sciendo

DOI 10.2478/bile-2024-0003

Biometrical Letters
Vol. 61 (2024), No. 1, 33-50

Incompatibility of Model Specifications in Generalized Linear Models

Changyong Feng[1,2], Hongyue Wang[1], Honghong Liu[1]

[1]Departments of Biostatistics and Computational Biology
[2]Department of Anesthesiology and Perioperative Medicine, University of Rochester, Rochester, NY 14642, USA, e-mail: changyong_feng@urmc.rochester.edu

Summary

The choice of regression models can vary depending on a study's objectives and the characteristics of the dataset. Yet it is crucial to recognize that the properties of conditional expectation impose a natural consistency requirement on the formulation of these models. This paper endeavors to demonstrate that, even under the most favorable assumptions concerning the covariates' structure, inconsistent specifications can arise in generalized linear models, particularly when nonlinear link functions are employed.

Key words: conditional expectation, linear regression, logistic regression

1. Introduction

The following notation will be used throughout this paper. Letting $(\Omega, \mathcal{F}, P)$ be a probability space, W be a random variable defined on it, and $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$, we denote the sub-σ-field generated by W as $\sigma(W)$, and the *conditional expectation* of W given $\mathcal{G}$, assuming $\mathrm{E}|W| < \infty$, as $\mathrm{E}(W \mid \mathcal{G})$. For any $A \subset \Omega$, we use $\mathbf{1}_A$ to denote its *indicator function*, i.e., $\mathbf{1}_A(\omega) = 1$ (or 0) if $\omega \in A$ (or $\omega \notin A$). Moreover, for two random variables U and V defined on the same probability space, we use $U \perp V$ to denote that they are independent.

Let Y represent the dependent variable (outcome variable), and $\mathbf{X}$ represent the vector of explanatory factors, also known as independent variables or covariates in some literature. Although the joint distribution function F perfectly characterizes the statistical relationship between Y and $\mathbf{X}$, it is often unknown and difficult to estimate when the dimension of $(Y, \mathbf{X})$ is high. This is known as the "curse of dimensionality" (Hastie et al., 2009).

In practice, regression analysis is a commonly used statistical tool for studying the association between the dependent variable Y and covariates $\mathbf{X}$. However, many researchers have misconceptions or misinterpretations about regression analysis. For example, some believe that the goal of regression analysis is to fit a linear function to the data using the least squares method, while others cannot even explain what regression analysis means.

William Feller, one of the greatest probabilists of the twentieth century, gave the following mathematically rigorous definition of regression in his two-volume textbook on probability theory (Feller, 1966, p. 71).

Definition 1. Given an integrable random variable Y and a random vector $\mathbf{X}$, the regression of Y on $\mathbf{X}$, denoted by $\mathrm{E}[Y \mid \mathbf{X}]$, is the conditional expectation of Y given $\sigma(\mathbf{X})$, i.e.,

$$\mathrm{E}[Y \mid \mathbf{X}] = \mathrm{E}[Y \mid \sigma(\mathbf{X})].$$

This definition is also used in Hastie, Tibshirani and Friedman (2009, p. 18). According to Theorem 4.2.8 in Dudley (2002), $\mathrm{E}[Y \mid \mathbf{X}] = h(\mathbf{X})$ for a unique measurable function h, which is also called the *regression function* (Hastie, Tibshirani and Friedman, 2009). However, the exact form of h is usually unknown in practice, and needs to be specified in data analysis. Any specification is an approximation of h.

When specifying regression models, it is important to ensure that they are consistent, which means that they do not make contradictory assumptions on the same data set. In practice, this issue has not been studied systematically in the literature, and many publications have specified inconsistent regression models in their data analysis. For example, Geng et al. (2005), Hole, Ball and Meads (2015), and Leon et al. (2016) have all used inconsistent regression models in their studies. Ensuring consistency in the specification of regression models is crucial for obtaining reliable and interpretable results. Inconsistency can lead to biased estimates, invalid inferences, and misleading conclusions.

In section 2 we discuss the importance of ensuring consistency in regression models. In sections 3 and 4 we use linear regression and logistic regression models to illustrate how inconsistency can arise in some simple cases. In section 5 we extend our discussion to generalized linear models, and in section 6 we examine the inconsistency presented in a renowned paper published in *Biometrika*. In section 7 we present a numerical example from logistic regression that demonstrates how inconsistent regression

models can lead to different conclusions about the relationship between the response variable and the covariates. Finally, we provide concluding remarks in section 8, highlighting the importance of checking the assumptions underlying the regression models to ensure consistency and obtain reliable results.

2. Compatibility of specification in regression models

In the following discussion we assume Y is a real-valued integrable dependent variable, and $\mathbf{X}$ is a p-dimensional random vector of covariates of interest.

For each subvector $\mathbf{X}^{(s)}$ of $\mathbf{X}$, let $d^{(s)}$ denote its dimension. There is a unique measurable function $f_s : \mathbb{R}^{d^{(s)}} \to \mathbb{R}$ such that

$$\mathrm{E}[Y \mid \mathbf{X}^{(s)}] = f_s(\mathbf{X}^{(s)}).$$

Generally, if $\mathbf{X}^{(s)} \neq \mathbf{X}^{(s')}$, then $f_s \neq f_{s'}$ for two possible reasons: (a) $d^{(s)} \neq d^{(s')}$, and (b) even if $d^{(s)} = d^{(s')}$, the functional forms of f_s and $f_{s'}$ may be different as the distribution of $(Y, \mathbf{X}^{(s')})$ may not be the same as $(Y, \mathbf{X}^{(s)})$. This means that if we want to use different subvectors of $\mathbf{X}$ in the regression analysis, we need to specify different regression functions.

