Errors-in-variables for binary regression models

By R. J. CARROLL[1], C. H. SPIEGELMAN[2], K. K. G. LAN[3],
K. T. BAILEY[3] & R. D. ABBOTT[4]

[1]University of North Carolina. Research supported by the Air Force Office of Scientific Research, Contract AFOSR-80-0080.

[2]Statistical Engineering Division, National Bureau of Standards, Washington, D.C. 20234.

[3]Mathematical and Applied Statistics Branch, National Heart, Lung, and Blood Institute of the National Institutes of Health, Bethesda, Maryland 20205.

[4]Biometrics Research Branch, National Heart, Lung, and Blood Institute of the National Institutes of Health, Bethesda, Maryland 20205.

Summary

We consider in detail probit and logistic regression models when some of the predictors are measured with error. For normal measurement errors, the functional and structural maximum likelihood estimates (MLE) are considered; in the functional case the MLE is not generally consistent. Non-normality in the structural case is also considered. By an example and a simulation, we show that if the measurement error is large, the usual estimate of the probability of the event in question can be substantially in error, especially for high risk groups.

Some key words: Probit regression; Logistic regression; Functional models; Structural models; Measurement errors.

I. Introduction

The Framingham Heart Study (Gordon & Kannel, 1968; Truett, Cornfield & Kannel, 1967) is an on-going prospective study of the development of cardiovascular disease. This study has been the basis for a considerable amount of epidemiologic research, much of it through the use of logistic regression. For example, there has been considerable emphasis on analyzing the probability of developing coronary heart disease (CHD). In this instance, the response is binary:

$$\begin{aligned} Y &= 1 \quad \text{means person develops CHD} \\ &= 0 \quad \text{means person does not develop CHD.} \end{aligned} \tag{1.1}$$

Many analyses have attempted to relate baseline risk factors to the probability of developing CHD; these risk factors include systolic and diastolic blood pressure, serum cholesterol, history of smoking, etc. Ordinarily, at some point in the analysis, multiple logistic regression is employed.

It is well-known that many of the baseline risk factors are measured with error; systolic blood pressure is a good example (Rosner & Polk, 1979). One of us was asked by a number of investigators and at least one referee whether such measurement errors could substantially affect the logistic regression estimates and, if so, what could be done to correct for the measurement error. The present study is an outgrowth of these questions, although there are many important practical facets of the problem yet to be investigated.

In an interesting paper, Michalek and Tripathi (1980) discuss the effect of measurement error on ordinary logistic regression; see also Ahmed and Lachenbruch (1975). Michalek and Tripathi conclude that ordinary logistic regression will not be too badly disturbed by measurement error as long as such error is moderate. We feel that our methods, in providing alternatives to ordinary logistic regression, will help the experimenter to get a more precise understanding of the effect of the measurement errors, especially if they are severe.

Our model is as follows. We have a sample of N persons from a particular population, e.g., males aged 45-54. The ith person in the sample is assumed to have a vector of baseline risk factors $\underline{x}_i$, with the probability of developing disease (CHD) given by

$$P(Y_i = 1|\underline{x}_i) = G(\underline{x}_i'\underline{\beta}_o), \; i = 1, \ldots, N, \tag{1.2}$$

where $G(\cdot)$ is a known distribution function such as

$$G(a) = \{1 + \exp(a)\}^{-1} \quad \text{(Logistic Regression)}$$
$$G(a) = \Phi(a), \quad \text{(Probit Regression)},$$

where $\Phi(\cdot)$ is the standard normal distribution function. We will return to probit regression later, but it is important to remember that probit and logistic regression usually give similar results (Halperin, Wu & Gordon, 1979; Gordon, et al., 1977).

We will partition the risk factors $\underline{x}_i$ into components observed without and with error, so that

$$\underline{x}_i' = (\underline{w}_i' \; \underline{z}_i') \qquad (1.3)$$
$$\underline{\beta}_0' = (\underline{\beta}_{01}' \underline{\beta}_{02}').$$

In (1.3), $\{\underline{w}_i\}$ can be observed at nearly exact levels; age and sex are examples. In (1.3), the $\{\underline{z}_i\}$ are measured with nontrivial error and cannot be observed; rather we observe only

$$\underline{Z}_i = \underline{z}_i + \underline{u}_i. \qquad (1.4)$$

To begin the discussion we are going to assume that the $\{\underline{u}_i\}$ are independently and normally distributed with mean zero and covariance matrix Σ_M assumed nonsingular.

When the risk factors $\{\underline{z}_i\}$ observed with error are assumed to be constants, the model is usually called the functional model (Kendall & Stuart, 1979). In this instance, $\underline{\beta}$ and the N values $\{\underline{z}_i\}$ are unknown parameters, and the number of these unknown parameters increases with the sample size N, so that classical maximum likelihood theory does not apply. In fact, in the next section we show that in a very simple logistic regression model, the functional maximum likelihood estimate (MLE) of $\underline{\beta}$ is not consistent when Σ_M is known. This is in contrast to the functional MLE for linear regression, which is generally consistent (Gleser, 1981).

In Section 3 we study the more tractable structural model, wherein the $\{\underline{z}_i\}$ are themselves independent with common distribution function F, which we will also initially suppose is that of a normal random vector with mean $\underline{\mu}_z$ and covariance Σ_z. In effect, we study a conditional likelihood, replacing (1.2) by

$$P(Y_i = 1|\underline{w}_i, \underline{z}_i).$$

In Section 4, the non-normal case is discussed. In Section 5, we present a small Monte-Carlo study. In Section 6, we analyze the effect of measurement error on predicting the probability of CHD on the basis of systolic blood pressure.

