The Pennsylvania State University
The Graduate School

VARIANCE-BIAS TRADE-OFF IN GENERALIZED LINEAR REGRESSION MODELS

A Thesis in
Statistics
by
Junyi Lin

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2011

The thesis of Junyi Lin was reviewed and approved* by the following:

Runze Li
Professor of Statistics
Thesis Advisor, Chair of Committee

Bruce Lindsay
Willaman Professor of Statistics
Head of the Department of Statistics

Naomi Altman
Associate Professor of Statistics

*Signatures are on file in the Graduate School.

Abstract

The variance-bias trade-off has been partially discussed for linear and logistic regression models, but not for generalized linear models as a whole. In this paper, we derive the bias of the treatment effect in covariate-unadjusted models, when some important covariates are omitted. This result encourages the use of the covariate-adjusted approach in general. On the other hand, we show that for a broad class of generalized linear models, estimation of the treatment effect obtained from covariate-adjusted models have larger variances compared to those obtained from covariate-unadjusted models. This result reveals the potential loss of efficiency related to the covariate-adjusted approach, particularly when sample size is not large. These theoretical results are illustrated through examples, a simulation study and a real data example.

Table of Contents

List of Tables

Acknowledgments

First of all, this thesis would not have been done without my advisor, Dr. Runze Li. I am grateful to my advisor for his many helpful ideas and discussions on this thesis. Secondly, I want to say thank you to Dr. Naomi Altman, who provided me insightful suggestions during my comprehensive exam. Last but not the least, I want to deeply thank Dr. Bruce Lindsay for his precious time and valuable suggestions in improving this thesis. I appreciate all faculties and students in my department who provided me helpful suggestions during the completion of this master thesis.

I also own my thanks to my dear parents and my beloved wife, Danqi Zhu. Without their love and support, I cannot finish this thesis by myself.

This thesis research was supported by grants from the National Institute on Drug Abuse (NIDA) grant P50-DA10075. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIDA.

Chapter 1

Introduction

1.1 Linear Regressions and Generalized Linear Regressions

In statistics, classical linear regression refers to modeling the relationship between the response variable and explanation variables such that the model linearly depends on some unknown parameters. Assume a vector of observations $\mathbf{y}$ with n components is a realization of a random variable $\mathbf{Y} = (y_1, \cdots, y_n)'$. Given the design matrix $\mathbf{X}$, the components of $\mathbf{Y}$ are assumed to be independently distributed with means $\mu = (\mu_1, \cdots, \mu_n)'$. For ordinary linear models, the specification for the vector μ in terms of a unknown vector of parameters $\beta = (\beta_1, \cdots, \beta_p)'$ takes the form

$$\mu = \mathbf{X}\beta, \tag{1.1}$$

where the design matrix $\mathbf{X} = (x_1, \cdots, x_n)'$ is $n \times p$.

More specifically, given the design matrix $\mathbf{X}$, the components of $\mathbf{Y}$ are independent normal variables with $E(\mathbf{Y}|\mathbf{X}) = \mu = \mathbf{X}\beta$ and $\text{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2 I_{n\times n}$, where σ^2 is constant and $I_{n\times n}$ is the identity matrix. In other words,

$$Y = \mathbf{X}\beta + \epsilon, \tag{1.2}$$

where $\epsilon \sim N(0, \sigma^2 I_{n\times n})$.

The ordinary least squares (OLS) method is used to estimate the unknown parameter β. The OLS estimator is a very common estimator of β since it is conceptually simple and computationally straightforward. The main idea of OLS is to minimize the sum of squared residuals and this leads to the following closed form for the estimator of β (under the assumption that $\mathbf{X}$ has the full column rank):

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}. \tag{1.3}$$

When the response variable has a conditional distribution other than normal, linear regression may not be suitable any more. For instance, if the response variable is a binary variable and if one is interested in the conditional probability of a specific event $(p_i, i = 1, \cdots, n)$ for given values of explanatory variables, it is a bad idea to regress the probability p_i on explanatory variables directly since: (1) a linear model may give predicted values of p_i outside $(0, 1)$; (2) the variance of p_i is $p_i(1-p_i)$, which is not constant. The latter one cannot be fixed by using weighted least squares because p_i is unknown.

This requires ones to use more flexible and more general regression models to fit outcomes like binary variables. Generalized linear models (GLM) were introduced by Nelder and Wedderburn (1972) and popularized by McCullagh and Nelder (1989). GLMs extend ordinary linear regression models in two ways. First, they allow the response variable to be modeled by a regular exponential families, which include some of the most commonly used statistical models. Second, they allow a monotone function of the mean response to vary linearly with covariates.

More precisely, a generalized linear model consists a three-part specification: a *random component*, a *systematic component*, and a *link function* which specifies the connection between the random and systematic components. Conditional on $\mathbf{X}$, the components of $\mathbf{Y}$ are assumed to have exponential family density

$$f(y;\theta,\phi) = \exp\Big\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\Big\}, \tag{1.4}$$

where θ is called the natural parameter or the canonical parameter and ϕ is the dispersion parameter. This is the random component of the model. The conditional expectation of $\mathbf{Y}$ given $\mathbf{X}$ has the form $g(\mu) = \eta$ where $\eta = \mathbf{X}\beta$. η is the systematic component, called the linear predictor and $g(\cdot)$ is the link function.

Following are the most common examples of GLMs.

- The normal linear regression model for continuous responses,

$$y_i \sim N(\mu_i, \sigma^2),$$
$$\mu_i = x_i'\beta.$$

- The logistic regression model for binary responses,

$$y_i \sim \text{Binomial}(1, p_i),$$
$$\log\big(\frac{p_i}{1-p_i}\big) = x_i'\beta.$$

- The Poisson regression model for count responses,

$$y_i \sim \text{Poisson}(\mu_i),$$
$$\log \mu_i = x_i'\beta.$$

Here are some advantages of GLMs over the classic linear regression models: (1) the response variable is not required to have a normal distribution; (2) the choice of the link function, $g(\cdot)$, is separate from the selection of random component thus allowing more flexibility in modeling; (3) the variance does not need to be constant; (4) models are fitted via maximum likelihood estimation (MLE), which has the optimal properties of estimators. However, GLMs have some limitations as well. The predictors are assumed to be linear and responses must be independent.

Using the first Bartlett identity, one can derive that the mean μ is related to the canonical parameter θ by $\mu = b'(\theta)$. Using the second Bartlette identity, one can confirm that the variance has the form $b''(\theta)a(\phi)$. Here $b''(\theta)$ is also called the variance function. One question is how to select the link function $g(\cdot)$? Since $\mu = b'(\theta)$ and $\eta = g(\mu)$, then $\eta = g \circ b'(\theta)$. An obvious choice is $g(\cdot) = b'^{-1}(\cdot)$, called the canonical link. With this link function, it follows that $\theta = \eta = \mathbf{X}\beta$. For normal distributions, Binomial distributions and Poisson distributions, the canonical links are identity function, logistic function and log function, respectively. Some non-

canonical links are also commonly used, e.g., *probit* link for Binomial distributional family.

It is convenient to define the model in terms of μ and $\eta = g(\mu)$, and thus θ does not play a role. Estimation of μ does not involve the dispersion parameter ϕ. For most cases, ϕ is not treated as a parameter in the same sense as μ. One carries out estimation and inference under an assumed value of ϕ. If ϕ needs to be estimated, one finds a way to estimate it first and then treats the estimate as if ϕ were fixed and known.

MLE for β may be carried out via Fisher scoring,

$$\beta^{(t+1)} = \beta^{(t)} + \big\{ - El''(\beta^{(t)})\big\}^{-1} l'(\beta^{(t)}), \tag{1.5}$$

where l is the log-likelihood function for the entire sample. The Fisher scoring procedure is the standard method to get MLEs, and it is a Newton-Raphson algorithm using the expected rather than observed information matrix. For an arbitrary link,

$$\frac{\partial l}{\partial \beta} = \mathbf{X}'\mathbf{A}(\mathbf{y} - \mu), \tag{1.6}$$

$$-E\big(\frac{\partial^2 l}{\partial \beta \partial \beta'}\big) = \mathbf{X}'\mathbf{W}\mathbf{X}, \tag{1.7}$$

where $\eta = (\eta_1, \cdots, \eta_n)'$, $\mathbf{A} = \text{Diag}\Big[\text{Var}(y_i)\big(\frac{\partial \eta_i}{\partial \mu_i}\big)\Big]^{-1}$, $\mathbf{W} = \text{Diag}\Big[\text{Var}(y_i)\big(\frac{\partial \eta_i}{\partial \mu_i}\big)^2\Big]^{-1}$. Then an iteration of Fisher scoring is

$$\beta^{(t+1)} = \beta^{(t)} + \big(\mathbf{X}'\mathbf{W}\mathbf{X}\big)^{-1}\mathbf{X}'\mathbf{A}(\mathbf{y} - \mu), \tag{1.8}$$

and $\mathbf{W}$, $\mathbf{A}$ and μ are calculated from $\beta^{(t)}$. The final value for $\left(\mathbf{X'WX}\right)^{-1}$ after convergence provides the estimated covariate matrix for $\hat{\beta}$, and the diagonal elements of this matrix are the squared standard error for the estimated coefficients.

(1.8) can be rearranged to get an equivalent procedure called *adjusted dependent variable regression*, which is of form of iterative-reweighted-least-squares (IRWLS). More precisely, (1.8) can be rewritten as

$$\beta^{(t+1)} = \left(\mathbf{X'WX}\right)^{-1}\mathbf{X'Wz}, \tag{1.9}$$

where

$$\mathbf{z} = \eta + \left(\frac{\partial \eta}{\partial \mu}\right)(\mathbf{y} - \mu).$$

The iterative procedure (1.9) can be viewed as regressing the adjusted dependent vector $\mathbf{z}$ on $\mathbf{X}$ with weight matrix $\mathbf{W}$ to obtain an updated estimate β.

1.2 Generalized Linear Regressions in Clinical Trials

Generalized linear models are widely used in randomized clinical trials and observational studies. Clinical trials are conducted to collect data for new drugs or devices. In designing a clinical trial, a sponsor usually has the goal of obtaining a statistically significant result showing a significant difference in outcomes between the groups who receive the study treatments. In a randomized clinical trial, each

subject is randomly assigned to a treated group or a control group before the experiment. When the number of subjects is sufficient large, this random allocation of treatments to subjects provides a balance between treatment groups for confounding variables (variables correlate with both the response variable and explanatory variables). Thus a randomized clinical trial can provide the most supportive evidence that the study treatment causes the expected effect. By contrast, in an observational study, the assignment of subjects into a treated group or a control group is beyond the control of the investigator. Thus, investigators only observe potential correlations between treatments and outcomes. More details about clinical trials are beyond the scope of this proposal, and readers who are interested in clinical trials can refer to Pocock (2004).

When applied to randomized clinical trials or observational studies, GLMs are frequently misspecified due to the fact that some important covariates are omitted. In randomized clinical trials, important covariates may be unknown or unmeasured. In observational studies, omission of covariates arises when either there is the inability to collect all relevant factors or potential incomplete understanding of the study. It is also common that some covariates are excluded from the final model due to the concern that an overfitted model may yield inefficient estimates of parameters of interest. As a result it is important to study the impact of omitted covariates on the estimated effects of included covariates in these cases.

It is well known that under assumptions of classic linear regression, randomization leads to unbiased estimates of treatment effect even when important covariates are omitted (Cox, 1958). Furthermore, adjustment for desired covariates can improve the precision (small variance) of the treatment effect (Fisher, 1932). The

improvement in the precision is due to the reduction of the residual variance.

However, it has been recognized that the property of unbiasedness cannot be extended to generalized linear models. For example, suppose (Cox, 1970) Y is a binary outcome, T and X are binary covariates, and that the true model for the probability of response is given by

$$\text{logit}\{\Pr(Y = 1|T, X)\} = \alpha + \beta T + \gamma X. \tag{1.10}$$

Suppose the 2×2 classification (Y, T) in sequence $(1, 1), (1, -1), (0, 1), (0, -1)$ has counts 900, 500, 100, 500 for $X = 1$, and counts 500, 100, 500, 900 for $X = -1$. The data are summarized in Table 1.1 and Table 1.2. In Table 1.1, it is easy to

Table 1.1. Synthetic Data for logistic regression regarding bias

	X=1				X=-1		
	T=1	T=-1	Total		T=1	T=-1	Total
Y=1	900	500	1400	Y=1	500	100	600
Y=0	100	500	600	Y=0	500	900	1400
Total	1000	1000	2000	Total	1000	1000	2000

Table 1.2. Synthetic Data from Table 1.1 after combination over levels of X

	T=1	T=-1	Total
Y=1	1400	600	2000
Y=0	600	1400	2000
Total	2000	2000	4000

verify that T and X are independent, and the true model leads to an estimate of

$\beta = \frac{1}{2}\log(9)$. In Table 1.2, the fitted logistic model is

$$\text{logit}\{\Pr(Y = 1|T)\} = \alpha^* + \beta^* T, \tag{1.11}$$

and omitting X yields an estimate of $\beta^* = \frac{1}{2}\log(49/9)$.

