,0.30,0.20)	0.91	0.83	0.73	0.95	0.92	0.89
	(0.60,0.20,0.20)	0.93	0.87	0.80	0.96	0.93	0.90
Exponential (λ)	(0.33,0.33,0.33)	0.91	0.85	0.77	0.94	0.90	0.86
	(0.20,0.20,0.60)	0.96	0.93	0.89	0.96	0.94	0.91
	(0.20,0.30,0.50)	0.94	0.89	0.84	0.95	0.91	0.87
	(0.50,0.30,0.20)	0.89	0.81	0.71	0.94	0.90	0.86
	(0.60,0.20,0.20)	0.92	0.86	0.78	0.95	0.91	0.88
Bernoulli (0.1)	(0.33,0.33,0.33)	0.93	0.87	0.79	0.95	0.92	0.89
	(0.20,0.20,0.60)	0.97	0.95	0.92	0.98	0.96	0.94
	(0.20,0.30,0.50)	0.95	0.91	0.86	0.96	0.94	0.91
	(0.50,0.30,0.20)	0.90	0.82	0.72	0.95	0.92	0.89
	(0.60,0.20,0.20)	0.93	0.87	0.80	0.95	0.93	0.90

[a] True exposure metameter (x) = linear (0.0.5,1).
[b] From Lagakos (*3*).

Table 3. AREs for alternative metameters (x^*) when the true metameter (x) is convex (0,0,1) in tests for trend with k = 3 levels.[a]

		ARE[(x^* ,z):(x,z)] when fitted exposure metameter (x^*) is					
		Linear	Convex		Concave		
Distribution of covariate z	Relative group sizes	(0,0.5,1)	(0,0.2,1)	(0,0.1,1)	(0,0.8,1)	(0,0.9,1)	(0,1,1)
No covariate[b]	(0.33,0.33,0.33)	0.75	0.96	0.99	0.42	0.33	0.25
	(0.20,0.20,0.60)	0.84	0.98	0.99	0.57	0.47	0.38
	(0.20,0.30,0.50)	0.80	0.98	0.99	0.47	0.35	0.25
	(0.50,0.30,0.20)	0.69	0.95	0.99	0.40	0.32	0.25
	(0.60,0.20,0.20)	0.77	0.96	0.99	0.52	0.44	0.38
Normal (0,σ^2)	(0.33,0.33,0.33)	0.80	0.96	0.99	0.61	0.55	0.50
	(0.20,0.20,0.60)	0.92	0.99	0.99	0.82	0.79	0.75
	(0.20,0.30,0.50)	0.87	0.98	0.99	0.72	0.67	0.63
	(0.50,0.30,0.20)	0.73	0.94	0.99	0.51	0.45	0.40
	(0.60,0.20,0.20)	0.80	0.96	0.99	0.61	0.55	0.50
Exponential (λ)	(0.33,0.33,0.33)	0.77	0.96	0.99	0.53	0.46	0.40
	(0.20,0.20,0.60)	0.89	0.98	0.99	0.75	0.70	0.64
	(0.20,0.30,0.50)	0.84	0.97	0.99	0.63	0.56	0.50
	(0.50,0.30,0.20)	0.71	0.94	0.99	0.46	0.39	0.33
	(0.60,0.20,0.20)	0.78	0.96	0.99	0.57	0.50	0.44
Bernoulli (0.1)	(0.33,0.33,0.33)	0.79	0.96	0.99	0.60	0.54	0.48
	(0.20,0.20,0.60)	0.92	0.98	0.99	0.81	0.77	0.73
	(0.20,0.30,0.50)	0.86	0.98	0.99	0.71	0.66	0.61
	(0.50,0.30,0.20)	0.72	0.94	0.99	0.50	0.44	0.39
	(0.60,0.20,0.20)	0.80	0.96	0.99	0.60	0.54	0.49

[a] True exposure metameter (x) = convex (0,0,1).
[b] From Lagakos (*3*).

z* approaches z at rate 0 ($1/\sqrt{n}$) as n goes to infinity. This latter assumption has no direct physical significance; it is merely a technique which guarantees a tractable result. It follows that this statistic also converges in distribution to a normally distributed random variable with mean μ^* and variance 1. The formula for μ^* is very complex and involves intricate expressions that depend on θ, α_0, β, and on the joint distribution of x, x^*, z, and z* (*2*).