2. The Functional Case

Consider logistic regression through the origin,

$$P(Y_i = 1|c_i) = \{1 + \exp(\alpha_0 c_i)\}^{-1}, \tag{2.1}$$

where α_o and $\{c_i\}$ are scalars. Because of measurement error, we observe

$$C_i = c_i + v_i, \tag{2.2}$$

where the errors $\{v_i\}$ are normally distributed with mean zero and variance σ_M^2 $(0 < \sigma_M^2 < \infty)$. For purposes of this example, we will assume σ_M^2 is known to the investigator.

In the circumstance that the measurement error variance is known, for linear regression the functional errors-in-variables maximum likelihood estimate of α is generally consistent and asymptotically normally distributed. We now outline why this happy circumstance does not carry over to logistic regression.

The maximum likelihood estimator (MLE) of α_0 for the functional model (2.1)-(2.2) with σ_M^2 known maximizes

$$\sum_{i=1}^{N}[Y_i \log G(\alpha c_i) + (1-Y_i)\log\{1-G(\alpha c_i)\}] - (2\sigma_M^2)^{-1} \sum_{i=1}^{N}(C_i-c_i)^2, \tag{2.3}$$

where

$$G(t) = \{1 + \exp(t)\}^{-1}$$

is the logistic distribution function. For this functional model, the parameters are $\{\alpha_0, (c_i)\}$. For given α, the estimates of $\{c_i\}$ satisfy

$$\hat{c}_i(\alpha) = C_i - \alpha\sigma_M^2 [G\{\alpha\hat{c}_i(\alpha)\} - Y_i], \quad i = 1, \ldots, N. \tag{2.4}$$

The MLE $\hat{\alpha}_0$ satisfies (2.4) and

$$N^{-1} \sum_{i=1}^{N} \hat{c}_i(\hat{\alpha}_0)[G\{\hat{\alpha}_0\hat{c}_i(\hat{\alpha}_0)\} - Y_i] = 0. \tag{2.5}$$

If the MLE exists and is unique, and if

$$N^{1/2}(\hat{\alpha}_0 - \alpha_0)$$

is asymptotically normally distributed with mean zero and positive, finite asymptotic variance, then one can prove (see Appendix) that

$$N^{-1} \sum_{i=1}^{N} \hat{c}_i(\alpha_0)[G\{\alpha_0 \hat{c}_i(\alpha_0)\} - Y_i] \xrightarrow{P} 0. \tag{2.6}$$

In (2.6), $\hat{c}_i(\alpha_0)$ satisfies (2.4). It turns out that (2.6) does not hold even in the following simple case: take $c_i = \pm 0.5$ (+ if i is odd, - otherwise) and $\sigma_M^2 = 1$ (we used numerical intergration to check this).

The preceding argument shows that even in the simplest of cases, the functional logistic errors-in-variables MLE will not be unique and asymptotically normal in the usual $\sqrt{N}$ sense. We believe this phenomenon carries over to other forms for the distribution function such as probit regression. In fact, for the model (2.1) - (2.2) with σ_M^2 known, we have been unable to construct any consistent and asymptotically normal estimate of α_0.

3. Structural Case: Normal Distribution

The model is given by (1.2) - (1.4), but in the structural case we eliminate the nuisance parameters $\{\underline{z}_i\}$ by assuming they are independent and normally distributed with mean vector $\underline{\mu}_z$ and covariance matrix Σ_z. The error vectors $\{\underline{u}_i\}$ are

also assumed to be independent (of one another and of $\{\underline{z}_i\}$) normal random vectors with mean $\underline{0}$ and covariance Σ_M. For the moment we shall assume that $\underline{\mu}_z$, Σ_z and Σ_M are known; we discuss more realistic cases near the end of the section. For a given general distribution function G in (1.2), we denote the marginal likelihood of the observed data by

$$L(G, \underline{\beta}_{01}, \underline{\beta}_{02}, \Sigma_M, \underline{\mu}_z, \Sigma_z).$$

Defining the dimension of $\underline{\beta}_0$ to be p, this marginal likelihood, which can be written as the product of the conditional likelihoods for Y_i given $\underline{Z}_i$, is proportional to

$$L(G, \underline{\beta}_{01}, \underline{\beta}_{02}, \Sigma_M, \underline{\mu}_z, \Sigma_z) = \prod_{i=1}^{N} S_{i+}^{Y_i} S_{i-}^{(1-Y_i)} \tag{3.1}$$

where

$$S_{i+} = A_1 \int G(\underline{w}_i' \underline{\beta}_{01} + \underline{z}' \underline{\beta}_{02}) \exp\{-0.5(\underline{Z}_i - \underline{z})' \Sigma_M^{-1} (\underline{Z}_i - \underline{z})\} \times \exp\{-0.5(\underline{z} - \underline{\mu}_z)' \Sigma_z^{-1} (\underline{z} - \underline{\mu}_z)\} dz \tag{3.2}$$

$$A_1 = (2\pi)^{-p} (|\Sigma_M|| \Sigma_z|)^{-1/2}, \tag{3.3}$$

and S_{i-} is defined by replacing $G(\cdot)$ by $1 - G(\cdot)$ in (3.2). Detailed calculations show that

$$S_{i+} = A_{3i}t_{i+},\ S_{i-} = A_{3i}(1 - t_{i+}),$$
$$t_{i+} = \int G\{\underline{w}_i'\underline{\beta}_{01} + (\underline{\beta}_{02}'A_2\,\underline{\beta}_{02})^{\frac{1}{2}}v + \underline{d}_{1i}'A_2\underline{\beta}_{02}\}\quad (v)dv, \qquad (3.4)$$

where

$$(v) = (2\pi)^{-\frac{1}{2}}\exp(-0.5\,v^2)$$
$$A_2 = (\Sigma_M^{-1} + \Sigma_z^{-1})^{-1}$$
$$\underline{d}_{1i} = \Sigma_M^{-1}\,\underline{z}_i + \Sigma_z^{-1}\,\underline{\mu}_z$$

$$d_{2i} = \underline{\mu}_z'\,\Sigma_z^{-1}\,\underline{\mu}_z + \underline{z}_i'\,\Sigma_M^{-1}\,\underline{z}_i$$
$$A_{3i} = A_1|A_2|^{\frac{1}{2}}(2\pi)^{p/2}\exp(-0.5d_{2i} + 0.5\underline{d}_{1i}'A_2\underline{d}_{1i}).$$