Gail et al. (1984) showed that the asymptotic bias from omitting covariates is zero if the link in the GLM is a linear function or a log function. Using Taylor series expansion, they presented an approximation of bias magnitude for regular cases. Neuhaus and Jewell (1993) proposed a geometric approach, which leads to analogous results as in Gail et al. (1984). A term called "Bias Factor" was developed in their paper by expanding around a different point from that in Gail et al. (1984). Drake and McQuarrie (1995) considered estimating the bias in observational studies of exposure effects, when some but not all covariates are omitted. All these approaches adopt the functional relationship between the parameters in the true model and the misspecified model.

In some situations, the gains in precision regarding covariate adjustment do not apply to generalized linear models, either. For example, a logistic regression is fitted to a specific $2 \times 2 \times 2$ contingency table (Robinson and Jewell, 1991). Using

Table 1.3. Synthetic Data for logistic regression regarding variance

	X=0				X=1		
	T=1	T=0	Total		T=1	T=0	Total
Y=1	10	20	30	Y=1	40	80	120
Y=0	20	64	84	Y=0	5	16	21
Total	30	84	114	Total	45	96	141

Table 1.4. Synthetic Data from Table 1.3 after combination over levels of X

	T=1	T=0	Total
Y=1	50	100	150
Y=0	25	80	105
Total	75	180	255

both T and X as covariates leads to $\beta = \log(1.6)$ with associated standard error 0.354 while using T only yields $\beta^* = \log(1.6)$ as well with associated standard error 0.286. Note that in this example, T and X are correlated. Robinson and Jewell (1991) compared the variances of the two estimators, $\hat{\beta}$ and $\hat{\beta}^*$, in a logistic regression when both T and X are dichotomous variables. Neuhaus (1998) presented expressions for the effect of omitted covariates on the efficiency of the estimated effects of the included covariates in testing the hypothesis of no treatment effect ($\beta = \beta^* = 0$).

Meanwhile, some experience shows that for some clinical trials, statistical models which adjust for covariates are in close agreement with the simpler unadjusted treatment comparisons. The properties of covariate-adjustment approach are not fully understood, and there remains confusion as to what is an appropriate statistical strategy. This thesis research is motivated by this confusion.

Previous theoretical results are obtained under the assumption that there are no interactions between treatment and covariates, while this assumption may not necessarily be realistic. In clinical trails, non-crossover (or quantitative) interactions between covariates and treatment are to be expected and crossover (or qualitative) interactions between covariates and treatment are possible (Peto, 1982; Gail and

Simon, 1985). In epidemiologic studies, analysis of interaction between gene and environment or interaction between two major exposure variables can be the major interest (Stümer and Brenner, 2002; Zou, 2008; Richardson and Kaufman, 2009). Interactions between predictive variables are also important in psychology and the social science (McClelland and Judd, 1993).

It is natural to ask whether those previous theoretical results still hold when there are treatment/covariates interactions. In a framework of generalized linear models which contain possible interactions between the main predictive variable (e.g. treatment) and covariates, we first investigate the bias caused by omitted covariates. Some bias approximations are appropriate when there are no interactions between the main predictive variable and covariates. However, they may not be meaningful in the presence of interactions. Consequently, we derive the relationship between the effects of the main predictive variable in the potentially misspecified model and the true model. Our results indicate that the difference between these two effects can be large in the presence of interactions. Specifically, omitting a covariate that interacts with the main predictive variable may induce large bias to the treatment main effect estimate if this covariate is not centered (by subtracting the overall mean). This result is different from the conclusion that omitting a covariate (that does not interact with the main predictive variable) with a small effect leads to negligible bias as in Gail et al (1984). Furthermore, we present a new approach to assess the variance comparison. We prove that covariate-adjusted approaches lead to variation inflation for a broad class of generalized linear models. Conclusions on logistic regression in Robinson and Jewell (1991) can be obtained from this class of regression models as a special case.

In our opinion, a model selection procedure is a good way to balance the accuracy and precision, although its practical implementation is itself an interesting and promising research area with many unsolved issues. How to incorporate variable selection procedures in formulating covariate-adjustment in large clinical trials is also an aim of our thesis research.

The rest of this thesis is organized as follow. Chapter 2 provides a brief review of empirical and theoretical results on accuracy and precision of treatment effect estimates of GLMs in the presence of model misspecification. Regularized regression is also reviewed in Chapter 2. Chapter 3 studies the accuracy and precision in the presence of treatment/covariates interactions, and presents a new approach to assess the precision for a broad class of GLMs. Chapter 4 illustrates our results through simulation studies and a real data example. Chapter 5 discusses the future work, and all technique results are left in Appendix A.

Chapter 2

Literature Review

This chapter mainly focuses on the literature review for this thesis research. We first briefly review the MLEs of misspecified models, and then summarize related empirical and theoretical results for biases and variances of misspecified models. The importance of treatment/covariates interactions is also involved in this chapter.

2.1 Basic Model Assumptions

Suppose Y is the response variable, X is a vector of covariates, and T is the main explanatory variable (e.g., a treatment indicator variable that takes values 1 or -1). Given X and T, the conditional expectation of Y satisfies

$$E(Y|T,X) = h(\eta)$$
$$\eta = \alpha + T\beta + X'\gamma, \tag{2.1a}$$

where $h(\cdot)$ is a known function.

Suppose also that it is mistakenly assumed that

$$E(Y|T,X) = E(Y|T) = h(\eta^*)$$
$$\eta^* = \alpha^* + T\beta^*. \tag{2.1b}$$

2.2 Maximum Likelihood Estimation of Misspecified Models

Maximum likelihood estimate (MLE) has become one of the most important approaches for statistical inference, after Fisher (1922, 1925) advocated the method of maximum likelihood. A key assumption on the properties of the maximum likelihood estimator (Wald, 1949; LeCam, 1953) is that the structure of the true parameter lies within a specified parametric family of models, i.e., the model must be correctly specified. However, in many cases, this assumption may not be satisfied.

Based on results of Wald (1949), Huber provided some general conditions, under which the MLE converges to a well-defined limit even when the model is not correctly specified. Akaike (1973) pointed out that when the true model is unknown, the MLE is a natural estimator for the parameters which minimize the Kullback-Leibler Information Criterion (Kullback and Leibler, 1951).

For model (2.1a) and model (2.1b), let $\theta' = (\alpha, \beta, \gamma')$, $(\theta^*)' = (\alpha^*, \beta^*)$. Assume $f(y|\theta, T, X)$ is the true (conditional) density function of Y, and $f(y|\theta^*, T)$ is the misspecified (conditional) density function. The MLE, $\hat{\theta}^*$, of θ^* under the mis-

specified model (2.1b) converges to the value that minimizes the Kullback-Leibler divergence between model (2.1a) and model (2.1b):

$$E\Big[\log\big\{f(y|\theta,T,X)/f(y|\theta^*,T)\big\}\Big]. \tag{2.2}$$

Here the expectation is taken with respect to the true model.

White (1982) gave the asymptotic normal distribution of $\hat{\theta}^*$ under certain regularity conditions. More precisely, he showed that

$$\sqrt{n}(\hat{\theta}^* - \theta^*) \xrightarrow{D} N\big(0, C(\theta^*)\big), \tag{2.3}$$

where $C(\theta^*) = A(\theta^*)^{-1}B(\theta^*)A(\theta^*)^{-1}$, and

$$\begin{aligned} A(\theta^*) &= E\Big\{\partial^2 \log f(y|\theta^*,T)/\partial\theta^*\partial(\theta^*)'\Big\}, \\ B(\theta^*) &= E\Big\{\partial \log f(y|\theta^*,T)/\partial\theta^* \cdot \partial \log f(y|\theta^*,T)/\partial(\theta^*)'\Big\}. \end{aligned}$$

Again, the expectation is taken with respect to the true model.

For the true model (2.1a), $A(\theta) = B(\theta)$, and (2.3) reduces to the classic asymptotic normal distribution of MLEs. In general, $A(\theta^*) \neq B(\theta^*)$, and thus special care must be taken for statistical inference in the presence of misspecification. Gail (1988) and Neuhaus (1998) observed that $A(\theta^*) = B(\theta^*)$ for all binary regression models. In other words, for binary regression models, the information matrix of the misspecified model provides the correct value of the covariance matrix when using the results of White (1982).

2.3 Covariate-Adjusted Approaches or Not?

The analysis of the primary objectives of randomized clinical trials often are not adjusted for covariates, except possibly for stratification variables. People often debate whether a covariate-adjusted approach should be adopted or not as the primary analysis. Grouin et al. (2004,2005) provided guidance: when subgroup analysis can be done; when they should be done; and their interpretations.

Hauck et al. (1998) reviewed the literature regarding logistic and Cox (proportional hazards) regression models and advocated the adjustment for important prognostic covariates in order to come as close as possible to the clinically most relevant subject-specific measure of treatment effect since omitting covariates from the analysis of randomized trials leads to a loss of efficiency as well as a change in the treatment effect being estimated. They pointed out that additional benefits from adjustment for important covariates would be an increase in efficiency of tests for no treatment effect and improved external validity, which is particularly relevant to meta-analyses.

Ford and Norrie (2002) reviewed previously published literature and showed that the impact of including covariates in models used to estimate the magnitude of treatment effects in long-term clinical trials is different from what would be predicted from results for the classic linear model. They used a data from clinical trials to evaluate the role of covariates in estimating treatment effects and risk in long-term clinical trials.

Based on a recent survey of 50 trial reports in four major journals, Pocock et

al (2002) examined how literature in the medical journals used baseline data on each patient at randomization to do (i) subgroup analysis; (ii) covariate-adjusted analysis; (iii) baseline comparions. Some major problems and key issues were highlighted, including inconsistencies in the use of covariate-adjustment and the lack of clear guidelines on covariate selection. They recommended to adjust for the appropriate covariates (that is, the strong predictors of outcome) and to make one's statistical policy for covariate adjustment completely objective. Pocock et al (2002) also advocated manipulation of variable selection procedure, and encouraged more methodological research on this topic. They believed that variable selection procedure is a useful role in formulating covariate-adjustment in larger trials.

As pointed out in Neuhaus et al. (1991), there exists a close connection between the bias analysis and the generalized linear mixed models. If we view the omitted covariate as an random intercept, then the bias due to omission of covariates can be used to explain the differences between cluster-specific models (or conditional models) and population-averaged models (or marginal models) (Zeger, et al. 1988). Cluster-specific models are similar to models without omissions ($\eta = \alpha + T\beta + X'\gamma$) while population-averaged models are analogous to ones with omitted covariates ($\eta^* = \alpha^* + T\beta^*$).

Senn (2004) discussed the existing controversy about the use of marginal and conditional models, particularly in the analysis of data from longitudinal studies. The seeming differences between marginal and conditional models were addressed to be caused by preimposed unidentifiable constraints on the random effects. Senn (2004) also discussed the advantages of conditional models over marginal models,

and regarded the conditional model as fundamental, from which marginal predictions can be made.

This debate over covariate-adjustment approach extends to other areas as well, e.g., epidemiology. Weng et al. (2009) generated twenty-five scenarios with case-control samples from 10 simulated populations to compare performances of 4 covariate selection approaches in the presence of confounders of various strengths.

2.4 Bias Approximations for Model Misspecification

Theoretical results show that omitting predictive covariates in GLMs often causes inaccurate treatment effect estimates. Assume that T , independent of X, is a treatment indicator variable that takes values 1 or -1 with probability p and $1-p$ respectively. Gail et al. (1984) confirmed that randomization does not always lead to asymptotically unbiased estimates of T when X is omitted. Let

$$\zeta_1 = E(Y|T=1) = E_X\{h(\alpha+\beta+\gamma'X)\}, \tag{2.4a}$$

$$\zeta_2 = E(Y|T=-1) = E_X\{h(\alpha-\beta+\gamma'X)\}. \tag{2.4b}$$

If h^{-1} exists and is well defined at ζ_1 and ζ_2, then

$$\beta^* = \frac{1}{2}\{h^{-1}(\zeta_1) - h^{-1}(\zeta_2)\}. \tag{2.5}$$

Using Taylor's expansion, they showed that

$$\beta^* - \beta \doteq \frac{1}{4}\gamma' Var(X)\gamma\big\{h''(\alpha+\beta)/h'(\alpha+\beta) - h''(\alpha-\beta)/h'(\alpha-\beta)\big\}. \qquad (2.6)$$

Note that the bias will be small if γ is small or the covariate X has little variability. This approximation provides reasonable magnitude of the bias. Particularly, they pointed out that $\beta^* = \beta$ when $h(\eta) = a\eta + b$ or $h(\eta) = c\exp(a\eta) + b$ for some real constants a, b and c. This can be verified through (2.5).

