ELSEVIER

Sample-Size Calculations for the Cox Proportional Hazards Regression Model with Nonbinary Covariates

F.Y. Hsieh, PhD, and Philip W. Lavori, PhD
CSPCC, Department of Veterans Affairs Palo Alto Health Care System, Palo Alto, California

ABSTRACT: This paper derives a formula to calculate the number of deaths required for a proportional hazards regression model with a nonbinary covariate. The method does not require assumptions about the distributions of survival time and predictor variables other than proportional hazards. Simulations show that the censored observations do not contribute to the power of the test in the proportional hazards model, a fact that is well known for a binary covariate. This paper also provides a variance inflation factor together with simulations for adjustment of sample size when additional covariates are included in the model. *Control Clin Trials* 2000;21:552–560 © Elsevier Science Inc. 2000

KEY WORDS: *Sample size, proportional hazards regression, nonbinary covariates, variance inflation factor*

INTRODUCTION

In survival analysis, the Cox proportional hazards (PH) regression model assumes that the hazard function $\lambda(t)$ for the survival time T given the predictors $X_1, X_2, \ldots, X_k$ has the following regression formulation:

$$\log[\lambda(t|X)/\lambda_0(t)] = \theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_k X_k$$

where $\lambda_0(t)$ is the baseline hazard. The survival analysis allows the response, the survival time variable t, to be censored. In the use of this model, one often wishes to test the effect of a specific predictor, X_1, possibly in the presence of other predictors or covariates, on the response variable. The null hypothesis on the parameter θ_1 is H_0: $[\theta_1, \theta_2, \ldots, \theta_k] = [0, \theta_2, \ldots, \theta_k]$ tested against alternative $[\theta^* \theta_2, \ldots, \theta_k]$. In the proportional hazards model, θ_1 represents the predicted change in log hazards at one unit change in X_1 when covariates X_2 to X_k are held constant.

When comparing two groups in a univariate model, the group indicator X_1 is binary, and $\theta_1 = \log\Delta$ is the log hazard ratio of the two groups. Cox proposed testing H_0 with the Rao-type statistic, also known as the score statistic [1]. When

Address reprint requests to: F.Y. Hsieh, PhD, Cooperative Studies Program Coordinating Center, VA Medical Center (151K), 795 Willow Road, Building 205, Menlo Park, CA 94025 (E-mail: fhsieh@mailsvr.icon.palo-alto.med.va.gov).

Received October 13, 1998; accepted August 9, 2000.

© Elsevier Science Inc. 2000
655 Avenue of the Americas, New York, NY 10010
0197-2456/00/$–see front matter
PII S0197-2456(00)00104-5

there is only one binary covariate X_1, the score test is the same as the Mantel-Haenszel test and the log-rank test if there are no ties in survival times. It is known that the power of the log-rank test depends on the sample size only through the number of deaths. This simplifies the sample-size formula. For comparing two groups, Schoenfeld derived the following formula:

$$D = (Z_{1-\alpha} + Z_{1-\beta})^2 \, [P(1-P) \, (\log\Delta)^2]^{-1} \tag{1}$$

where D is the total number of deaths, P is the proportion of the sample assigned to the first treatment group, and $Z_{1-\alpha}$ and $Z_{1-\beta}$ are standard normal deviates at the desired one-sided significance level α and power $1 - \beta$, respectively [2]. Formula (1) was designed for a randomized comparison of two groups using survival analysis with covariates, although it applies to nonrandomized comparisons as well. In addition to formula (1), there are other sample-size formulas, such as Freedman's and Lakatos' formulas, proposed for the log-rank test to compare two survival distributions [3, 4]. Zhen and Murphy derived a formula for a nonbinary covariate assuming exponential survival time [5]. In this paper we derive a sample-size formula for a nonbinary covariate X_1 without assuming exponential survival time by generalizing formula (1). However, in our experience, a nonbinary covariate occurs most often in a nonexperimental context such as an epidemiologic study. In this context, one must often adjust for other confounding covariates to appropriately model the effect of X_1, the covariate of greatest interest.

