The Annals of Statistics
1989, Vol. 17, No. 3, 1009–1052

REGRESSION ANALYSIS UNDER LINK VIOLATION

BY KER-CHAU LI[1] AND NAIHUA DUAN[2]

University of California, Los Angeles and RAND Corporation

We study the behavior of regression analysis when there might be some violation of the assumed link function, the functional form of the model which relates the outcome variable y to the regressor variable $\mathbf{x}$ and the random error. We allow the true link function to be completely arbitrary, except that y depends on $\mathbf{x}$ only through a linear combination $\beta\mathbf{x}$. The slope vector β is identified only up to a multiplicative scalar. Under appropriate conditions, any maximum likelihood-type regression estimate is shown to be consistent for β up to a multiplicative scalar, even though the estimate might be based on a misspecified link function. The crucial conditions are (1) the estimate is based on minimizing a criterion function $L(\theta, y)$ which is convex in θ, where $\theta = a + b\mathbf{x}$, (2) the expected criterion function $E[L(a + b\mathbf{x}, y)]$ has a proper minimizer and (3) the regressor variable $\mathbf{x}$ is sampled randomly from a probability distribution such that $E(b\mathbf{x}|\beta\mathbf{x})$ is linear in $\beta\mathbf{x}$ for all linear combinations $b\mathbf{x}$. The least squares estimate, the GLM estimates and the M-estimates for robust regression are discussed in detail.

These estimates are asymptotically normal. With the assumption that the regressor variable has an elliptically symmetric distribution, we show that under a scale-invariant null hypothesis of the form H_0: $\beta W = 0$, the asymptotic covariance matrix for $\hat{\beta}W$ is proportional to the one derived by treating the assumed link function as being true. The Wald test as well as the likelihood ratio test for a scale-invariant null hypothesis has the correct asymptotic null distribution after an appropriate rescaling of the test statistic to account for the proportionality constant between the two asymptotic covariance matrices. For normally distributed $\mathbf{x}$, the rescaling factor for M-estimates is the same as the one used in robust regression, while the rescaling factor for GLM estimates is related to adjustment for overdispersion. Confidence sets can be constructed by inverting Wald's tests.

The impact of the violation of linear conditional expectation condition 3 is discussed. A new dimension is added to the regression diagnostics by exploring the elliptical symmetry of the design distribution.

A connection between this work and adaptive estimation is briefly discussed.

1. Introduction.

Regression analysis is probably the most widely used statistical method other than simple descriptive statistics such as means and frequency tables. Usually we assume a parametric model and then choose an

Received April 1987; revised October 1988.

[1]Part of this work was done while the author was visiting the Mathematics Research Center, University of Wisconsin, Madison. Part of the research was supported by NSF Grants MCS-83-01587 and DMS-86-02018.

[2]Part of this work was done while the author was visiting the Mathematics Research Center, University of Wisconsin, Madison. Part of the work was done at the RAND Corporation, supported in part by SIMS, EPA and RAND research funds. Part of the work was done while the author was visiting the Department of Mathematics, UCLA.

AMS 1980 *subject classifications.* 62J99, 62F35.

Key words and phrases. Adaptive estimation, robustness, link violation, Fisher consistency, exponential family, GLM, M-estimate, semiparametric models, elliptically symmetric design, overdispersion.

estimation method appropriate for this model. However, for empirical applications, the assumed model is unlikely to be exactly true and the specification of the model can be somewhat arbitrary. A well-known example is the choice between the logistic and the probit regression. When the true model deviates from the assumed model, the regression analysis based on the assumed model might be inappropriate.

There has been a good deal of research on the behavior of regression analysis under deviations from the assumed model. Quite often, the assumed model (or the ideal model) is a linear one, $y = \alpha + \beta\mathbf{x} + \varepsilon$, with Gaussian error ε. Distribution robustness, as reviewed in Huber (1981), concerns the violation of the assumed error distribution (cf. Remark 1.2 at the end of this section). On the other hand, data may have to be transformed to follow a linear model and the correct transformation may be misspecified. Thus, for instance, the correct model may be

$$\log y = \alpha + \beta\mathbf{x} + \varepsilon,$$

but we may misspecify the transformation and assume

$$y^{1/2} = \alpha + \beta\mathbf{x} + \varepsilon.$$

Under the misspecified model, we might use the least squares regression of $y^{1/2}$ on $\mathbf{x}$ to estimate α and β. Does this apparently fallacious regression tell us anything? Under appropriate conditions on the regressor variable $\mathbf{x}$, the answer is yes; see Sections 2 and 5.

In the above example, we have misspecified the functional form of the model. More generally an assumed model can take a general form

$$y = g(\alpha + \beta\mathbf{x}, \varepsilon), \qquad \varepsilon \sim F(\varepsilon), \tag{1.1}$$

where g is a given bivariate function, the *link function*, which relates the outcome variable y to the regressor variable $\mathbf{x}$ and the stochastic error ε, and F is the error distribution.

In this paper we study the behavior of regression analysis when the assumed link function might be incorrect. We allow the true model to be completely arbitrary, except that the outcome variable y depends on the explanatory variable $\mathbf{x}$ only through a linear combination $\beta\mathbf{x}$. (See Remark 1.1 at the end of this section for more discussion.) The conditional distribution of y given $\beta\mathbf{x}$ is allowed to be completely arbitrary. This is equivalent to allowing g and F to be both arbitrary and unknown, which implies that β can be identified only up to a multiplicative scalar (see Observation 1 in Section 2).

The population case is studied in Section 2 where we establish a general result (Theorem 2.1) that any maximum likelihood-type regression estimate is Fisher consistent for the slope vector β up to a multiplicative scalar, even though the estimate might be based on a misspecified link function, provided that (1) the regression is based on minimizing a criterion function $L(\theta, y)$ which is convex in θ, with $\theta = a + \mathbf{bx}$, (2) the expected criterion function $E[L(a + b\mathbf{x}, y)]$ has a proper minimizer and (3) the regressor variable $\mathbf{x}$ is sampled randomly from a distribution such that the conditional expectation $E(b\mathbf{x}|\beta\mathbf{x})$ is linear in $\beta\mathbf{x}$ for any linear combination $b\mathbf{x}$.

This result indicates that many maximum likelihood-type regression estimates are "robust" in the sense that even when the assumed model is grossly misspecified, the result can still be meaningful. In the minimum, we can estimate the ratios β_j/β_k consistently; those ratios measure the substitutability of different components $\mathbf{x}_j$ and $\mathbf{x}_k$ of the regressor variable, and are the key quantities of interest in many studies.

Condition 1 is satisfied for many important estimation methods, including least squares, M-estimates with nondecreasing influence functions and generalized linear model (GLM) estimates with canonical link: The linear model is specified for the natural parameter (see also Section 3.5). The convexity property of the criterion is crucial here. Without the convexity, we may have inconsistency (see Section 2.4). On the other hand, sometimes the regression may be based on a more complicated type of criterion so our general result does not apply immediately. Such cases may be studied on an individual basis. We demonstrate one important case, namely the Cox regression estimate, which turns out to be consistent as well again, due to some convexity property of the criterion (see Section 2.5). Condition 2 is usually satisfied, but not always; see Sections 3 and 4. Condition 3 looks rather restrictive. It is satisfied when the regressor variable is normally distributed or is elliptically symmetric. The impact of violations of condition 3 is studied in Section 6. It is interesting to observe that Stein's necessary condition for adaptive estimation, simplified by Bickel (1982), holds under condition 3; see Section 7, where adaptive estimation is briefly discussed.

Condition 3 has important implications in data collection and analysis. At the design stage when the levels of $\mathbf{x}$ can be chosen by the statistician, elliptically symmetric designs are favorable from the viewpoint of providing protection against link violations according to Theorem 2.1. On the other hand, if the data have already been collected and the distribution of $\mathbf{x}$ is not close to being elliptically symmetric, we may still conduct meaningful regression analysis on those subsamples of the data with the $\mathbf{x}$ distribution being closer to the elliptic symmetry. This is particularly attractive at the exploratory stage of data analysis; specific proposals to implement this are still under investigation. A simulation study is conducted to illustrate the role of elliptic symmetry in Section 6.4. There we see that a new dimension is added to the existing regression diagnostics by exploring this design condition. Bias bound and other asymptotic aspects are discussed in Sections 6.1–6.2.

In Sections 3 and 4 we will discuss the GLM estimates and the M-estimates in detail. Under appropriate regularity conditions and some a priori verifiable conditions, we show that the existence condition (A.2) in Theorem 2.1 is valid for these estimates. However, detailed study on the likelihood equation (Section 3.3) reveals some inherent dangers in applying GLM with the natural parameter space being restricted (for example, the gamma family).

Sampling behavior and inference are studied in Section 5. First, we establish strong consistency. Then we derive the asymptotic distribution for the regression estimates. Under the assumption (A.3′): the distribution of the regressor variable is elliptically symmetric and the asymptotic covariance matrix for the estimated

slope $\hat{\beta}$ can be written as the sum of two matrices, the first one being proportional to the one derived by treating the assumed link function as being true and the second one being proportional to $\beta'\beta$. For all inference problems about β that are identifiable, the second matrix can be neglected. The proportionality constant for the first matrix can be estimated consistently. It follows that for any scale-invariant null hypothesis H_0: $\beta W = 0$, the standard Wald and likelihood ratio tests based on the assumed link function have the correct asymptotic null distributions after being rescaled to account for the proportionality constant. Inother words, those procedures are robust in validity after the rescaling. Note that non-scale-invariant hypotheses such as H_0: $\beta W = 1$ are not identifiable because β is identified only up to a multiplicative scalar. We can also invert the Wald test to construct confidence sets; they have to be cone-shaped.

Under the stronger assumption (A.3″): The distribution of the regressor variable is normal and the above rescaling has interesting interpretations. For GLM estimates, the proportionality constant is analogous to the generalized X^2 usually used to adjust for overdispersion. For M-estimates, the rescaled asymptotic covariance matrix coincides with what is usually used for the linear model in robust regression. In other words, the inference for robust regression is robust in validity not only against distribution violations but also against link violations. For the least squares estimate, the rescaled asymptotic covariance matrix also coincides with the one based on the standard linear model. In other words, the standard linear model inference is robust in validity against link violations.

We will postpone the review of related literature [in particular, Brillinger (1977, 1983)] until Section 2.6, after introducing the necessary notations and terminologies. Technical proofs are given in the Appendix.

REMARK 1.1. The assumption about the relationship between y and $\mathbf{x}$ made in this paper can be called a general regression model with one component. It can be generalized to allow k components: The conditional distribution of y given $\mathbf{x}$ depends on $\mathbf{x}$ only through k linear combinations $\beta_1\mathbf{x}, \ldots, \beta_k\mathbf{x}$. The single component model can be viewed as a general form of additivity model. Multicomponent models allow nonadditivity.

REMARK 1.2. In contrast to distribution robustness, "model robustness" is usually used when the deterministic part of the linear model is incorrect. Thus the true model may take the form $y = \beta\mathbf{x} + g(\mathbf{x}) + \varepsilon$, where $g(\mathbf{x})$ is an unknown function incorporated into the assumed model $y = \beta\mathbf{x} + \varepsilon$ to allow for model violation. Box and Draper (1959) assumed that $g(\mathbf{x})$ can be parametrized by a linear model to account for higher order interactions or nonlinearity. See Kiefer (1973) and Galil and Kiefer (1977) for more discussion on this finite dimensional violation approach. On the other hand, infinite dimensional models for $g(\mathbf{x})$ are studied in Huber (1981), Marcus and Sacks (1977), Li (1982, 1984), Sacks and Ylvisaker (1984) and Speckman (1979). The model violations considered in these papers are different from the link violation that we consider in this paper.

2. Population case: Fisher consistency. We shall first describe the estimation methods considered in this paper. Then in Section 2.2, we discuss the role

of link violation in our robustness consideration. The main result of Fisher consistency for regression estimates is given in Section 2.3. Section 2.4 addresses the convexity condition required for the criterion functions used in deriving the regression estimates. Section 2.5 studies the Cox regression model and the partial likelihood estimate, which does not belong to the general type of regression estimates considered in the main result. Section 2.6 reviews related works.

2.1. *Estimation methods.* We consider maximum likelihood-type regression estimates based on a specified one-parameter family of probability distributions $\{K_\theta, \theta \in \Theta\}$ for y and a linear relationship between θ and $\mathbf{x}$,

$$\begin{aligned} y &\sim K_\theta(y), \\ \theta &= \alpha + \beta\mathbf{x}. \end{aligned} \tag{2.1}$$

We assume throughout this paper that the regressor variable $\mathbf{x}$ is sampled randomly from a nondegenerate probability distribution $Q(\mathbf{x})$ in R^p. Suppose K_θ has a density $k_\theta(y)$ with respect to an appropriate carrier measure $\nu(y)$. Then the maximum likelihood estimate of (α, β) is a solution of the following minimization problem:

$$\text{minimize } n^{-1} \sum_{i=1}^{n} L(a + b\mathbf{x}_i, y_i), \tag{2.2}$$

where

$$L(\theta, y) = -\log k_\theta(y). \tag{2.3}$$

The consistency of m.l.e. when the assumed model (2.1) is true can be shown under certain regularity conditions.

We shall consider the regression estimate based on minimizing (2.2) for any criterion $L(\theta, y)$ that is convex in θ.

An important special case which we shall study in detail is the class of estimates based on the generalized linear models [GLM; see, e.g., Nelder and Wedderburn (1972)] with canonical link: We assume that $\{K_\theta\}$ is a natural exponential family (NEF). The criterion function (which will be called the NEF criterion) can be written as

$$L(\theta, y) = -y\theta + \psi(\theta), \tag{2.4}$$

where $\psi(\theta)$, the cumulant generating function for y, is strictly convex. The estimate $(\hat{\alpha}, \hat{\beta})$ based on this criterion function will be referred to as the GLM estimate. In Sections 3 and 5, we shall study the behavior of the GLM estimate when the true model may deviate from the assumed GLM. Note that the squared error criterion

$$L(\theta, y) = -y\theta + \theta^2/2 = (y - \theta)^2/2 - y^2/2 \tag{2.5}$$

is a special case of the NEF criterion.

It is not always necessary to formulate criterion functions via probability distributions. In particular, we will also consider location invariant criterion functions

$$L(\theta, y) = \rho(y - \theta) - \rho(y) \tag{2.6}$$

for some convex function ρ which might not correspond to a proper probability distribution. These criterion functions result in the M-estimates for robust regression, whose behavior under distribution violation has received extensive study; see, e.g., Huber (1981) or Portnoy (1985). We give results in Sections 4 and 5 for the behavior of M-estimates under link violation, which is a more general form of violation than distribution violations. Note that in (2.5) and (2.6), we have subtracted $\rho(y)$ from $\rho(y - \theta)$ to eliminate unnecessary moment conditions; see Section 4.

Of course some nonconvex criteria have also been used in many situations. An example is the Cauchy distribution for the location family. Nonconvex criteria can have undersirable numerical properties, such as multiple local minimizers. Furthermore, the asymptotic behavior for the resulting estimates may also be undesirable; see, e.g., Diaconis and Freedman (1982) for results in the robust location estimation problems and see Section 4 for more discussions.