Neuhaus and Jewell (1993) presented a geometric approach to assess the direction of the bias resulting from omitted covariates in GLMs. Under their assumptions, T is not required to be an indicator variable. Let $\delta = X'\gamma$ and $\mu_k = E(Y|T+k,\delta) = h\{\alpha + (T+k)\beta + \delta\}$, then $\beta = h^{-1}(\mu_1) - h^{-1}(\mu_0)$. When T and X are independent and the covariate vector X is omitted, one obtains the marginal effect of a unit increase in the covariate T by considering

$$\beta^* = h^{-1}(\mu_1^*) - h^{-1}(\mu_0^*) = H(\beta), \qquad (2.7)$$

where $\mu^* = E_\delta(\mu_k) = \int h\{\alpha + (T+k)\beta + \delta\} f(\delta) d\delta$.

Note that $H(0) = 0$. Thus, when $\beta = 0$ in the true model, unbiased estimate of the treatment effect can be obtained through a model with omitted covariates. Furthermore, $H(\beta)$ can be approximated by expanding about $\beta = 0$:

$$\beta^* \doteq \beta H'(0) = \beta g'[E\{g^{-1}(\delta)\}]E[1/g'\{g^{-1}(\delta)\}], \qquad (2.8)$$

where $g(\cdot) = h^{-1}(\cdot)$. The direction of the bias can be determined by checking

whether $1/g'(\cdot)$ is concave or convex. Table 2.1 gives the bias factor $H'(0)$ for several popular link functions.

Table 2.1. Bias factors for generalized linear model parameter estimates obtained with omitted covariates

Link function	$g(\mu)$	Bias factor $H'(0)$ in (2.8)
Linear	$a + b\mu$	1
Log	$(1/b)\log(a + b\mu)$	1
Logistic	$\log\{\mu/(1-\mu)\}$	$1 - \frac{Var(\mu_0)}{E(\mu_0)\{1-E(\mu_0)\}}$
Complementary log-log	$\log\{-\log(1-\mu)\}$	$\frac{E\{(1-\mu_0)\log(1-\mu_0)\}}{\{1-E(\mu_0)\}\log\{1-E(\mu_0)\}}$

Note that (2.5) and (2.7) are essentially the same. (2.6) is obtained by a Taylor expansion about $\gamma = 0$ while (2.8) is obtained by expanding about $\beta = 0$.

Drake and McQuarrie (1995) considered estimating the bias in observational studies of exposure effects for generalized linear models with canonical link. In their case, the bias is due to omission of some but not all confounders. They pointed out that the bias approximation can be derived from a system of linear equations after using first order Taylor expansion.

2.5 Variance and Asymptotic Relative Efficiency for Model Misspecification

Let's first consider a simple classic linear regression. Suppose the structure of a population is described by the following two linear models:

$$E(Y|T, X) = \eta, \ \eta = \alpha + T\beta + X\gamma, \tag{2.9a}$$

$$E(Y|T) = \eta^*,\ \eta^* = \alpha^* + T\beta^*. \tag{2.9b}$$

Here T is not required to be an indicator and X is a scalar. Denote $\hat{\beta}$ and $\hat{\beta}^*$ to be MLEs of β and β^*, respectively. The asymptotic relative efficiency of $\hat{\beta}$ to $\hat{\beta}^*$ is defined to be

$$\text{ARE}(\hat{\beta} \text{ to } \hat{\beta}^*) = \frac{\text{Var}(\hat{\beta}^*)}{\text{Var}(\hat{\beta})}. \tag{2.10}$$

For $\hat{\beta}$ and $\hat{\beta}^*$ associated with (2.9a) and (2.9b), the formula for ARE($\hat{\beta}$ to $\hat{\beta}^*$) is given by (Robinson and Jewell, 1991)

$$\text{ARE}(\hat{\beta} \text{ to } \hat{\beta}^*) = \frac{1 - \rho_{TX}^2}{1 - \rho_{YX \cdot T}^2}, \tag{2.11}$$

where ρ_{TX} is the simple correlation between T and X, and $\rho_{YX \cdot T}$ is the partial correlation between Y and X conditional on T.

Under assumptions of classic linear regression, if $\rho_{TX} = 0$, then $\beta = \beta^*$. In this case, we can see that ARE($\hat{\beta}$ to $\hat{\beta}^*$)≥ 1. This explains why adjusting for desired covariates improves the precision of the estimates of treatment effects. Note also that $\rho_{YX \cdot T} = 0$ is equivalent to $\gamma = 0$ (hence $\beta = \beta^*$). In this case, ARE($\hat{\beta}$ to $\hat{\beta}^*$)≤ 1. This explains why it is not wise to adjust for non-predictive covariates.

For generalized linear models (2.1a) and (2.1b), the above conclusions do not always hold. Neuhaus (1998) derived the estimation efficiency with omitted covariates in generalized linear models. Using the results of White (1982) on estimation in misspecified models, Neuhaus (1998) showed that when T and X are indepen-

dent, the Pitman efficiency of $\hat{\beta}$ to $\hat{\beta}^*$ at $\beta = 0$ is

$$\begin{aligned}&\text{ARE}(\hat{\beta} \text{ to } \hat{\beta}^* \text{ at } \beta = 0)\\ &= \frac{E\{V + \phi^{-1}(\mu_0 - E\mu_0)^2\}E[1/\{V\{g'(\mu_0)\}^2\}]}{[E\{1/g'(\mu_0)\}]^2} \geq 1. \end{aligned} \tag{2.12}$$

Here $\mu_0 = h(\alpha + X\gamma)$, $g(\cdot) = h^{-1}(\cdot)$, V is the variance function and ϕ is the dispersion parameter. Particularly, when X is a nonconfounding covariate ($\beta = \beta^*$), then

$$\begin{aligned}&\text{ARE}(\hat{\beta} \text{ to } \hat{\beta}^* \text{ at } \beta = 0)\\ &= \frac{\text{Var}(\hat{\beta}^*|\beta = 0)}{\text{Var}(\hat{\beta}|\beta = 0)} = E\{V + \phi^{-1}(\mu_0 - E\mu_0)^2\} \cdot \{g'(E\mu_0)\}^2 \cdot E[1/\{V\{g'(\mu_0)\}^2\}]. \end{aligned} \tag{2.13}$$

Robinson and Jewell (1991) presented ARE($\hat{\beta}$ to $\hat{\beta}^*$ at $\beta = 0$) for the logistic regression model with T and X both being binary. They showed that

$$\text{ARE}(\hat{\beta} \text{ to } \hat{\beta}^* \text{ at } \beta = 0) = \frac{E(p_{0j})E(q_{0j})}{E(p_{0j}q_{0j})} \geq 1, \tag{2.14}$$

with equality occurring if and only if X is independent of (Y, T). Here, $p_{ij} = \Pr(Y = 1|T = i, X = j)$ for $i, j = 0, 1$, $q_{ij} = 1 - p_{ij}$.

2.6 The Presence of Interactions between Treatment and Covariates

Previous theoretical results are obtained under the assumption that there are no interactions between treatment and covariates, while this assumption may not always hold. The importance of treatment/covariate interactions has been recognized recently, and the evaluation of evidence of that treatment effects vary among different subsets of patients has gained increasing attention in the analysis of large clinical trials. Qualitative (or crossover) interactions occur when one treatment is superior for some subsets of patients while the alternative treatment is superior for other subsets. Quantitative (or non-crossover) interactions arise when there is variation in the magnitude, but not in the direction, of treatment effects among subsets.

Gail and Simon (1985) developed a likelihood ratio test for qualitative interactions. Let ω_i denote the true difference in treatment effects within disjoint patient subset i for $i = 1, \cdots, I$. Assume estimates D_i of ω_i have independent normal distribution with mean ω_i and known variance σ_i^2. In case of large samples, consistent estimates of σ_i^2 can be used instead. The hypothesis of no crossover interactions is equivalent to the vector $\Omega = (\omega_1, \cdots, \omega_I)$ satisfying either $\omega_i \geq 0$ for all i or $\omega_i \leq 0$ for all i. Let $\mathbf{O}^+ = \{\Omega : \ \omega_i \geq 0 \text{ for all } i\}$, and $\mathbf{O}^- = \{\Omega : \ \omega_i \leq 0 \text{ for all } i\}$. The likelihood ratio test of this hypothesis is based on the test statistic

$$\frac{\max_{\Omega \in \mathbf{O}^+ \cup \mathbf{O}^-} \exp[\sum_{i=1}^{I} \{-(D_i - \delta_i)^2/(2\sigma_i^2)\}]}{\max_{\Omega} \exp[\sum_{i=1}^{I} \{-(D_i - \delta_i)^2/(2\sigma_i^2)\}]}. \tag{2.15}$$

Since the maximum value of the denominator is 1, the likelihood ratio test is thus

$$\max_{\Omega \in \mathbf{O}^+ \cup \mathbf{O}^-} \exp[\sum_{i=1}^{I} \{-(D_i - \delta_i)^2/(2\sigma_i^2)\}] < k, \tag{2.16}$$

or equivalently

$$\min_{\Omega \in \mathbf{O}^+} \sum_{i=1}^{I} \{(D_i - \delta_i)^2/\sigma_i^2\} > c, \tag{2.17}$$

and

$$\min_{\Omega \in \mathbf{O}^-} \sum_{i=1}^{I} \{(D_i - \delta_i)^2/\sigma_i^2\} > c, \tag{2.18}$$

with $c = -2 \log k$. Let

$$Q^- \equiv \sum (D_i^2/\sigma_i^2) I(D_i > 0), \tag{2.19a}$$

$$Q^+ \equiv \sum (D_i^2/\sigma_i^2) I(D_i < 0), \tag{2.19b}$$

where $I(\cdot)$ is an indicator function. The rejection region is given by $\min(Q^+, Q^-) > c$. Gail and Simon (1985) also provided a table of values of c for different significance levels.

Stürner and Brenner (2002) extended the concept of flexible matching strategies to the field of gene-environment interactions. They assessed the power and efficiency of such studies to detect and estimate gene-environment interactions under a variety of assumptions regarding the prevalence and effects of the environmental exposure and the genetic susceptibility as well as their association in the population. Zou (2008) presented a new way that uses the conventional asymmetric

intervals for risk ratios to set confidence limits for measures of additive interaction in a four-by-two table. A four-by-two table, with its four rows representing the presence and absence of gene and environmental factors, has been suggested as the fundamental unit in the assessment of gene-environment interaction.

Chapter 3

Bias and Variance of Estimation for Covariate-adjusted and -unadjusted Approaches

We consider a study whose major objective is to assess the treatment effect. Suppose Y is a response variable, X and Z are two sets of covariates whose roles will be clarified shortly, and T is a treatment indicator variable that takes values 1 or -1 with probabilities p and $1-p$ respectively. We assume that the response variable Y follows a generalized linear model (McCullagh and Nelder, 1983). Specifically, Y has an exponential family probability density function

$$\exp[\{y\theta - b(\theta)\}/a(\phi) + c(u,\phi)]$$

for some functions $a(), b(), c()$, and a scalar ϕ. The mean of Y, $\mu = b'(\theta)$ depends on the covariates through:

$$\mu = h(\eta), \quad \eta = \alpha + T\beta + X'\gamma \cdot T + X'\xi_1 + Z'\xi_2,$$

where $h^{-1}(\cdot)$ is the link function and ϕ is the dispersion parameter. Note that, although X and Z are both vectors of covariates, X interacts with T while Z does not. In this framework, the variance, $b''(\theta)a(\phi)$, can be defined as a function of the mean

$$var(Y) = V_\phi(\mu).$$

The above generalized linear model with

$$\mu = h(\alpha + T\beta + X'\gamma \cdot T + X'\xi_1 + Z'\xi_2), \tag{3.1}$$

is often called a covariate-adjusted model because it examines the treatment effect adjusting the impact (e.g. confounding and interaction effects) of the covariates. When $\gamma = 0$ (i.e. there is no treatment and covariate interaction), then this model reduces to the model considered by Gail and Simon (1984), among others.

In contrast, the following model estimates a crude treatment effect which does not adjust the impact of covariates,

$$\mu^* = h(\eta^*), \quad \eta^* = \alpha^* + T\beta^*. \tag{3.2}$$

This model is often called the covariate-unadjusted generalized linar model.

The focus of the paper will be on comparing β, the treatment effect estimated from covariate-adjusted model (3.1) with β^*, the treatment effect estimated from the covariate-unadjusted model (3.2). In order to make sure β and β^* represent the same overall (average) treatment effect, throughout this paper, we assume that X

and Z are all centralized so that $E(X)$=0 and $E(Z) = 0$, similar to the assumption made in Gail and Simon (1984). In practice, when $E(X)$ and $E(Z)$ are not all zero, we transform X to $X - E(X)$, Z to $Z - E(Z)$. Consequently, we re-parameterize α to $\alpha + E(X)'\xi_1 + E(Z)'\xi_2$ and β to $\beta + E(X)'\gamma$.

Comparing model (3.1) with model (3.2) is not straightforward because we do not know the true underlying model. In some situations, covariates do not influence the response variable except through the treatment; in other words, covariates are causal intermediaries. In this case, model (3.2) is correct and (3.1) is theoretically correct if all γ, ξ_1 and ξ_2 are equal to 0. Therefore, the estimated treatment effects obtained from both models (3.1) and (3.2) are asymptotically unbiased.