In effect, the calculation of the likelihood depends only on being able to evaluate (3.4). This is no easy matter in general for the logistic function $G(t) = \{1 + \exp(t)\}^{-1}$, although if the number of variables measured with error is small, (3.4) can in principle be evaluated by numerical integration. For probit regression, (3.4) can be evaluated explicitly; in fact,

$$t_{i+} = \Phi\{(\underline{w}_i'\underline{\beta}_{01} + \underline{d}_{1i}'A_2\,\underline{\beta}_{02})(1 + \underline{\beta}_{02}'A_2\,\underline{\beta}_{02})^{-\frac{1}{2}}\}. \qquad (3.5)$$

Since logistic and probit regression generally give similar estimates of event probabilities (Halperin, Wu & Gordon, 1979), in the rest of the paper we confine our discussion to probit regression.

In most instances, the nuisance parameters $\underline{\mu}_z$, Σ_z and Σ_M will be unknown. Joint estimation of these

parameters and $\underline{\beta}_0$ through the likelihood (3.1) may be computationally feasible by such devices as the E-M algorithm (Dempster, Laird & Rubin, 1977), although this remains to be explored. A simpler and reasonable alternative is through the method of pseudo maximum likelihood estimation (PMLE - see Gong & Samaniego, 1981). Computing PMLE's for $\underline{\beta}_0$ simply consists of finding estimates of $\underline{\mu}_Z$, Σ_Z and Σ_M and plugging these estimates into (3.1). One obvious estimate for $\underline{\mu}_Z$ is

$$\hat{\underline{\mu}}_Z = N^{-1} \sum_{i=1}^{N} \underline{Z}_i, \tag{3.6}$$

while an estimate for $\Sigma_Z + \Sigma_M$ is

$$\widehat{(\Sigma_Z + \Sigma_M)} = N^{-1} \sum_{i=1}^{N} (\underline{Z}_i - \hat{\underline{\mu}}_Z)(\underline{Z}_i - \hat{\underline{\mu}}_Z)'. \tag{3.7}$$

One common way to estimate Σ_M is by replication. For example, suppose that each variable subject to error but with unknown covariance is measured twice. Call these replicates $\underline{Z}_{i1}$, $\underline{Z}_{i2}$. Then, in terms of the earlier notation,

$$\underline{Z}_i = (\underline{Z}_{i1} + \underline{Z}_{i2})/2. \tag{3.8}$$

Since $\{\underline{Z}_i\}$ have common covariance Σ_M, we can compute the estimates

$$\hat{\Sigma}_M = \text{sample covariance of } \{(\underline{Z}_{i1} - \underline{Z}_{i2})/2\}, \tag{3.9}$$

$$\hat{\Sigma}_z = \widehat{(\Sigma_M + \Sigma_z)} - \hat{\Sigma}_M. \tag{3.10}$$

The substitutions (3.6)-(3.10) provide an easy way to obtain consistent and asymptotically normal estimates of $\underline{\beta}_0$, say $\hat{\underline{\beta}}_0$. There are many ways to estimate the covariance matrix of $\hat{\underline{\beta}}_0$. One method is the bootstrap (Efron, 1979; 1981); this cousin to the jackknife merely requires having enough computer time to calculate $\hat{\underline{\beta}}_0$ for sufficiently many randomly drawn (with replacement) samples of size N from the original data. Alternatively, one could use the theory of PMLE's given by Gong and Samaniego (1981) (actually, one must generalize their equations (2.5) and (2.6) slightly). The difficulty with this approach is also computational, as it involves taking derivatives of the log of (3.1) with respect to $(\underline{\beta}_{01}, \underline{\beta}_{02}, \underline{\mu}_Z, \Sigma_Z, \Sigma_M)$.

4. Structural Case: Non-Normal Distributions

In the previous section we have made the assumption that both the measurement errors $\{\underline{u}_i\}$ and the structural parameters $\{z_i\}$ are normally distributed. One may wish to take a more nonparametric view and not assume that either $\{\underline{u}_i\}$ or $\{\underline{z}_i\}$ are normal random variables. We will outline a method for this problem, retricting ourselves to the following situation:

The random variable z_i subject to measurement error is scalar. (4.1a)

The variable subject to measurement error is (4.1b)

replicated as in (3.8).

Of course, if the common density $h(z)$ of the $\{z_i\}$ were known, conditional likelihood methods could be used as in the previous Section. However, we are interested in situations for which $h(z)$ is not completely known. One very simple device is to assume that $h(z)$ has a simple two-term Edgeworth expansion, e.g.,

$$h(z) = (2\pi\sigma_z)^{-\frac{1}{2}} \exp[-0.5\{(z-\mu_z)/\sigma_z\}^2] \{1 - c_3(z^3-3z)/6 + c_4(z^4 - 6z^2 + 3)/24\}, \quad (4.2)$$

where μ_z and σ_z are the mean and variance of the $\{z_i\}$ and c_3 and c_4 are standard measures of skewness and kurtosis. Because of the replication assumed in (4.1b), these four parameters are easily estimated, giving us a sample based density with which to work. The multivariate case can also be handled, see Johnson and Kotz (1972).