In many situations the criterion function has nuisance parameters. For example, we might have a dispersion parameter

$$L(\theta, \sigma, y) = -\sigma^{-1} \log k_\theta(y),$$

where σ is an unknown scalar which does not depend on $\mathbf{x}$. The quasilikelihood functions of the above form are studied extensively in the GLM and related literature; see, e.g., Wedderburn (1974), McCullagh (1983), Nelder and Pregibon (1986) and Efron (1986). The nuisance parameters might not affect the ranking of the criterion in terms of θ. (The dispersion parameter above is an example.) In this case we can use any admissible values of the nuisance parameters to derive point estimates for α and β. However, the nuisance parameters might affect the Fisher information and need to be considered in making inference. We will discuss GLM with a dispersion parameter in Section 5.

2.2. *Link misspecification.* In empirical applications, it is rare for the specified model (2.1) to hold exactly. In this paper, we assume that the true model has the same form (2.1), but with a different one-parameter family:

$$\begin{aligned} y &\sim H_\theta(y), \\ \theta &= \alpha + \beta\mathbf{x}. \end{aligned} \tag{2.7}$$

The family $\{H_\theta\}$ is allowed to be arbitrary and unknown. (The specified family $\{K_\theta\}$ is usually our speculation about what $\{H_\theta\}$ should be.) This is equivalent to assuming that the conditional distribution of y given $\mathbf{x}$ depends only on $\beta\mathbf{x}$ and is arbitrary and unknown otherwise. We will refer to models of form (2.7) as the *general regression models*.

The class of general regression models is very rich, including transformation models [see, e.g., Box and Cox (1964) and Bickel and Doksum (1981)], Efron's (1982) general scaled transformation family (GSTF), dichotomous regression models, censored regression models, projection pursuit regression with one ridge component [Friedman and Stuetzle (1981)] and the generalized linear models (GLM).

Any given general regression model of form (2.7) can be expressed in the link function form (1.1) as

$$(2.8) \qquad y = g(\theta, \varepsilon) = H_\theta^{-1}(\varepsilon) = \inf\{\tilde{y}: \varepsilon \le H_\theta(\tilde{y})\}, \qquad \varepsilon \sim U(0,1).$$

Specifying a model of form (2.1) is equivalent to specifying a link function g and an error distribution F in (1.1). The specified model is subject to link violation: The true model has the same form (1.1), but with a different link function and/or a different error distribution. If the specified link function g is believed to be correct, while the specified error distribution F might be wrong, we have distributional violation.

The correspondence between models of form (2.7) and (1.1) is not unique: The same one-parameter family $\{K_\theta\}$ can correspond to different pairs (g, F) of link functions and error distributions. For models of form (1.1), we can always absorb the error distribution F into the link function using the inverse c.d.f.

$$g(\theta, \varepsilon) = g\big(\theta, F^{-1}(u)\big) = \tilde{g}(\theta, u), \qquad u \sim U(0,1).$$

Therefore we can always assume that the error distribution is uniform over $(0,1)$. It follows that we need only specify the link function $\tilde{g}$ and link violation is equivalent to the misspecification of the link function $\tilde{g}$. (Distributional violation can therefore be viewed as a special type of link violation.)

Under link violation, we allow both the link function g and the error distribution F to be arbitrary and unknown. In particular, g need not be monotonic or invertible and F need not be symmetric. The following is an important observation.

OBSERVATION 1. When the link function g is unspecified, the intercept α is not identified and the slope vector β is identified only up to a multiplicative scalar. (Any location-scale change in $\alpha + \beta\mathbf{x}$ can be absorbed into the link function.)

When g is unspecified, the best we can achieve is to estimate the direction of the slope vector β. (In other words, we can estimate the ratios β_j/β_k, but not the magnitudes of the components β_j.) For power transformation models, there have been some controversial views on the interpretation of slope vector β [see Hinkley and Runger (1984), with discussion]. However, the ratios β_j/β_k do have a simple interpretation: They measure the amount of treatment $\mathbf{x}_k$ required to match the effect of a unit of treatment $\mathbf{x}_j$. Duan (1986) studied the power transformation models in future detail.

REMARK 2.1. Discrete $Q(\mathbf{x})$ will not be considered in this section; the direction of β is not identifiable in this case.

2.3. *Fisher consistency.* We shall study the large sample behavior of regression estimates based on a specified model (2.1) when the true link function is unknown. Thus our observations $(y_i, \mathbf{x}_i)$ are i.i.d. with the conditional distribution of y given $\mathbf{x}$ determined by (2.7), or equivalently (1.1), and the marginal

distribution

$$\mathbf{x} \sim Q(\mathbf{x}). \tag{2.9}$$

By the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} L(a + b\mathbf{x}_i, y_i) \to EL(a + b\mathbf{x}, y) \quad \text{(a.s.)}$$

if the expectation is well-defined. Call the right-side term the expected criterion and write

$$R(a, b) = EL(a + b\mathbf{x}, y).$$

In this and the next two sections, we shall consider the minimization of the expected criterion. In other words, we shall first demonstrate the Fisher consistency property of the estimate based on L. Under appropriate regularity conditions Fisher consistency often implies strong consistency, which is studied further in Section 5.

In order to consider the minimization of the expected criterion, we need to consider the domain for (a, b). For convenience, we now restrict to a domain that is stricter than necessary. Later at the end of this section we shall discuss this domain problem in further detail.

Define $\Omega = \{(a, b)$: $R(a, b)$ is well-defined and is finite$\}$. Assume that

(A.0) Ω is a nonempty convex set in R^{p+1}.

THEOREM 2.1. *Under* (1.1) *and* (2.9), *the minimization problem*

$$\text{minimize } R(a, b) \text{ over } (a, b) \in \Omega \tag{2.10}$$

has a solution (α^*, β^*) *such that* β^* *is proportional to* β:

$$\beta^* = \gamma\beta \tag{2.11}$$

for some scalar γ, *provided that* (A.0) *and the following conditions hold*:

(A.1) *The criterion function* $L(\theta, y)$ *is convex in* θ *with probability* 1.

(A.2) *The conditional expectation* $E(b\mathbf{x}|\beta\mathbf{x})$ *exists and is linear in* $\beta\mathbf{x}$ *for all* $b \in \mathbf{R}^p$.

(A.3) *There exists a proper solution for* (2.10).

PROOF. Let a superscript over the expectation sign E denote conditioning in taking the expectation. Then by Jensen's inequality,

$$\begin{aligned} R(a, b) &= EE^{\beta\mathbf{x}, \varepsilon}L(a + b\mathbf{x}, g(\alpha + \beta\mathbf{x}, \varepsilon)) \geq EL(a + E^{\beta\mathbf{x}, \varepsilon}b\mathbf{x}, y) \\ &= EL(a + (c + d\beta\mathbf{x}), y) \end{aligned}$$

for some $c, d \in \mathbf{R}$. Here we have used (A.2) to obtain the last equality. Now the theorem follows immediately. □

If the inequality in the proof of this theorem can be replaced by a strict inequality for all b not proportional to β, then all minimizers of (2.10) must fall

along the direction of β. Thus any regression slope estimate based on minimizing the criterion function $L(\theta, y)$ is Fisher consistent for β up to a multiplicative scalar. This is the case, for example, when the criterion is strictly convex in θ. When the convexity of $L(\theta, y)$ is not strict, we need additional assumptions to reach the same conclusion. Nonstrict convexity will be studied further for the location invariant criteria in Section 4.

The existence condition (A.3) does not always hold. An example is in the application of logistic regression where we have a degenerate population with a perfect discriminant function $\beta\mathbf{x}$: $y \equiv 1$ if $\beta\mathbf{x} > -\alpha$ and $y \equiv 0$ if $\beta\mathbf{x} < -\alpha$.

A direct verification of the existence condition (A.3) may be complicated since the dimension of the minimization domain is high. A careful examination of the proof of Theorem 2.1 reveals that we may cut down the dimension to 2 by dealing only with the minimization problem given in the following condition which may replace condition (A.3) in Theorem 2.1:

(A.3′) There exists a proper solution to the minimization problem

$$\text{minimize } \tilde{R}(a, c) \text{ over } (a, c) \in \tilde{\Omega}, \tag{2.12}$$

where a and c are real numbers, $\tilde{R}(a, c) = R(a, c\beta)$ and $\tilde{\Omega} = \{(a, c): (a, c\beta) \in \Omega\}$.

The following lemma further reduces this two dimensional minimization problem to a one dimensional minimization problem, which is easier to verify and will be used in Sections 3 and 4.

LEMMA 2.1. *Assume that $\tilde{\Omega}$ is open and contains the origin. The following condition implies* (A.3′):

(A.3″) *For any $(a, c) \in \tilde{\Omega}$, the solution set for the minimization problem*

$$\text{minimize } \tilde{R}(at, ct) \text{ over } t \in \{t: (at, ct) \in \tilde{\Omega}\}, \tag{2.13}$$

is nonempty and is bounded away from the boundary of $\{t: (at, ct) \in \tilde{\Omega}\}$.

PROOF. First observe that $\tilde{R}(a, c)$ is a convex function of (a, c) over the open domain $\tilde{\Omega}$ and is therefore continuous. Suppose that (2.12) does not have a proper solution. We can take a sequence (a_n, c_n) such that $\tilde{R}(a_n, c_n)$ tends to $\inf\{\tilde{R}(a, c): (a, c) \in \tilde{\Omega}\}$, but $\{(a_n, c_n): n = 1, 2, \dots\}$ does not have an accumulation point in $\tilde{\Omega}$ (otherwise the lemma is proved). By compactness, we can find a subsequence, also denoted by (a_n, c_n) for convenience, such that the unit vector $a_n^2 + c_n^2)^{-1/2} \cdot (a_n, c_n)$ converges to some vector (a_0, c_0). Let T be the solution set for (2.13) with $a = a_0$, $c = c_0$. Since T is bounded, we may take three points $t_1 < t_0 < t_2$, such that $t_0 \in T$ and $t_1, t_2 \notin T$. By the continuity of $\tilde{R}(a, c)$, we can take two small open balls with centers at $(a_0 t_i, c_0 t_i)$, $i = 1, 2$, such that for any point (a, c) in each ball, $\tilde{R}(a, c) > \tilde{R}(a_0 t_0, c_0 t_0)$. Now consider the line segment connecting (a_n, c_n) to $(a_0 t_0, c_0 t_0)$. For large n, this line segment intersects one of the two open balls. Therefore by convexity, $\tilde{R}(a_n, c_n) > \tilde{R}(a_0 t_0, c_0 t_0)$. It follows that $(a_0 t_0, c_0 t_0)$ should be a minimizer for (2.12). □

The domain condition in this lemma is usually satisfied for many important estimation methods, including the M-estimates and the GLM estimates; see Sections 3 and 4. Furthermore, Lemma 2.1 does not depend on the fact that $\tilde{\Omega}$ is two dimensional, and can be generalized to higher dimensions with essentially the same proof. In particular, the existence condition (A.3) may also follow from (A.3″) with $\tilde{\Omega}$ replaced by Ω, if Ω is open and contains the origin.

The rest of this section will be devoted to the discussion of the domain condition (A.0). First we define the domains on which the expected criterion can be defined.

DEFINITION 2.1. The extended domain $\Omega_0 \subseteq R^{p+1}$ is the set of (a, b) for which $L(a + b\mathbf{x}, y)$ is defined with probability 1. The proper domain $\Omega_1 \subseteq \Omega_0$ is the set of (a, b) for which the expected criterion is well-defined, i.e., the positive and the negative parts of L cannot both have infinite expectations. The integrable domain $\Omega \subseteq \Omega_1$ is the set of (a, b) for which the expected criterion is finite [this is the domain considered in condition (A.0)].

We shall assume that $L(\cdot, y)$ is defined on an interval (may be unbounded). From this and (A.1) it follows that Ω_0 is convex. To avoid trivial cases we further assume that Ω_0 is indeed $p + 1$ dimensional, i.e., it is not contained in any affine subspaces with dimensionality less than $p + 1$. Otherwise, we can express $\alpha + \beta\mathbf{x}$ differently and thus reduce the dimensionality of $\mathbf{x}$.

LEMMA 2.2. *The expected criterion* (2.10) *is always well-defined, i.e.,* $\Omega_1 = \Omega_0$, *and cannot assume the value* $-\infty$, *provided that* (A.1) *and the following condition hold*:

(A.0′) *There exists at least one interior point in* Ω.

The proof of this lemma is given in the Appendix. It follows from the lemma that Theorem 2.1 can be extended.

COROLLARY 2.1. *Theorem* 2.1 *is valid with condition* (A.0) *replaced by* (A.0′) *and the integrable domain* Ω *in the minimization problem* (2.10) *replaced by the extended domain* Ω_0 (*see Definition* 2.1).

PROOF. Under (A.1), (A.0′) and the fact that $R(a, b)$ is a convex function, Ω is a convex set. Now the proof of Theorem 2.1 applies. □

2.4. *Nonconvexity and normality.* The conclusion of Theorem 2.1 may still be true without convexity condition (A.1), providing that (A.2) is replaced by the much stronger condition:

(A.2′) The regressor variable $\mathbf{x}$ is normally distributed.

THEOREM 2.2. *Under* (1.1), (2.9), (A.0), (A.2′) *and* (A.3), *the minimization problem* (2.10) *has a solution* (α^*, β^*) *such that* (2.11) *holds.*

PROOF. Because of (A.2′), for any b, we may write $b\mathbf{x} = t\beta\mathbf{x} + \varepsilon'$, with ε' independent of $\beta\mathbf{x}$ and ε. Then

$$\begin{aligned} R(a, b) &= EL(a + b\mathbf{x}, g(\alpha + \beta\mathbf{x}, \varepsilon)) \\ &= EE^{\varepsilon'}L(a + t\beta\mathbf{x} + \varepsilon', g(\alpha + \beta\mathbf{x}, \varepsilon)) \\ &\geq \min_{a} EL(a + t\beta\mathbf{x}, g(\alpha + \beta\mathbf{x}, \varepsilon)) \\ &\geq \min_{a,\gamma} EL(a + \gamma\beta\mathbf{x}, g(\alpha + \beta\mathbf{x}, \varepsilon)). \end{aligned}$$

This proves the theorem. □

It can be seen from the proof that condition (A.2′) can be replaced by the weaker condition:

(A.2″) For each $b \in R^p$, there exists some constant t such that $b\mathbf{x} - t\beta\mathbf{x}$ is independent of $\beta\mathbf{x}$.

The only case that (A.2″) will hold without knowing β is when we have (A.2′).

The following is an example where the conclusion of Theorem 2.1 is false due to the violation of (A.1) and (A.2′).

EXAMPLE 1. Suppose $p = 2$ and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)'$ follows a uniform distribution on the circle $x_1^2 + x_2^2 = 1$. Assume that $y = x_1$ (so $\varepsilon = 0$) and the criterion $L(\theta, y) \geq 0$ with the equality holding only for $\theta = \pm\sqrt{1 - y^2}$. Now it is clear that $EL(a + b\mathbf{x}, y) \geq 0$ with equality holding only for $a = 0$, $b = (0, \pm 1)$. Thus the minimizer of (2.10) is $a^* = 0$, $\beta^* = (0, \pm 1)$, but the true β is $(1, 0)$.

2.5. *Cox regression.* The conclusion of Theorem 2.1 may still be true for other types of regression. We illustrate this point by studying the case of Cox regression, a widely used model in survival analysis.