If model (3.1) is the true model, then covariates do impact the response variable and are not causal intermediaries. When $\gamma = 0$ (i.e. none of the covariates interact with treatment), the results by Gail et al. (1984) indicate that the estimated treatment effect obtained from model (3.2) can be asymptotically biased in the sense that the expected values of the estimated β and β^* differ.

The first goal of this paper is to examine the bias when $\gamma \neq 0$ as in Section 2.1 under the assumption that covariates are independent of the treatment, a condition held in randomized studies. The second goal is to compare the variances of β^* and β as in Section 2.2. In contrast to Section 2.1, the assumption that covariates are independent of the treatment is not a general requirement here. In particular, if y follows a linear model, then the variance of β^* is not less than that of β. On the other hand, if y follows a logistic model, then the variance of β^* is not greater than that of β; see Robinson and Jewell (1991). The comparison of the variances

of β^* and β are unknown for generalized linear models other than linear or logistic regression models even when $\gamma = 0$ (i.e. none of the covariates interact with treatment). Section 2.2 will reveal the comparison in generalized linear models.

3.1 Bias for covariate-adjusted and -unadjusted approaches

Following Gail et al. (1984), we let

$$\zeta_1 = E(Y|T=1) = E_{X,Z}\Big\{h\Big(\alpha+\beta+X'\gamma+X'\xi_1+Z'\xi_2\Big)\Big\},$$
$$\zeta_2 = E(Y|T=-1) = E_{X,Z}\Big\{h\Big(\alpha-\beta-X'\gamma+X'\xi_1+Z'\xi_2\Big)\Big\}.$$

Here $E_{X,Z}$ is the expectation with respect to X and Z. Let $\kappa(\eta) = \frac{\partial\mu}{\partial\eta}\cdot\frac{1}{V_\phi(\mu)}$ and recall that $\eta = \alpha + T\beta + X'\gamma\cdot T + X'\xi_1 + Z'\xi_2$.

Similar to Gail et al (1984), the maximum likelihood equations divided by the sample size converge to

$$E\big[\kappa(\eta)\cdot\big\{Y-h(\eta)\big\}\big]=0,$$
$$E\big[\kappa(\eta)\cdot T\cdot\big\{Y-h(\eta)\big\}\big]=0,$$
$$E\big[\kappa(\eta)\cdot X'\cdot T\cdot\big\{Y-h(\eta)\big\}\big]=0,$$
$$E\big[\kappa(\eta)\cdot X'\cdot\big\{Y-h(\eta)\big\}\big]=0,$$
$$E\big[\kappa(\eta)\cdot Z'\cdot\big\{Y-h(\eta)\big\}\big]=0.$$

Hereafter this thesis, we denote $\dot{h}(\cdot)$ and $\ddot{h}(\cdot)$ to be the first and second deriva-

tives of function $h(\cdot)$, respectively. The relationship between β and β^* is given in the following result.

Theorem 1. *Under the following assumptions*

(a) T is independent of all covariate variables;

(b) $E\big[\kappa(\eta)\cdot\big\{Y-h(\eta)\big\}\big]$, $E\big[\kappa(\eta)\cdot T\cdot\big\{Y-h(\eta)\big\}\big]$, ζ_1, and ζ_2 exist;

(c) if $h(\cdot)$ has a unique inverse $h^{-1}(\cdot)$ which is well defined at ζ_1 and ζ_2; $\dot{h}$ and $\ddot{h}$ exist;

(d) $h^{-1}(\cdot)$ is nonsingular at $h\big(\alpha+\beta\big)$ and $h\big(\alpha-\beta\big)$;

(e) $\kappa(\alpha^*+\beta^*)$ and $\kappa(\alpha^*-\beta^*)$ do not vanish,

we have the following results

$$\alpha^* = \frac{1}{2}\{h^{-1}(\zeta_1)+h^{-1}(\zeta_2)\}, \tag{3.3}$$

$$\beta^* = \frac{1}{2}\{h^{-1}(\zeta_1)-h^{-1}(\zeta_2)\}, \tag{3.4}$$

and for small γ, ξ_1 and ξ_2,

$$\beta^* \doteq \beta + \left\{ \frac{E_{X,Z}(X'\xi_1+Z'\xi_2+X'\gamma)^2}{4}\frac{\ddot{h}\big(\alpha+\beta\big)}{\dot{h}\big(\alpha+\beta\big)} - \frac{E_{X,Z}(X'\xi_1+Z'\xi_2-X'\gamma)^2}{4}\frac{\ddot{h}\big(\alpha-\beta\big)}{\dot{h}\big(\alpha-\beta\big)} \right\}. \tag{3.5}$$

Proof of Theorem 1. Parameters in the misspecified model satisfy the following equations:

$$\begin{aligned} &E\Big[\kappa(\eta^*)\cdot\Big\{Y-h(\eta^*)\Big\}\Big] \\ =&p\cdot\kappa(\alpha^*+\beta^*)\cdot\zeta_1+(1-p)\cdot\kappa(\alpha^*-\beta^*)\cdot\zeta_2 \end{aligned}$$

$$- p \cdot \kappa(\alpha^* + \beta^*) \cdot h\big(\alpha^* + \beta^*\big) - (1-p) \cdot \kappa(\alpha^* - \beta^*) \cdot h\big(\alpha^* - \beta^*\big) = 0, \quad (3.6a)$$

$$\begin{aligned} & E\Big[\kappa(\eta^*) \cdot T \cdot \Big\{Y - h(\eta^*)\Big\}\Big] \\ =& p \cdot \kappa(\alpha^* + \beta^*) \cdot \zeta_1 - (1-p) \cdot \kappa(\alpha^* - \beta^*) \cdot \zeta_2 \\ & - p \cdot \kappa(\alpha^* + \beta^*) \cdot h\big(\alpha^* + \beta^*\big) + (1-p) \cdot \kappa(\alpha^* - \beta^*) \cdot h\big(\alpha^* - \beta^*\big) = 0. \end{aligned} \quad (3.6b)$$

Since $\kappa(\alpha^* + \beta^*)$ and $\kappa(\alpha^* - \beta^*)$ do not vanish, solutions to (3.6a) and (3.6b) are given by

$$\begin{cases} h(\alpha^* + \beta^*) = \zeta_1 \\ \\ h(\alpha^* - \beta^*) = \zeta_2 \end{cases}, \text{ or equivalently } \begin{cases} \beta^* = \frac{1}{2}\{h^{-1}(\zeta_1) - h^{-1}(\zeta_2)\} \\ \\ \alpha^* = \frac{1}{2}\{h^{-1}(\zeta_1) + h^{-1}(\zeta_2)\} \end{cases}.$$

By Taylor's expansion, for small γ, ξ_1 and ξ_2, ζ_1 and ζ_2 can be approximated by

$$\begin{aligned} \zeta_1 \doteq\, & h\big(\alpha + \beta\big) + \dot{h}\big(\alpha + \beta\big) E_{X,Z}(X'\xi_1 + Z'\xi_2 + X'\gamma) \\ & + \frac{1}{2}\ddot{h}\big(\alpha + \beta\big) E_{X,Z}(X'\xi_1 + Z'\xi_2 + X'\gamma)^2, \\ \zeta_2 \doteq\, & h\big(\alpha - \beta\big) + \dot{h}\big(\alpha - \beta\big) E_{X,Z}(X'\xi_1 + Z'\xi_2 - X'\gamma) \\ & + \frac{1}{2}\ddot{h}\big(\alpha - \beta\big) E_{X,Z}(X'\xi_1 + Z'\xi_2 - X'\gamma)^2. \end{aligned}$$

Note that

$$\begin{aligned} h^{-1}(\zeta_1) \doteq & h^{-1}\big\{h\big(\alpha + \beta\big)\big\} + \dot{h}^{-1}\big\{h(\alpha + \beta)\big\}\Big\{E_{X,Z}\big(X'\xi_1 + Z'\xi_2 + X'\gamma\big)\dot{h}\big(\alpha + \beta\big) \\ & + \frac{E_{X,Z}\big(X'\xi_1 + Z'\xi_2 + X'\gamma\big)^2}{2}\ddot{h}\big(\alpha + \beta\big)\Big\} \end{aligned}$$

$$=\alpha+\beta+\Bigg\{E_{X,Z}(X'\xi_1+Z'\xi_2+X'\gamma) + \frac{E_{X,Z}(X'\xi_1+Z'\xi_2+X'\gamma)^2}{2}\frac{\ddot{h}\big(\alpha+\beta\big)}{\dot{h}\big(\alpha+\beta\big)}\Bigg\}.$$

Similarly,

$$h^{-1}(\zeta_2)\doteq\alpha-\beta+\Bigg\{E_{X,Z}(X'\xi_1+Z'\xi_2-X'\gamma) + \frac{E_{X,Z}(X'\xi_1+Z'\xi_2-X'\gamma)^2}{2}\frac{\ddot{h}\big(\alpha-\beta\big)}{\dot{h}\big(\alpha-\beta\big)}\Bigg\}.$$

Therefore for small γ, ξ_1 and ξ_2, the asymptotic bias is

$$E(X'\gamma)+\Bigg\{\frac{E_{X,Z}(X'\xi_1+Z'\xi_2+X'\gamma)^2}{4}\frac{\ddot{h}\big(\alpha+\beta\big)}{\dot{h}\big(\alpha+\beta\big)} - \frac{E_{X,Z}(X'\xi_1+Z'\xi_2-X'\gamma)^2}{4}\frac{\ddot{h}\big(\alpha-\beta\big)}{\dot{h}\big(\alpha-\beta\big)}\Bigg\}.$$

This completes the proof. □

Results (3.3)-(3.5) are extensions of (2.5), (2.6) and (2.9) in Gail et al. (1984), which correspond to the case $\gamma = 0$. As in Gail et al. (1984), the result (3.5) is derived for small γ, ξ_1 and ξ_2.

Corollary 1. *Under conditions of Theorem 1,*

- *If $h(\eta)=\eta$ (e.g. the linear model and the Poisson model with identity link), $\beta^*=\beta$.*

- *If* $h(\eta) = \exp(\eta)$ *(e.g. the Poisson model with canonical link),*

$$\beta^* = \beta + \frac{1}{2}\log\Big[E_{X,Z}\Big\{\exp(X'\xi_1 + Z'\xi_2 + X'\gamma)\Big\}\Big] - \frac{1}{2}\log\Big[E_{X,Z}\Big\{\exp(X'\xi_1 + Z'\xi_2 - X'\gamma)\Big\}\Big].$$

Particularly, with small γ, ξ_1 *and* ξ_2,

$$\beta^* \doteq \beta + E(X'\xi_1 \cdot X'\gamma) + E_{X,Z}(Z'\xi_2 \cdot X'\gamma).$$

- *If* $h(\eta) = \exp(\eta)/\{1 + \exp(\eta)\}$ *(e.g. the binary model with canonical link), and if* γ, ξ_1 *and* ξ_2 *are small, then*

$$\beta^* \doteq \beta + \left\{ \frac{E_{X,Z}(X'\xi_1 + Z'\xi_2 + X'\gamma)^2}{4} \frac{1 - \exp\big(\alpha + \beta\big)}{1 + \exp\big(\alpha + \beta\big)} - \frac{E_{X,Z}(X'\xi_1 + Z'\xi_2 - X'\gamma)^2}{4} \frac{1 - \exp\big(\alpha - \beta\big)}{1 + \exp\big(\alpha - \beta\big)} \right\}. \tag{3.7}$$

Proof of Corollary 1. (1). If $h(\eta) = \eta$, then

$$h^{-1}(\zeta_1) = E_{X,Z}(\alpha + \beta + X'\gamma + X'\xi_1 + Z'\xi_2),$$
$$h^{-1}(\zeta_2) = E_{X,Z}(\alpha - \beta - X'\gamma + X'\xi_1 + Z'\xi_2).$$

Substituting those values into (3.3), hence $\beta^* = \beta + E(X'\gamma)$.