In general, $Q(x^*,z^*)$ is not asymptotically valid. Clearly this result is reasonable, since we would expect mismodeling a covariate z that is not balanced across exposure groups to introduce bias. When $Q(x^*,z^*)$ is valid we can consider its asymptotic efficiency relative

to $Q(x,\mathbf{z}^*)$. As we would expect, misspecification of covariates causes a loss in asymptotic efficiency. However, formulas for the ARE$[(x^*,\mathbf{z}^*):(x,\mathbf{z}^*)]$ do not readily simplify. In general, numerical techniques are needed to evaluate these expressions and quantify the extent of power loss.

Results do simplify, to some extent, for the case of omitted covariates (*6*). Since it is well known that excluding covariates that are related to exposure alters test size and efficiency, we restrict attention to the case where the omitted covariates are independent of exposure. It has been shown (*6*) that omitting an important covariate will not distort test size; hence, the test statistic $Q(x^*,\mathbf{0})$ retains asymptotic validity. Covariate omission does, however, reduce efficiency. We have the following expression for the ARE of a misspecified test which excludes $\mathbf{z}$ versus a misspecified test which includes $\mathbf{z}$:

$$\text{ARE}[(x^*,\mathbf{0}):(x^*,\mathbf{z})] = 1 - \left\{\frac{\mathbf{E}[(p(\mathbf{z}) - \mathbf{E}[p(\mathbf{z})])^2]}{\mathbf{E}[p(\mathbf{z})](1 - \mathbf{E}[p(\mathbf{z})])}\right\}$$

where

$$p(\mathbf{z}) = \frac{e^{\theta+\beta'\mathbf{z}}}{1 + e^{\theta+\beta'\mathbf{z}}}$$

Unless $\mathbf{z}$ is degenerate, the term in brackets is always positive; hence the ARE is always less than one. Therefore, omitting important covariates causes a loss in asymptotic efficiency; this loss can be measured by evaluating the expression above for ARE. It can also be shown that the ARE of a test that omits $\mathbf{z}$ versus a complete test is the same, whether or not x has been correctly specified:

$$\text{ARE}[(x^*,\mathbf{0}):(x^*,\mathbf{z})] = ARE[(x,\mathbf{0}):(x,\mathbf{z})]$$

For further details, see Begg and Lagakos (*6*).

This result addresses the issue of covariate omission in a general way. Earlier results, however, have dealt with the consequences of omitting important covariates in particular applications. We present two such special cases as examples. The first result examines the consequences of omitting a covariate on estimating treatment effect. The second result, taken from the field of animal carcinogenicity experiments, studies the loss in efficiency incurred by omitting an important covariate from the model. For other examples, see the references in the papers by Gail et al. (*7*) and Ryan (*8*).

Estimation of Treatment Effect. Suppose that an important scalar covariate has been excluded from the fitted model, but that this covariate is balanced across exposure groups; that is, x and x^* are independent of z and z^*. Such is the situation in a randomized clinical trial where covariates are balanced across treatment groups. It is well known that the omission of a balanced covariate will not bias the estimate of treatment effect in the setting of linear models. However, Gail et al. (*7*) have shown that this is not necessarily the case with nonlinear regression. The authors show that when treatment x is binary, the omission of a balanced covariate $\mathbf{z}$ in logistic regression causes the estimate of treatment effect to be biased towards the null hypothesis. This result emphasizes the fact that for logistic models, randomization cannot guarantee unbiased estimates of treatment effect when important covariates are omitted.

Animal Carcinogenicity Experiments. As an example, let us consider a bioassay experiment in which a control group of animals is compared with an exposed group with respect to the development of a nonlethal tumor. One approach for analysis is the lifetime incidence test, which compares the proportions of tumor-bearing animals. This test is valid, provided that the compound in question does not alter longevity in the exposed group. However, Ryan (*8*) notes that this method is inefficient relative to other methods that adjust for age-at-death. One of these tests is the Hoel-Walburg test (*9*).

Dinse and Lagakos have shown (*10*) that the Hoel-Walburg test arises as a likelihood score test from a logistic model. There is one covariate, $\mathbf{z}$, in this model; it is a step function representing the logit of tumor prevalence in the control group. Similarly, the lifetime incidence test is just a special case of the Hoel-Walburg test, where z is simply a constant representing the constant logit of tumor prevalence in the control group. Hence the lifetime incidence test can be viewed as a misspecified model from which an important covariate (i.e., the logit of tumor prevalence) has been omitted. Ryan (*8*) has studied this problem in detail and has derived an expression for the ARE of the lifetime incidence test versus the Hoel-Walburg test. [It can be shown (*6*) that Ryan's formula follows as a special case of the general result for omitted covariates discussed earlier.] Ryan has evaluated this expression for the ARE when the prevalence function for the control group animals is assumed to be zero during the first year and linear thereafter. She shows that the lifetime incidence test can become very inefficient relative to the Hoel-Walburg test. When the slope of the prevalence function is close to zero, the lifetime incidence test almost matches the Hoel-Walburg test in efficiency. But as the slope increases, the ARE falls off precipitously.