In practice, we may be interested in different subsets of covariates $\mathbf{X}^{(s)}$ for various reasons, such as reducing model complexity, testing specific hypotheses, or accounting for potential interactions between covariates. However, it is important to keep in mind that the choice of $\mathbf{X}^{(s)}$ affects the regression function and the resulting inference about the relationship between Y and $\mathbf{X}$. Careful consideration should be given to the choice of covariates in regression analysis.

Define

$$\mathcal{M}_{\mathbf{X}} = \left\{ \mathrm{E}[Y \mid \mathbf{X}^{(s)}] : \ \mathbf{X}^{(s)} \text{ is a subvector of } \mathbf{X} \right\}.$$

Since there are p elements in $\mathbf{X}$, the total number of subvectors (including the null vector) of X is 2^p. Hence, the cardinality of $\mathcal{M}_{\mathbf{X}}$ is $\leqslant 2^p$. This means that when using the regression method to study the association between Y and some components in $\mathbf{X}$, we can specify up to 2^p regression models. If more than 2^p regression models are specified given X, then either the specification is redundant, i.e., some specified models are exactly the same, or some specified models are not in $\mathcal{M}_{\mathbf{X}}$.

From the properties of conditional expectation, it is always true that

$$\begin{aligned} \mathrm{E}[f_s(\mathbf{X}^{(s)})] &= \mathrm{E}\left\{\mathrm{E}[Y \mid \mathbf{X}^{(s)}]\right\} = \mathrm{E}[Y] \\ &= \mathrm{E}\left\{\mathrm{E}[Y \mid \mathbf{X}^{(s')}]\right\} = \mathrm{E}[f_{s'}(\mathbf{X}^{(s')})] \end{aligned} \tag{1}$$

for all $\mathbf{X}^{(s)}$ and $\mathbf{X}^{(s')}$ of subvectors of X.

Aside from (1), there is a natural *consistency* constraint on regression functions. If $\mathbf{X}^{(s')}$ is a subvector of $\mathbf{X}^{(s)}$ (we say $\mathbf{X}^{(s')}$ is *nested* in $\mathbf{X}^{(s)}$), then we say that the regression model $\mathrm{E}[Y \mid \mathbf{X}^{(s')}]$ is nested within the model $\mathrm{E}[Y \mid \mathbf{X}^{(s)}]$. Then we should have

$$f_{s'}(\mathbf{X}^{(s')}) = \mathrm{E}[f_s(\mathbf{X}^{(s)}) \mid \mathbf{X}^{(s')}]. \tag{2}$$

This is the minimum consistency requirement for regression models. Any specification not satisfying (2) is said to be inconsistent. Inconsistency means that at least one of the regression functions is misspecified.

Regression analysis involves modeling the relationship between one or more covariates and a dependent variable. The choice of covariates to be included in the model depends on the specific research question or purpose of the analysis.

Since regression models are used to estimate the conditional expectation of a dependent variable, the choice of covariates will depend on the factors that are believed to be relevant to the dependent variable. In some cases, a researcher may include only a few key covariates that are hypothesized to be most important in explaining the variation in the dependent variable. In other cases, a researcher may include a large number of covariates to account for a wide range of potential factors that could be influencing the dependent variable.

Ultimately, the choice of covariates should be based on a combination of theoretical knowledge, prior research, and empirical evidence. Different regression models can be constructed depending on the purpose of the study, and the appropriate model will depend on the research question and available data.

Example 1. Suppose we are interested in studying the relationship between blood pressure (Y) and three factors, namely age (X_1), gender (X_2), and family history of hypertension (X_3), in a population of interest. To investigate the relationship between blood pressure and all three factors, we can construct the regression model:

$$\mathrm{E}[Y \mid X_1, X_2, X_3].$$

However, if we are only interested in the relationship between blood pressure and age, we can construct the regression model

$$\mathrm{E}[Y \mid X_1].$$

Both simple and multiple regression models are plausible methods for studying the relationship between dependent variable and one or more covariates. However, the specific research question and the nature of the data will determine which regression model is most appropriate for the analysis. For example, if the research question involves examining the effect of only one covariate on the dependent variable, a univariate regression model may be used. On the other hand, if the research question involves examining the effect of multiple covariates on the dependent variable, a multiple regression model may be more appropriate. Therefore, the purpose of the study and the research question being addressed should guide the selection of the appropriate regression model.

The form of the regression function is generally unknown and needs to be appropriately specified in order to satisfy the consistency condition as described in equation (2). Failure to satisfy this condition can result in inconsistent and uninterpretable analysis results.

Unfortunately, inconsistent specifications are not uncommon in the statistical literature, as evidenced by Neuhaus and Jewell (1993). Therefore, it is crucial to carefully consider the choice of regression function specification and ensure that it satisfies the consistency condition to ensure the validity and interpretability of the analysis results.