(2). If $h(\eta) = \exp(\eta)$, then

$$h^{-1}(\zeta_1) = \log\Big[E_{X,Z}\Big\{\exp(\alpha + \beta + X'\gamma + X'\xi_1 + Z'\xi_2)\Big\}\Big],$$
$$h^{-1}(\zeta_2) = \log\Big[E_{X,Z}\Big\{\exp(\alpha - \beta - X'\gamma + X'\xi_1 + Z'\xi_2)\Big\}\Big].$$

Substituting those values into (3.3), it follows that

$$\beta^* = \beta + \log\Big[E_{X,Z}\Big\{\exp(X'\xi_1 + Z'\xi_2 + X'\gamma)\Big\}\Big] - \log\Big[E_{X,Z}\Big\{\exp(X'\xi_1 + Z'\xi_2 - X'\gamma)\Big\}\Big].$$

Note that $\ddot{h} = \dot{h} = h$, then

$$\Big\{\frac{E_{X,Z}(X'\xi_1 + Z'\xi_2 + X'\gamma)^2}{4}\frac{\ddot{h}\big(\alpha+\beta\big)}{\dot{h}\big(\alpha+\beta\big)} - \frac{E_{X,Z}(X'\xi_1 + Z'\xi_2 - X'\gamma)^2}{4}\frac{\ddot{h}\big(\alpha-\beta\big)}{\dot{h}\big(\alpha-\beta\big)}\Big\}$$
$$= E(X'\xi_1 \cdot X'\gamma) + E_{X,Z}(Z'\xi_2 \cdot X'\gamma).$$

Therefore, when γ, ξ_1 and ξ_2 are small, the approximate bias in (3.5) is

$$E(X'\gamma) + E(X'\xi_1 \cdot X'\gamma) + E_{X,Z}(Z'\xi_2 \cdot X'\gamma).$$

(3). If $h(\eta) = \exp(\eta)/\{1+\exp(\eta)\}$, then $\dot{h} = h(1-h)$, $\ddot{h} = h(1-h)(1-2h)$ and $0 < h(.) < 1$. Therefore, when γ, ξ_1 and ξ_2 are small, $|\beta^* - \beta - E(X'\gamma)|$ is bounded by

$$\frac{E_{X,Z}(X'\xi_1 + Z'\xi_2 + X'\gamma)^2}{4} + \frac{E_{X,Z}(X'\xi_1 + Z'\xi_2 - X'\gamma)^2}{4}.$$

This completes the proof. □

The approximation in Theorem 1 is derived based on Taylor's expansion at $\xi_1 = \xi_2 = \gamma = 0$. This is similar to that in Gail and Simon (1984). A different approximation is also derived based on Taylor's expansion at $\beta = 0$ below, as in

Neuhaus and Jewell (1993). Let $k = 1, -1$, and

$$\eta_k = \alpha + \beta k + X'\gamma k + X'\xi_1 + Z'\xi_2;$$
$$\eta_k^* = \alpha^* + \beta^* k.$$

Assume model (3.1) is the correct model; we know that

$$E(y|T = k, X, Z) = h(\eta_k).$$

Denote by $f(X, Z|T = k)$ the probability density function of X and Z given k; then

$$\begin{aligned} E(y|T = k) &= \int E(y|T = k, X, Z) f(X, Z|T = k)\ dXdZ \\ &= \int E(y|T = k, X, Z) f(X, Z)\ dXdZ. \end{aligned} \tag{3.8}$$

On the other hand, if model (3.2) is used to make inferences, we would have misspecified the expectation of Y

$$E(y|T = k) = h(\eta_k^*). \tag{3.9}$$

It follows by (3.8) and (3.9) that

$$\alpha^* + \beta^* k = h^{-1}\Big(\int h\big(\alpha + \beta k + X'\gamma k + X'\xi_1 + Z'\xi_2\big) f(X, Z)\ dXdZ \Big),$$

and

$$\beta^* \;=\; \frac{1}{2}\Big\{\alpha^* + \beta^*\Big\} - \frac{1}{2}\Big\{\alpha^* - \beta^*\Big\}$$

$$
\begin{aligned}
&= \frac{1}{2}h^{-1}\Big(\int h\big(\alpha+\beta+X'\gamma+X'\xi_1+Z'\xi_2\big)f(X,Z)\ dXdZ\Big)\\
&- \frac{1}{2}h^{-1}\Big(\int h\big(\alpha-\beta-X'\gamma+X'\xi_1+Z'\xi_2\big)f(X,Z)\ dXdZ\Big).
\end{aligned}
$$

We denote the function in the right side of the last equation as $H(\beta)$ for simplicity. By expanding $H(\beta)$ in Taylor series around $\beta = 0$, we have

$$\beta^* = H(\beta) \approx H(0) + \beta \cdot \dot{H}(0), \tag{3.10}$$

where

$$
\begin{aligned}
H(0) = \frac{1}{2}h^{-1}\Big[E_{X,Z}\{h(\alpha+X'\gamma+X'\xi_1+Z'\xi_2)\}\Big]&\\
- \frac{1}{2}h^{-1}\Big[E_{X,Z}\{h(\alpha-X'\gamma+X'\xi_1+Z'\xi_2)\}\Big],&
\end{aligned}
$$

and

$$
\begin{aligned}
&\dot{H}(0) =\\
&\frac{1}{2}\dot{h}^{-1}\Big[E\{h(\alpha+X'\gamma+X'\xi_1+Z'\xi_2)\}\Big] \times E\Big[1/\dot{h}^{-1}\{h(\alpha+X'\gamma+X'\xi_1+Z'\xi_2)\}\Big]\\
&+\frac{1}{2}\dot{h}^{-1}\Big[E\{h(\alpha-X'\gamma+X'\xi_1+Z'\xi_2)\}\Big] \times E\Big[1/\dot{h}^{-1}\{h(\alpha-X'\gamma+X'\xi_1+Z'\xi_2)\}\Big].
\end{aligned}
$$

Note that $\dot{h}^{-1}$ is the first derivative of h^{-1}.

When $\gamma = 0$, $H(0) = 0$. However $H(0)$ may be nonzero when $\gamma \neq 0$. Therefore, the existence of treatment and covariate interaction further complicates the relationship between β^* and β.

We now illustrate the results through two simple examples. The first example

suggests that the covariate-unadjusted linear regression model provides an unbiased estimate despite the existence of treatment and covariate interaction. The second example, on the other hand, illustrates the important role of interactions between treatment and covariates in a nonlinear regression model.

Example 1. Suppose that the true model is $\theta = \eta = \alpha + T\beta + X'\gamma \cdot T + X'\xi_1$ and $h(.)$ is the identity function. In this case, $\beta = \beta^*$; therefore the covariate-unadjusted model, even omitting true interaction between covariate and treatment, still provides an unbiased treatment estimate. We also note that $H(0) = 0$ and $H'(0) = 1$.

Example 2. Let us consider a logistic regression model, in which the exponential family probability density function is

$$\exp[\{y\theta - \log(1 + \exp(\theta))\} + c].$$

Here c is a constant. The mean of Y, $\mu = b'(\theta) = \exp(\theta)\{1 + \exp(\theta)\}$ and

$$\mu = h(\theta) = \exp(\theta)\{1 + \exp(\theta)\}, \quad \theta = \alpha + T\beta + X'\gamma \cdot T + X'\xi_1.$$

We consider a few different choices for ξ and γ. The treatment indicator $T(1, -1)$ and $X(1, -1)$ have independent binomial distribution Binomial(1,1/2). It can be shown that

$$\zeta_1 = 1/2\Big\{h\Big(\alpha + \beta + \gamma + \xi_1\Big)\Big\} + 1/2\Big\{h\Big(\alpha + \beta - \gamma - \xi_1\Big)\Big\}$$
$$\zeta_2 = 1/2\Big\{h\Big(\alpha - \beta - \gamma + \xi_1\Big)\Big\} + 1/2\Big\{h\Big(\alpha - \beta + \gamma - \xi_1\Big)\Big\}.$$

First we consider a special case with $\beta = 0$,

$$\zeta_1 = 1/2\Big\{h\Big(\alpha + \gamma + \xi_1\Big)\Big\} + 1/2\Big\{h\Big(\alpha - \gamma - \xi_1\Big)\Big\}$$
$$\zeta_2 = 1/2\Big\{h\Big(\alpha - \gamma + \xi_1\Big)\Big\} + 1/2\Big\{h\Big(\alpha + \gamma - \xi_1\Big)\Big\}.$$

It becomes straightforward from equation (3.4) that $\beta^* = 0$ when $\gamma = 0$; that is there is no interaction. On the other hand, if there is treatment and covariate interaction, β^* is generally not 0 unless $\xi_1 = 0$. For example, $\beta^* = -0.035$ when $\alpha = 1$, $\gamma = 0.2$ and $\xi_1 = 0.4$.

Next, we consider a different case with $\alpha = \beta = 1$ and $\gamma = 0$ and $\xi_1 = 1$. According to Neuhause and Jewell (1993), $0 < \beta^* < 1$. Actually, we obtain $\beta^* = 0.836$. However, if $\gamma \neq 0$, we no longer expect $0 < \beta^* < 1$. In fact, let $\alpha = \beta = \xi_1 = 1$ and $\gamma = -1$; then $\beta^* = 1$. In this case, the interaction $\gamma = -1$ seems to correct the difference of -0.164 between β and β^* when $\gamma = 0$ and $\xi_1 = 1$.

Now, we consider another case in which $\alpha = \beta = 1$. It can be shown that $\beta^* = 0.954$ if $\gamma = 0$ and $\xi_1 = 0.5$, and $\beta^* = 0.836$ if $\gamma = 0.5$ and $\xi_1 = 05$. In this case, the interaction shifted the difference -0.046 between β and β^* when $\gamma = 0$ and $\xi_1 = 0.5$ even further to -0.164.

Finally, we let $\alpha = 2$, $\beta = \xi = 1$, and $\gamma = -1$. It can be shown that $\beta^* = 1.275$ which is greater than $\beta = 1$. This is not consistent with the results described in Neuhause and Jewell (1993), indicating a difference occurs when treatment and covariate interact.

3.2 Variance of Estimation for Covariate-adjusted and -unadjusted Approaches

To present general results in a simple form, we first consider the following two models:

$$\eta = \alpha + T\beta + X'\gamma, \tag{3.11a}$$

and

$$\eta = \alpha^* + T\beta^*, \tag{3.11b}$$

where X and γ are $p \times 1$ vectors. Note that, here X is not necessarily independent of T. In other words, we allow components of X to contain interactions between T and covariates; therefore model (3.11a) and model (3.1) are similar.

Let t_i, x_i and y_i be the observed values of T, X and Y, respectively, $i = 1, \cdots, n$. Denote $\widetilde{1}$ the $n \times 1$-vector with all components equal to 1, t the $n \times 1$-vector $[t_1, \cdots, t_n]'$ and x' the $p \times n$-matrix $[x_1, \cdots, x_n]$. Recall the procedure of fitting generalized linear models via the Fisher scoring algorithm. The estimated covariance matrix is the inverse of the expected Hessian matrix $M'WM$, where M is the corresponding design matrix and W is the $n \times n$ diagonal matrix of iterative weights with diagonal terms $w_i = \{V_\phi(\mu_i)(\frac{\partial \eta_i}{\partial \mu_i})^2\}^{-1}$, $i = 1, \cdots, n$.

Let $\mu(T) = E(Y|T)$, $\mu(T, X) = E(Y|T, X)$ and $g(\mu) = \left\{V_\phi(\mu)\left(\frac{\partial \eta}{\partial \mu}\right)^2\right\}^{-1}$. We also let W and W^* be the iterative weights matrix for models (3.11a) and (3.11b),

respectively.

Lemma 1. *Suppose that* $\lim_{n\to\infty}\frac{1}{n}(\widetilde{1},t,x)'W(\widetilde{1},t,x)$ *and* $\lim_{n\to\infty}\frac{1}{n}(\widetilde{1},t)'W^*(\widetilde{1},t)$ *exist, and* $E_T\Big[g\big\{\mu(T)\big\}\Big]<\infty$, $E_{T,X}\Big[g\big\{\mu(T,X)\big\}\Big]<\infty$. *Further assume that the dispersion parameter* ϕ *is fixed and known; we have the following results:*

- *If* $g(\mu)$ *is a strictly concave (convex) function of* μ, *then* Δ, *defined as*

$$\lim_{n\to\infty}\frac{1}{n}\begin{pmatrix}\widetilde{1}'W^*\widetilde{1} & \widetilde{1}'W^*t\\ t'W^*\widetilde{1} & t'W^*t\end{pmatrix}-\lim_{n\to\infty}\frac{1}{n}\begin{pmatrix}\widetilde{1}'W\widetilde{1} & \widetilde{1}'Wt\\ t'W\widetilde{1} & t'Wt\end{pmatrix}$$

  is either a positive (negative) definite matrix or a zero matrix. It is a zero matrix if and only if $\gamma=0$.

- *If* $g(\mu)$ *is a linear function of* μ, *then* Δ *is a zero matrix.*

Proof of Lemma 1. By the weak law of large numbers, it follows that

$$\lim_{n\to\infty}\frac{1}{n}\widetilde{1}'W^*\widetilde{1}=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}g\{\mu(t_i)\}=E_T\Big[g\big\{\mu(T)\big\}\Big].$$

Similarly, it follows that

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n}\widetilde{1}'W\widetilde{1}=E_{T,X}\Big[g\{\mu(T,X)\}\Big], && \lim_{n\to\infty}\frac{1}{n}\widetilde{1}'W^*t=E_T\Big[T\cdot g\big\{\mu(T)\big\}\Big],\\
&\lim_{n\to\infty}\frac{1}{n}\widetilde{1}'Wt=E_{T,X}\Big[T\cdot g\{\mu(T,X)\}\Big], && \lim_{n\to\infty}\frac{1}{n}t'W^*t=E_T\Big[T^2\cdot g\big\{\mu(T)\big\}\Big],\\
&\lim_{n\to\infty}\frac{1}{n}t'Wt=E_{T,X}\Big[T^2\cdot g\{\mu(T,X)\}\Big].
\end{aligned}$$

(1). If $g(\mu)$ is a strictly concave function of μ. By Jensen's inequality,

$$E_{T,X}\big[g\big\{\mu(T,X)\big\}\big]=E_T\Big[E_X\Big\{g\Big(\mu(T,X)\Big)\Big|T\Big\}\Big]$$

$$\leq E_T\Big[g\Big\{E_X\Big(\mu(T,X)\Big|T\Big)\Big\}\Big] = E_T\big[g\big\{\mu(T)\big\}\big].$$

The equality occurs if and only if X is independent of Y given T (i.e., $\gamma = 0$).