Suppose y is the survival time. Assume no censoring now (see Remark 2.2 for the censoring case). Cox (1972) considered the following model for y:

$$\lambda(y|\mathbf{x}) = \lambda_0(y)\exp(\beta\mathbf{x}), \tag{2.14}$$

where $\lambda(y|\mathbf{x})$ denotes the hazard function given $\mathbf{x}$ and $\lambda_0(y)$ is the baseline hazard. Cox proposed estimating β by maximizing the partial likelihood

$$L(b) = \prod_{i=1}^{n} \left\{ \exp(b\mathbf{x}_i) \Big/ \sum_{j \in R(y_i)} \exp(b\mathbf{x}_j) \right\}, \tag{2.15}$$

where $R(y_i)$ denotes the risk set at time y_i, namely, $R(y_i) = \{j\colon y_j \geq y_i\}$.

The population version of maximizing the logarithm of (2.15) is

$$\max_{b} Eb\mathbf{x} - E\{\log E^{y}\exp(b\tilde{\mathbf{x}})I(\tilde{y} \geq y)\}, \tag{2.16}$$

where $(\tilde{y}, \tilde{\mathbf{x}})$ denotes an independent replicate of $(y, \mathbf{x})$ and I is an indicator function (taking values 0 or 1, depending on $\tilde{y} < y$ or $\tilde{y} \geq y$).

THEOREM 2.3. *Under* (1.1), (2.9), (A.2) *and*
(A.3‴) *there exists a proper maximizer for* (2.16),
the maximization problem (2.16) *has a solution* β^* *satisfying* (2.11).

PROOF. This follows easily from the observations that for some c, d,

$$\begin{aligned} E^y \exp(b\tilde{\mathbf{x}}) I(\tilde{y} \geq y) &= E^y E^{y, \tilde{y}, \beta\tilde{\mathbf{x}}} \exp(b\tilde{\mathbf{x}}) I(\tilde{y} \geq y) \\ &\geq E^y \exp(E^{\beta\tilde{\mathbf{x}}} b\tilde{\mathbf{x}}) \cdot I(\tilde{y} \geq y) \\ &= E^y \exp(c + d\beta\tilde{\mathbf{x}}) \cdot I(\tilde{y} \geq y) \end{aligned}$$

and that

$$Ebx = E(c + d\beta\tilde{\mathbf{x}}).$$

□

REMARK 2.2. Partial likelihood estimates based on a specified model taking the same form as (2.14) but with the exponential function being replaced by some other function may not share the same Fisher consistency property we demonstrate in Theorem 2.3.

REMARK 2.3. When there is a censoring process involved, Tsiatis (1981) proved that the partial likelihood estimate is consistent providing that the censoring time is independent of the survival time given the covariate $\mathbf{x}$ and that the specified model (2.14) is true. However, we are unable to extend the result of Theorem 2.3 to this general case.

2.6. *Related work.* The proportionality result (2.11) in Theorem 2.1 has been known for various special cases. In this section, we give a brief review of related published works. The earliest related result that we know about is Fisher's (1936) work on the relationship between the discriminant function and logistic regression. Haggstrom (1983) gave a comprehensive discussion of the OLS estimation when the true model is the logistic regression model, a special case of the general regression model (2.1).

Brillinger (1977, 1983) gave a general result for the OLS estimates. Under the assumption that $\mathbf{x}$ is normally distributed, Brillinger showed that (1) the OLS slope vector $\hat{\beta}$ is strongly consistent for the true slope vector β up to a multiplicative scalar, when the true model has the additive-error form $y = g(\alpha + \beta\mathbf{x}) + \varepsilon$ and (2) $\sqrt{n}(\hat{\beta} - \beta^*)$ is asymptotically normal with mean 0 and covariance matrix

$$\sigma^2 \hat{\Sigma}^{-1} + \hat{\Sigma}^{-1} E\{h(\mathbf{x})^2 (\mathbf{x} - \mu)(\mathbf{x} - \mu)'\} \hat{\Sigma}^{-1}, \tag{2.17}$$

where $h(\mathbf{x}) = g(\alpha + \beta\mathbf{x}) - \alpha^* - \beta^*\mathbf{x}$, $\mathbf{x} \sim N(\mu, \hat{\Sigma})$ and $\sigma^2 = \mathrm{Var}(\varepsilon)$. Brillinger also noted that the key to (1) is that $\mathbf{x}$ has linear conditional expectations, and the strong consistency for OLS estimate holds under more general models such as the Cox regression model, censored regression, etc. Brillinger also made an interesting discussion on conditional inference (see Remark 6.2 for more details). In addition, similar results were shown to hold in some important time series problems.

Our interest in this area was motivated in part by surprise on learning of Brillinger's results. Theorems 2.1, 2.2, 2.3 and Theorem 5.1 extend Brillinger's result 1 to estimation methods other than OLS. We also extend Brillinger's result 2 in Section 5.3 [see (5.2.4)] and give a useful new expression for the asymptotic covariance matrix (Theorems 5.3.1 and 5.3.2) which relates to the usual asymptotic covariance matrix based on the assumption that the specified model is correct.

Goldberger (1981) derived the result (2.11) for a truncated linear model, assuming that the ideal data follow a linear model $y = \alpha + \mathbf{x}\beta + \varepsilon$, with both $\mathbf{x}$ and ε being normally distributed, but the datum $(y, \mathbf{x})$ is observable only if the dependent variable y falls inside a known subset Q of the real line. Chung and Goldberger (1984) generalized this result to a broader context in which the underlying model is not necessarily linear and the explicit selection rule is extended to allow either an arbitrary transformation (including censoring) of the dependent variable or a probabilistic selection rule. Without any assumptions of normality, Chung and Goldberger obtained (2.11) for the OLS estimates for the case of an arbitrary transformation under the assumption that $E(\mathbf{x}|y)$ is linear in y and for the case of probabilistic selection under the additional assumption that $\mathrm{Var}(\mathbf{x}|y)$ is constant.

Greene (1981, 1983) derived the same result for the OLS estimates for the Tobit model, the truncated regression model and the probit model, under the same normality assumption in Goldberger (1981), Ruud (1983) derived the same result for the maximum likelihood estimates in discrete choice models under the weaker assumption that $\mathbf{x}$ has linear conditional expectations. Ruud also argued that the failure to identify the absolute magnitude of the slope vector is unimportant and that "the ratios of the slopes yield the correct, relevant economic information about welfare trade-offs." Both Brillinger and Greene demonstrated by empirical studies that for some cases the proportionality result may still approximately hold under a modest violation of (A.2).

3. GLM estimates. Consider the NEF criterion (2.4). Since $\psi(\theta)$ is strictly convex in θ, condition (A.1) is satisfied; thus the solution of (2.10), if any, is unique. We need to verify the existence condition (A.3). The following moment condition implies that $\Omega_0 = \Omega$ [see Definition 2.1 and (A.0)]:

$$\text{(B.1)} \qquad \begin{aligned} &E|y| < \infty, \qquad E\|\mathbf{x}\| < \infty, \qquad E\|\mathbf{x}\, y\| < \infty, \\ &E|\psi(a + b\mathbf{x})| < \infty \quad \text{for all } (a, b) \in \Omega_0. \end{aligned}$$

3.1. *Unrestricted natural parameter space.* When the natural parameter space Θ is the whole real line, we have $\Omega = R^{p+1}$ and $\tilde{\Omega} = R^2$. Thus we can apply Lemma 2.1 to verify (A.3) using (A.3″).

LEMMA 3.1. *The existence condition* (A.3″) *holds for the NEF criterion* (2.4) *with an unrestricted natural parameter space* $\Theta = R$ *provided that* (B.1) *and the*

following condition hold:

(B.2) *With probability* 1, *the conditional expectation* $E(y|\beta\mathbf{x})$ *belongs to the set* $\{\psi'(\theta)\colon \theta \in \Theta\}$.

PROOF. For any (a, c), denote $U = a + c\beta\mathbf{x}$. Since $\tilde{R}(at, ct)$ is convex in t, it has a proper solution if the equation

$$\frac{\partial}{\partial t}\tilde{R}(at, ct) = 0$$

has a proper solution. Interchange the expectation and differentiation to get

$$EUy = E\psi'(tU)U.$$

Since the right-side term is nondecreasing in t, it suffices to show that

$$\lim_{t\to -\infty} E\psi'(tU)U < EUy < \lim_{t\to \infty} E\psi'(tU)U. \tag{3.1}$$

Denote the limits of $\psi'(t)$ as $t \to \pm\infty$ by $\psi'(\pm\infty)$ (they can be infinite). Clearly,

$$\lim_{t\to \infty} E\psi'(tU)U = \psi'(+\infty)EU_+ + \psi'(-\infty)EU_-,$$

where U_+ and U_- are the positive and the negative parts of U, respectively (i.e., $U_+ = U$ if $U > 0$ and $U_+ = 0$ if $U \le 0$; $U_- = U - U_+$). On the other hand,

$$\begin{aligned} EUy &= E\left[UE^U y\right] = E\left[U_+E^U y\right] + E\left[U_-E^U y\right] \\ &< EU_+\psi'(+\infty) + EU_-\psi'(-\infty), \end{aligned}$$

proving the second inequality in (3.1). In a similar way, the first inequality in (3.1) can also be verified. □

Applying the lemma to Theorem 2.1, we have

THEOREM 3.1. *The GLM estimate, based on a NEF criterion* (2.7) *with an unrestricted natural parameter space* $\Theta = R$ *is Fisher consistent in estimating the slope vector* β *up to a multiplicative scalar* [*i.e.*, (2.10) *has a unique solution and* (2.11) *is satisfied*] *under* (A.2), (B.1) *and* (B.2).

Note that $\psi'(\theta)$ is the expectation for the natural exponential family. Thus condition (B.2) requires that the true conditional expectation $E(y|\mathbf{x})$ be inside the range of the expectations specified by the assumed GLM. In empirical applications we usually have the prior knowledge about the range of the outcome variable; thus we can make an appropriate choice of GLM which satisfies this condition.

REMARK 3.1. The moment condition $E|y| < \infty$ and $E\|\mathbf{x}y\| < \infty$ in (B.1) is necessary. In particular, it is well-known that the least squares estimate is not consistent if $E|y| = \infty$ or $E\|\mathbf{x}y\| = \infty$. To accommodate the possibility $E|y| = \infty$ or $E\|\mathbf{x}y\| = \infty$, see the discussion on the M-estimate in next section.

REMARK 3.2. It can be seen from the proof of Lemma 3.1 that (A.3″) holds for any direction (a, b) [with $\tilde{R}(a, c)$ defined as $R(a, cb)$; b may be different from β] provided that (B.2) is replaced by the slightly stronger condition:

(B.2′) With probability 1, the conditional expectation $E(y|\mathbf{x})$ belongs to $\{\psi'(\theta): \theta \in \Theta\}$.

REMARK 3.3. When the range of t is an interval $(\underline{t}, \bar{t})$, the proof still holds provided that (3.1) holds with $-\infty, +\infty$ replaced by $\underline{t}, \bar{t}$.

3.2. *Restricted natural parameter space.* When the natural parameter space Θ is restricted, the domains Ω_0 [$= \Omega$ under (B.1)] and $\tilde{\Omega}$ are also restricted. Denote the lower and the upper bounds of Θ by $\underline{\theta}$ and $\bar{\theta}$, respectively, and assume at least one of them is finite. Then we have

$$\Omega_0 = \Omega = \{(a, b): \underline{\theta} < a + b\mathbf{x} < \bar{\theta} \text{ with probability } 1\} \tag{3.2.1}$$

and

$$\tilde{\Omega} = \{(a, c): \underline{\theta} < a + c\beta\mathbf{x} < \bar{\theta} \text{ with probability } 1\}. \tag{3.2.2}$$

Let $\underline{B}$ and $\bar{B}$ be the essential lower and upper bounds of $\beta\mathbf{x}$. If $\underline{B}$ and $\bar{B}$ are both infinite, the inequality in (3.2.2) cannot be satisfied unless $c = 0$, an uninteresting degenerated case. Thus we assume that at least one of the bounds is finite. In addition, we assume $\underline{B} < 0 < \bar{B}$ without loss of generality.

If $\beta\mathbf{x}$ has a point mass at each of the boundary points $\bar{B}$ and $\underline{B}$ ($\tilde{\Omega}$ contains no boundary points), we can apply Lemma 2.1. Lemma 3.1 can be shown to hold for this case by a similar proof (see Remark 3.3). Therefore the conclusion of Theorem 3.1, namely the Fisher consistency result, is also true. The result remains true when $\tilde{\Omega}$ contains some boundary points; the details are omitted.

3.3. *Likelihood equations.* The GLM estimate is usually obtained by solving the sample likelihood equations

$$n^{-1} \sum_{i=1}^{n} [\mathbf{x}_i y_i - \mathbf{x}_i \psi'(a + b\mathbf{x}_i)] = 0, \tag{3.3.1}$$

$$n^{-1} \sum_{i=1}^{n} [y_i - \psi'(a + b\mathbf{x}_i)] = 0. \tag{3.3.2}$$

The population version of (3.3.1) and (3.3.2) is

$$E\mathbf{x}y - E\mathbf{x}\psi'(a + b\mathbf{x}) = 0, \tag{3.3.3}$$

$$Ey - E\psi'(a + b\mathbf{x}) = 0, \tag{3.3.4}$$

which are based on the partial derivatives of the expected criterion $R(a, b)$ with respect to b and a. [It follows from (B.1) that these expectations exist.] In this section we study the relationship between the minimization problem (2.10) and the likelihood equations (3.3.3) and (3.3.4).

Throughout this section we assume conditions (B.1) and (B.2). In particular we have by Lemma 3.1 that the solution (α^*, β^*) to the minimization problem

(2.10) exists. If the domain Ω $(= \Omega_0)$ is open, we have by convexity of $R(\cdot, \cdot)$ that the population minimizer (α^*, β^*) is also the unique solution to the population likelihood equations (3.3.3) and (3.3.4). On the other hand, if the domain Ω is not open, the population minimizer might be a boundary point of Ω, in which case the likelihood equations (3.3.3) and (3.3.4) might not have a solution. We now discuss this problematic case in some detail.

In order for the domain Ω not to be open, it is necessary for the natural parameter space Θ to be restricted and the random variable $\beta \mathbf{x}$ not to have a probability mass at both essential bounds $\underline{B}$ and $\bar{B}$. To be more specific, we focus on the case that Θ is a half line, assumed to be $(-\infty, 0)$ without loss of generality. We also assume that $\beta \mathbf{x}$ does not have a probability mass at $\bar{B}$. The domain $\tilde{\Omega}$ given in (3.2.2) is then a cone with the vertex at $(0, 0)$ and it contains the edge $\{(a, c)\colon (c > 0,\ a + c\bar{B}) = 0\}$.

THEOREM 3.2. *Assume that* $(y, \mathbf{x})$ *follows the general regression model* (1.1) *and* (2.9) *with unknown link function* g *and unknown error distribution* F. *Assume that* $\beta \mathbf{x}$ *does not have a probability mass at its essential upper bound* $\bar{B}$. *For the GLM estimate based on a NEF criterion with the restricted natural parameter space* $\Theta = (-\infty, 0)$ *there exist* g *and* F *such that the minimization problem* (2.10) *has a solution but the population likelihood equations* (3.3.3) *and* (3.3.4) *do not have a solution, even though* (A.2), (B.1) *and* (B.2) *hold.*

The proof is given at the end of this section. We can characterize the conditions on g and F for the existence of a solution to the population likelihood equation (see Lemma 3.2), but unlike (B.2), these conditions cannot be verified a priori.