Schoenfeld extended the multivariate model power calculation for binary X_1 to the case that additional covariates $X_2 \ldots X_k$ are included [2]. His argument depends on the assumption that X_1 is independent of $X_2 \ldots X_k$, as would occur if X_1 were randomly assigned in a controlled experiment. We show that Schoenfeld's argument also works when X_1 is nonbinary and is independent of $X_2 \ldots X_k$. This could be relevant in a randomized study if, for example, X_1 records dose levels to which the study subjects are randomized. In epidemiologic studies, X_1 is often a measure of a risk factor, such as numbers of cigarettes smoked per day, of interest to the investigators, and $X_2 \ldots X_k$ are possible confounders such as age and sex. By definition, covariates $X_2 \ldots X_k$ are correlated with the main factor of interest, X_1, and formula (1) doesn't apply. We describe a method for adjusting sample sizes to preserve power when X_1 is correlated with $X_2 \ldots X_k$.

SAMPLE-SIZE METHOD FOR NONBINARY COVARIATES

In a univariate model, without making assumptions about the distributions of covariate X_1 and survival time T, the total number of deaths required is given by the following formula, derived in Appendix A:

$$D = (Z_{1-\alpha} + Z_{1-\beta})^2 \, [\sigma^2 \, (\log\Delta)^2]^{-1} \tag{2}$$

where σ^2 is the variance of X_1 and $\log\Delta = \theta^*$ is the log hazards ratio associated with a one-unit change in X_1. Formula (2) is similar to formula (1) except that the variance of X_1, $P(1 - P)$ in formula (1) is now replaced by a more general term, σ^2. The required sample size is then equal to the number of deaths divided by the overall proportion of death. In practice, investigators may have a good idea of the overall death rate. In clinical trials with specific numbers of years of patient recruitment and follow-up, the overall death rate can also be approximately calculated [2, 5].

In deriving formula (2), we assumed that either (1) X_1 is the only covariate and PH holds or (2) there are additional covariates, PH holds for the full model with all covariates, and X_1 is independent of other covariates. These assumptions are likely to be justified in an experiment where X_1 has been randomized and the other covariates are introduced specifically to improve the fit of the data to the PH assumption. In a later section, we relax the independence assumption.

POWER SIMULATIONS (FOR A SIMPLE COVARIATE X_1)

Formula (2) is based on asymptotic arguments, in the limit of small effect size θ^*, for a proportional hazards model with a possibly nonbinary predictor. In this section, these estimated sample sizes are compared with power simulations. For each set of design parameters, enough simulations (6200, 3500, and 1000 for powers of 80, 90, and 95%, respectively) were generated to calculated power, with simulation errors within 1%, using a published simulation program [6]. The simulations of survival times use a simple random censoring pattern, and the single predictor variable X_1 has either a normal distribution $N(2,1)$ or a gamma distribution $G(4, 0.5)$. The latter is used to get an idea of the robustness of the results to non-normality in X_1. Without loss of generality, the predictor X_1 in both distributions has a mean of 2 and variance of 1. The results of simulations, tabulated in Table 1, indicate that formula (2) provides sample sizes in most designs with either a normal or gamma variate, within 1% of the expected power. The results also show that the power remains unchanged when the number of deaths remains constant while the number of patients entered varies.

EFFECT OF ADJUSTMENT FOR COVARIATES ON POWER

In randomized trials, the baseline variables have no population correlation with the treatment variables. Therefore, the inclusion of baseline variables or nonconfounding covariates in a correctly specified linear model usually improves the precision of the estimate of the treatment parameter by reducing the residual variance in the model. The adjustment for baseline variables in a linear model thus increases the power of the analysis. The covariates also adjust for chance confounding (despite random allocation of X_1 the observed conditional distribution of the other covariate depends on X_1).