Since $E_T\big[g\big\{\mu(T)\big\}\big] = E_{T,X}\big[g\big\{\mu(T)\big\}\big]$,

$$\begin{aligned}
\Delta &= \lim_{n\to\infty}\frac{1}{n}\begin{pmatrix}\widetilde{1}'W^*\widetilde{1} & \widetilde{1}'W^*t\\ t'W^*\widetilde{1} & t'W^*t\end{pmatrix} - \lim_{n\to\infty}\frac{1}{n}\begin{pmatrix}\widetilde{1}'W\widetilde{1} & \widetilde{1}'Wt\\ t'W\widetilde{1} & t'Wt\end{pmatrix}\\
&= \begin{pmatrix}E_T\big[g\big\{\mu(T)\big\}\big] & E_T\big[T\cdot g\big\{\mu(T)\big\}\big]\\ E_T\big[T\cdot g\big\{\mu(T)\big\}\big] & E_T\big[T^2\cdot g\big\{\mu(T)\big\}\big]\end{pmatrix}\\
&\qquad - \begin{pmatrix}E_{T,X}\big[g\big\{\mu(T,X)\big\}\big] & E_{T,X}\big[T\cdot g\big\{\mu(T,X)\big\}\big]\\ E_{T,X}\big[T\cdot g\big\{\mu(T,X)\big\}\big] & E_{T,X}\big[T^2\cdot g\big\{\mu(T,X)\big\}\big]\end{pmatrix}\\
&= E_{T,X}\Big[\begin{pmatrix}1\\ T\end{pmatrix}\cdot\Big\{g\Big(\mu(T)\Big) - g\Big(\mu(T,X)\Big)\Big\}\cdot\begin{pmatrix}1 & T\end{pmatrix}\Big],
\end{aligned}$$

which is either a positive definite matrix or a zero matrix. The latter occurs if and only if $\gamma = 0$.

(2). If $g(\mu)$ is a strictly convex function of μ, then following a proof similar to the one above,

$$E_{T,X}\big[g\{\mu(T,X)\}\big] \geq E_T\big[g\big\{\mu(T)\big\}\big].$$

Therefore,

$$\Delta = \lim_{n\to\infty}\frac{1}{n}\begin{pmatrix}\widetilde{1}'W^*\widetilde{1} & \widetilde{1}'W^*t\\ t'W^*\widetilde{1} & t'W^*t\end{pmatrix} - \lim_{n\to\infty}\frac{1}{n}\begin{pmatrix}\widetilde{1}'W\widetilde{1} & \widetilde{1}'Wt\\ t'W\widetilde{1} & t'Wt\end{pmatrix}$$

$$=E_{T,X}\Big[\begin{pmatrix}1\\T\end{pmatrix}\cdot\Big\{g\Big(\mu(T)\Big)-g\Big(\mu(T,X)\Big)\Big\}\cdot\begin{pmatrix}1 & T\end{pmatrix}\Big],$$

which is either a negative definite matrix or a zero matrix. It is a zero matrix if and only if $\gamma = 0$.

(3). If $g(\mu)$ is a linear function of μ, then

$$E_{T,X}\Big[T^i\cdot g\Big\{\mu(T,X)\Big\}\Big]=E_T\Big[T^i\cdot g\Big\{\mu(T)\Big\}\Big],\ i=0,1,2.$$

Thus,

$$\Delta=\lim_{n\to\infty}\frac{1}{n}\begin{pmatrix}\widetilde{1}'W^*\widetilde{1} & \widetilde{1}'W^*t\\ t'W^*\widetilde{1} & t'W^*t\end{pmatrix}-\lim_{n\to\infty}\frac{1}{n}\begin{pmatrix}\widetilde{1}'W\widetilde{1} & \widetilde{1}'Wt\\ t'W\widetilde{1} & t'Wt\end{pmatrix}=0.$$

This completes the proof. □

Remark 1. Lemma 1 imposes an important assumption: the dispersion parameter ϕ is the same in both covariate-adjusted and covariate-unadjusted models. Consequently, Lemma 1 does not apply to some models, including linear regression models, where the dispersion parameter ϕ may differ in (3.11a) and (3.11b).

Denote $I_1=\lim_{n\to\infty}\frac{1}{n}(\widetilde{1},t,x)'W(\widetilde{1},t,x)$, $I_0=\lim_{n\to\infty}\frac{1}{n}(\widetilde{1},t)'W^*(\widetilde{1},t)$. Let $\hat{\alpha}^*$, $\hat{\beta}^*$, $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$ be MLE estimators of α^*, β^*, α, β, γ in (3.11a) and (3.11b), respectively.

If (3.11a) is the true model, then

$$\sqrt{n}\left\{ \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\gamma} \end{pmatrix} - \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \right\} \xrightarrow{D} N\left(0, I_1^{-1}\right),\ \sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N(0, I_{1,22}^{-1}). \tag{3.12a}$$

If (3.11b) is the true model, then

$$\sqrt{n}\left\{ \begin{pmatrix} \hat{\alpha}^* \\ \hat{\beta}^* \end{pmatrix} - \begin{pmatrix} \alpha^* \\ \beta^* \end{pmatrix} \right\} \xrightarrow{D} N\left(0, I_0^{-1}\right),\ \sqrt{n}(\hat{\beta}^* - \beta^*) \xrightarrow{D} N(0, I_{0,22}^{-1}). \tag{3.12b}$$

Here $I_{i,22}^{-1}$ are the $(2,2)$ entries of I_i^{-1}, $i = 0, 1$.

We note that with misspecification, the asymptotic covariance matrix of the maximum likelihood estimation may not be the inverse of Fisher's information matrix. White (1982) showed that the covariance matrix is given by a sandwich form of matrix products. However, people often assume the model that they select to make inferences is the correct one, and statistical packages automatically construct the covariance matrix by the inverse of the information matrix. Therefore, it is interesting to compare $I_{1,22}^{-1}$ and $I_{0,22}^{-1}$. Furthermore, as shown in Neuhaus (1998), for binary regression models, the information matrix of the misspecified model provides the correct value of the covariance matrix when using the results of White (1982).

Theorem 2. *Under conditions of Lemma 1,*

- *If $g(\mu)$ is a strictly concave function of μ, then*

$$\frac{I_{1,22}^{-1}}{I_{0,22}^{-1}} > 1 \text{ if } \gamma \neq 0, \text{ and } \frac{I_{1,22}^{-1}}{I_{0,22}^{-1}} \geq 1 \text{ if } \gamma = 0.$$

  In addition, $I_{1,22}^{-1} = I_{0,22}^{-1}$ if X is independent of (Y, T).

- *If $g(\mu)$ is constant, then*

  $\frac{I_{1,22}^{-1}}{I_{0,22}^{-1}} \geq 1$. *The equality occurs if and only if X and T are uncorrelated.*

Proof of Theorem 2. Partition I_1 as

$$I_1 = \begin{pmatrix} A & b' \\ b & c \end{pmatrix},$$

where

$$A = \begin{pmatrix} E_{T,X}\big[g\{\mu(T,X)\}\big] & E_{T,X}\big[T \cdot g\{\mu(T,X)\}\big] \\ E_{T,X}\big[T \cdot g\{\mu(T,X)\}\big] & E_{T,X}\big[T^2 \cdot g\{\mu(T,X)\}\big] \end{pmatrix} \text{ is } 2 \times 2,$$

$b = \Big(E_{T,X}\big[X \cdot g\{\mu(T,X)\}\big] \quad E_{T,X}\big[X \cdot T \cdot g\{\mu(T,X)\}\big]\Big)$ is a $p \times 2$ matrix and c is a $p \times p$ matrix. Let $\Delta = I_0 - A$.

(1). First consider the case where $g(\mu)$ is a strictly concave function of μ. If $\gamma \neq 0$, by Lemma 1, A, I_0 and Δ are all positive definite matrices. Thus,

$$\begin{aligned} \Big(A - b'c^{-1}b\Big)^{-1} &= A^{-1} + A^{-1}b'\Big(c - bA^{-1}b'\Big)^{-1}bA^{-1} \\ &= (I_0 - \Delta)^{-1} + A^{-1}b'\Big(c - bA^{-1}b'\Big)^{-1}bA^{-1} \\ &= I_0^{-1} + I_0^{-1}\big(\Delta^{-1} - I_0^{-1}\big)^{-1}I_0^{-1} + A^{-1}b'\Big(c - bA^{-1}b'\Big)^{-1}bA^{-1}. \end{aligned}$$

Note that $\left(\Delta^{-1} - I_0^{-1}\right)^{-1} > 0$. This results in

$$I_{1,22}^{-1} - I_{0,22}^{-1} = \begin{pmatrix} 0 & 1 \end{pmatrix} I_0^{-1}\left(\Delta^{-1} - I_0^{-1}\right)^{-1} I_0^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 & 1 \end{pmatrix} A^{-1}b'\Big(c - bA^{-1}b'\Big)^{-1} bA^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix} > 0.$$

If $\gamma = 0$, by Lemma 1, $I_0 = A$. This leads to

$$I_{1,22}^{-1} - I_{0,22}^{-1} = \begin{pmatrix} 0 & 1 \end{pmatrix} A^{-1}b'\Big(c - bA^{-1}b'\Big)^{-1} bA^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \geq 0,$$

or equivalently,

$$\frac{I_{1,22}^{-1}}{I_{0,22}^{-1}} \geq 1$$

Note that under the condition of $\gamma = 0$, $g\big\{\mu(T,X)\big\} = g\big\{\mu(T)\big\}$ almost surely; then

$$\begin{pmatrix} 0 & 1 \end{pmatrix} A^{-1}b' = 0$$

$$\Leftrightarrow$$

$$\begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} E_T\big[T^2 \cdot g\big\{\mu(T)\big\}\big] & -E_T\big[T \cdot g\big\{\mu(T)\big\}\big] \\ -E_T\big[T \cdot g\big\{\mu(T)\big\}\big] & E_T\big[g\big\{\mu(T)\big\}\big] \end{pmatrix} \times \begin{pmatrix} E_{T,X}\big[X' \cdot g\big\{\mu(T)\big\}\big] \\ E_{T,X}\big[X' \cdot T \cdot g\big\{\mu(T)\big\}\big] \end{pmatrix} = 0$$

$$\Leftrightarrow$$

$$E_T\big[g\{\mu(T)\}\big] \cdot E_{T,X}\big[X' \cdot T \cdot g\{\mu(T)\}\big]$$
$$= E_T\big[T \cdot g\{\mu(T)\}\big] \cdot E_{T,X}\big[X' \cdot g\{\mu(T)\}\big]. \quad (3.13)$$

The last identity (3.13) holds if X and T are independent. Hence, $I_{1,22}^{-1} = I_{0,22}^{-1}$ if $\gamma = 0$ and X is independent of T.

(2). We then consider the case where $g(\mu)$ is constant with respect to μ, by Lemma 1, $I_0 = A$. Following the proof above, $I_{1,22}^{-1} \geq I_{0,22}^{-1}$. The equality occurs if and only if $\begin{pmatrix} 0 & 1 \end{pmatrix} A^{-1} b' = 0$. Since $g(\mu)$ is constant with respect to μ,

$$\begin{pmatrix} 0 & 1 \end{pmatrix} A^{-1} b' = 0$$
$$\Leftrightarrow \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} E(T^2) & -E(T) \\ -E(T) & 1 \end{pmatrix} \begin{pmatrix} E(X') \\ E_{T,X}(X' \cdot T) \end{pmatrix} = 0$$
$$\Leftrightarrow E_{T,X}(X' \cdot T) = E(T) \cdot E(X') \Leftrightarrow \ X \text{ and } T \text{ are uncorrelated.}$$

This completes the proof. □

In general, the relative efficiency function $I_{1,22}^{-1}/I_{0,22}^{-1}$ is complicated. To gain insights into Theorem 2, we compute this function for a Gamma distribution $\Gamma(\phi, \mu/\phi)$ $(\phi > 0)$ with log link.

Example 3. Consider a setting where Y follows a Gamma distribution $\Gamma(\phi, \mu/\phi))$ with log link, where $\phi > 0$. For simplicity, we also assume X is a scalar in (3.11a). If ϕ is the same in (3.11a) and (3.11b), then $g(\mu) = \phi$. Following the proof of

Theorem 2,

$$\begin{aligned}\frac{I_{1,22}^{-1}}{I_{0,22}^{-1}} =& 1 + \frac{E(XT) - E(X)E(T)}{Var(T)} \times \Upsilon \times \frac{E(XT) - E(X)E(T)}{Var(T)} \\ =& 1 + \frac{\rho_{T,X}^2}{1 - \rho_{T,X}^2},\end{aligned}$$

where

$$\Upsilon = \frac{Var(T)}{E(X^2) - \frac{1}{Var(T)}\left[\{E(XT) - E(X)E(T)\}^2 + E^2(X)Var(T)\right]}$$

and $\rho_{T,X}$ is the simple correlation between T and X.