Although the population likelihood equations (3.3.3) and (3.3.4) might not have a solution, the sample likelihood equations (3.3.1) and (3.3.2) always have a solution, provided that all observed y_i's fall inside the range $\{\psi'(\theta) = \theta \in \Theta\}$. To see this, consider the minimization problem (2.10) with the random vector $(y, \mathbf{x})$ being uniformly distributed over the observed vectors $\{(y_i, \mathbf{x}_i)\colon i = 1, \ldots, n\}$. The extended domain Ω_n over which the sample criterion function

$$R_n(a, b) = \frac{1}{n} \sum_{i=1}^{n} L(a + b\mathbf{x}_i, y_i)$$

is defined is the intersection of n open sets,

$$\Omega_n = \bigcap_{i=1}^{n} \{(a, b)\colon \underline{\theta} < a + b\mathbf{x}_i < \bar{\theta}\}.$$

Therefore the domain Ω_n is open. Hence if the sample minimization problem has a solution, it must satisfy the sample likelihood equations (3.3.1) and (3.3.2). Following from Remarks 3.2 and 3.3 and the discussion that immediately follows the proof of Lemma 2.1, the sample minimization problem has a solution. [To see that Remark 3.3 applies, note that if $\bar{t}$ is finite, then for some i, $\bar{t}(a + cb\mathbf{x}_i) = \bar{\theta}$ or $\underline{\theta}$, depending on whether $a + cb\mathbf{x}_i$ is positive or negative. It follows that $\psi'(\underline{\theta})$ or $\psi'(\theta)$ is infinite; thus the second inequality in (3.1) holds. Details are omitted.]

The sample domain Ω_n defined above converges to Ω, although for each n, $\Omega \subset \Omega_n$. This means that the solution for (3.3.1) and (3.3.2) may occur in the thin strip $\Omega_n - \Omega$ if the population equations (3.3.3) and (3.3.4) do not have a solution. Hence the numerical solution to (3.3.1) and (3.3.2) is highly unstable. The curvature of the sample criterion function $R_n(a, b)$ is very large in the strip $\Omega_n - \Omega$.

Now we shall prove Theorem 3.2 by characterizing the conditions on g and F for the existence of the solution for (3.3.3) and (3.3.4). Write $A = \beta \mathbf{x}$. Assume that $(y, \mathbf{x})$ follows the general regression model (2.8) and (2.9). First note that by multiplying β to (3.3.3), we see that a necessary condition for (3.3.3) and (3.3.4) to have a solution along the direction β is to have

$$EyA = E\psi'(a + cA)A, \tag{3.3.5}$$

$$Ey = E\psi'(a + cA). \tag{3.3.6}$$

For a given marginal distribution $Q(\mathbf{x})$, whether a solution to the population likelihood equations (3.3.5) and (3.3.6) exists or not might depend on the true link function $g(\theta, \varepsilon)$ and the true error distribution $F(\varepsilon)$. We define

$$D_1 = \{(Ey, EyA): y \text{ follows (2.8) for some } g, F \text{ and satisfies (B.1), (B.2)}\},$$

$$D_2 = \{(E\psi'(a + cA), E\psi'(a + cA)A): (a, c) \in \tilde{\Omega}\}.$$

It is clear that D_2 is a subset of D_1. If D_1 and D_2 are the same, then for any true model of the form (2.8), the population likelihood equations have a solution. This would be the case, for example, if the domain Ω is open. However, if the domain is not open, D_2 might be a proper subset of D_1. The following lemma confirms this statement and therefore proves Theorem 3.2. For convenience, we assume $Ex = 0$ without loss of generality.

LEMMA 3.2. *Assume* (1.1), (B.1) *and* (B.2), $\Theta = (-\infty, 0)$, A $(= \beta \mathbf{x})$ *satisfies the assumption in Theorem* 3.2 *and* $EA = 0$. *Then the domains* D_1 *and* D_2 *can be characterized as follows*:

(i) $D_1 = \{(\eta, \zeta)$: $\eta > \psi'(-\infty)$, $\underline{B}(\eta - \psi'(-\infty)) < \zeta < \overline{B}(\eta - \psi'(-\infty))\}$ [*which equals* R^2 *if* $\psi'(-\infty) = -\infty$].

(ii) $D_2 = \{(\eta, \zeta)$: $\eta > \psi'(-\infty)$, $E\psi'(\underline{c}_\eta(A - \underline{B}))A \le \zeta \le E\psi'(\bar{c}_\eta(A - \overline{B}))A\}$, *where* $\underline{c}_\eta$ ($\bar{c}_\eta$, *respectively*) *is the solution of* c *such that* $E\psi'(c(A - \underline{B})) = \eta$ [*respectively*, $E\psi'(c(A - \overline{B})) = \eta$] *is satisfied.*

Moreover, D_2 *is a proper subset of* D_1.

The proof of the lemma is given in the Appendix. The following example illustrates the use of this lemma.

EXAMPLE 2. Consider the gamma family. The natural parameter space is $\Theta = (-\infty, 0)$ with the mean $\psi'(\theta) = -\theta^{-1}$ and $\psi'(-\infty) = 0$. Thus $\bar{c}_\eta = -\eta^{-1}E(A - \overline{B})^{-1}$ and $\underline{c}_\eta = -\eta^{-1}E(A - \underline{B})^{-1}$. A simple calculation shows that

$$D_2 = \left\{(\eta, \zeta): \eta > 0, \eta\left(\underline{B} + 1/E(A - \underline{B})^{-1}\right) < \zeta < \eta\left(\overline{B} + 1/E(A - \overline{B})^{-1}\right)\right\}.$$

Compared with

$$D_1 = \{(\eta, \zeta)\colon \eta > 0,\ \eta\underline{B} < \zeta < \eta\overline{B}\},$$

we see that D_2 is a proper subset of D_1. Hence if the true model has a mean-covariance pair (Ey, EyA) falling outside D_2, then the population likelihood equations (3.3.3) and (3.3.4) do not have a solution.

REMARK 3.4. Using conditional expectation argument, it is also easy to verify directly that the solution of (3.3.5) and (3.3.6) will yield a solution of (3.3.3) and (3.3.4). However, it would not be immediately clear whether or not the nonexistence of the solution for (3.3.5) and (3.3.6) implies the nonexistence of the solution for (3.3.3) and (3.3.4) unless we have shown that the solution for the minimization problem (2.10) must take the form (a, cA), which we have done in Theorem 3.1 and the discussion in Section 3.2.

3.4. *Noncanonical link.* We have restricted our discussion of GLM's to those with canonical link (the natural parameter θ is related linearly to the regressor $\mathbf{x}$), mainly because this results in a convex criterion function. Alternatively, one might specify a GLM with a noncanonical link: The natural parameter θ is related linearly to the regressor $\mathbf{x}$ after a nonlinear reparametrization $h(\cdot)$,

$$\theta' = \alpha + \beta\mathbf{x} = h(\theta). \tag{3.4.1}$$

The reparametrization is usually taken to be invertible. The criterion function is then

$$L(\theta', y) = -yh^{-1}(\theta') + \psi\big(h^{-1}(\theta')\big). \tag{3.4.2}$$

If the range of y is unbounded both from below and from above, the criterion function cannot be convex in θ' for all y. If the range is bounded from at least one end, the criterion function (3.4.2) may or may not be convex in θ', as illustrated in the following example.

EXAMPLE 3. A common reparametrization for the gamma family in Example 2 is

$$\theta' = \log(-\theta).$$

The criterion function for this parametrization is

$$L(\theta', y) = y\exp(\theta') - \theta',$$

which is strictly convex in θ' provided that $y > 0$, a condition which should hold for any reasonable application of the gamma family.

Alternatively, if we take the reparametrization

$$\theta'' = \theta^2,$$

the criterion function is

$$L(\theta'', y) = y\sqrt{\theta''} - (\log\theta'')/2,$$

which is concave in θ'' for $y^2 > 1/\theta''$.

4. *M*-estimates. The M-estimate, based on the minimization of the location invariant criterion (2.9), is usually proposed to guard against derivations from the assumed error distribution in the linear model $y = \alpha + \beta\mathbf{x} + \varepsilon$. The criterion ρ is usually chosen to be convex and to have a bounded influence function. Asymptotic results can be found in Huber (1981), Yohai and Marrona (1979), Cheng and Li (1984), Portnoy (1985), etc. For nonconvex ρ, the M-estimate can be inconsistent even for the location models [see Diaconis and Freedman (1982)].

We now study the behavior of the M-estimate under (2.1) which allows for deviations from the linear model both in the error distribution and the link function. We shall assume the following conditions:

(C.1) ρ is convex on R such that $\lim_{\mathbf{x} \to \pm\infty} \rho(\mathbf{x}) = +\infty$.

(C.2) The (one-sided) derivative ρ' of ρ satisfies the condition that there exist positive constants K_1 and K_2 such that for any θ, θ',

$$|\rho'(\theta) - \rho'(\theta')| \leq K_1(|\theta - \theta'| + K_2).$$

(C.3) $E\|\mathbf{x}\|^2 < \infty$ and $E|\rho'(y)|^2 < \infty$.

Condition (C.2) means that the tails of ρ do not go to the infinity faster than the squared error criterion (2.5). The conditions are commonly assumed in the robustness literature. It can be shown that (C.1)–(C.3) imply that Ω is the entire R^{p+1}; thus (A.0) is satisfied. It follows that $\tilde{\Omega} = R^2$, so Lemma 2.1 is applicable. We need to verify (A.3″).

LEMMA 4.1. *Conditions* (C.1)–(C.3) *imply condition* (A.3″).

PROOF. For any fixed (a, c), let $U = a + c\beta\mathbf{x}$. The one-sided right derivative of

$$\tilde{R}(ta, tc) = E[\rho(y - tU) - \rho(y)]$$

with respect to t can be written as

$$-E\rho'(y - tU)U_+ - E\rho'(y - tU)U_-, \tag{4.1}$$

where U_+ and U_- are the positive and negative parts of U, respectively ($U_+ + U_- = U$). Here ρ' can be treated as the left and the right derivatives for the first and the second terms, respectively. From (C.1), we may find positive constants a, a', M such that $\rho'(t) > a$ for $t > M$ and $\rho'(t) < -a'$ for $t < -M$. Thus applying the monotone convergence theorem we have

$$\lim_{t \to +\infty} -E\rho'(y - tU)U_+ \geq a'EU_+.$$

Similarly we can also show that

$$\lim_{t \to +\infty} -E\rho'(y - tU)U_- \geq -aEU_-.$$

Therefore the derivative (4.1) is positive for $t \to +\infty$. The same argument also shows that (4.1) is negative for $t \to -\infty$. This implies that $E[\rho(y - tU) - \rho(y)]$ has a minimizer, completing the proof of our lemma. □

Now applying Theorem 2.1, we see that at least one solution of (2.10) will be in the right direction. For a strictly convex ρ, due to the uniqueness of the solution, we may conclude that the corresponding M estimate is Fisher consistent for estimating the slope vector β up to a multiplicative scalar. For the case of nonstrict convexity, we have the following theorem.

THEOREM 4.1. *Assume conditions* (A.2), (C.1)–(C.3) *and the following conditions*:

(C.4) *For each* b *not proportional to* β, *with probability* 1, *the conditional distribution of* $b\mathbf{x}$, *given* $\beta\mathbf{x}$ *is nondegenerate.*

(C.5) *For any real number* d, *the support of the random variable* $y - d\beta\mathbf{x}$ *is an interval* (*could be infinite*) *with length larger than the length of the interval of the minimizers for* $\rho(\cdot)$.

Then, the minimizer for (2.10) *is unique and satisfies* (2.11).

PROOF. Suppose (a, b) is a minimizer of (2.10) such that b is not proportional to β. Then the inequality in the proof of Theorem 2.1 is an identity. By (C.4) and convexity, with probability 1, ρ is a straight line in some neighborhood of $y - (a + c) - d\beta\mathbf{x}$. Since the support of $y - (a + c) - d\beta\mathbf{x}$ is an interval we see that ρ is a straight line on this interval. On the other hand, by convexity the set of minimizers of ρ is also an interval, but with smaller length due to (C.5). Therefore these two intervals must be disjoint; ρ is a straight line on each interval. This is contradictory to the assumption that (a, b) is a minimizer because we can always shift (a, b) to some (a', b) so that the resulting interval for the support of $y - (a' + c) - d\beta\mathbf{x}$ is closer to the set of the minimizers for $\rho(\cdot)$ and hence reduces the $R(a, b)$. Therefore we have shown that any minimizer should be of the form $(a, c\beta)$. Now suppose there is more than one solution. Since the solution set must be convex, we can choose a solution not in the boundary, say $(a, c\beta)$ again, and conclude that ρ must be a straight line on the support of $y - a - c\beta\mathbf{x}$. The rest of the proof is straightforward and is omitted. □

REMARK 4.1. If (C.4) is replaced by the stronger condition that

(C.4′) for any b not proportional to β, the conditional distribution of $b\mathbf{x}$ given $\beta\mathbf{x}$ does not have a finite essential maximum or minimum,

then without (C.5) we can still prove that any minimizers for (2.10) are proportional to β.

REMARK 4.2. The following example shows that when both (C.4′) and (C.5) are violated, we may find some solution of (2.10) that does not fall on the correction direction. Take

$$\rho(\theta) = \begin{cases} 0, & \text{for } |\theta| \le 1, \\ |\theta| - 1, & \text{otherwise,} \end{cases}$$

$(x_1, x_2) \sim$ uniform on the unit ball $x_1^2 + x_2^2 \leq 1$ and $y = x_1$. Thus we may take $\alpha = 0$ and $\beta = (1, 0)$. But $a = 0$, $b = (0, 1)$ is also a solution of (2.10) since $R(a, b) = 0$ for this choice of (a, b). But if $x_1 \sim$ uniform on $(-1, 1)$ and $x_2 \sim N(0, 1)$, x_1, x_2 independent, then although there is more than one solution for (2.10), they all fall on the direction of β.

5. Sampling properties. In this section we study the asymptotic properties of the maximum likelihood-type estimates $(\hat{\alpha}, \hat{\beta})$ based on the sample minimization problem (2.2). First we establish the strong consistency and asymptotic normality of $(\hat{\alpha}, \hat{\beta})$. We then discuss how link violation affects the asymptotic covariance matrix. The results are applied to inference problems in Section 5.4.

5.1. *Strong consistency.* Fisher consistency usually implies strong consistency under suitable regularity conditions. A typical case is the maximum likelihood estimate for parametric models [see, e.g., Cramér (1946), Lehmann (1983) and Le Cam (1953)]. The results in Huber (1967), with applications to the robust estimation problems, might also be applicable in our case here [see, also, White (1981)]. However, instead of verifying or modifying Huber's conditions, it is easier to derive our results directly.

THEOREM 5.1. *Assume that* (A.1) *and the following additional conditions hold*:

(D.1) *The minimization problem* (2.10) *has a unique solution* (α^*, β^*).
(D.2) (α^*, β^*) *is an interior point of* Ω.

Then the set of estimates $(\hat{\alpha}, \hat{\beta})$ *which solves the sample minimization problem* (2.2) *converges almost surely to* (α^*, β^*).

PROOF. Since $R(a, b)$ and $L(a + b\mathbf{x}, y)$ are convex in (a, b), they are continuous in a neighborhood of (α^*, β^*). Let B be a closed hypercube contained in Ω with center (α^*, β^*) and width $2\gamma > 0$ on each side. Denote the sup-norm of a continuous function on B by $\| \cdot \|_B$. Using Mourier's (1953) theorem for the strong law of large numbers in Banach space, we have

$$\left\| \frac{1}{n} \sum_{i=1}^{n} L(a + b\mathbf{x}_i, y_i) - R(a, b) \right\|_B \to 0 \quad \text{a.s.}, \tag{5.1}$$

provided that

$$E\| L(a + b\mathbf{x}, y) - L(\alpha^* + \beta^*\mathbf{x}, y) \|_B < \infty. \tag{5.2}$$

The verification of (5.2) is given in the Appendix.