Given that we either know or can estimate $h(z)$, the method of estimation we propose is based on nonlinear regression. It has the appealing feature that we do not need to know the distribution of the measurement errors $\{u_i\}$ in defining the estimator. In particular, we will turn the problem around and consider the distribution of z_i given Y_i and $\underline{w}_i$ {recall, $\underline{x}_i' = (\underline{w}_i', z_i)$}. Let $h(z|Y_i, \underline{w}_i)$ be the conditional density of z_i given Y_i and $\underline{w}_i$. This is a complicated but easily computed function of $h(z)$, Y_i, w_i, $\underline{\beta}_{01}$ and β_{02}. Define the conditional means of $\{z_i\}$ by

$$r(\underline{\beta}_{01}, \beta_{02}, y, \underline{w}_i) = E(z_i | Y_i = y, \underline{w}_i). \tag{4.3}$$

In analogy with nonlinear regression, for weights wgt_1 and wgt_2, we propose minimizing

$$wgt_1 \Sigma\{z_{i*} - r(\underline{\beta}_{01}, \beta_{02}, 1, \underline{w}_i)\}^2 Y_i \tag{4.4}$$

$$+ wgt_2 \Sigma\{z_{i*} - r(\underline{\beta}_{01}, \beta_{02}, 0, \underline{w}_i)\}^2 (1 - Y_i).$$

Actually computing the estimates of $\underline{\beta}_{01}$ and β_{02} is quite feasible because it relies only on nonlinear regression. Inference based on the estimates is complex; we have no simple large sample theory and suggest that bootstrap methodology be used.

5. A Monte-Carlo Study

A simulation study was performed for the probit model. Specifically,

$$P(Y_i = 1 | z_i) = \Phi(z_i - 1), \ (i = 1, \ldots, 200)$$

and we observe

$$z_{ij} = z_i + u_{ij}, \ j = 1, 2.$$

Here $\{z_i\}$ and $\{u_{ij}\}$ were independent normal random variables with mean zero and variances $\sigma_z^2 = \sigma_M^2 = 0.25$. Thus, the simulation concerns a situation in which the measurement error is large, as

is the sample size N = 200. All computations were performed at the National Institutes of Health Computing Center using the SAS statistical package, specifically the procedure NLIN. The experiments were replicated 100 times. The estimates of μ_z, Σ_z and Σ_M were obtained as described by (3.6) - (3.10).

In Table 1, we list the means and mean square errors for the estimates of intercept (= -1.0) and slope (= 1.0) obtained by the usual probit regression ($\hat{\beta}_{0P}$, $\hat{\beta}_{1P}$) and probit errors-in-variables (EIV) regression ($\hat{\beta}_{0E}$, $\hat{\beta}_{1E}$). This table is a classical expression of the trade-off between bias and variance, especially for the slopes. The usual probit slopes are badly biased but not particularly variable. The probit EIV slopes are relatively unbiased but quite variable; overall, they result in an approximately 23% gain over the usual probit regression in terms of mean square error.

Often more important than the estimates of individual parameters is the behavior of the estimated risk or probability function as a function of the true value of the predictor:

$$\text{Probit:} \quad \Phi(\hat{\beta}_{0P} + \hat{\beta}_{1P} z)$$
$$\text{Probit EIV:} \quad \Phi(\hat{\beta}_{0E} + \hat{\beta}_{1E} z).$$

In Fig. 1, we plot the average values of the risk or probability as a function of z, as well as the true risk function; these were averages over the 100 simulations for different values of z, smoothed by spline interpolation. Note that the probit EIV risk

function is approximately unbiased while the usual probit risk function is badly biased for those at highest risk.

In estimating the risk function, it turns out that there is not nearly the trade-off between bias and variance as there is for estimating individual parameters. In Fig. 2, we plot the mean square error functions as a function z; again, mean square errors were calculated for various z and then the function was interpolated by a spline available in the SAS procedure GPLOT. For the high risk cases, the probit EIV is noticeably better than the usual probit risk function. In Fig. 3, the ratios of mean square errors for the probit versus probit EIV risk functions are plotted.

We also experimented with the nonlinear least squares methodology of Section 4. We followed the suggestions of Section 4 with the exception that we assume normality. The resulting estimates had almost the same mean square error properties as the probit EIV estimators, a fact which we found both surprising and encouraging.

6. An Example

To get some idea of the possible effects of measurement error in a more realistic context, we considered some of the data from the Framingham Heart Study (Gordon & Kannel, 1968). The Framingham Study has followed a sample of the male and female population of Framingham (Massachusetts) biennially since around 1950 in order to study the development of cardiovascular disease. For purposes of this paper, data used here were on men

aged 45-54, systolic blood pressures being taken at exam four. Individuals were called diseased cases if they developed coronary heart disease within the six year interval after exam four. There were 513 cases, of whom 66 were eventually considered to be diseased cases.

For the average of the two systolic blood pressures, we estimated

$$\hat{\sigma}_z^2 = 1.14$$
$$\hat{\sigma}_M^2 = 0.10 = 0.09\ \hat{\sigma}_z^2$$

Hence, the apparent measurement error was quite small, with the usual probit and probit EIV estimates of slope, intercept and risk being only minimally different. At this point, we realized that we were ignoring other sources of variation which might be more appropriately classified as "measurement error." Specifically, one might think of the variance of systolic blood pressure as

$$\sigma_S^2 + \sigma_T^2 + \sigma_{ME}^2,$$

where

σ_S^2 = population variance of the "true" systolic blood pressures calculated at a fixed time, say 9:00 am,

σ_T^2= exam time of day effect; within individuals there is a diurnal effect for blood pressure, see Comstock (1957) and Gould, et al. (1981). Other effects may also be noted, e.g., those which could be attributed to nurse or physician reading the blood pressure or to the subject's physical or psychological disposition.