For logistic regression models, the first part of Theorem 2 is consistent with and extends results in Robinson and Jewell (1991) and Robinson et al.(1998) where treatment and covariates do not interact. Our results are applicable to many commonly used models for which $g(\mu)$ is strictly concave, such as binomial distribution with logit link, probit link, log-log link. The second part of Theorem 2 applies to some distributions (e.g. Gamma distribution with log link) in which $g(\mu)$ is a constant.

For Poisson regression with the log link, $g(\mu) = \mu$, we have the following interesting result.

Corollary 2. *If the underlying model is Poisson regression with the log link, and if T is independent of X, then under conditions of Lemma 1, $I_{1,22}^{-1}/I_{0,22}^{-1} = 1$.*

Proof of Corollary 2. Partition I_1 in the same way as in the proof of Theorem

2,

$$I_1 = \begin{pmatrix} A & b' \\ b & c \end{pmatrix}.$$

Since $g(\mu) = \mu$, by Lemma 1, $I_0 = A$. Hence

$$I_{1,22}^{-1} - I_{0,22}^{-1} = \begin{pmatrix} 0 & 1 \end{pmatrix} A^{-1}b'\Big(c - bA^{-1}b'\Big)^{-1} bA^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

However

$$\begin{pmatrix} 0 & 1 \end{pmatrix} A^{-1}b' = 0$$

$$\Leftrightarrow$$

$$\begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} E_{T,X}\big\{T^2 \cdot \mu(T,X)\big\} & -E_{T,X}\big\{T \cdot \mu(T,X)\big\} \\ -E_{T,X}\big\{T \cdot \mu(T,X)\big\} & E_{T,X}\big\{\mu(T,X)\big\} \end{pmatrix} \begin{pmatrix} E_{T,X}\big\{X' \cdot \mu(T,X)\big\} \\ E_{T,X}\big\{X' \cdot T \cdot \mu(T,X)\big\} \end{pmatrix} = 0$$

$$\Leftrightarrow$$

$$E_{T,X}\big\{\mu(T,X)\big\} \cdot E_{T,X}\big\{X' \cdot T \cdot \mu(T,X)\big\} = E_{T,X}\big\{T \cdot \mu(T,X)\big\} \cdot E_{T,X}\big\{X' \cdot \mu(T,X)\big\}. \quad (3.14)$$

Recall that $\mu(T,X) = \exp(\alpha) \cdot \exp(T\beta) \cdot \exp(X'\gamma)$, hence the left part of (3.14) is

$$e^{2\alpha} \cdot E\big(e^{T\beta}\big) \cdot E\big(T \cdot e^{T\beta}\big) \cdot E\big(e^{X'\gamma}\big) \cdot E\big(X' \cdot e^{X'\gamma}\big),$$

and the right part of (3.14) is

$$e^{2\alpha} \cdot E\big(e^{T\beta}\big) \cdot E\big(T \cdot e^{T\beta}\big) \cdot E\big(e^{X'\gamma}\big) \cdot E\big(X' \cdot e^{X'\gamma}\big).$$

This completes the proof. □

Now let us revisit models (3.1) and (3.2). Again, let $\hat{\beta}$, $\hat{\gamma}$, $\hat{\beta}^*$ be maximum likelihood estimates of β, γ, β^* in (3.1) and (3.2). The relative magnitude between $Var(\hat{\beta})$ and $Var(\hat{\beta}^*)$ can be obtained directly through Theorem 2. Let $I_{1,22}^{-1}$ be $\lim_{n\to\infty} n \cdot Var(\hat{\beta})$ under model (3.1). Let $I_{0,22}^{-1}$ be $\lim_{n\to\infty} n \cdot Var(\hat{\beta}^*)$ under model (3.2). Similar to the proof of Theorem 2, we can show the following results.

Theorem 3. *Suppose that T is independent of all covariate variables, and the conditions of Lemma 1 hold.*

- *If $g(\mu)$ is a strictly concave function of μ, then $I_{1,22}^{-1}/I_{0,22}^{-1} \geq 1$.*
- *If $g(\mu)$ is a constant function of μ, then $I_{1,22}^{-1}/I_{0,22}^{-1} = 1$.*

The results suggest that the variance of $\hat{\beta}$ is no smaller than that of $\hat{\beta}^*$. That is, the covariate-adjusted model provides an estimate of treatment effect which usually has a larger variance compared to that obtained from a covariate-unadjusted model.

Chapter 4

Simulation Studies and Real Data Illustration

4.1 Simulation Studies

The main purpose of these simulation studies are to confirm the main results developed in Section 3.

4.1.1 Bias Derived from Small γ, ξ_1 and ξ_2

In our simulation studies, we focus on the logistic regression model:

$$Y \sim \text{Binomial}(1, p),\ \text{logit}(p) = \alpha + \beta T + \xi X + \gamma XT.$$

Assuming $\alpha = 1$ and $\beta = 1$, we consider three different scenarios: (1) $\xi = 0$, $\gamma = 0$; (2) $\xi = 0.5$, $\gamma = 0$; and (3) $\xi = 0.5$, $\gamma = 0.5$. In each scenario, 1000 data sets of 200 per treatment group were generated to mimic a moderate clinical trial, in which T and X are independently generated. Specifically, each treatment ($T = 1, -1$) and covariate ($X = 1, -1$) combination will have the same number of patients to

achieve the balance. Since $P(T = i, X = j) = .25$, for $i, j = 1, -1$, X and T are independent. Each of the 1000 data sets was fitted through three different models, either correct or misspecified:

$$\text{logit}(p) = \alpha + \beta T, \ \text{logit}(p) = \alpha + \beta T + \xi X, \ \text{and logit}(p) = \alpha + \beta T + \xi X + \gamma XT.$$

The parameters $\alpha = 1$, $\beta = 1$, and $\xi = 0.5$ are the same as they were in Gail et al. (1984). Estimates of β and γ, when applicable, are reported in Table 4.1.

In Table 4.1, the true model refers to the model that we used to generate the simulated datasets. The fitted models refer to the models that were used to fit the simulated datasets and the estimates of β and γ were results of model fitting. In some setups, the true model is simpler than the fitted model; that is, the fitted models included some variables which are not necessary. In this case, as we discussed in Section 2.1, the fitted model still provides an unbiased estimate. However, the estimates have larger variances. For example, the first true model in Table 4.1 is $\eta = 1 + T$. When the true model is fitted, the standard error of $\hat{\beta}$ is 0.131. When the model $\eta = \alpha + \beta T + \xi X + \gamma XT$ is fitted, the standard error of $\hat{\beta}$ is 0.134.

In some other setups, the fitted models are simpler than the true model; therefore the fitted models are misspecified. The second true model in Table 4.1 is $\eta = 1 + T + 0.5X$. When the model $\eta = \alpha + \beta T$ (more formally, $\eta = \alpha^* + \beta^* T$) is fitted, the estimated β^* is 0.963 with standard error 0.128 when the sample size is 200 per treatment group. Table 4.1 also provides the prediction of β, which is computed through equation (3.4) (i.e. $\beta^* = \frac{1}{2}\{h^{-1}(\zeta_1) - h^{-1}(\zeta_2)\}$). The predicted value is 0.954, which is close to 0.963 the sample mean. When the true model is fitted, $\hat{\beta}$ does not bear bias, but it has a larger standard error 0.148.

Table 4.1 also presents results for a smaller sample size, where 1000 data sets of 100 per treatment group were generated. Each treatment ($T = 1, -1$) and covariate ($X = 1, -1$) combination will have the same number of patients, which gives similar conclusions.

Table 4.1 demonstrates that an unadjusted model could lead to results that are difficult to interpret if we ignore a true interaction. When treatment and covariate do not interact, our simulation results are consistent with what Gail et al. (1984)'s result suggested, namely that treatment effect bears small bias when the covariate effect is small. For example, with $(\xi, \gamma) = (0.5, 0)$ in the true model and 200 subjects per treatment group, the estimate of β through the unadjusted model is 0.963.

On the other hand, when treatment and covariate do interact, our results suggest that smaller covariate effects may lead to large treatment effect, different than the phenomenon described in Gail et al. (1984). For example, with $(\xi, \gamma) = (0.5, 0.5)$ in the true model and 200 subjects per treatment group, the estimate of β through the unadjusted model is 0.843. For the same case when the sample is 100 subjects per treatment group, the estimate of β through the unadjusted model is 0.842. We notice the magnitude of bias is not very small despite the small effects of $(\xi, \gamma) = (0.5, 0.5)$. When $(\xi, \gamma) = (1, 0)$ in the true model, the estimate of β through the unadjusted model is 0.842, in which the bias increases considerably from the one that appeared in the case where $(\xi, \gamma) = (0.5, 0)$. Provided that $(\xi, \gamma) = (1, 1)$ in the true model, the estimate of β through the unadjusted model is 0.527, which is again very different from the case in which $(\xi, \gamma) = (0.5, 0.5)$.

The results are generally consistent when $\xi = \gamma = 1$ (i.e. ξ and γ have larger effects). However, we note that $P(Y = 0) = 1/(\exp(4)) = 2\%$ for $X = T = 1$ when data are generated through model $\text{logit}(Y) = 1 + T + X + XT$. Therefore, estimation may not realizable when sample size is small. Consequently, when the sample size is 100 per group, the estimation appeared to be poor. This phenomenon could happen in clinical trials when the primary endpoint is a rare event.

Table 4.1 also confirms the theoretical results in Theorem 2. When the unadjusted model is correct, the estimate of β through an adjusted model is valid with almost the same standard error. However, when an adjusted model is correct with $\xi \neq 0$, the standard error of the treatment effect estimate obtained form the unadjusted model is always smaller.

4.1.2 Bias Derived from Small β

To evaluate the bias approximation (3.10), some simulation studies are conducted. The true model is assume to be

$$\eta = \alpha + \beta T + \gamma_1 X + \gamma_2 XT.$$

Two scenarios are considered: (1) $X = X_1$ and T are independent; (2) $X{=}T \cdot X_1 + X_2$. Here X_1 and X_2 are independent standard normal random variables, and are independent of T. We use $\alpha = -1$ and $\gamma_1 = 1$. Each data set contains 500 observations, and for each simulation, we generate 1000 data sets. Table 4.2 summarizes the approximation results. We can see that (3.10) performs reasonably well even for moderate magnitude of β, though it is derived on the assumption of small β. Also the value of γ_2 influences β^* substantially. For example, in the

Table 4.1. Logistic regression model with E(X)=0

True Model	Fitted Model	Estimate of β	Prediction of β
Sample Size: 200 per treatment group			
$1+T+0X+0XT$	$\alpha+\beta T$	1.010(0.131)	1.000
	$\alpha+\beta T+\xi X$	1.012(0.132)	1.000
	$\alpha+\beta T+\xi X+\gamma XT$	1.021(0.134)	1.000
$1+T+0.5X+0XT$	$\alpha+\beta T$	0.963(0.128)	0.954
	$\alpha+\beta T+\xi X$	1.011(0.135)	1.000
	$\alpha+\beta T+\xi X+\gamma XT$	1.024(0.148)	1.000
$1+T+0.5X+0.5XT$	$\alpha+\beta T$	0.843(0.121)	0.836
	$\alpha+\beta T+\xi X$	0.859(0.121)	.
	$\alpha+\beta T+\xi X+\gamma XT$	1.030(0.159)	1.000
$1+T+X+XT$	$\alpha+\beta T$	0.527(0.096)	0.526
	$\alpha+\beta T+\xi X$	0.566(0.100)	.
	$\alpha+\beta T+\xi X+\gamma XT$	0.996(0.157)	1.000
Sample Size: 100 per treatment group			
$1+T+0X+0XT$	$\alpha+\beta T$	1.018(0.182)	1.000
	$\alpha+\beta T+\xi X$	1.024(0.183)	1.000
	$\alpha+\beta T+\xi X+\gamma XT$	1.042(0.192)	1.000
$1+T+0.5X+0XT$	$\alpha+\beta T$	0.967(0.183)	0.954
	$\alpha+\beta T+\xi X$	1.019(0.191)	1.000
	$\alpha+\beta T+\xi X+\gamma XT$	1.036(0.203)	1.000
$1+T+0.5X+0.5XT$	$\alpha+\beta T$	0.842(0.170)	0.826
	$\alpha+\beta T+\xi X$	0.860(0.170)	.
	$\alpha+\beta T+\xi X+\gamma XT$	1.019(0.195)	1.000
$1+T+X+XT$	$\alpha+\beta T$	0.515(0.136)	0.526
	$\alpha+\beta T+\xi X$	0.552(0.142)	.
	$\alpha+\beta T+\xi X+\gamma XT$	0.891(0.162)	1.000

logistic regression with $\beta = 4$ and $\gamma_2 = 5$, the bias due to independent covariate can be as large as 2.4, about 60% of the true value. Similar findings exist for Poisson regression.