In nonlinear models like PH, adjusting for covariates such as gender, race, or disease group may improve the fit of the data to the model. The hazard functions of two treatment groups may not be proportional if the covariates are not included in the model. For a conservative sample-size calculation, one can use formula (2) in designing an experiment since the use of covariates in the analysis can substantially increase power when the PH model holds in the full model.

However, in epidemiologic studies or nonrandomized trials where X_1 will not usually be independent of $X_2 \ldots X_k$, the covariate adjustments necessary for PH will not always increase the power. Note also that in this context θ^* is the "correct" alternative for X_1, the effect of X_1 adjusted for the confounders. Hsieh et al. [7–9] proposed increasing the sample size for linear regression,

Table 1 Powers of Sample Sizes Calculated from Formula (2) Compared with Simulations, One–Sided Significant Level 5%

Design Parameter	Death Rate	Number of Deaths	Sample Size	Expected Power (%)	Number of Simulations	Simulated Power (%)	
						Normal	Gamma
$\theta^* = 0.2$	0.3	215	717	90	3500	89.1 ± 0.53	88.5 ± 0.54
$\theta^* = 0.2$	0.5	215	430	90	3500	90.6 ± 0.49	90.7 ± 0.49
$\theta^* = 0.2$	1.0	215	215	90	3500	91.2 ± 0.48	86.7 ± 0.57
$\theta^* = 0.35$	0.1	51	510	80	6200	78.1 ± 0.53	83.8 ± 0.47
$\theta^* = 0.35$	0.3	51	170	80	6200	81.3 ± 0.50	79.2 ± 0.52
$\theta^* = 0.35$	0.5	51	102	80	6200	79.8 ± 0.51	78.8 ± 0.52
$\theta^* = 0.35$	1.0	51	51	80	6200	81.3 ± 0.50	74.5 ± 0.55
$\theta^* = 0.35$	1.0	70	70	90	3500	90.2 ± 0.50	87.0 ± 0.57
$\theta^* = 0.35$	1.0	89	89	95	1000	94.3 ± 0.73	91.0 ± 0.90
$\theta^* = 0.50$	0.1	25	250	80	6200	82.0 ± 0.49	83.7 ± 0.47
$\theta^* = 0.50$	0.3	25	84	80	6200	80.5 ± 0.50	82.3 ± 0.48
$\theta^* = 0.50$	0.5	25	50	80	6200	83.1 ± 0.48	81.1 ± 0.50
$\theta^* = 0.50$	1.0	25	25	80	6200	76.2 ± 0.54	83.0 ± 0.48
$\theta^* = 0.50$	1.0	34	34	90	3500	87.0 ± 0.57	88.1 ± 0.55
$\theta^* = 0.50$	1.0	44	44	95	1000	93.7 ± 0.77	93.3 ± 0.79

logistic regression, and Cox PH regression by a variance inflation factor when covariates are included in the model. In this context, the use of covariates (which are necessary for the model assumption to hold) increases the true variance of the estimate of the parameter, and hence formula (2) may overestimate the power. Robinson and Jewell showed that adjustment for covariates in logistic regression increases the variance of the estimate of a parameter and results in a loss of precision of the estimate [10]. However, they suggested that when testing for a treatment effect in randomized studies, it is always more efficient to adjust for covariates in the logistic model. Whitehead suggested increasing sample size to preserve power when covariates are adjusted in polytomous logistic regression with proportional odds model [11]. Ford et al. showed that in an exponential regression model, inclusion of important covariates can only increase the variance of the estimates of the covariates already in the model [12]. Lagakos and Schoenfeld discussed the effects on hazard ratios when the PH assumption no longer holds in a reduced model [13]. Here we suppose that the PH assumption holds in the full model and show that one can apply a variance inflation factor (VIF) to the estimate of sample size obtained by using formula (2) as if the reduced model were correct, and obtain good estimates of the true required number of deaths.