In the following sections, we will demonstrate how inconsistency can easily occur in regression model specifications using linear and logistic regression models. By highlighting examples of inconsistent specifications, we aim to illustrate the importance of carefully considering the choice of regression function specification in order to ensure the validity and interpretability of the analysis results.

3. Inconsistency in linear regression model specification

In this section we use examples in linear regression to show the inconsistency between univariate and multiple regressions.

Given two covariates X_1 and X_2, suppose the dependent variable Y is generated in the following way:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

where $\beta_1\beta_2 \neq 0$, $\varepsilon \perp (X_1, X_2)$, $\mathrm{E}(\varepsilon) = 0$, and $\mathrm{Var}(\varepsilon) = \sigma^2 < \infty$. Without loss of generality, assume $\mathrm{E}[X_1] = \mathrm{E}[X_2] = 0$. The regression of Y on (X_1, X_2) is

$$\mathrm{E}[Y \mid X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

The univariate regression of Y on X_1 is

$$\begin{aligned} \mathrm{E}[Y \mid X_1] &= \mathrm{E}\left\{\mathrm{E}[Y \mid X_1, X_2] \mid X_1\right\} \\ &= \beta_0 + \beta_1 X_1 + \beta_2 \mathrm{E}[X_2 \mid X_1] \\ &= \beta_0 + \beta_1 X_1 + \beta_2 h(X_1) \end{aligned} \tag{3}$$

for some measurable function h.

From (3) we can see that the univariate regression of Y on X_1 depends on the conditional expectation of X_2 given X_1. Generally, from (3) it is not true that

$$\mathrm{E}[Y \mid X_1] = \alpha_0 + \alpha_1 X_1 \tag{4}$$

for some constants α_0 and α_1, unless $h(X_1)$ is a linear function of X_1.

If the function $h(X_1)$ is nonlinear, the univariate regression of Y on X_1 is not a *simple linear regression*, as in the form of the right-hand side of (4). This implies that, given a *multiple linear regression* model, we cannot automatically assume that the *univariate regression* of the outcome on each individual covariate has the form of a simple linear regression. This phenomenon is illustrated in Examples 2–4.

In the special case that $X_1 \perp X_2$, $\mathrm{E}[X_2 \mid X_1] = \mathrm{E}[X_2] = 0$, and the univariate regression of Y on X_1 is a simple linear regression,

$$\mathrm{E}[Y \mid X_1] = \beta_0 + \beta_1 X_1.$$

This may not hold if X_1 and X_2 are correlated.

In the following three examples, the multiple linear regressions have exactly the same form. However, the univariate regression of Y on X_1 is completely different.

Example 2. Assume the following multiple linear regression model:

$$Y = 1 + 3X_1 - X_2 + \varepsilon, \tag{5}$$

where (X_1, X_2, ε) has standard multivariate normal distribution. Then

$$\mathrm{E}[Y \mid X_1] = 1 + 3X_1.$$

The regression of Y on X_1 is a simple linear regression.

Example 3. Assume the same multiple linear regression model as (5), with $X_1 \sim N(0,1)$, $\mathrm{E}[X_2 \mid X_1] = X_1^2 - 1$, $\varepsilon \sim N(0,1)$, and $(X_1, X_2) \perp \varepsilon$. Then

$$\mathrm{E}[Y \mid X_1] = 2 + 3X_1 - X_1^2.$$

The regression of Y on X_1 is no longer a simple linear regression, as it cannot be written in the form of the right-hand side of (4).

Example 4. Assume the same multiple linear model as (5), with $X_1 \sim N(0,1)$, $\mathrm{E}[X_2|X_1] = X_1^3$, $\varepsilon \sim N(0,1)$, and $(X_1, X_2) \perp \varepsilon$. Then

$$\mathrm{E}[Y|X_1] = 1 + 3X_1 - X_1^3.$$

The regression of Y on X_1 no longer satisfies a simple linear regression model. It is easy to prove that $\mathrm{Cov}(Y, X_1) = 0$. This means that even if a covariate is known to be uncorrelated with the outcome, we cannot simply exclude it from the multiple regression.

Remark 1. These three examples highlight how the same multiple linear regression model can result in entirely different univariate regression models, depending on the correlations between the covariates. Given a multiple linear regression model, simply assuming a simple linear regression for each individual covariate can lead to inconsistent specifications. Unfortunately, this issue has often been overlooked in biomedical publications, as evidenced by studies such as Geng et al. (2005), Hole, Ball and Meads (2015), and Leon et al. (2016).

From these three examples, we can see that we need to make a distinction between *univariate regression* and *simple linear regression*:

- Even if both Y and X_1 are continuous random variables, the univariate regression of Y on X_1 is not necessarily a simple linear regression. In other words, the relationship between Y and X_1 may not be adequately captured by a linear equation of the form $Y = \beta_0 + \beta_1 X_1 + \epsilon$.
- Even if the regression of Y on (X_1, X_2) satisfies a multiple linear regression model, the univariate regression of Y on X_1 is not necessarily a simple linear regression.