It follows from (5.1) that the set S_n of (a, b) that minimizes

$$n^{-1} \sum_{i=1}^{n} L(a + b\mathbf{x}_i, y_i)$$

over B must converge to (α^*, β^*) almost surely. Since (α^*, β^*) is an interior point of B, this means that with probability arbitrarily close to 1, the set S_n is in the

interior of B for sufficiently large n. Thus each point of S_n is a local minimizer; hence it is also a global minimizer due to the convexity condition (A.1). Again because of convexity, no other points outside B can be a global minimizer. This shows that the set of $(\hat{a}, \hat{\beta})$ defined in this theorem converges to (α^*, β^*) almost surely. □

This theorem can be applied to the GLM and the M-estimates, yielding the following results.

THEOREM 5.2. *Assume that* $(y, \mathbf{x})$ *follows the general regression model* (1.1) *and* (2.9). *The GLM estimate* $\hat{\beta}$ *based on the NEF criterion* (2.4) *with the natural parameter space* $\Theta = R$, *is strongly consistent for* β *up to a multiplicative scalar, under conditions* (A.2), (B.1) *and* (B.2).

THEOREM 5.3. *Assume that* $(y, \mathbf{x})$ *follows model* (1.1) *and* (2.9). *The M-estimate* $\hat{\beta}$, *based on the location-invariant criterion* (2.6), *is strongly consistent for* β *up to a multiplicative scalar, under conditions* (A.2) *and* (C.1)–(C.5).

REMARK 5.1. Theorem 5.2 still holds for the case that Θ is restricted, provided that (α^*, β^*) is in the interior of Ω. This would be true, e.g., if $\beta\mathbf{x}$ has a positive probability mass at both of its essential bounds.

5.2. *Asymptotic normality.* We assume that $L(\cdot, y)$ is smooth enough to allow the usual Taylor expansion derivation for asymptotic normality. Take

$$M = \begin{pmatrix} 1 & \mathbf{x}' \\ \mathbf{x} & \mathbf{x}\mathbf{x}' \end{pmatrix},$$

$$l_n(a, b) = \sum_{i=1}^{n} L(a + b\mathbf{x}_i, y_i),$$

$s_n(a, b)$ = the $(p + 1)$ column vector of all partial derivatives of $l_n(a, b)$,

$i_n(a, b)$ = the $(p + 1) \times (p + 1)$ matrix of second order partial derivatives of $l_n(a, b)$.

Clearly

$$s_n(a, b) = \sum_{i=1}^{n} L_1(a + b\mathbf{x}_i, y_i)(1\mathbf{x}_i)'$$

and

$$i_n(a, b) = \sum_{i=1}^{n} L_{11}(a + b\mathbf{x}_i, y_i)M_i,$$

where $L_1(\cdot, \cdot)$ is the partial derivative $\partial L(\theta, y)/\partial\theta$, $L_{11}(\cdot, \cdot)$ is $\partial^2 L(\theta, y)/\partial\theta^2$ and M_i is M with $\mathbf{x}$ replaced by $\mathbf{x}_i$.

We expand $s_n(a, b)$ as

$$s_n(a, b) = s_n(\alpha^*, \beta^*) + \left[\int_0^1 i_n(\alpha^* + \lambda(a - \alpha^*), \beta^* + \lambda(b - \beta^*))\, d\lambda\right](a - \alpha^*, b - \beta^*)'.$$

Since $s_n(\hat{\alpha}, \hat{\beta}) = 0$, we see that

$$-n^{-1/2}s_n(\alpha^*, \beta^*) = \left[\int_0^1 n^{-1} i_n\big(\alpha^* + \lambda(\hat{\alpha} - \alpha^*), \beta^* + \lambda(\hat{\beta} - \beta^*)\big)\, d\lambda\right] \times n^{1/2}(\hat{\alpha} - \alpha^*, \hat{\beta} - \beta^*)'. \tag{5.2.1}$$

By the central limit theorem,

$$n^{-1/2}s_n(\alpha^*, \beta^*) \to N(0, C),$$

where

$$C = EL_1(\alpha^* + \beta^*\mathbf{x}, y)^2 M. \tag{5.2.2}$$

On the other hand, the term inside the square brackets in (5.2.1) converges to

$$\Lambda = EL_{11}(\alpha^* + \beta^*\mathbf{x}, y)M. \tag{5.2.3}$$

Therefore, we have

$$\sqrt{n}\,(\hat{\alpha} - \alpha^*, \hat{\beta} - \beta^*) \to N(0, \Lambda^{-1}C\Lambda^{-1}). \tag{5.2.4}$$

The result (5.2.4) can be made rigorous under conditions (A.1), (D.1), (D.2) and the conditions:

(E.1) $L_{11}(\theta, y)$ exists and is continuous in θ with probability 1.

(E.2) $E\|L_{11}(a + b\mathbf{x})\|_B < \infty$ and $E\|L_{11}(a + b\mathbf{x}, y)\|_B \cdot \|\mathbf{x}\|^2 < \infty$ for some closed hypercube B in Ω with center (α^*, β^*).

Details are omitted.

5.3. *Asymptotic covariance.* The asymptotic covariance matrix for $\hat{\beta}$ takes a much simpler form under the assumption:

(A.2‴) The regressor variable $\mathbf{x}$ has an elliptically symmetric distribution with mean μ and a nonsingular covariance matrix. V.

THEOREM 5.3.1. *Assume* (5.2.4) *holds. Then under* (A.2‴), *the asymptotic covariance for* $\hat{\beta}$ (*i.e., the* $p \times p$ *submatrix obtained by deleting the first row and column vectors from* $\Lambda^{-1}C\Lambda^{-1}$) *has the form*

$$\lambda\eta V^{-1} + k\beta^{*\prime}\beta^*, \tag{5.3.1}$$

where λ, η, k *are scalars such that*

$$\lambda = \frac{EL_1(\alpha^* + \beta^*\mathbf{x}, y)^2\Gamma(\mathbf{x})}{EL_{11}(\alpha^* + \beta^*\mathbf{x}, y)\Gamma(\mathbf{x})}, \tag{5.3.2}$$

$$\eta = (p - 1)/\big(EL_{11}(\alpha^* + \beta^*\mathbf{x}, y)\Gamma(\mathbf{x})\big), \tag{5.3.3}$$

$$\Gamma(\mathbf{x}) = (\mathbf{x} - \mu)'V^{-1}(\mathbf{x} - \mu) - \big(\beta^*(\mathbf{x} - \mu)\big)^2/(\beta^*V\beta^{*\prime}). \tag{5.3.4}$$

In addition, if $\mathbf{x}$ *is normally distributed, then*

$$\lambda = \frac{EL_1(\alpha^* + \beta^*\mathbf{x}, y)^2}{EL_{11}(\alpha^* + \beta^*\mathbf{x}, y)}, \tag{5.3.5}$$

$$\eta = 1/EL_{11}(\alpha^* + \beta^*\mathbf{x}, y). \tag{5.3.6}$$

The proof of this theorem is based on the following lemma.

LEMMA 5.1. *Under* (A.2‴), *for any real-valued function* ψ, *we have*

$$E\psi(\beta\mathbf{x})(\mathbf{x} - \mu)' = c_1\beta V,$$

$$E\psi(\beta\mathbf{x})(\mathbf{x} - \mu)(\mathbf{x} - \mu)' = c_2 V + c_3 V\beta'\beta V,$$

where c_1, c_2, c_3 *are scalars.*

The proofs for Lemma 5.1 and Theorem 5.3.1 are given in the Appendix. In the following sections, we discuss the implication of this theorem.

5.3.1. *Maximum likelihood estimate.* It is well known that for a regular parametric setting, the asymptotic covariance matrix of the m.l.e. is the inverse of the Fisher information matrix, under the assumption that the parametric model is correct. But under link violation, the asymptotic covariance matrix takes the form (5.3.1). We now compare these two matrices to understand the effect of link violation on the asymptotic covariance.

COROLLARY 5.3.1. *Suppose the criterion function* $L(\theta, y)$ *in* (2.3) *is indeed the negative of the log-likelihood function for the true model. The asymptotic covariance matrix for the m.l.e.* $\hat{\beta}$ *based on* L *has the form*

$$\eta V^{-1} + k'\beta^{*\prime}\beta^*, \tag{5.3.7}$$

where η *is given by* (5.3.3) *and* k' *is another scalar, under conditions* (A.1), (A.2‴) *and the usual regularity conditions for asymptotic normality.*

PROOF. The Fisher information matrix for (α, β) is Λ. Therefore the asymptotic covariance for $(\hat{\alpha}, \hat{\beta})$ is Λ^{-1}, which takes the form of (5.3.7); see the proof of Theorem 5.3.1 in the Appendix. □

Now comparing (5.3.1) and (5.3.7), we have the following important theorem, which is useful for making inference about β.

THEOREM 5.3.2. *Under* (A.1), (A.2‴) *and the regularity conditions for asymptotic normality, the asymptotic covariance matrix of* $\hat{\beta}W$ *is changed under link violation only by the multiplicative scalar* λ *for any matrix* W *such that* $\beta W = 0$.

For GLM estimates with $\Theta = \mathbf{R}$, we have

$$L_1(\alpha^* + \beta^*\mathbf{x}, y) = -y + \psi'(\alpha^* + \beta^*\mathbf{x}) \equiv -\gamma(y, \mathbf{x})$$

and

$$L_{11}(\alpha^* + \beta^*\mathbf{x}, y) = \psi''(\alpha^* + \beta^*\mathbf{x}) \equiv \sigma^2(\mathbf{x}),$$

where $\gamma(y, \mathbf{x})$ and $\sigma^2(\mathbf{x})$ are the residual and variance operators for the specified GLM. Thus if $\mathbf{x}$ is normal, then the rescaling factor is simply

$$\lambda = E\gamma^2(y, \mathbf{x})/E\sigma^2(\mathbf{x}),$$

the ratio of true residual variance $E\gamma^2(y, \mathbf{x})$ to the model-based average variance $E\sigma^2(\mathbf{x})$. [Note that $E\gamma(y, \mathbf{x}) = 0$, as can be seen from the likelihood equation (3.3.4).]

For the least squares estimate, we have $\psi'' \equiv 1$; thus the rescaling factor is reduced to $E\gamma^2(y, \mathbf{x})$. Strictly speaking, (2.5) misspecifies the variance as exactly 1; therefore the covariance matrix Λ^{-1} based on the error distribution $\varepsilon \sim N(0, 1)$ needs to be rescaled by the residual variance $E\gamma^2(y, \mathbf{x})$. When the linear model $y = \mathbf{x}\beta + \varepsilon$ holds, we have $E\gamma^2(y, \mathbf{x}) = E\varepsilon^2$ as usual.

5.3.2. *Location-invariant criterion.* When the location-invariant criterion (2.6) is used, it is well-known in robust regression that the asymptotic covariance matrix for $\hat{\beta}$ is given by

$$\frac{E\rho'(y - \alpha - \beta\mathbf{x})^2}{(E\rho''(y - \alpha - \beta\mathbf{x}))^2} V^{-1}, \tag{5.3.8}$$

provided that the linear model

$$y = \alpha + \beta\mathbf{x} + \varepsilon, \qquad \varepsilon \sim F(\varepsilon), \tag{5.3.9}$$

holds [see, e.g., Huber (1981), Chapter 7]. In other words distribution violation changes the asymptotic covariance matrix by the multiplicative scalar

$$\lambda = \frac{E\rho'(y - \alpha - \beta\mathbf{x})^2}{E\rho''(y - \alpha - \beta\mathbf{x})}.$$

[If $-\rho$ is indeed the log-likelihood function for the true model, the asymptotic covariance matrix would be $V^{-1}/E\rho''(y - \alpha - \beta\mathbf{x})$.]

If we allow for link violation, so that the linear model assumption (5.3.9) might be false, the asymptotic covariance matrix should be given by (5.3.1). Since now we have

$$L_1(\alpha^* + \beta^*\mathbf{x}, y) = \rho'(y - \alpha^* - \beta^*\mathbf{x})$$

and

$$L_{11}(\alpha^* + \beta^*\mathbf{x}, y) = \rho''(y - \alpha^* - \beta^*\mathbf{x}),$$

it follows that the first term in (5.3.1) is the same as (5.3.8) (with α, β replaced by α^*, β^*) if $\mathbf{x}$ is normally distributed.

COROLLARY 5.3.2. *If* $\mathbf{x}$ *is normally distributed, the asymptotic covariance matrix for* $\hat{\beta}W$ *based on* (5.3.8) *for robust regression is robust in validity against link violation for any matrix* W *such that* $\beta W = 0$.

5.4. *Statistical inference.* In this section, we discuss how to modify standard parametric inference methods to accommodate possible link violations. In view of Observation 1, only the scale-invariant inference problems are of interest to us. For hypothesis testing problems, we consider scale-invariant null hypotheses of the form H_0: $\beta W = 0$. We consider Wald's test and the likelihood ratio test. We also invert Wald's test to construct confidence regions. We assume (A.2‴) throughout this section.

5.4.1. *Wald's test.* Consider the hypothesis testing problem with H_0: $\beta W = 0$ against H_1: $\beta W \neq 0$, where W is a $p \times r$ matrix with rank $r \leq p$. Under link violation, $\hat{\beta}$ converges to β^*, which is proportional to β. Hence under H_0, we have $\beta^* W = 0$. In view of Theorem 5.3.2, this implies that $\sqrt{n}\hat{\beta}W$ converges to $N(0, \lambda U)$, where

$$U = (0 \quad W')\Lambda^{-1}(0 \quad W')' \tag{5.4.1}$$

can be estimated consistently by $\hat{U}$, replacing Λ with a method of moments estimate $\hat{\Lambda}$. Therefore if we divide the usual Wald test statistic

$$\chi^2 = n(\hat{\beta}W)\hat{U}^{-1}(\hat{\beta}W)' \tag{5.4.2}$$

by a consistent estimate $\hat{\lambda}$ of λ, then we would have the valid null distribution for the test statistic. Under (A.2‴), we can use the method of moment estimate for λ based on (5.3.2). Under (A.2′), we can use the method of moment estimate for λ based on (5.3.5). In either case, the rescaled Wald test

$$\text{accept } H_0 \text{ if } \chi^2/\hat{\lambda} < \chi^2_r(1-\alpha) \tag{5.4.3}$$

is robust in validity against link violation.

For the GLM estimate, the rescaling in (5.4.3) can be used to protect against the misspecification of both the link function and the variance function. McCullagh (1983) used the generalized X^2,

$$\hat{\lambda}' = n^{-1}\sum_{i=1}^{n}\frac{\left(y_i - \psi'(\hat{\alpha} + \hat{\beta}\mathbf{x}_i)\right)^2}{\psi''(\hat{\alpha} + \hat{\beta}\mathbf{x}_i)},$$

to adjust for the dispersion parameter when the NEF log-likelihood function is correct only up to a multiplicative scalar. Note that the generalized X^2 is analogous to the rescaling factor $\hat{\lambda}$; for the NEF criterion, we have

$$\hat{\lambda} = \frac{(n-p-1)^{-1}\sum_{i=1}^{n}\left(y_i - \psi'(\hat{\alpha} + \hat{\beta}\mathbf{x}_i)\right)^2}{n^{-1}\sum_{i=1}^{n}\psi''(\hat{\alpha} + \hat{\beta}\mathbf{x}_i)}$$

if $\mathbf{x}$ has a normal distributions ($\hat{\lambda}$ can be interpreted as a ratio estimate, while $\hat{\lambda}'$ can be interpreted as a regression estimate). Therefore a minor modification of the generalized X^2 gives link robustness under the normality assumption (A.2′). If we have (A.2‴) instead of (A.2′), a different rescaling factor based on (5.3.2) would be required.