VARIANCE INFLATION FACTOR

In a regression model, the variance of the estimate b_1 of the parameter θ_1 is inversely related to the variance of the corresponding covariate X_1. For example, if we increase the scale of X_1 by a factor of 10, the variance of X_1 will increase by 100 and the variance of b_1 will decrease by 100. If covariates explain some of the variance of X_1, the same effect results. Let R be the multiple correlation coefficient $\rho_{1.23\ldots k}$ relating X_1 with $X_2, \ldots, X_k$. Then R^2 is the proportion of variance explained by the regression of X_1 on $X_2, \ldots, X_k$. In a multiple regression model with covariates $X_1, X_2, \ldots, X_k$, the conditional variance of $X_1 | X_2, \ldots, X_k$ is smaller than the marginal variance of X_1 by a factor of $1 - R^2$. Therefore the variance of b_1 estimated from the multiple regression model will increase by a factor of $1/(1 - R^2)$. In multiple linear regression the variance inflation factor can also be shown directly from the ratio

$$\mathrm{Var}_k(b_1)/\mathrm{Var}_1(b_1^*) = 1/(1 - R^2)$$

where $\mathrm{Var}_k(b_1)$ and $\mathrm{Var}_1(b_1^*)$ are the variances of the parameter estimate b_1 and b_1^* obtained from multiple regression models with k and 1 covariates, respectively [8]. In the PH context, to preserve the power we propose that the required number of deaths be calculated as if there were only one predictor and then inflated by the same proportion that the variance of the estimate of the effect of the predictor has been inflated by the adjustment for the other covariates. That is, $D = D_1/(1 - R^2)$, where we define $1/(1 - R^2)$ as the VIF and D_1 is the required number of deaths calculated from formula (2). In Appendix B, we present simulations that allow θ_2 to vary from small to large and to be statistically significantly different from 0, and demonstrate that this only increases the variation of the approximation.

EXAMPLE

A multiple myeloma data set from an example in the SAS PHGLM procedure is used to illustrate the sample-size calculation [14, 15]. In this data set, 65 patients were treated with alkylating agents and, during the study, 17 of the 65 survival times were censored. The data was fitted into a PH model to identify which of the nine prognostic factors are significant predictors.

Let us assume that LOGBUN is the variable X_1 of interest. The standard deviation of LOGBUN is $\sigma = 0.3126$ and the R^2 obtained from the regression of X_1 on $X_2, \ldots, X_9$ is 0.1837. The overall death rate is $1 - 17/65 = 0.738$. Suppose one wanted to have 80% power with a one-sided significance level of 5% to detect a log hazards ratio of $\log\Delta = \theta^* = 1$. By applying formula (2), with no other covariates, the required number of deaths would be $(1.645 + 0.842)^2/(0.3126 \times 1)^2 = 64$. Equivalently, a sample size of $64/0.738 = 87$ is needed. Now suppose that it was considered necessary to adjust for the other covariates, perhaps because of confounding. The approximated VIF for the full model obtained from $1/(1 - \rho_{1.23\ldots9}^2) = 1/(1 - R^2)$ is 1.225. The required total sample size for a full model can be approximated by $(87 \times 1.225) = 107$.

DISCUSSION AND CONCLUSION

If the PH assumption holds with respect to the full model with k covariates, it may no longer hold if some of the k covariates are left out of the model [12]. In addition, the inclusion of confounding variables may be necessary on scientific grounds to produce meaningful estimates of the effects of the covariate X_1. We assume that the model with k covariates is valid with a PH assumption, and we would like to estimate the required sample size for this model. The "naive" sample-size calculation begins with a reduced model with only one covariate. We then inflate the sample size to the correct power for the model with k covariates. We only use formula (2) as a start to the variance calculation to estimate the power in the full model. The VIF is obtained from the multiple correlation of X_1 with $X_2, \ldots, X_k$, which will often be available from prior studies, even when estimates of the parameters relating the covariates to survival are not available.