When both the dependent variable and covariates are continuous, we can use the least squares method to fit a linear function to the data. The

idea behind the least squares method is to find the values of the parameters that minimize the sum of the squared residuals, which is the difference between the observed values of the dependent variable and the values predicted by the linear function. In the case of fitting a linear function to the data $(Y_i, X_{1i}), i = 1, 2, \ldots, n$, we want to find the values of the intercept α_0 and the slope α_1 such that the sum of the squared residuals is minimized. The estimate of α_1 can be found using the formula:

$$\hat{\alpha}_1 = \frac{\sum_i^n (Y_i - \bar{Y})(X_{1i} - \bar{X}_1)}{\sum_i^n (X_{1i} - \bar{X}_1)^2},$$

where $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$ and $\bar{X}_1 = n^{-1} \sum_{i=1}^n X_{1i}$ are the sample means of the dependent variable and covariate, respectively. The formula shows that $\hat{\alpha}_1$ is the ratio of the sample covariance between Y and X_1 and the sample variance of X_1. As n goes to infinity, $\hat{\alpha}_1$ converges almost surely to $\mathrm{Cov}(Y, X_1)/\mathrm{Var}(X_1)$. If X_1 and Y are uncorrelated, then $\hat{\alpha}_1$ will be close to zero. Sometimes, even if Y is perfectly determined by X, they may be uncorrelated. For example, assume $X \sim N(0, 1)$ and $Y = X^2$; then $\mathrm{Cov}(Y, X) = 0$, which implies that X and Y are uncorrelated, despite Y being a deterministic function of X.

4. Inconsistency in logistic regression model specification

Given a 0-1-valued dependent variable Y and two covariates X_1 and X_2, assume the regression of Y on (X_1, X_2) is of the form

$$\mathrm{E}[Y \mid X_1, X_2] = P\{Y = 1 \mid X_1, X_2\} = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)},$$

where $\beta_1 \beta_2 \neq 0$.

Consider the univariate regression of Y on X_1. We want to know if there always exist constants β_0^* and β_1^* such that

$$\mathrm{E}[Y \mid X_1] = P\{Y = 1 \mid X_1\} = \frac{\exp(\beta_0^* + \beta_1^* X_1)}{1 + \exp(\beta_0^* + \beta_1^* X_1)}.$$

To this end, we first present two lemmas, whose proofs are given in the Appendix.

Lemma 1. *Given $p \in (0,1)$, real variables x, y, and z satisfy the following two equations:*

$$\frac{p}{1+\exp(y)} + \frac{1-p}{1+\exp(z)} = \frac{1}{1+\exp(x)} \quad \textit{and} \tag{6}$$

$$\frac{p}{\exp(y)} + \frac{1-p}{\exp(z)} = \frac{1}{\exp(x)} \tag{7}$$

if and only if $x = y = z$.

Lemma 2. *Given constants $p, \beta_{01}, \beta_{02}, \beta_1, \beta_2$, with $0 < p < 1$, $\beta_1\beta_2 \neq 0$, and $(\beta_{01}, \beta_1) \neq (\beta_{02}, \beta_2)$, define a function*

$$f(x) = \frac{p}{1+\exp(\beta_{01}+\beta_1 x)} + \frac{1-p}{1+\exp(\beta_{02}+\beta_2 x)}, \quad x \in \mathbb{R}.$$

Then there do not exist constants α_0 and α_1 such that $f(x)$ can be written in the form

$$f(x) = \frac{1}{1+\exp(\alpha_0+\alpha_1 x)}, \quad x \in \mathbb{R}. \tag{8}$$

The main result of this section now follows.

Theorem 1. *Suppose Y and Z are 0-1-valued random variables, X is a continuous random variable with range $\mathbb{R}$, $X \perp Z$, and $0 < P\{Z=1\} < 1$. If the regression of Y on (X,Z) is of the form*

$$E[Y \mid X, Z] = \frac{\exp(\beta_0+\beta_1 X+\beta_2 Z)}{1+\exp(\beta_0+\beta_1 X+\beta_2 Z)}, \tag{9}$$

where $\beta_1\beta_2 \neq 0$, then we cannot find constants α_0 and α_1 such that the univariate regression of Y on X is of the form

$$E[Y \mid X] = \frac{\exp(\alpha_0+\alpha_1 X)}{1+\exp(\alpha_0+\alpha_1 X)}. \tag{10}$$

Proof. Since $X \perp Z$, we have

$$\begin{aligned} P\{Y=0 \mid X\} &= \mathrm{E}[P\{Y=0 \mid X,Z\}|X] \\ &= \frac{p}{1+\exp(\beta_0+\beta_2+\beta_1 X)} + \frac{1-p}{1+\exp(\beta_0+\beta_1 X)}. \end{aligned}$$

Since $\beta_2 \neq 0$, Lemma 2 indicates that $P\{Y=0 \mid X\}$ cannot be written in the form of (10). □

Remark 2. Theorem 1 reveals that even when the covariates are independent and the distribution of Z is simple, it is not appropriate to assume that the univariate regression of Y on X follows the form of (10) as derived from the multiple logistic regression in (9). This is in contrast to the linear regression setting discussed in section 3. In cases where X and Z are correlated, the univariate regression of Y on X is expected to be much more complex. Additionally, if Z has non-degenerate distributions other than the assumed simple distribution, we cannot expect (10) to hold. On the other hand, assuming that the univariate regression of Y on X is in the form of a univariate logistic regression (10), we cannot assume that a multiple logistic regression (9) holds, even if X and Z are independent.