σ_{ME}^2 = "mechanical" measurement error as seen by differences in two readings.

In the analysis based on (3.6) - (3.10), we had

$$\hat{\sigma}_z^2 \doteq \sigma_s^2 + \sigma_T^2$$

$$\hat{\sigma}_M^2 \doteq \sigma_{ME}^2$$

when we actually should have had

$$\hat{\sigma}_z^2 \doteq \sigma_s^2$$

$$\hat{\sigma}_M^2 \doteq \sigma_T^2 + \sigma_{ME}^2.$$

We have no estimate of σ_T^2 for the Framingham males aged 45-54, so we decided upon the following device. Let $0<PVAR<1$ and define

$$\hat{\sigma}_z^2(new) = PVAR\ \hat{\sigma}_M^2 + (1-PVAR)\hat{\sigma}_z^2$$

$$\hat{\sigma}_M^2(new) = (1-PVAR)\hat{\sigma}_M^2 + PVAR\ \hat{\sigma}_z^2.$$

Basically, PVAR is something like the proportion of variance due to diurnal or other unmeasured effects.

In Fig. 4, we plot the probit EIV risk functions for the cases PVAR = 0, 0.2, 0.4, representing no, moderate and substantial time of day effects respectively. What is clear from Fig. 4 is that, if there is a large time of day effect, our estimate (PVAR = 0.0) of the relationship of risk for CHD and "true" systolic blood pressure could be badly biased for high risk patients.

References

Ahmed, S. & Lachenbruch, P. A. (1975). Discriminant analysis when one or both of the initial samples is contaminated: large sample results. EDV in Medizin und Biologie 6, 35-42.

Armitage, P. & Rose, G. A. (1966). The variability of measurements of casual blood pressure, I. A laboratory study. Clinical Sciences 30, 325-366.

Comstock, G. W. (1957). An epidemiologic study of blood pressure levels in a biracial community in the southern United States. The American Journal of Hygiene 65, 271-315.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood with incomplete data via the E-M algorithm. Journal of the Royal Statistical Society, Series B 39, 1-38.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. Annals of Statistics 7, 1-26.

Efron, B. (1981). Censored data and the bootstrap. Journal of the American Statistical Association 76, 312-319.

Gleser, L. J. (1981). Estimation in a multivariate "errors in variables" regression model: large sample results Annals of Statistics 9, 24-44.

Gong, G. & Samaniego, F. J (1981). Pseudo maximum likelihood estimation: theory and applications. Annals of Statistics 9, 861-869.

Gordon, T., Castelli, W. P., Hjortland, M. C., Kannel, W. B. & Dawber, T. R. (1977). Predicting coronary heart disease in

middle-aged and older persons: The Framingham Study. Journal of the American Medical Association 238, 497-499.

Gordon, T., & Kannel, W. E. (1968). Introduction and general background in the Framingham study - The Framingham Study, Sections 1 and 2. National Heart, Lung, and Blood Institute, Bethesda, Maryland.

Gould, B. A., Mann, S., Davies, A. B., Altman, D. G. & Raferty, E. B. (1981). Does placebo lower blood pressure? Lancet, 2, 1377-1381.

Halperin, M., Wu, M. & Gordon, T. (1979). Genesis and interpretation of differences in distribution of baseline characteristics between cases and non-cases in cohort studies. Journal of Chronic Diseases 32, 483-491.

Johnson, N. L. & Kotz, S. (1972) Distributions in Statistics: Continuous Multivariate Distributions. Wiley, New York.

Kendall, M. & Stuart, A. (1979). The Advanced Theory of Statistics, Volume 2, pp. 399-443. Macmillan Publishing Co, New York.

Michalek, J. E. & Tripathi, R. C. (1980). The effect of errors in diagnosis and measurement on the estimation of the probability of an event. Journal of the American Statistical Association 75, 713-721.

Rosner, B. & Polk, B. F. (1979). The implications of blood pressure variability for clinical and screening purposes. Journal of Chronic Diseases 32, 451-461.

Appendix

Proof of (2.6)

Assume as in Section 2 that

$$N^{1/2}(\hat{\alpha}_0 - \alpha_0) = O_p(1). \tag{A1}$$

Assume also the normalizing conditions

$$N^{-1}\sum_{i=1}^{N} c_i \to A \tag{A2}$$

$$N^{-1}\sum_{i=1}^{N} c_i^2 \to B. \tag{A3}$$

Then (A2) and (A3) imply

$$\max\{c_i^2/N:\ 1\le i\le N\} \to 0. \tag{A4}$$

From (2.4) and the definition (2.2), it follows that

$$\lim_{\varepsilon\to 0} \max_{1\le i\le N} \sup_{|\alpha-\alpha_0|\le\varepsilon} |\hat{c}_i(\alpha)-v_i|/(1+|\alpha_0|+|c_i|) = O_p(1). \tag{A5}$$

Further, by normality of $\{v_i\}$,

$$\max\{|v_i|/N^{1/2}:\ 1\le i\le N\} \xrightarrow{p} 0. \tag{A6}$$

Lemma A1 It follows that

$$\max\{|\hat{c}_i(\hat{\alpha}_0) - \hat{c}_i(\alpha_0)|:\ 1\le i\le N\} \xrightarrow{p} 0. \tag{A7}$$

Proof of Lemma A1 Make the definitions

$$H_i(u,\alpha) = u - c_i - v_i - \alpha\{G(\alpha u) - Y_i\},$$

so that

$$H_i\{c_i(\alpha),\alpha\} = 0.$$

The partial derivatives of H_i are

$$D_1 H_i(u,\alpha) = \frac{\partial}{\partial u} H_i(u,\alpha) = 1 + \alpha^2 G(\alpha u)\{1-G(\alpha u)\},$$
$$D_2 H_i(u,\alpha) = \frac{\partial}{\partial \alpha} H_i(u,\alpha) = -\{G(\alpha u) - Y_i\} + \alpha u G(\alpha u)\{1-G(\alpha u)\}.$$