The proposed sample-size formula is useful for a nonrandomized study or for a randomized study with a nonbinary predictor. This formula does not assume anything about the distributions of the survival time and the covariates other than proportional hazards. The simulations so far attempted show that the power remains unchanged when the number of deaths or events remains constant while the number of patients entered varies under random censoring. In other words, the censored observations do not contribute to the power. The simulations also show that the proposed formula provides sample sizes, in a range of designs with either a normal or gamma variate, within 5% of the expected power. In PH regression, the adjustment for covariates in epidemiologic analyses may result in a true precision of an estimate of the correct parameter that is lower than naively estimated. We use simulation to show the VIF for confounders with various effect in mortality (Appendix B). When the coefficient of the confounder increases, the variation of the VIF approximations also increases. The simulations show that in general the proposed VIF approximates the ratio of estimated variances well.

This work was supported by the Department of Veterans Affairs Cooperative Studies Program. The authors wish to thank an associate editor for suggesting the formula to generate survival times from bivariate normal variates.

REFERENCES

1. Miller RG Jr. *Survival Analysis.* New York: Wiley; 1981.
2. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics* 1983;39:499–503.
3. Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. *Stat Med* 1982;1:121–129.
4. Lakatos E. Sample sizes based on the logrank statistics in complex clinical trials. *Biometrics* 1988;44:229–241.
5. Zhen B, Murphy J. Sample size determination for an exponential survival model with an unrestricted covariate. *Stat Med* 1994;13:391–397.
6. Akazawa K, Nakamura T, Moriguchi S, Shimada M, Nose Y. Simulation program for estimating statistical power of Cox's proportional hazards model assuming no specific distribution for the survival time. *Computer Methods and Programs in Biomedicine* 1991;35:203–212.
7. Hsieh FY. Sample size tables for logistic regression. *Stat Med* 1989;8:795–802.
8. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. *Stat Med* 1998;17:1623–1634.
9. Hsieh FY, Lavori WP, Bloch DA. A Simple Method of Sample Size Calculation for Multiple Linear, Logistic and Proportional Hazards Regressions. Presentation at Workshop in Biostatistics, Division of Biostatistics, Stanford University, May 14, 1998.
10. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* 1991;58:227–240.
11. Whitehead J. Sample size calculation for ordered categorical data. *Stat Med* 1993; 12:2257–2271.
12. Ford IF, Norrie J, Ahmadi S. Model inconsistency, illustrated by the Cox proportional hazards model. *Stat Med* 1995;14:735–746.
13. Lagakos SW, Schoenfeld DA. Properties of proportional-hazards scores tests under misspecified regression models. *Biometrics* 1984;40:1037–1048.
14. SAS Institute Inc. SAS Technical Report P-229, SAS/STAT software: Changes and Enhancements. Gary, NC, pp. 458–460, 1992.
15. Krall JM, Uthoff VA, Harley JB. A step-up procedure for selecting variables associated with survival. *Biometrics* 1975;31:49–57.
16. Tsiatis AA. A large sample study of Cox' regression model. *Ann Stat* 1981;9:93–108.

APPENDIX A

To simplify the derivations, we first assume that there is only one predictor in the model. The PH model assumes that the hazard function has the following relationship

$$\log[\lambda(t|X)/\lambda_0(t)] = \theta X$$

To test the null hypothesis $H_0 : \theta = 0$, the score statistic for the Cox proportional hazards model can be written in a simple form

$$S^2 = [\Sigma_D(X_i - E_i)]^2/\Sigma_D V_i$$

where i indexes the ordered death times; $R(i)$ and $D(i)$ identify the risk set and the death set, respectively, at the ith death time; $E_i = \Sigma_R X_j/n_i$, $V_i = \Sigma_R (X_j -$