5. Inconsistency in generalized linear model specification

In the last two sections, it was demonstrated that inconsistency of model specification can occur in both linear and logistic regressions, even in simple cases. A multiple linear regression cannot necessarily imply a simple linear regression for each covariate unless the covariates are independent. Similarly, a multiple logistic regression may not imply a univariate logistic regression, even if the covariates are independent.

This section is devoted to discussing model specification within the framework of Generalized Linear Models (GLMs) (McCullagh and Nelder, 1989). GLMs extend the well-developed theory and practice of linear models (Rao, 1973). To simplify the notation, we include 1 as the first element of X and define $\mathbf{X} = (1, X_1, X_2, \ldots, X_p)$. GLMs assume that the regression of Y on $\mathbf{X}$ takes the form

$$\mathrm{E}[Y \mid \mathbf{X}] = g^{-1}(\eta), \tag{11}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a strictly increasing function (called *link function*), and

$$\eta = \mathbf{X}\boldsymbol{\beta}$$

is called the *linear predictor*, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$, with $\beta_1\beta_2 \cdots \beta_p \neq 0$. Therefore, the regression function in GLMs is of the form

$$f(\mathbf{X}) = g^{-1}(\mathbf{X}\boldsymbol{\beta}).$$

In comparison with linear regression models, the "generalized" qualifier in GLMs refers to the link function, which is simply the identity function in

linear regressions. Although the link function can be highly flexible, GLMs remain relatively restrictive, as they stipulate that the regression of Y on X can only be accomplished through the linear predictor.

Suppose we have a regression of Y on $\mathbf{X}$ in the form of (11). If we augment the covariate list to include additional variables, denoted by $\mathbf{X}^* = (1, X_1, \ldots, X_p, X_{p+1}, \ldots, X_r)$ where $r > p$, we can consider the regression of Y on X^*. Let $f^*(\mathbf{X}^*) = \mathrm{E}[Y \mid \mathbf{X}^*]$. As discussed above, f^* is not necessarily equal to f. We want to determine whether f^* also has the form

$$f^*(\mathbf{X}^*) = g^{-1}(\eta_a), \tag{12}$$

where

$$\eta_a = \mathbf{X}^* \boldsymbol{\alpha}, \tag{13}$$

and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p, \alpha_{p+1}, \ldots, \alpha_r)'$, with $\alpha_1 \alpha_2 \cdots \alpha_r \neq 0$.

To ensure that equation (12) is valid, the following must hold:

- The regression of Y on $\mathbf{X}^*$ is solely determined by a linear predictor given by equation (13). This implies that the relationship between Y and $\mathbf{X}^*$ should satisfy a generalized linear model.
- The link function used in equation (12) is the same as the one used in equation (11).

Are the specifications in (11) and (12) consistent? In other words, can we assume that (11) is the regression of Y on $\mathbf{X}$ and (12) is the regression of Y on $\mathbf{X}^*$, and vice versa? If the answer is yes, then we should have:

$$g^{-1}(\mathbf{X}\boldsymbol{\beta}) = \mathrm{E}[g^{-1}(\mathbf{X}^*\boldsymbol{\alpha}) \mid \mathbf{X}] \tag{14}$$

To show that the models in (11) and (12) are inconsistent, we only need to demonstrate that (14) does not hold in general.

The findings presented in sections 3 and 4 suggest that a multiple generalized linear model may not necessarily correspond to a generalized linear model for a subset of the covariates, especially when a nonlinear link function is employed in the generalized linear model. In the following section, we will consider the case of a prominent paper in a leading statistical journal in which this fact was overlooked.

6. An example from a well-cited work

Neuhaus and Jewell (1993) conducted a study on the bias of estimating the coefficient of a covariate in a GLM when some component covariates are omitted. Their setting is as follows:
(i) Assume that the regression of Y on covariate vector (X, Z) is of the form

$$\mathrm{E}[Y \mid X, Z] = g^{-1}(\beta_0 + \beta X + \gamma Z), \tag{15}$$

where X and Z are independent random variables with arbitrary distributions, g is a strictly increasing link function, and $\beta\gamma \neq 0$. Then the *univariate regression* of Y on X is

$$\mathrm{E}[Y \mid X] = \mathrm{E}\{\mathrm{E}[Y \mid X, Z] \mid X\} = \int g^{-1}(\beta_0 + \beta X + \gamma z) dF_Z(z),$$

where F_Z is the distribution function of Z.
(ii) If Z is omitted in the multiple regression, regardless of the distribution of Z, they assume that there exist constants β_0^* and β^* such that the following equation holds:

$$\int g^{-1}(\beta_0 + \beta X + \gamma z) dF_Z(z) = g^{-1}(\beta_0^* + \beta^* X). \tag{16}$$

(iii) They prove that $0 < |\beta^*| < |\beta|$.
(iv) They assume that the estimator $\hat{\beta}_n^*$ of the coefficient of X in the univariate regression using only the data of (Y, X) will converge to β^*. Combining this with the result in (iii), they show that the estimate of β is biased if Z is omitted in the multiple regression analysis.