By the chain rule,

$$\frac{\partial}{\partial \alpha}\hat{c}_i(\alpha) = -[D_1 H_i\{\hat{c}_i(\alpha),\alpha\}]^{-1} D_2 H_i\{\hat{c}_i(\alpha),\alpha\}. \qquad \text{(A8)}$$

By (A4), (A6), (A7) and (A8), it follows that for every $M>0$,

$$N^{-\frac{1}{2}} \max_{1\le i\le N} \sup_{|\alpha-\alpha_0|\le M/N^{\frac{1}{2}}} \left|\frac{\partial}{\partial \alpha}\hat{c}_i(\alpha)\right|$$
$$= O_p\{\max_{1\le i\le N} \sup_{|\alpha-\alpha_0|\le M/N^{\frac{1}{2}}} N^{\frac{1}{2}} |\hat{c}_i(\alpha)|/N^{\frac{1}{2}}\} \xrightarrow{P} 0.$$

This means that for every $M>0$,

$$\max_{1\le i\le N} \sup_{|\alpha-\alpha_0|\le M/N^{\frac{1}{2}}} N^{\frac{1}{2}} |\hat{c}_i(\alpha) - \hat{c}_i(\alpha_0)| \xrightarrow{P} 0, \qquad \text{(A9)}$$

which by (A3) completes the proof of Lemma A1.

The term on the left side of (2.6) can be written as

$A_{1N} + A_{2N} + A_{3N}$, where

$$A_{1N} = N^{-1} \sum_{i=1}^{N} \{\hat{c}_i(\alpha_0) - \hat{c}_i(\hat{\alpha}_0)\} [G\{\alpha_0 \hat{c}_i(\alpha_0)\} - Y_i],$$

$$A_{2N} = N^{-1} \sum_{i=1}^{N} \hat{c}_i(\hat{\alpha}_0)[G\{\alpha_0 \hat{c}_i(\alpha_0)\} - G\{\hat{\alpha}_0 \hat{c}_i(\hat{\alpha}_0)\}],$$

$$A_{3N} = N^{-1} \sum_{i=1}^{N} \hat{c}_i(\hat{\alpha}_0)[G\{\hat{\alpha}_0 \hat{c}_i(\hat{\alpha}_0)\} - Y_i].$$

By (2.5), $A_{3N} = 0$ and, since G is bounded, (A7) gives

$$A_{1N} \xrightarrow{P} 0.$$

Because G and its derivative are bounded, Lemma A1 says that

$$A_{2N} \xrightarrow{P} 0$$

as long as

$$N^{-1} \sum_{i=1}^{N} \{\hat{c}_i(\hat{\alpha}_0)\}^2 = O_p(1),$$

which follows from (A3) - (A5). This proves (2.6).

Note that in (2.6) we are essentially stating that we can replace $\hat{\alpha}_0$ by α_0 in (2.5) as long as we replace "=" to "$\xrightarrow{P}$".

This crucial substitution is true in fairly general circumstances. It does not follow from ordinary likelihood calculations because, in the functional case, the number of parameters increases with the sample size.