For the M-estimate, (5.4.3) can be used as a link-robust test. Huber (1981) suggested a distribution-robust test based on (5.3.8), which coincides with (5.4.3)

if we estimate λ and η using (5.3.5) and (5.3.6). Under the linear model (5.3.9), Huber's result does not depend on the normality assumption (A.2′). But if (A.2′) does hold, then the test is also robust under link violation. If we have (A.2‴) instead of (A.2′), the test has to be modified by a multiplicative scalar [i.e., using (5.3.2) and (5.3.3)] in order to be link robust.

REMARK 5.2. Motivated by the usual ANOVA for the linear model, we may consider the use of F-test instead of χ^2-test in (5.4.3) by replacing $\chi^2_r(1-\alpha)$ by $r \cdot F_{r,n-p-1}(1-\alpha)$.

5.4.2. *Confidence region.* We may invert the Wald test to construct confidence regions for β. Due to identifiability (Observation 1), we consider only cone-shaped confidence sets.

For any nonzero vector $\mathbf{v}$, consider testing $H_{\mathbf{v}}$: $\beta \propto \mathbf{v}$. The $(1-\alpha)\times 100\%$ confidence region for β can be constructed by finding the set of $\mathbf{v}$ such that $H_{\mathbf{v}}$ is accepted at α level. This can be viewed as the Scheffé method for constructing confidence sets for the direction of β. We now derive a simple expression for these confidence sets.

For any vector $\mathbf{e}$ with unitary length, let $\pi_{\mathbf{e}}$ be any $p \times (p-1)$ matrix such that $\pi_{\mathbf{e}}'\pi_{\mathbf{e}} = I_{p-1}$ and $\mathbf{e}\pi_{\mathbf{e}} = 0$. Consider the testing problem $H_{\mathbf{e}}$: $\beta V^{1/2}\pi_{\mathbf{e}} = 0$. This is equivalent to testing whether β is proportional to $V^{-1/2}\mathbf{e}$. Therefore the $(1-\alpha)\times 100\%$ confidence set for β based on inverting Wald's test is the cone spanned by

$$\left\{\mathbf{e}V^{-1/2}\colon \|\mathbf{e}\| = 1 \text{ and } n\hat{\lambda}^{-1}\hat{\eta}^{-1}\hat{\beta}V^{1/2}\pi_{\mathbf{e}}\pi_{\mathbf{e}}'V^{1/2}\hat{\beta}' \le \chi^2_{p-1}(1-\alpha)\right\},$$

because the U in (5.4.1) with $W = V^{1/2}\pi_{\mathbf{e}}$ equals η^{-1} times the identity matrix under $H_{\mathbf{e}}$. Since $\pi_{\mathbf{e}}$ is a projection, we see that

$$\left\{\mathbf{e}V^{-1/2}\colon \|\mathbf{e}\| = 1 \text{ and } n\hat{\lambda}^{-1}\hat{\eta}^{-1}\left(\hat{\beta}V\hat{\beta}' - \left(\mathbf{e}V^{1/2}\hat{\beta}'\right)^2\right) \le \chi^2_{p-1}(1-\alpha)\right\}.$$

Replace V by $\hat{V}$ and rearrange the inequality. Then we have the following $(1-\alpha)\times 100\%$ confidence region for β:

$$\left\{\beta\colon \left(\beta\hat{V}\hat{\beta}'\right)^2/\beta\hat{V}\beta' \ge \hat{\beta}\hat{V}\hat{\beta}' - n^{-1}\hat{\lambda}\hat{\eta}\chi^2_{p-1}(1-\alpha)\right\}. \tag{5.4.4}$$

In terms of the geometry based on the inner produce $\langle \mathbf{v}, w\rangle = \mathbf{v}\hat{V}w'$, the confidence set given by (5.4.4) can be interpreted as the cone consisting of the vectors having an angle with $\hat{\beta}$ of no more than

$$\sin^{-1}\left(\sqrt{\chi^2_{p-1}(1-\alpha)\hat{\lambda}\hat{\eta}/n\hat{\beta}\hat{V}\hat{\beta}'}\right).$$

Note that when $(\hat{\lambda}\hat{\eta})^{-1}\hat{\beta}\hat{V}\hat{\beta}'$ is small, the confidence set might be the entire R^p.

The same technique can also be applied to construct confidence intervals for ratios β_j/β_k, based on inverting the Wald tests for H_c: $c\beta_k - \beta_j = 0$. Tukey type confidence sets for $(\beta_2/\beta_1, \ldots, \beta_p/\beta_1)$ can also be obtained.

5.4.3. *Likelihood ratio test.* In addition to the Wald test given in Section 5.4.1, we may also consider the likelihood ratio test based on twice the difference

between the maximized criteria under H_1 and H_0, where H_0 and H_1 are scale-invariant hypotheses with $H_0 \subset H_1$. For simplicity, we assume that H_1 is the unrestricted hypothesis H_1: $\beta \in R^p$. Suppose H_0: $\beta = \mathbf{v}A$ for some $\mathbf{v} \in R^h$, where A is an $h \times p$ matrix with rank h, $h < p$.

The likelihood ratio test is then based on

$$Q = 2\sum_{i=1}^{n} L(\hat{a} + \hat{\mathbf{v}}A\mathbf{x}_i, y_i) - 2\sum_{i=1}^{n} L(\hat{\alpha} + \hat{\boldsymbol{\beta}}\mathbf{x}_i, y_i),$$

where $(\hat{a}, \hat{\mathbf{v}})$ denotes the estimate for $(a, \mathbf{v})$ by minimizing $n^{-1}\sum_{i=1}^{n} L(a + \mathbf{v}A\mathbf{x}_i, y_i)$ over $a \in R$ and $\mathbf{v} \in R^h$. The following theorem shows that Q can be rescaled to give the asymptotic χ^2 test under link violation.

THEOREM 5.4.1. *Under link violation, the rescaled likelihood ratio* $Q/\hat{\lambda}$ *converges to the* χ^2 *distribution with* $p - h$ *degrees of freedom under* H_0, *provided that the conditions* (A.1), (A.2‴), (D.1), (D.2), (E.1) *and* (E.2) *hold.*

The proof is given in the Appendix.

REMARK 5.3. We may use F with $p - h$, $n - p - 1$ degrees of freedom instead of χ^2_{p-h} to determine the significance level.

6. Design condition. The most restrictive assumption we have made in this paper is the design condition (A.2). When (A.2) is violated, the consistency result may be invalid. This raises at least three important issues:

1. How serious is the inconsistency if (A.2) is only slightly violated?
2. How to empirically "verify" (A.2) to the extent that the resulting bias would not be serious?
3. How to reduce the bias when it is necessary?

These issues are discussed in Sections 6.1–6.3. Generally speaking, the vulnerability to link violation increases as the design distribution departs further away from elliptic symmetry. To help the illustration of this phenomenon, a global measure of elliptic asymmetry (EASY) based on a crucial concept, the ICE curve, is introduced. A practical implication from our discussion is that when conducting a regression analysis, it is worthwhile to take a closer look at the distribution of the explanatory variable to make sure it does not bluntly deviate from elliptic symmetry (cf. Remark 6.4). This aspect of design robustness may have escaped most statisticians' attention. Our point is further illustrated by a simulation study which is reported in Section 6.4.

6.1. *Seriousness of inconsistency.* We denote any distribution satisfying (A.2) by Q_0. To emphasize the dependence of the minimizer of (2.10) on Q, the distribution of $\mathbf{x}$, we shall write $\beta^*(Q)$ and $\alpha^*(Q)$ for β^* and α^*, respectively. We assume the condition (D.1) in Section 5 for simplicity. Theorem 2.1 implies $\beta^*(Q_0) \propto \beta$.

The issue of inconsistency is discussed in two phases: (i) Examine the continuity property of the function $\beta^*(\cdot)$ at the point Q_0 and (ii) bound the bias with a measure of elliptic asymmetry for Q.

6.1.1. *Continuity.* To simplify the discussion, we shall concentrate on the class of Q's with support in a bounded Borel set B in R^p. The case of unbounded support will be briefly discussed at the end.

Assume the general regression model (2.8). Define the bivariate function $\mathscr{L}(\theta_1,\theta_2) = EL(\theta_1, g(\theta_2, \varepsilon))$ and let $\Theta \subset R^2$ be the domain of this function. Observe that $E_Q\mathscr{L}(a + b\mathbf{x}, \alpha + \beta\mathbf{x}) = R(a, b)$. Assume the regularity condition

$$\mathscr{L}(\theta_1, \theta_2) \text{ is continuous in } \Theta. \tag{6.1.1}$$

THEOREM 6.1. *Under conditions* (A.1), (6.1.1) *and* (D.2) *in Theorem* 5.1 *for* Q_0, $\beta^*(\cdot)$ *is weakly continuous at* Q_0 *with respect to the class of distributions with support in* B.

PROOF. Due to (D.2), it is possible to take a small open ball B_0 in R^{p+1} with center $(\alpha^*(Q_0), \beta^*(Q_0))$ such that the closure Θ_0 of the set $\{(a + b\mathbf{x}, \alpha + \beta\mathbf{x})$: $(a, b) \in B_0$ and $x \in$ support of $Q_0\}$ is in Θ. By convexity of $\mathscr{L}(\theta_1, \theta_2)$ in θ_1, it suffices to show that for any sequence Q_n that converges weakly to Q_0,

$$\sup_{(a, b)\in B_0} \left|E_{Q_n}\mathscr{L}(a + b\mathbf{x}, \alpha + \beta\mathbf{x}) - E_{Q_0}\mathscr{L}(a + b\mathbf{x}, \alpha + \beta\mathbf{x})\right| \to 0. \tag{6.1.2}$$

Now suppose (6.1.2) is false. We may find a subsequence (a_n, b_n) converging to some point, say (a_0, b_0) such that

$$\left|E_{Q_n}\mathscr{L}(a_n + b_n\mathbf{x}, \alpha + \beta\mathbf{x}) - E_{Q_0}\mathscr{L}(a_n + b_n\mathbf{x}, \alpha + \beta\mathbf{x})\right| \tag{6.1.3}$$

does not converge to 0.

Since Q_n converges to Q_0 weakly, by Slutsky's theorem the joint distribution of $(a_n + b_n\mathbf{x}, \alpha + \beta\mathbf{x})$ under Q_n converges to the distribution of $(a_0 + b_0\mathbf{x}, \alpha + \beta\mathbf{x})$ under Q_0. But $\mathscr{L}$ is continuous and bounded in Θ_0; a contradiction to (6.1.3) is obtained, proving the theorem. □

REMARK 6.1. For the case of unbounded support, the weak continuity result no longer holds. In order to have (6.1.2), we need some uniform integrability condition. In practical terms, one has to be more cautious when some design points are remote from the center.

6.1.2. *Bias bound.* In this section, a bound for the bias will be derived from the likelihood equation. We shall assume (D.1), (D.2) and that the function $L(\cdot, y)$ is smooth enough to allow the usual Taylor expansion. Thus we require that

$$\mathscr{L}(\theta_1, \theta_2) \text{ is twice differentiable in } \theta_1 \text{ for each } \theta_2.$$

Let $\mathscr{L}_1$ and $\mathscr{L}_{11}$ be the first and the second derivatives of $\mathscr{L}$ with respect to θ_1.

The likelihood equation

$$E_Q \mathscr{L}_1(a + b\mathbf{x}, \alpha + \beta\mathbf{x})(1 \quad \mathbf{x}') = 0$$

has a solution at $(\alpha^*(Q), \beta^*(Q))$, to be abbreviated as (α_Q^*, β_Q^*). Taylor expansion at $(\alpha^*(Q_0), \beta^*(Q_0))$, abbreviated as (α_0^*, β_0^*), leads to

$$\begin{aligned} &E_Q \mathscr{L}_1(\alpha_0^* + \beta_0^* \mathbf{x}, \alpha + \beta\mathbf{x})(1 \quad \mathbf{x}') \\ &\quad = -\left[\int_0^1 i_Q\big(\alpha_Q^* + \lambda(\alpha_Q^* - \alpha_0^*), \beta_Q^* + \lambda(\beta_Q^* - \beta_0^*)\big)\, d\lambda\right] \\ &\qquad \times (\alpha_Q^* - \alpha_0^*, \beta_Q^* - \beta_0^*)', \end{aligned} \tag{6.1.4}$$

where $i_Q(a, b)$ is defined to be $E_Q \mathscr{L}_{11}(a + b\mathbf{x}, \alpha + \beta\mathbf{x})M$ and M is defined in Section 5.2.

The convexity of $L(\cdot, y)$ implies that $\mathscr{L}_{11}$ is nonnegative. If we further assume that

$$\mathscr{L}_{11} \geq c_0 \text{ for some positive constant } c_0, \tag{6.1.5}$$

then the matrix inside the brackets in (6.1.4) is bounded away from 0 by $c_0 E_Q M$. Therefore a bound for the Euclidean norm of the bias is

$$\begin{aligned} &\left\| (\alpha_Q^* - \alpha_0^*, \beta_Q^* - \beta_0^*) \right\| \\ &\quad \leq \left\| c_0^{-1} (E_Q M)^{-1} E_Q \mathscr{L}_1(\alpha_0^* + \beta_0^* \mathbf{x}, \alpha + \beta\mathbf{x})(1 \quad \mathbf{x}') \right\|. \end{aligned} \tag{6.1.6}$$

When the minimum eigenvalue of the design matrix $E_Q M$, denoted by $\lambda_{\min}(Q)$, is bounded away from 0, the most crucial factor in this bound is the length of the vector $E_Q \mathscr{L}_1(\alpha_0^* + \beta_0^* \mathbf{x}, \alpha + \beta\mathbf{x})(1 \quad \mathbf{x}')$. The next lemma helps interpret this vector. We need the notation

$$Q_\beta(\cdot) = \text{the cumulative distribution of } \beta\mathbf{x} \text{ under } Q,$$

$$H_{Q,\beta}(t) = \int_{\beta\mathbf{x} \leq t} \mathbf{x}'\, dQ(\mathbf{x}).$$

LEMMA 6.1. *The following identity holds*:

$$\begin{aligned} &E_Q \mathscr{L}_1(\alpha_0^* + \beta_0^* \mathbf{x}, \alpha + \beta\mathbf{x})(1 \quad \mathbf{x}') \\ &\quad = \int \mathscr{L}_1(\alpha_0^* + \beta_0^* \mathbf{x}, \alpha + \beta\mathbf{x}) \big(d(Q_\beta - Q_{0\beta})(\beta\mathbf{x}), d(H_{Q,\beta} - H_{Q_0,\beta})(\beta\mathbf{x})\big). \end{aligned}$$

PROOF. Observe that β_0^* is proportional to β and that the left side of the identity equals 0 for $Q = Q_0$. The rest of the proof is trivial. □

How does the departure of Q from Q_0 affect the bias bound? Lemma 6.1 together with (6.1.6) provides a clear picture. The integrand on the right side of the equality in Lemma 6.1 is seen to be fixed. Thus the main factor for the bias bound comes from two functions: (i) $Q_\beta(\cdot) - Q_{0\beta}(\cdot)$ and (ii) $H_{Q,\beta}(\cdot) - H_{Q_0,\beta}(\cdot)$. The first function is just the difference in the marginal cumulative distributions

along the true direction β. To interpret the second function, consider the curve of the conditional mean $E_Q(\mathbf{x}|\beta\mathbf{x} = t)$ denoted as $\xi(t; \beta, Q)$. Express $H_{Q,\beta}(t)$ as $\int_{-\infty}^{t}\xi(t'; \beta, Q)\, dQ_\beta(t')$. Thus the second term is the difference between the two *integrated conditional expectation* curves (ICE curves, for short) of $\mathbf{x}$ along the direction β.