Discussion

We have considered the consequences of misspecification in logistic regression. Types of misspecification can include mismodeling the functional form of a variable, mismeasuring a continuous variable, discretizing a continuous variable, misspecifying dose metameter in trend tests, or omitting an important covariate from the model. Our treatment of this problem has allowed misspecification to be arbitrary, but it has always required that the distribution of outcome Y conditional on $x,x^*,\mathbf{z}$, and $\mathbf{z}^*$ be equal to the distribution of Y conditional on x and $\mathbf{z}$ alone. We have explored the likelihood score test's validity and efficiency subject to mismodeling. Its bias and power characteristics were investigated for

cases with a misspecified exposure and no covariates, cases with misspecified exposure and correctly specified covariates, and cases with misspecified exposure and misspecified covariates.

The case with a single exposure variable has already been researched extensively. When there are no covariates, the misspecified score test is always valid. Its efficiency was evaluated by computing the ARE of a test based on the misspecified exposure variable versus a test based on the correctly specified exposure variable. The simple result is that the ARE[x^*:x] is equal to the square of the correlation between the fitted exposure variable and the correct exposure variable.

When there are other covariates besides the exposure, the score test retains its validity. However, when an independent scalar covariate is present, the ARE differs slightly from the ARE in the absence of covariates. The formula for ARE[(x^*,$\mathbf{z}$):(x,$\mathbf{z}$)] is approximately equal to the square of the correlation between $x\mathbf{z}$ and $x^*\mathbf{z}$. This formula resembles the formula for the ARE when there are no covariates, but takes into account the presence of z.

Finally, we considered cases in which there has been misspecification of both exposure and covariates. As we would expect, this case gives the most complex results. We find that bias is indeed of concern here. The score test is no longer valid in general. We also find that expressions for the ARE[(x^*,$\mathbf{z}^*$):(x,$\mathbf{z}^*$)] become extremely complicated in this setting. Evaluation of the ARE will usually require numerical techniques for the general case. Of particular interest in this setting is the question of the omitted covariate. It can be shown that the omission of a needed covariate causes biased estimates of treatment effect, and reduced efficiency in tests of association between exposure and response.

The methods given here for evaluating bias and efficiency prove to be quite flexible. They allow for misspecification of the exposure, the covariates, or both simultaneously. These results derive from the likelihood score test from a logistic model, but also apply to tests based on the maximum likelihood estimator of α and the likelihood ratio statistic, since all three tests are asymptotically equivalent. It has been beyond the scope of this paper to consider all possible types of misspecification of the exposure, all types of misspecification of the covariates, and all combinations thereof. Our purpose has been to provide the machinery for doing so and to give a few illustrative examples. The generality of the results allows us to think more generally about the effects of misspecification, but their ultimate value depends on detailed numerical evaluations to develop simple rules of thumb.

We thank John Orav and Louise Ryan for their helpful comments. This research was supported by grant CA39640 from the National Cancer Institute.

REFERENCES

1. Cox, D. R., and Hinkley, D. V. Theoretical Statistics. Chapman and Hall, London, 1974.
2. Begg, M. D., and Lagakos, S. W. Effects of mismodeling on tests of association based on logistic regression models. Technical report, Department of Biostatistics, Harvard School of Public Health, Boston, MA, 1989.
3. Lagakos, S. W. Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. Stat. Med. 7: 257–274 (1988).
4. Tosteson, T. D., and Tsiatis, A. A. The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates. Biometrika 75(3): 507–14 (1988).
5. Connor, R. J. Grouping for testing trends in categorical data. J. Am. Stat. Assoc. 67: 601–604 (1972).
6. Begg, M. D., and Lagakos, S. W. Loss in efficiency caused by omitting covariates from a logistic regression model. Technical report, Department of Biostatistics, Harvard School of Public Health, Boston, MA, 1989.
7. Gail, M. H., Wieand, S., and Piantadosi, S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika 71(3): 431–444 (1984).
8. Ryan, L. M. Efficiency of age-adjusted tests in animal carcinogenicity experiments. Biometrics 41: 525–531 (1985).
9. Hoel, D., and Walburg, H. E. Statistical analysis of survival experiments. J. Natl. Cancer Inst. 49: 361–372 (1972).
10. Dinse, G. E., and Lagakos, S. W. Regression analysis of tumor prevalence data. Appl. Stat. 32(3): 236–248 (1983).