Let us assess the validity of the argument presented in Neuhaus and Jewell (1993). Their work evidently hinges on the assumption that if (15) represents the regression of Y on X and Z, then the regression of Y on X should adhere to the format on the right-hand side of (16) for any distribution of Z, provided Z and X are independent. If (15) and (16) are inconsistent, as discussed in section 2, then their argument loses validity.

We use an example to demonstrate that (15) and (16) are inconsistent. Assume $g^{-1}(u) = \frac{e^u}{1+e^u}$. Let us consider a scenario where Y is a 0-1-valued random variable, X and Z are independent, X follows a standard normal distribution, and Z is uniformly distributed on the interval $[0, 1]$. Assume that

$$\mathrm{E}[Y \mid X, Z] = \frac{e^{X+Z}}{1 + e^{X+Z}} = g^{-1}(X + Z).$$

Then

$$\mathrm{E}[Y \mid X] = \log \frac{1 + e^{X+1}}{1 + e^{X}}.$$

If (16) held, then we should have

$$\log \frac{1 + e^{x+1}}{1 + e^{x}} = \frac{\exp(\beta_0^* + \beta_1^* x)}{1 + \exp(\beta_0^* + \beta_1^* x)}. \tag{17}$$

By letting $x = -1$, 0, and 1, respectively, we prove that we cannot find β_0^* and β^* such that equation (17) holds.

Since the validity of equation (16) cannot be established, the hypothesized value of β^* is questionable. Therefore, their analysis of bias is not applicable, as comparing β with a non-existent quantity β^* lacks meaningful interpretation.

Despite this flaw, Neuhaus and Jewell's (1993) work has been widely cited in the statistical literature without the validity of their argument being questioned. For example, Drake and McQuarrie (1995), Wang and Louis (2003), and Wan, Small and Mitra (2018) have cited their work. However, it is essential for researchers to critically evaluate the assumptions and methods used in previous studies to ensure the validity of their conclusions.

7. Numerical example

In this section, we present a numerical example to demonstrate the impact of inconsistent specifications in logistic regression.

Let Y and Z be 0-1-valued random variables, and let X follow a standard normal distribution, where X and Z are independent and $P\{Z = 1\} = 1/2$. Suppose the conditional distribution of Y given (X, Z) is given by the logistic regression model

$$P\{Y = 0 \mid X, Z\} = \frac{1}{1 + \exp(0.2 + X + Z)}.$$

The conditional distribution of Y given X is

$$P\{Y = 0 \mid X\} = \frac{0.5}{1 + \exp(1.2 + X)} + \frac{0.5}{1 + \exp(0.2 + X)}.$$

Define the function

$$\mathrm{OR}(x) = \frac{P\{Y = 1 \mid X = x + 1\}/P\{Y = 0 \mid X = x + 1\}}{P\{Y = 1 \mid X = x\}/P\{Y = 0 \mid X = x\}}, \quad x \in \mathbb{R}.$$

Figure 1 shows the graph of $\mathrm{OR}(x)$. Since it is not a constant function, we cannot define a general marginal odds ratio for X. This also shows that the regression of Y on X does not take the form of a univariate logistic regression.

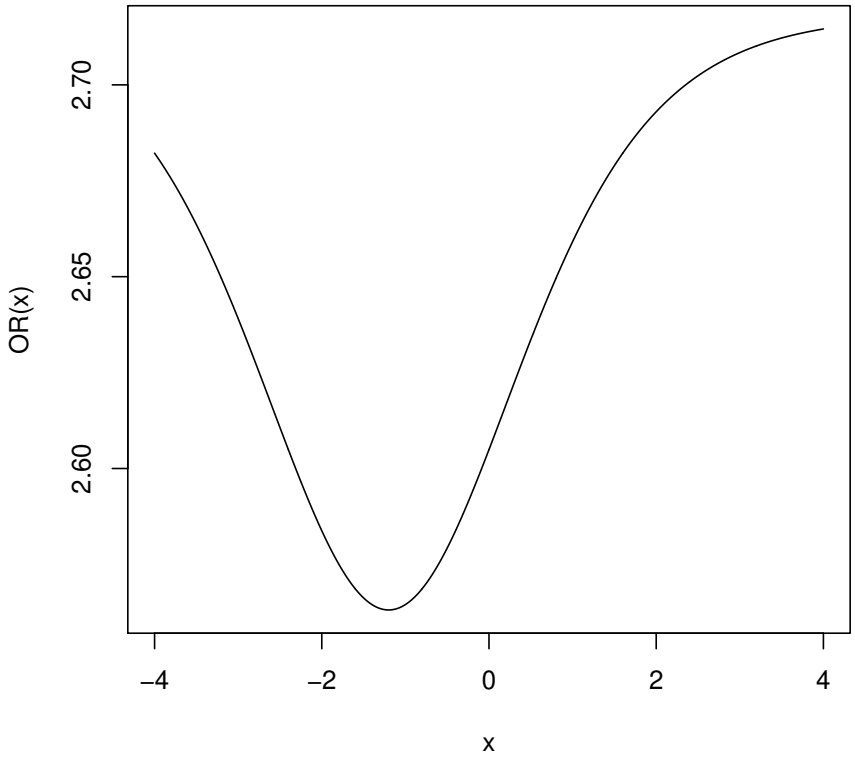


Figure 1. Plot of $\mathrm{OR}(x)$

Suppose (Y_i, X_i, Z_i), $i = 1, \ldots, n$, is a random sample from the joint distribution of (Y, X, Z). Let

$$p_i(\theta_0, \theta_1) = \frac{\exp(\theta_0 + \theta_1 X_i)}{1 + \exp(\theta_0 + \theta_1 X_i)}, \quad \theta_0, \theta_1 \in \mathbb{R}, \tag{18}$$

$$L(\theta_0, \theta_1) = \prod_{i=1}^{n} \left\{ p_i(\theta_0, \theta_1)^{Y_i} \left(1 - p_i(\theta_0, \theta_1)\right)^{1-Y_i} \right\}. \tag{19}$$