$E_i)^2/n_i$, where n_i = number of patients in $R(i)$; and Σ_D and Σ_R are summations over $D(i)$ and $R(i)$, respectively [1]. Under the null hypothesis, S^2 is treated as chi-square distributed with one degree of freedom, or equivalently, S is standard normal with mean 0 and variance 1. To follow the derivation of Schoenfeld [2], we define

$$e_i = [\Sigma_R X_j \exp(\theta X_j)]/[\Sigma_R \exp(\theta X_j)]$$

The numerator of S can be written as

$$\Sigma_D(X_i - E_i) = \Sigma_D(X_i - e_i) + \Sigma_D(e_i - E_i)$$

The first term is asymptotically normal [16] with mean 0 and variance $\Sigma_D \Sigma_R (X_j - e_i)^2/n_i$. Since $\theta \to 0$, e_i approaches E_i, and the variance $\Sigma_D \Sigma_R (X_j - e_i)^2/n_i$ approaches $\Sigma_D V_i$. The first term has an asymptotic normal distribution $N(0, \Sigma_D V_i)$. Expanding the second term in a Taylor series about $\theta = 0$, this term approaches $\theta [\Sigma_D V_i]$. By adding the two terms together and dividing by the denominator $[\Sigma_D V_i]^{1/2}$, S is asymptotically normal with unit variance and mean equal to $D^{1/2} \theta \sigma$ where σ^2 is the variance of X and D is the expected number of deaths on the trial. Formula (2) follows this result.

To adjust for other covariates in an experiment where X is randomly assigned, we can assume that X is independent of other covariates and follow the derivation of Schoenfeld to obtain formula (2). Without the independence assumption, the estimate of the required number of deaths from formula (2) is in general too small and should be increased by a variance inflation factor for adjustment for other covariates (see text).

APPENDIX B

In a Cox proportional hazards regression model with two normally distributed covariates, we used simulations to demonstrate how well the ratio of the estimated variances $\text{Var}_2(b_1)/\text{Var}_1(b_1^*)$ is approximated by $1/(1 - \rho_{12}^2)$ where ρ_{12} is the Pearson correlation coefficient between X_1 and X_2.

The 100 computer simulations each used a sample size of 1000 generated from SAS [14]. First we generated variates X_1 and Z from a normal generator RANNOR and variate U from a uniform generator RANUNI. We then obtained variate $X_2 = \rho_{12} * X_1 + (1 - \rho_{12}^2)^{1/2} * Z$ where the two bivariate normal variables X_1 and X_2 had ten different correlation coefficients ρ_{12} ranging from 0.01 to 0.46. The survival times were generated from $T = -\log(U)/\exp(b_1X_1 + b_2X_2)$ where b_1 had a fixed value at 0.01 and b_2 had ten values ranging from 0.01 to 1.81. As a result, each of the 100 computer simulations consisted of a different pair of ρ_{12} and b_2. In each simulation, the survival time T was first regressed with X_1 alone and then with X_1 and X_2 using the SAS PHGLM procedure. The ratio of the variances, $\text{Var}_2(b_1)/\text{Var}_1(b_1^*)$, obtained from the two PHGLM procedures was compared with $1/(1 - \rho_{12}^2)$. The estimates of $\text{Var}_2(b_1)/\text{Var}_1(b_1^*)$ versus $1/(1 - \rho_{12}^2)$ from the simulations are plotted in Figure 1. The results show that the estimates of $1/(1 - \rho_{12}^2)$ closely approximate the ratio $\text{Var}_2(b_1)/\text{Var}_1(b_1^*)$. As expected, with the increases of the values of b_2, the variation of the approximation of the VIF to the variance ratio also increases (Figure 2).