Suppose we measure the size of the two functions discussed above by the supremum norm. We further take the supremum over any direction β. This leads to two metrics: (i) $d(Q, Q_0) = \sup_{t\in R,\beta\in R^p}|Q_\beta(t) - Q_{0\beta}(t)|$ and (ii) $\tilde{d}(Q, Q_0) = \sup_{t\in R,\beta\in R^p}\|H_{Q,\beta}(t) - H_{Q_0,\beta}(t)\|$. The first one, called the half space metric, is popular in the study of the Vapnik and Červonenkis type problems and has been considered in the robust and nonparametric statistics. The second metric is only defined for those distributions with finite means and may be called the half space linear metric to emphasize the linear term $\mathbf{x}$ in the integrand of the definition of $H_{Q,\beta}(\cdot)$.

For the case that Q_0 has bounded support, using integration by part for the right side of the identity in Lemma 6.1, we see that the bias bound is at most of the same magnitude as $d(Q, Q_0) + \tilde{d}(Q, Q_0)$.

THEOREM 6.2. *Assume* (6.1.1) *and* (D.2) *for* Q_0 *and that*

(6.1.7) $\qquad \mathscr{L}_{11}$ *and* $\mathscr{L}_{12}$ *is continuous in* Θ,

where $\mathscr{L}_{12}$ *is the derivative of* $\mathscr{L}_1$ *with respect to the second argument. Then for an elliptically symmetric* Q_0 *with bounded support, we have* $\|\beta_Q^* - \beta_0^*\| = O(d(Q, Q_0) + \tilde{d}(Q, Q_0))$ *as* $d(Q, Q_0) + \tilde{d}(Q, Q_0)$ *converges to* 0.

PROOF. By Theorem 6.1, for any sequence Q_n such that $d(Q_n, Q_0) + \tilde{d}(Q_n, Q_0)$ converges to 0, $\alpha_{Q_n}^*$ and $\beta_{Q_n}^*$ converge to α_0^* and β_0^*. The matrix inside the brackets of the Taylor expansion (6.1.4) converges to $E_{Q_0}\mathscr{L}_{11}(\alpha_0^* + \beta_0^*\mathbf{x}, \alpha + \beta\mathbf{x})M$, the Hessian matrix, which is positive definite. The rest of the argument is trivial. □

6.2. *Empiric view of the bias bound.* For a given data set, the empiric distribution of $\mathbf{x}$, denoted by $\hat{Q}_n$, is available. If we may find an elliptically symmetric distribution Q_0 with bounded support such that $d(\hat{Q}_n, Q_0)$ and $\tilde{d}(\hat{Q}_n, Q_0)$ are small (say, of order $n^{-1/2}$), then by Theorem 6.2 we see that the bias bound is also small (of order at most $n^{-1/2}$). Note that when $\hat{Q}_n$ is generated from Q_0, both the half space distance and the half space linear distance between Q_0 and $\hat{Q}_n$ are $O_p(n^{-1/2})$. Thus if we conduct a significance test for $Q = Q_0$ based on these distances, then the acceptance of the null hypothesis leads to a bias $\beta^*(\hat{Q}_n) - \beta_0^*$ of order no greater than $n^{-1/2}$.

REMARK 6.2. Brillinger (1982) discussed the difference between the conditional inference (conditional on the observed $\mathbf{x}_i$'s) and the unconditional inference for the least squares estimation. In order to obtain a bias bound of order smaller than $n^{-1/2}$, he considered the case that $\mathbf{x}_i$ are generated from a deterministic quasi-Gaussian sequence, based on Halton (1960). Adopting the

same idea, we may show that $\mathbf{x}_i$ can be carefully designed so that $d(\hat{Q}_n, Q_0) + \tilde{d}(\hat{Q}_n, Q_0)$ is of order $n^{-1}(\log n)^p$, for any nondegenerate elliptically symmetric distribution Q with bounded support. For such design sequences, the bias is negligible compared with the variance (note that the conditional variance is typically smaller than the unconditional variance and can be calculated in a way similar to Section 5 before). Theoretically, we have some reservations for using Q_0 with an unbounded support (including the normal one), which seems to create a problem for obtaining the identity (4.10) in Brillinger (1982). However, this may not be a problem in practice.

6.3. *Measuring elliptic asymmetry and bias reduction.* We shall introduce a measure of elliptic asymmetry, called EASY, for $\hat{Q}_n$, which indicates how vulnerable $\hat{Q}_n$ may be to link violation. Bias reduction will be based on subsampling, using EASY as a criterion.

Throughout this section, we assumed that $\mathbf{x}_i$'s are normalized so that $\hat{Q}_n$ has mean 0 and covariance I.

To begin with, recall from the discussion in Section 6.1.2 that the major factors for the bias bound are $\hat{Q}_{n\beta} - Q_{0\beta}$ and $H_{\hat{Q}_{n,\beta}} - H_{Q_{0,\beta}}$. Since Q_0 is arbitrary as long as it satisfies (A.2), we may choose Q_0 to minimize the effect of these factors, hoping that it may lead to a sharper bias bound.

We shall consider only those Q_0's that have the same marginal cumulative distributions as $\hat{Q}_n$ along the direction β. This restriction exterminates the first factor. Furthermore, to eliminate the boundary effect (due to integration by part), we require Q_0 to have the same mean as $\hat{Q}_n$, implying that $H_{Q_{0,\beta}}(\infty) = H_{\hat{Q}_{n,\beta}}(\infty) = 0$. Let $\hat{\mathscr{C}}_{n,\beta}$ be the class of all Q_0's satisfying these constraints and the condition [implied by (A.2)] that $E(\mathbf{x}|\beta\mathbf{x}) \propto \beta$. From $\hat{\mathscr{C}}_{n,\beta}$, we shall choose a Q_0 to minimize the supremum norm of the difference between ICE curves $H_{\hat{Q}_{n,\beta}}$ *and* $H_{Q_{0,\beta}}$ *up to a constant vector* (note that the derivative of a constant is 0).

Denote the projection of a vector v onto the orthogonal complement of a unitary vector e by v_e, i.e., $v_e = v - \langle v, e\rangle e$. Then we have

$$\inf_{Q_0 \in \hat{\mathscr{C}}_{n,\beta}} \inf_{v \in R^p} \|H_{\hat{Q}_{n,\beta}} - H_{Q_{0,\beta}} - v\|_\infty = \inf_{v \in R^p} \left\| \left(H_{\hat{Q}_{n,\beta}}(\cdot) - v \right)_\beta \right\|_\infty. \tag{6.3.1}$$

Geometrically, the right side of (6.3.1) is the minimum of the radii of the cylinders with axes parallel to β that are circumscribed outside the empirical ICE curve. This quantity provides a measure of straightness for the empirical ICE curve.

Now since β is unknown, conservatively we take the supremum over β to get the measure of elliptic asymmetry EASY:

$$\mathrm{EASY}(\hat{Q}_n) = \sup_{\beta \in R^p} \inf_{v \in R^p} \left\| \left(H_{\hat{Q}_{n,\beta}}(\cdot) - v \right)_\beta \right\|_\infty.$$

The domain of EASY can be extended to any Q. It is not hard to see that $\mathrm{EASY}(Q) = 0$ if and only if Q is elliptically symmetric in R^p or its linear subspaces (including straight lines). The larger the value $\mathrm{EASY}(\hat{Q}_n)$ is, the more

serious the bias in estimating the direction of β under link violation may be. Thus EASY may be viewed as a design diagnosis criterion which serves the purpose of alerting us against the possible ill effects due to the link violation.

When $\text{EASY}(\hat{Q}_n)$ is large, we know that our design is vulnerable to link violation. Thus we should be more concerned about model checking or model searching to ensure that our final regression model is highly accurate (but unfortunately the effectiveness of such efforts may be limited). In addition to this, we may search for subsamples (or more generally, distributions $\tilde{Q}_n$ with supports contained in the support of $\hat{Q}_n$) that have lower EASY and run the regression analysis on these subsamples [for instance, to get a point estimate we may minimize $E_{\tilde{Q}_n} L(a + b\mathbf{x}, y)$], hoping for the results to be reliable. Moreover, different estimates obtained from different subsamples can either be combined to increase efficiency or be compared to check if the one component model assumption is violated or not. Further works need to be done in order to offer a practical guidance on this last aspect.

REMARK 6.3. A less conservative measure of the vulnerability to link violation is simply to estimate the quantity in (6.3.1) by plugging in $\beta = \hat{\beta}$.

REMARK 6.4. All discussions in this section deal with the "worst" case situation. However, empirical study by Brillinger and others suggests that quite often the bias may be negligible even for a moderate violation of the design condition.

6.4. *A simulation study.* As an illustration of the impact of elliptical asymmetry on the regression estimates, we have conducted the following simulation study. We take the regressor to be bivariate, which we denote as $\mathbf{x} = (x_1, x_2)'$. We take the design points to be distributed evenly over the square $-0.5 \le x_1 \le 0.5$, $-0.5 \le x_2 \le 0.5$:

$$(x_{1ij}, x_{2ij}) = \left(\frac{2i - m - 1}{2(m-1)}, \frac{2j - m - 1}{2(m-1)} \right), \qquad i = 1, \ldots, m,\ j = 1, \ldots, m.$$

We take the true model to be

$$y = k(\beta\mathbf{x}) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2),$$

where $\beta = (\cos\theta, \sin\theta)$ and $k(\cdot)$ is antisymmetric, continuous and piecewise linear:

$$k(t) = \begin{cases} 0.374t, & \text{if } 0 \le t \le 0.135, \\ 0.0477 + 0.0211t, & \text{if } 0.135 \le t \le 0.321, \\ 0.2554 - 0.626t, & \text{if } 0.321 \le t \le 0.468, \\ -0.0477 + 0.0211t, & \text{if } t \ge 0.468, \\ -k(-t), & \text{if } t \le 0. \end{cases}$$

This design satisfies condition (A.2) only for $\theta = 0$, $\pi/4$, $\pi/2$ and $3\pi/4$, i.e., $\beta = (1,0)$, $(1/\sqrt{2}, 1/\sqrt{2})$, $(0,1)$ and $(1/\sqrt{2}, -1/\sqrt{2})$. For our illustration, we

TABLE 1
Distribution of $\hat{\theta}$ and $\tilde{\theta}$

	Mean	Standard deviation	Quantiles						
			0.05	0.10	0.25	0.50	0.75	0.90	0.95
$\hat{\theta}$	−0.282	0.084	−0.420	−0.339	−0.339	−0.283	−0.226	−0.174	−0.144
$\tilde{\theta}$	0.419	0.060	0.322	0.342	0.378	0.420	0.460	0.496	0.518

take $\theta = 0.416$, i.e., $\beta = (0.915, 0.404)$, for which the symmetry condition (A.2) is violated. We take $m = 20$ and $\sigma = 0.025$.

Under the given assumptions, the least squares estimate $\hat{\beta}$ is normally distributed with expectation

$$\beta^* = E(\hat{\beta}) = (0.0477, -0.0138)$$

and a diagonal covariance matrix

$$\text{Cov}(\hat{\beta}) = \tau^2 I, \qquad \tau = 0.00412.$$

The direction of $E(\hat{\beta})$ differs from that of β by 0.697 rad ($\sim 40°$). The first row of Table 1 summarizes the distribution of the direction of $\hat{\beta}$, i.e., $\hat{\theta} = \tan^{-1}(\hat{\beta}_2/\hat{\beta}_1)$, which is estimated from 10,000 draws from the bivariate normal distribution for $\hat{\beta}$, using the RANNOR function in SAS. The estimated direction is practically always in the fourth quadrant, which is at least 0.416 rad away from β.

To reduce the bias, consider the following subsampling procedure: Ignore the design points in the four corners of the square and apply the linear regression to the subdesign $\{X: x_1^2 + x_2^2 \leq 0.25\}$. The subdesign is very close to being spherically symmetric. The least squares estimate $\tilde{\beta}$ based on this subdesign is normally distributed with expectation

$$\beta^{**} = E(\tilde{\beta}) = (0.0904, 0.0403)$$

and a diagonal covariance matrix

$$\text{Cov}(\tilde{\beta}) = \tilde{\tau}^2 I, \qquad \tilde{\tau} = 0.00593.$$

The direction of the expectation β^{**} is 0.420 rad, which is, as expected, almost identical to that of the true direction θ; the two are different by less than 0.004 rad.

The second row of Table 1 summarizes the distribution of the estimated direction $\tilde{\theta} = \tan^{-1}(\tilde{\beta}_2/\tilde{\beta}_1)$ based on this subdesign. There is a probability of about 0.9 for the estimated direction to be within 0.1 rad ($\sim 6°$) from the true direction $\theta = 0.416$ rad. Furthermore, the estimated direction is practically always in the first quadrant, which is at least 0.282 rad away from of β^*, the expected estimated direction based on the complete design.

Despite the substantial bias of the full design estimate $\hat{\beta}$, it is not easy to detect the model violation by most standard diagnostic tools. As an illustration,

TABLE 2
Linear regression, $R^2 = 0.1036$

Variable	d.f.	$\hat{\beta}$	t
Intercept	1	−0.0022	−0.99
x_1	1	0.0484	6.67
x_2	1	−0.0087	−1.20
Source	**d.f.**	**Mean Square**	F
Model	2	0.0445	22.94
Error	397	0.0019	

we generate a data set according to the above specifications, using the RANNOR function in SAS to generate the ε's. Table 2 summarizes the linear regression of y on x_1 and x_2. The estimated direction $\hat{\theta} = -0.178$ differs from the true direction by 0.594 rad ($\sim 34°$). We carried out some standard diagnostic techniques but failed to detect the model violation. For instance, the linear regression of y on x_1, x_2, x_1^2, x_2^2 and x_1x_2 shows none of the three quadratic terms is significant at the nominal 5% level; the combined F-test has a nominal P-value of 0.985. The more parsimonious test, Tukey's 1 degree-of-freedom test, has a nominal one-sided P-value of 0.432. Figure 1 is the usual residual plot (residual versus prediction scatter diagram) for regressing y on x_1, x_2. There does not appear to be any interesting patterns in the plot. If we treat the residuals as being independent, we can partition the residuals by the ranks of the predicted values, then use a one-way ANOVA to test for the presence of patterns. Using 20 partitions with 20 observations in each, the test has a nominal P-value of 0.943. All of these residual-based diagnostic tools suggest that the fitted least squares model is satisfactory (cf. Remark 6.5 below).