As shown in Theorem 1, the conditional distribution of Y_i given X_i cannot be in the form of (18), i.e., there do not exist constants θ_0 and θ_1 such that

$$P\{Y_i = 1 \mid X_i\} = p_i(\theta_0, \theta_1).$$

Therefore, $L(\theta_0, \theta_1)$ in (19) is not the likelihood function based on (Y_i, X_i), $i = 1, \ldots, n$.

We study the behavior of the maximizer of $L(\theta_0, \theta_1)$. Let $\ell(\theta_0, \theta_1) = \log L(\theta_0, \theta_1)$. Let $(\hat{\theta}_0, \hat{\theta}_1)$ be the solution of

$$\frac{1}{n}\frac{\partial \ell(\theta_0, \theta_1)}{\partial \theta_0} = 0, \quad \frac{1}{n}\frac{\partial \ell(\theta_0, \theta_1)}{\partial \theta_1} = 0,$$

and (θ_0^*, θ_1^*) be the solution of

$$\mathrm{E}[Y] = \mathrm{E}[p_1(\theta_0^*, \theta_1^*)], \quad \mathrm{E}[YX] = \mathrm{E}[p_1(\theta_0^*, \theta_1^*)X].$$

It is easy to prove that $\ell(\theta_0, \theta_1)$ is a concave function. Hence,

$$(\hat{\theta}_0, \hat{\theta}_1) = \mathrm{argmax}_{\theta_0, \theta_1} L(\theta_0, \theta_1).$$

Using the same method as in the Appendix of Andersen and Gill (1982), we can prove that $(\hat{\theta}_0, \hat{\theta}_1) \to (\theta_0^*, \theta_1^*)$ almost surely as $n \to \infty$. We obtained $\theta_1^* = 0.957$ by Monte Carlo simulation with a sample size of $n = 1,000,000$.

Remark 3. While $(\hat{\theta}_0, \hat{\theta}_1)$ represent the values that maximize the function $L(\theta_0, \theta_1)$, there is no clear interpretation for these estimated parameters. Similarly, (θ_0^*, θ_1^*) also lack an obvious interpretation. It is important to note that we cannot interpret $\exp(\theta^*)$ as the marginal odds ratio of X. See Matthews and Badi (2015) for further discussion.

8. Conclusion

In a GLM, the conditional expectation of the outcome is assumed to be a monotone function of a linear combination of covariates. However, if the set of covariates is changed, the conditional expectation of the outcome given the new list of covariates may change as well. Firstly, the new conditional expectation may depend on the covariates in a much more complicated way than as a linear combination of them. Secondly, even if it is still a function of the linear combination of the covariates, the new link function may differ from the original one.

Our research has shown that model inconsistency can arise in logistic regression even when covariates are independent, and the issue becomes more complex when covariates are correlated. This issue also applies to other link functions, except for the identity link. Many publications have

employed both univariate and multiple logistic regression models in their analyses without verifying the consistency of their models; Florez, Jaoude, Patel et al. (2023) is one such example.

One limitation of our study is that we have only examined the consistency of two nested regression models, which can be justified based on the properties of conditional expectation. However, if two regression models are not nested, it can be challenging to check their consistency, and research is ongoing in this area.

As Box and Luceco (1997, p. 6) famously stated, "all models are wrong, but some models are useful." Models are employed to approximate the true relationship between variables, but logically inconsistent regression models should not be specified on the same dataset. Interpreting the outcomes from inconsistent models is difficult, and currently, there is no universal theory to guide the specification of consistent models.

References

Andersen P., Gill R. (1982): Cox's Regression Model for Counting Processes: A Large Sample Study. The Annals of Statistics 10(4): 1100–1120.

Box G.E.P., Luceño A. (1997): Statistical Control: By Monitoring and Feedback Adjustment. New York: John Wiley.

Drake C., McQuarrie A. (1995): A Note on the Bias due to Omitted Confounders. Biometrika 82(3): 633–638.

Dudley R.M. (2002): Real Analysis and Probability. New York: Cambridge University Press.

Feller W. (1966): An Introduction to Probability Theory and Its Applications (Vol. II). New York: Wiley.

Florez M.A., Jaoude J.A., Patel R.R. et al. (2023): Incidence of Primary End Point Changes Among Active Cancer Phase 3 Randomized Clinical Trials. JAMA Netw Open 6(5): e2313819.

Geng E., Kreiswirth B., Burzynski J., Schluger N.W. (2005): Clinical and radiographic correlates of primary and reactivation tuberculosis: a molecular epidemiology study. JAMA 293(22): 2740–2745.