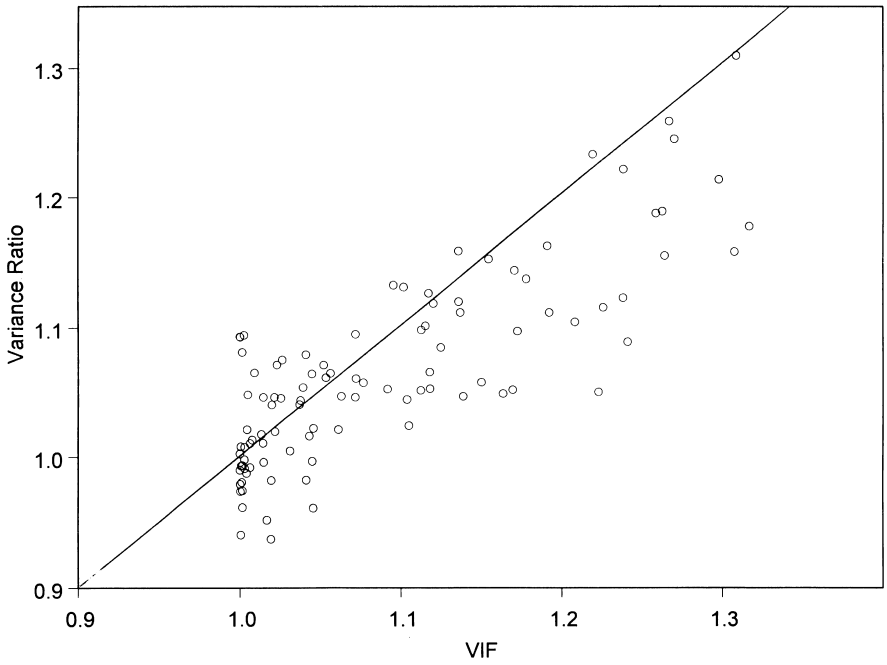

Figure 1 Results of 100 simulations: variance ratio versus VIF.

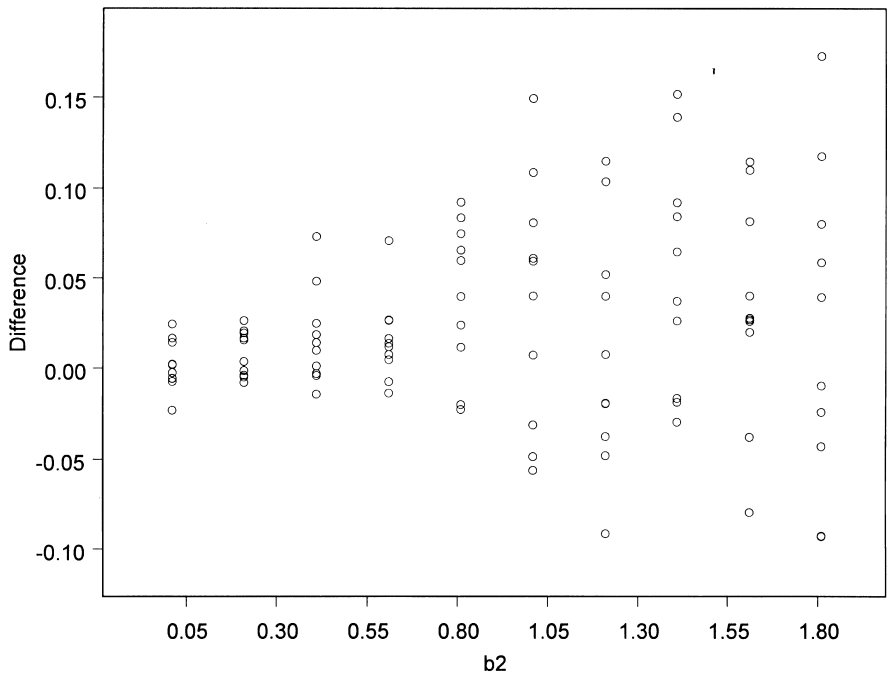

Figure 2 Difference of VIF and variance ratio versus b_2.