Table 4.2. Estimated covaraite effect $\hat{\beta}^*$ (standard deviation) and predicted β^* for logistic regressions and Poisson regressions.

Logistic Regression				
	$Z = X_1$		$Z = T \cdot X_1 + X_2$	
γ_2	Estimated	Predicted	Estimated	Predicted
		$\beta = 2$		
1	1.443360(0.191288)	1.45804	1.316943(0.186801)	1.320386
2	1.304652(0.199228)	1.30566	1.181350(0.183441)	1.202560
5	1.097617(0.184414)	1.08381	1.022826(0.189531)	1.008172
		$\beta = 4$		
1	2.743310(0.239851)	2.68512	2.331781(0.213346)	2.279147
2	2.272628(0.211908)	2.23166	1.901979(0.195592)	1.874906
5	1.593876(0.195371)	1.59686	1.386834(0.190386)	1.384709
Poisson Regression				
	$Z = X_1$		$Z = T \cdot X_1 + X_2$	
γ_2	Estimated	Predicted	Estimated	Predicted
		$\beta = 1$		
0.1	1.101648(0.152721)	1.10487	1.700137(0.224806)	1.697815
0.5	1.612100(0.223788)	1.62796	2.677884(0.392929)	2.728037
1	2.440693(0.337080)	2.49086	4.130599(0.705567)	4.479328
		$\beta = 2$		
0.1	2.102946(0.158514)	2.10472	2.718248(0.215063)	2.694547
0.5	2.619232(0.213241)	2.62375	3.667898(0.378647)	3.71778
1	3.446755(0.349116)	3.51611	5.183614(0.680485)	5.369677

4.2 Real Data Illustration

A Phase III randomized, double-blind, placebo-controlled clinical trial was conducted from 1996 to 1997 to evaluate the safety and efficacy of prophylaxis with

Palivizumab in reduction of respiratory syncytial virus (RSV) infection in high-risk infants. A total of 1502 children with prematurity or bronchopulmonary dysplasia (BPD) were randomized to receive either palivizumab or placebo intramuscularly. The primary endpoint was RSV-related hospitalization within 150 days after administration of the first dose of treatment. For more information about this trial please refer to IMpact-RSV Study Group (1998).

Among the 500 subjects who received placebo, 53 (10.6%) had an RSV-related hospitalization; among the 1002 subjects who received palivizumab, 48 (4.8%) had an RSV-related hospitalization.

Without considering any confounders, we use model (3.1) to estimate the odds ratio of Placebo vs Palivizumab. The status of RSV-related hospitalization for the i^{th} subject, y_i, is modeled through the following logistic regression model:

$$y_i \sim \text{Binomial}(1, p_i),\ \text{logit}(p_i) = \alpha^* + \beta^* T_i,\ \ i = 1, \cdots, n$$

with $n = 1502$. Here $T_i = 1$ or -1 if the ith subject took Palivizumab or Placebo, respectively. Note that the parameter β^* is half of the log odds ratio.

Maximum likelihood estimates and standard errors of the parameters are $\hat{\alpha}^* = -2.561(0.104)$ and $\hat{\beta}^* = -0.429(0.104)$. Therefore, the response rates for experimental and standard treatment differ significantly with a two sided p-value smaller than 0.0001.

It worth noting that the population of this trial included exclusively two disjointed subgroups: 1) children 24 months old or younger with a clinical diagnosis

of BPD requiring ongoing medical treatment; and 2) children with 35 weeks gestation or less and 6 months old or younger, who did not have a clinical diagnosis of BPD. Among patients enrolled with a diagnosis of BPD, the incidence rate of RSV-related hospitalization is 12.8% (34/266) in the placebo arm and 7.9% (39/496) in the Palivizumab arm. Among patients enrolled without a diagnosis of BPD, the the incidence rate of RSV-related hospitalization is 8.1% (19/234) in the placebo arm and 1.8% (9/506) in the Palivizumab arm.

Understanding the heterogeneity of treatment effect size plays important role in improving treatments and developing new targeted treatments. For this purpose, we continue to analyze the data to incorporate each subject's BPD status. Let $X_i = 1$ or 0 if the i^{th} subject had a diagnosis of BPD or not, respectively. As we explained before, in order to make β and β^* have the same meaning, we centralize X so that $E(X) = 0$. In this example, we therefore subtract X_i from 0.507, the sample mean of X. Let $X_i^* = 0.493$ or -0.507 if the i^{th} subject had a diagnosis of BPD or not, respectively. The following model is therefore fitted:

$$y_i \sim \text{Binomial}(1, p_i),\ \text{logit}(p_i) = \alpha + \beta T_i + \xi X_i^* + \gamma X_i^* T_i,\ i = 1, \cdots, n.$$

Maximum likelihood estimates and standard errors of the parameters are $\hat{\alpha} = -2.697(0.120)$, $\hat{\beta} = -.528(0.120)$, $\hat{\xi} = 1.028(0.241)$ and $\hat{\gamma} = 0.522(0.241)$. The interaction between X and T is statistically significant with the p-value 0.03.

In addition to the logistic model, we also used log and probit link functions for the binomial distribution. The results are given in Table 4.3. From the results, we notice that the covariate adjusted and unadjusted approaches usually give different point estimates. In addition, the results confirm that $var(\hat{\beta}^*)$ is smaller than

$var(\hat{\beta})$, which is always true whether the covariate does or does not interact with treatment.

Table 4.3. Treatment effects under different models and link functions for Palivizumab study

Model	Log Link	Logit Link	Probit Link
$\eta = \alpha^* + \beta^* T_i$	-.397 (.096)	-.429(.104)	.-209(.051)
$\eta = \alpha + \beta T_i + \xi X_i$	-.375 (.095)	-.418(.104)	-.212(.052)
$\eta = \alpha + \beta T_i + \xi X_i^* + \gamma X_i^* T_i$	-.497 (.113)	-.528(.120)	.244(.055)

Chapter 5

Future Work

5.1 Discussions

In this thesis, we theoretically compare covariate-adjusted and -unadjusted approaches for generalized linear models. Simulation studies confirm these theoretical results and suggest the results are applicable to cases with moderate sample size. In general, a covariate-adjusted model allows correct interpretations of the treatment effect, while an unadjusted model can lead to misleading results. On the other hand, a covariate-adjusted model is associated with reduced precision. When the covariate effects are small, then an unadjusted approach might still be preferable in some cases. Thus, there is a trade-off between accuracy and precision. Researchers should choose suitable approaches to achieve the goal, keeping in mind their advantages and disadvantages.

When the treatment interacts with covariates, β^* is usually biased when the covariate-adjusted model is true. The interaction leads to two different scenarios: 1) treatment A is always superior to treatment B, while the magnitude of difference

between A and B is different in different subgroups; 2) treatment A is superior to B in on some subgroups and is inferior to treatment B in other subgroups. In scenario 2, although both β^* and β both represent the "average" treatment effect, they may not the best way to define the treatment effect. In this case, treatment effects may be better quantified in subgroups. Consequently, the results in Section 2.2 may not be very relevant in scenario 2.

We only compare adjusted and unadjusted models. At this point, we are not able to provide general results to compare all possible covariate-adjusted models. Nevertheless, we believe all comparisons involve trade-off between accuracy and precision. In our opinion, a model selection procedure is a good way to balance accuracy and precision, although its practical implementation is itself an interesting and promising research area with many unsolved issues. The results form a basis to evaluate a covariate-adjusted approach in terms of bias and precision. However, we acknowledge that further research is needed.

We compare covariate-adjusted and -unadjusted generalized linear models with the same dispersion parameter. Extension to cases with different dispersion parameters are of interest for future research.

Bibliography

[1] Akaike, H. (1973). Information theory and an extension of the likelihood principle. Proceedings of the Second International Symposium of Information Theory, ed. B. N. Petrov and F. Csáki. Budapest: Akadémiai Kiado.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, **19**, 716-723.

[3] Behrendt, C. E. and Gehan, E. A. (2009). Treatment-subgroup interaction: An example from a published, phase II clinical trial. *Contemporary Clinical Trials*, **30**, 279-281.

[4] Cox, D. R. (1958). Planning of Experiments. New York: Wiley.

[5] Drake, C. and McQuarrie, A. (1995). A note on the bias due to omitted confounders. *Biometrika*, **82**, 633-638.

[6] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, **222**, 309-368.

[7] Fisher, R. A. (1925). Theory of Statistical Estimation. *Proceedings of the Cambridge Philosophical Society*, **22**, 700-725.

[8] Ford, I. and Norrie, J. (2002). The role of covariates in estimating treatment effects and risk in long-term clinical trials. *Statistics in medicine*, **21** 2899 – 2908.

[9] Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109-148.

[10] Friedman, J. H. (2008). Fast sparse regression and classification. http://stat.stanford.edu/ jhf/#reports

[11] Gail, M. and Simon, R. (1985). Testing for qualitative interaction between treatment effects and patient subsets. *Biometrics*, **41**, 361-372.

[12] Gail, M., Wieand, S. and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with non-linear regressions and omitted covariates. *Biometrika*, **71**, 431-44.

[13] Grouin, J. M., Coste, M. and Lewis, J. (2005). Subgroup Analyses in Randomized Clinical Trials: Statistical and Regulatory Issues. *Journal of Biopharmaceutical Statistics*, **15**, 869-882.

[14] Grouin, J. M., Day, S. and Lewis, J. (2004). Adjustment for baseline covariates: an introductory note. *Statistics in medicine*, **23** 697-699.

[15] Hauck, W. W., Anderson, S. and Marcus, S. M. (1998). Should We Adjust for Covariates in Nonlinear Regression Analyses of Randomized Trials? *Controlled Clinical Trials*, **19**, 249 - 256.

[16] Knol, M. J., van der Tweel, I., Grobbee, D. E., Numans, M. E. and Geerlings, M. I. (2007). Estimating interaction on an additive scale between continuous determinants in a logistic regression model. *Int J Epidemiol*, **36**, 1111-1118.

[17] Koch, G. G., Tangen, C. M., Jung, J. W. and Amara, I. A. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine*, **17**, 1863-1892.

[18] Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79-86.

[19] LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related bayes' estimates. *University of California Publications in Statistics*, **1**, 277-330.

[20] McClelland, G. H. and Judd, C. M. (1993). Statistical Difficulties of Detecting Interactions and Moderator Effects. *Psychological Bulletin*, **114**, 376-390.

[21] McCullagh, P., Nelder, J. (1989). Generalized linear models, second edition. Chapman and Hall.

[22] Nelder, J., Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, **135**, 370-384.

[23] Neuhuas, J. M. (1998). Estimation efficiency with omitted covariates in generalized linear models. *Journal of American Statistical Association*, **93**, 1124-1129.

[24] Neuhaus, J. M. and Jewell, N. P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, **80**, 807-815.

[25] Neuhaus, J. M., Kalbfleisch, J. D., Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, **59**, 25-35.

[26] Peto, R. (1982). Statistical aspects of cancer trials. *In Treatment of Cancer*, K. E. Halnan (ed.), 867-871. London: Chapman and Hall.

[27] Pocock, S. J. (2004). Clinical trials: a practical approach. John Wiley.

[28] Pocock, S. J., Assmann, S. E., Enos, L. E. and Kasten, L. E. (2002). Subgroup analysis, covariate adjusted and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in medicine*, **21** 2917 - 2930.

[29] Richardson, D. B. and Kaufman, J. S. (2009). Estimation of the Relative Excess Risk Due to Interaction and Associated Confidence Bounds. *American Journal of Epidemiology*, **169**, 756-760.

[30] Robinson, L. D., Dorroh, J. R., Lien, D. and Tiku, M. L. (1998). The effects of covariate adjusted in generalized linear models. *Communications in Statistics - Theory and Methods*, **27**,1653-1675.

[31] Robinson, L. D. and Jewell, N. P. (1991). Some Surprising Results About Covariate Adjusted in Logistic Regression Models. *International Statistical Review*, **58**, 227-240.

[32] Rosset, S and Zhu J. (2007). Piecewise linear regularized solution paths. *Annual of Statistics*, **35**, 1012-1030.

[33] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **19**, 461-464.

[34] Senn, S. (2004). Conditional and Marginal Models: Another View. *Statistical Science*, **19**, 228-230.

[35] Senn, S. (2007). *Statistical issues in Drug development*, John Wiley and Sons.

[36] Stürmer, T. and Brenner, H. (2002). Flexible Matching Strategies to Increase Power and Efficiency to Detect and Estimate Gene-Environment Interactions in Case-Control Studies. *American Journal of Epidemiology*, **155**, 593-602.

[37] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **60**, 595-603.

[38] Weng, H. Y., Hsueh, Y. H. and Messam, L. L. and Hertz-Picciotto, I. (2009). Methods of covariate selection: directed acyclic graphs and the change-in-estimate procedure. *Am J Epidemiol*, **169**, 1182-1190.

[39] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1-25.

[40] Zeger, S. L., Liang, L. Y., Albert, P. A. (1998). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049-60.

Zou, G. Y. (2008). On the estimation of additive interaction by use of the four-by-two table and beyond. *Am J Epidemiol*, **168**, 212-224.