Hastie, T., Tibshirani, R., Friedman, J. (2009): The Elements of Statistical Learning (2nd Ed.). New York: Springer.

Hole J., Hirsch M., Ball E., Meads C. (2015): Music as an aid for postoperative recovery in adults: a systematic review and meta-analysis. Lancet 386: 1659–1671.

Leon M.B., Smith C.R., Mack M.J. et al. (2016): Transcatheter or Surgical Aortic-Valve Replacement in Intermediate-Risk Patients. New England Journal of Medicine 374(17): 1609–1620.

Matthew, J.N.S. Badi N.H. (2015): Inconsistent treatment estimates from misspecified logistic regression analyses of randomized trials. Statistics in Medicine 34: 2681–2694.

McCullagh P., Nelder J. R. (1989): Generalized Linear Models (2nd Ed.). Boca Raton, FL: Chapman & Hall/CRC.

Neuhaus J., Jewell N. (1993): A Geometric Approach to Assess Bias Due to Omitted Covariates in Generalized Linear Models. Biometrika 80(4): 807–815.

Rao C. R. (1973): Linear Statistical Inference and its Applications (2nd Ed). New York: Wiley.

Wan F., Small D., Mitra N. (2018): A general approach to evaluating the bias of 2-stage instrumental variable estimators. Statistics in Medicine 37: 1997–2015.

Wang Z., Louis T. (2003): Matching Conditional and Marginal Shapes in Binary Random Intercept Models Using a Bridge Distribution Function. Biometrika 90(4): 765–775.

Appendix

Proof of Lemma 1. From (7) we have

$$\exp(x) = \frac{1}{p\exp(-y) + (1-p)\exp(-z)}.$$

Substituting into (6) we have

$$\exp(y-z) + \exp(z-y) = 2$$

which is true if and only if $y = z$. Hence we have $x = y = z$. □

Proof of Lemma 2. Suppose there exist constants α_0 and α_1 so that (8) holds. Let

$$g(x) = \frac{p[1+\exp(\alpha_0+\alpha_1 x)]}{1+\exp(\beta_{01}+\beta_1 x)} + \frac{(1-p)[1+\exp(\alpha_0+\alpha_1 x)]}{1+\exp(\beta_{02}+\beta_2 x)}. \tag{20}$$

Then $g(x) = 1$ for all $x \in \mathbb{R}$ and hence $\alpha_1 \neq 0$.

Case 1: $\beta_1 = \beta_2$. From the hypothesis, $\beta_{01} \neq \beta_{02}$. We first prove that to make (8) true, we should have $\alpha_1 = \beta_1$. Suppose not. Consider the following situations:
(i) $\alpha_1 < \beta_1$. If $\beta_1 > 0$, then $\lim_{x\to\infty} g(x) = 0$. If $\beta_1 < 0$, then $\lim_{x\to-\infty} g(x) = \infty$.
(ii) $\alpha_1 > \beta_1$. If $\alpha_1 > 0$, then $\lim_{x\to\infty} g(x) = \infty$. If $\alpha_1 < 0$, then $\lim_{x\to-\infty} g(x) =$

0.
We obtain contradictions in both situations. Hence $\alpha_1 = \beta_1$.

Letting $x = 0$ in (20), we have

$$\frac{p[1+\exp(\alpha_0)]}{1+\exp(\beta_{01})} + \frac{(1-p)[1+\exp(\alpha_0)]}{1+\exp(\beta_{02})} = 1.$$

Letting $x \to \infty$ if $\beta_1 > 0$ and $x \to -\infty$ if $\beta_1 < 0$in (20), we have

$$p\exp(\alpha_0 - \beta_{01}) + (1-p)\exp(\alpha_0 - \beta_{02}) = 1.$$

From Lemma 1 we have $\alpha_0 = \beta_{01} = \beta_{02}$.

Case 2: $\beta_1 \neq \beta_2$. WLOG, assume $\beta_1 > \beta_2$. It is obvious that $\alpha_1 \neq \beta_1$ and $\alpha_1 \neq \beta_2$. Consider the following situations:
(i) $\alpha_1 > \beta_1$. If $\alpha_1 > 0$, then $\lim_{x\to\infty} g(x) = \infty$. If $\alpha_1 < 0$, then $\lim_{x\to-\infty} g(x) = 0$.
(ii) $\beta_2 < \alpha_1 < \beta_1$. If $\beta_1 < 0$, then $\lim_{x\to-\infty} g(x) = \infty$. If $\beta_2 < \alpha_1 < 0 < \beta_1$, then $\lim_{x\to\infty} g(x) = (1-p)$. If $\beta_2 < 0 < \alpha_1 < \beta_1$, then $\lim_{x\to\infty} g(x) = \infty$. If $\beta_2 > 0$, then $\lim_{x\to\infty} g(x) = \infty$.
(iii) $\alpha_1 < \beta_2$. If $\beta_2 > 0$, then $\lim_{x\to\infty} g(x) = 0$. If $\beta_1 < 0$, then $\lim_{x\to-\infty} g(x) = \infty$. If $\beta_2 < 0 < \beta_1$, then $\lim_{x\to\infty} g(x) = 1-p$. □