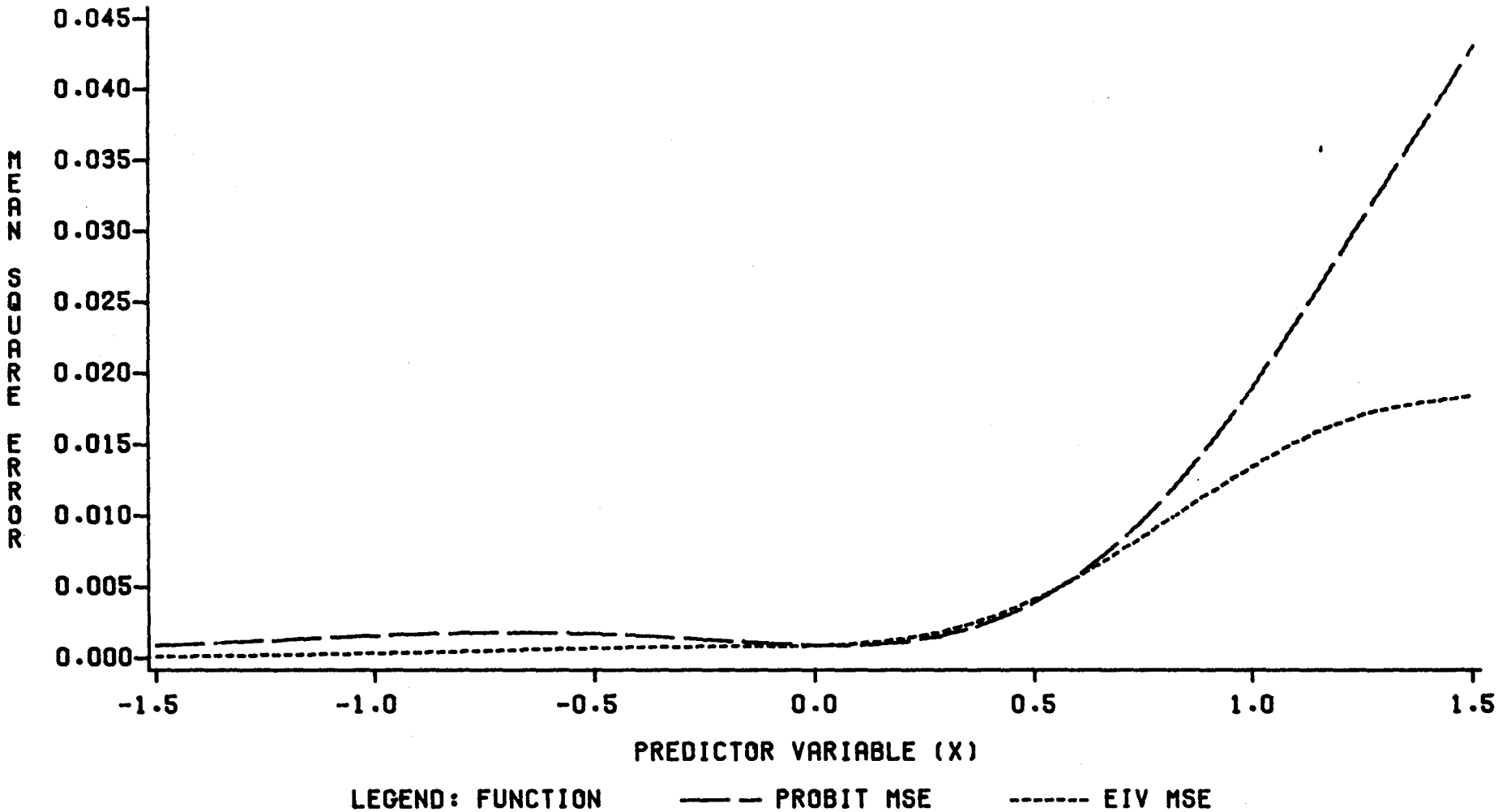
Table 1

	Usual Probit Regression		Errors-in-Variables Probit Regression	
	Intercept (= -1.0)	Slope (=1.0)	Intercept (= - 1.0)	Slope (=1.0)
Mean	-0.963	0.663	-1.011	1.070
Mean Square Error	0.0136	0.142	0.0155	0.110
Minimum	-1.246	0.324	-1.371	0.454
Maximum	-0.625	1.208	-0.663	2.368
Interquartile Range	0.148	0.244	0.174	0.403

Fig. 1 Average risk for simulation data.

Fig. 2. Average MSE for simulation data.

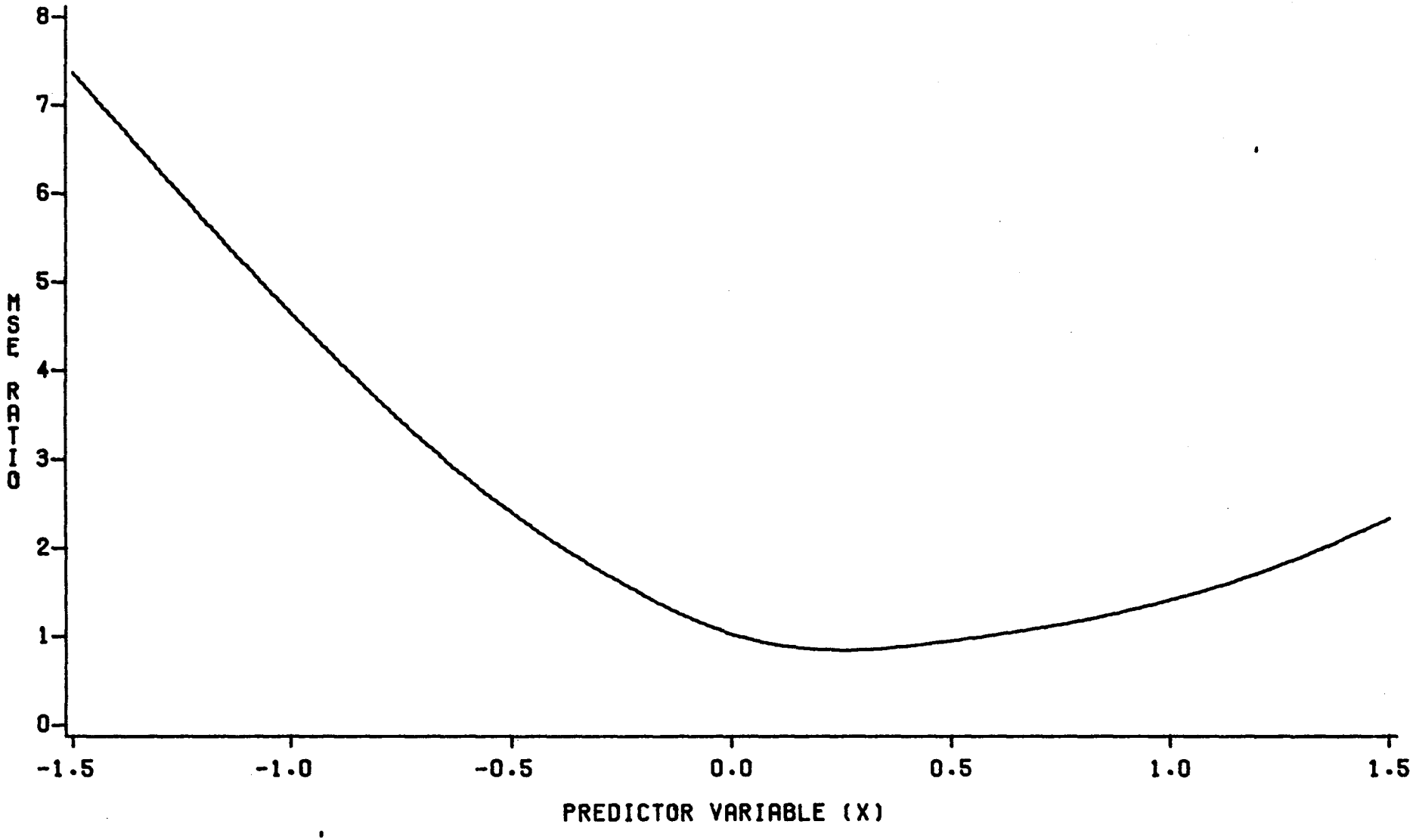
AVERAGE MSE FOR SIMULATION DATA



THE PREDICTOR VARIABLE X IS NORMAL, MEAN ZERO, VARIANCE 0.25
THE MEASUREMENT ERRORS ARE NORMAL, MEAN ZERO, VARIANCE 0.25
THE SAMPLE SIZE IS N=200
THERE WERE 100 MONTE-CARLO REPLICATIONS
THE TRUE PROBIT RISK PARAMETERS ARE INTERCEPT=-1, SLOPE=1

Fig. 3. EIV risk function efficiency for simulated data.

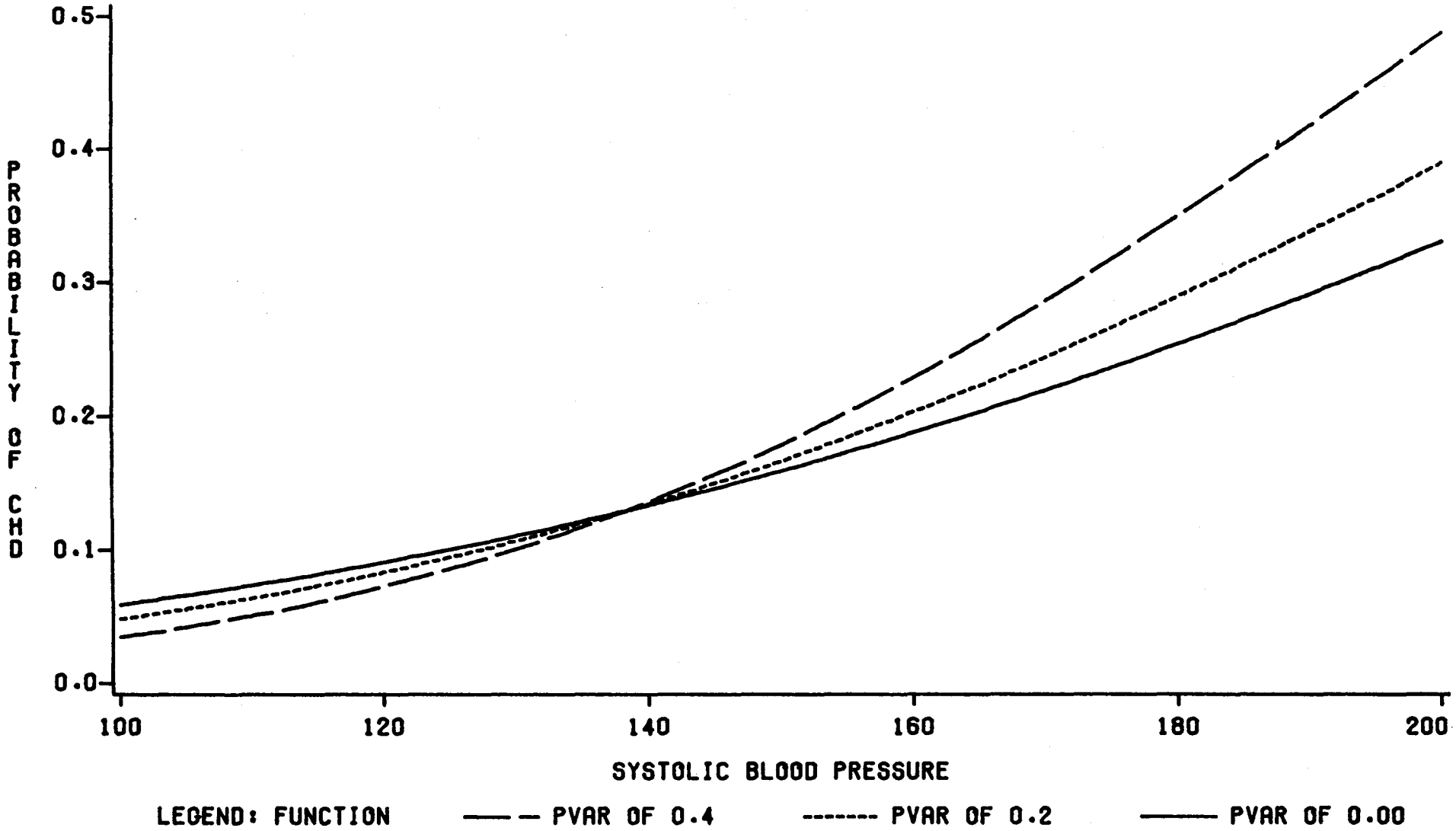
EIV RISK FUNCTION EFFICIENCY SIMULATED DATA



THE PREDICTOR VARIABLE X IS NORMAL, MEAN ZERO, VARIANCE 0.25
THE MEASUREMENT ERRORS ARE NORMAL, MEAN ZERO, VARIANCE 0.25
THE SAMPLE SIZE IS N=200
THERE WERE 100 MONTE-CARLO REPLICATIONS
THE TRUE PROBIT RISK PARAMETERS ARE INTERCEPT=-1, SLOPE=1

Fig. 4. Framingham data risks: Mixed variances.

FRAMINGHAM DATA RISKS: MIXED VARIANCES



THESE ARE PLOTS OF PROBIT RISK FUNCTIONS ON FRAMINGHAM DATA.
THE CASE PVAR OF 0.00 IS THE ORDINARY PROBIT EIV RISK FUNCTION.
PVAR IS IN GENERAL THE MIXING PROPORTION USED IN APPORTIONING
THE VARIANCES. TO TAKE INTO ACCOUNT TIME OF DAY VARIATION.