On the other hand, Table 3 summarizes the least squares regression of y on x_1 and x_2, restricted to the subdesign, for the same data set. The estimated direction, $\tilde{\theta} = 0.502$ rad, different from the true direction by only 0.086 rad ($\sim 5°$), is dramatically different from the original estimate $\hat{\theta}$. Thus the consideration of elliptical symmetry of the design distribution adds a new dimension to regression diagnostics. This new viewpoint is rather different from the more popular one based on the concept of the leverage point; the latter is a local sensitivity analysis, while the former is a global violation analysis.

REMARK 6.5. It might be possible to detect model violation in this example by examining the residuals more thoroughly. For example, if we plot the residuals against a number of possible linear combinations of x_1 and x_2, rather than just the fitted combination $\hat{\beta}\mathbf{x}$, it is possible to detect nonlinear patterns from those residual plots against the linear combinations near $\beta\mathbf{x}$. However, such procedures might lead to spurious patterns [cf. Huber (1985), Section 21] and

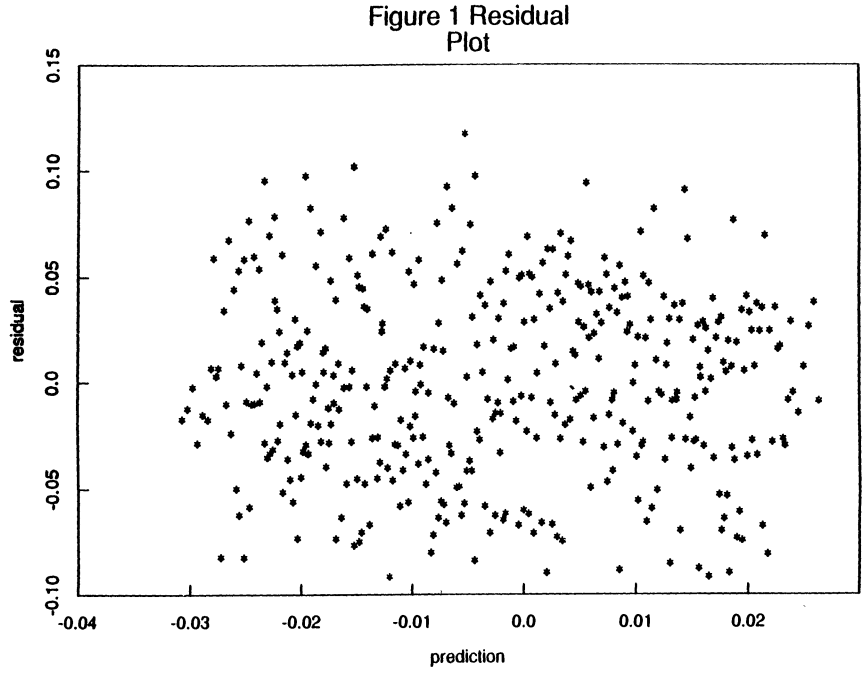

FIG. 1. *Residual plot.*

TABLE 3
Linear regression for subdesign, $R^2 = 0.2664$

Variable	**d.f.**	$\hat{\beta}$	t
Intercept	1	−0.0014	−0.58
x_1	1	0.0850	8.86
x_2	1	0.0466	4.86
Source	**d.f.**	**Mean Square**	F
Model	2	0.0835	51.01
Error	281	0.0016	

might not be easy to apply when the dimensionality of $\mathbf{x}$ is higher. Other possibilities include nonparametric regression techniques such as kernel estimates, thin plate spline, partial spline [Engle, Granger, Rice and Weiss (1986) and Wahba (1984)] and projection pursuit regression [Friedman and Stuetzle (1981)]. Eubank (1988) provides a nice account on nonparametric regression.

REMARK 6.6. For the full design, we have a probability of $\Phi(4.20) = 1.000$ to reject the hypothesis that the slope vector β is proportional to (0.915, 0.404), the

true direction, using a standard two-sided 5% level test. Therefore the full design almost certainly misleads us to the false conclusion that β is significantly different from (0.915, 0.404). On the other hand, if the regression analysis is restricted to the subdesign, the probability of rejecting this hypothesis is only 0.0014. The test based on the subdesign also has a probability of $\Phi(7.57) = 1.000$ to reject the false hypothesis that the slope vector is proportional to β^*.

7. Adaptive estimation. The discussion so far has been from the viewpoint of robustness, so the criterion function used for regression is already given. We may also ask the question of how to obtain an efficient estimate of β (up to a multiplicative scalar) under the general regression model (2.8) with g, F being unknown. It turns out that under the elliptical symmetry condition for $\mathbf{x}$, (A.2‴), we may estimate β (up to a proportionality scalar) as well as if g and F are known. In other words, adaptive estimation is possible here.

The main tool we shall use here is the device given by Bickel (1982), a paper which the reader is expected to be familiar with in order to follow the discussion below.

Bickel (1982) discovered that for many semiparametric problems (i.e., those with both parametric and nonparametric components), there is a convexity structure for the nonparametric component. Utilizing convexity, he simplifed Stein's necessary condition for adaptive estimation. In the following we shall demonstrate that the convexity condition and the simplified Stein necessary condition [called the generalized S^* condition in Bickel (1982)] hold. We also briefly discuss how to obtain an adaptive estimate without giving the regularity conditions and the proofs. The distribution of $\mathbf{x}$ will be assumed known in our discussion although for constructing an adaptive estimate it may be unknown as well.

7.1. *Convexity.* The convexity condition in Bickel (1982) amounts to the following: For each fixed (α, β), the set of distributions

$$\{\text{distribution of } (y, \mathbf{x}) \text{: (1.1) holds, } g, F \text{ arbitrary}\}$$

is a convex set in the sense that any mixture of two distributions in this set also belongs to this set.

To see why this condition holds for our case, we need only to observe that under (1.1) the conditional distribution of y given $\mathbf{x}$ takes the form $H(\alpha + \beta\mathbf{x}, y)$ where $H(t, \cdot)$ is an arbitrary distribution function for each t.

7.2. *Generalized S^*.* Let $l(y, \mathbf{x}, \alpha, \beta, h)$ be the logarithm of the density for $(y, \mathbf{x})$ when the conditional density of y given $\mathbf{x}$ is $h(\alpha + \beta\mathbf{x}, y)$. Clearly, the vector of partial derivatives of l with respect to α and β is

$$\dot{l}(y, \mathbf{x}, \alpha, \beta, h) = \left(\frac{h'}{h}(\alpha + \beta\mathbf{x}, y), \quad \mathbf{x}\frac{h'}{h}(\alpha + \beta\mathbf{x}, y)\right),$$

where $h'(t, y) = \partial h(t, y)/\partial t$. Write the information matrix $I = E\dot{l}'\dot{l}$ and its

inverse in block form,

$$I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}, \qquad I^{-1} = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix},$$

where I_{11} and I^{11} are scalars. Suppose we are interested in estimating $\varphi(\beta)$ where φ is any differentiable function (could be vector-valued) such that $\varphi(c\beta) = \varphi(\beta)$ for any scalar c.

Now define $\tilde{l}(y, \mathbf{x}, \alpha, \beta, h) = \dot{l}(y, \mathbf{x}, \alpha, \beta, h)I^{-1}(0, \dot{\varphi}(\beta))$, where $\dot{\varphi}(\beta)$ denote the matrix of partial derivatives of φ. The generalized S^* condition is

$$E_{h^*}\tilde{l}(y, \mathbf{x}, \alpha, \beta, h) = 0, \quad \text{for any } h, h^*, \tag{7.2.1}$$

where the subscript of E indicates the true conditional distribution.

The proof of (7.2.1) is given below. First using Lemma 5.1 we see that

$$I_{12} \propto \beta V \quad \text{and} \quad I_{22} \propto V + cV\beta'\beta V$$

for some scalar c. Then using the identities

$$I^{22} = \left(I_{22} - I_{21}(I^{11})^{-1}I_{12}\right)^{-1} \tag{7.2.2}$$

and

$$I_{11}I^{12} + I_{12}I^{22} = 0 \tag{7.2.3}$$

we see that

$$I^{22} \propto V^{-1} + c'\beta'\beta \tag{7.2.4}$$

and

$$I^{12} \propto \beta \tag{7.2.5}$$

for some scalar c'. In addition, the homogeneous restriction of φ implies

$$\beta\dot{\varphi}(\beta) = 0. \tag{7.2.6}$$

Finally using Lemma 5.1 again, we see that

$$E_{h^*}\dot{l}(\mathbf{x}, y, \alpha, \beta, h) = (*, c''\beta V), \tag{7.2.7}$$

where $*$ and c'' are scalars. Putting together (7.2.4)–(7.2.7), we easily obtain (7.2.1) as desired.

7.3. *Adaptive estimation.* A general recipe of constructing an adaptive estimator is given in Bickel (1982). To apply his method, all we need is to be able to find a consistent estimate of $\tilde{l}(y, \mathbf{x}, \alpha, \beta)$, which is seen to be proportional to

$$\frac{h'}{h}(\alpha + \beta\mathbf{x}, y) \cdot \mathbf{x}V^{-1}\dot{\varphi}(\beta).$$

Thus, we have to estimate $h'/h(t, y)$. In principle this is not hard, using techniques from the nonparametric density (and its partial derivative) estimation based on the data $(\hat{\alpha} + \hat{\beta}\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$. Here we need to impose the identifiability constraint $\alpha^* = \alpha$, $\beta^* = \beta$, as was done in Bickel (1982). The details on the mode of consistency and regularity conditions as well as practical

guidance for choosing the appropriate smoothing parameter involved in the density estimation will be examined more closely in the future.

APPENDIX

PROOF OF LEMMA 2.2. For almost any $(\mathbf{x}, y)$, we can take a supporting hyperplane for the criterion function L at an interior point (a^*, b^*) in Ω. The hyperplane can be taken as

$$H(a, b; \mathbf{x}, y) = L(a^* + b^*\mathbf{x}, y) + ((a - a^*) + (b - b^*)\mathbf{x})L_1(a^* + b^*\mathbf{x}, y),$$

where L_1 is the right-side derivative of L with respect to θ. Take a closed cube $B \subset \Omega$ centered at (a^*, b^*). For any (a, b) in B, the supporting hyperplane is bounded from above by $L(a + b\mathbf{x}, y)$, therefore its expectation is either finite or $-\infty$. However, the case of $-\infty$ can be ruled out by considering the two points $(a, b) \in B$ and $2(a^*, b^*) - (a, b) \in B$ together: If the expectation of the supporting hyperplane is $-\infty$ at one of the two points, then expectation at the other will be $+\infty$. Therefore $EH(a, b; \mathbf{x}, y)$ is finite for all $(a, b) \in B$ and hence for all $(a, b) \in R^{p+1}$. This is a finite lower bound for all $(a, b) \in \Omega_0$. The proof of this lemma is complete. □

PROOF OF LEMMA 3.2. (i) Observe that $Ey = EE^A g(\alpha + A, \varepsilon) = h(A)$, where $h(A)$ is defined by $h(A) = E^A g(\alpha + A, \varepsilon)$. Similarly, $EyA = Eh(A)A$. Thus D_1 consists of those pairs $(Eh(A), Eh(A)A)$ for some h such that $h(A) > \psi'(-\infty)$. We shall consider only the case that $\psi(-\infty) > -\infty$ [the case $\psi(-\infty) = -\infty$ is trivial].

By the assumption $EA = 0$, we have

$$Eh(A)A = E(h(A) - \psi'(-\infty))A < E(h(A) - \psi'(-\infty))\overline{B}.$$

Therefore $(Eh(A), Eh(A)A)$ belongs to the set described in (i) of Theorem 3.2.

On the other hand, for any point (η, ζ) in the set described in (i), we can find corresponding h by considering only those h which take the form $h(A) = a + \psi(-\infty)$, $\psi(-\infty)$ or $a_2 + \psi(-\infty)$ depending on $A \geq b_1$, $b_2 < A < b_1$ or $A \leq b_2$, where a_1, a_2, b_1, b_2 are constants with $b_1 > \zeta(\eta - \psi(-\infty))^{-1} > b_2$. This completes the proof of part (i).

(ii) For a fixed $c > 0$, the largest a such that $(a, c) \in \tilde{\Omega}$ is $a = -c\overline{B}$. Now since $E\psi'(a + cA)$ is increasing in c, we have

$$E\psi'(a + cA) \leq E\psi'(c(A - \overline{B})).$$

On the other hand, since $E\psi'(c(A - \overline{B}))$ is decreasing in c, we have for $c > \bar{c}_\eta$,

$$E\psi'(c(A - \overline{B})) < E\psi'(\bar{c}\eta(A - \overline{B})) = \eta.$$

Therefore $\bar{c}_\eta$ is the largest c such that there exists some a to satisfy

$$E\psi'(a + cA) = \eta. \tag{A.1}$$

Now we need the following lemma.

LEMMA A. *Subject to* (A.1), $E\psi'(a + cA)A$ *is an increasing function of* c.

The proof of this lemma is given later. Continuing the proof of part (ii), we see that by this lemma, the largest value of $E\psi'(a + cA)A$, subject to (A.1), is achieved at $c = \bar{c}_\eta$. Similarly, we can show that the lower bound is also as specified in (ii) of Theorem 3.2, completing this part of the proof. □

PROOF OF LEMMA A. First observe that due to the constraint (A.1), a is a function of c. Now taking the derivative with respect to c on both sides of (A.1), we get

$$E\psi''(a + cA)(a' + A) = 0,$$

leading to

$$a' = -E\psi''(a + cA)A/E\psi''(a + cA).$$

On the other hand,

$$\begin{aligned}\frac{d}{dc}E\psi'(a + cA)A &= E\psi''(a + cA)A^2 + a'E\psi''(a + cA)A \\ &= \frac{[E\psi''(a + cA)A^2][E\psi''(a + cA)] - [E\psi''(a + cA)A]^2}{E\psi''(a + cA)} \\ &\geq 0,\end{aligned}$$

where the last inequality is due to the Cauchy–Schwarz inequality. This completes the proof of Lemma A. □

PROOF OF (5.2). The proof is essentially the same as the proof for Lemma 2.2, noting that over a closed cube, the convex function $L(a + b\mathbf{x}, y) - L(\alpha^* + \beta^*\mathbf{x}, y)$ assumes its sup at one of the 2^{p+1} vertices of B and the supporting hyperplane also assumes its inf at one of the vertices. □

PROOF OF LEMMA 5.1. First take $\tilde{\mathbf{x}} = V^{-1/2}(\mathbf{x} - \mu)$, $\tilde{\beta} = \beta V^{1/2}$ and $\tilde{\psi}(\cdot) = \psi(\cdot + \beta\mu)$. It suffices to show that for some scalars c_1, c_2, c_3, we have

$$E\tilde{\psi}(\tilde{\beta}\tilde{\mathbf{x}})\tilde{\mathbf{x}}' = c_1\tilde{\beta} \tag{A.2}$$

and

$$E\tilde{\psi}(\tilde{\beta}\tilde{\mathbf{x}})\tilde{\mathbf{x}}\tilde{\mathbf{x}}' = c_2 I + c_3\tilde{\beta}'\tilde{\beta}. \tag{A.3}$$

Note that $\tilde{\mathbf{x}}$ is spherically symmetric with $E\tilde{\mathbf{x}} = 0$ and $\operatorname{Var}\tilde{\mathbf{x}} = I$. Let $\mathbf{u}$ be the projection of $\tilde{\mathbf{x}}'$ to the orthogonal complement of $\tilde{\beta}$. Clearly, given $\tilde{\beta}\tilde{\mathbf{x}}$, $\mathbf{u}$ is spherically symmetric in the orthogonal complement of $\tilde{\beta}$. Therefore,

$$\begin{aligned}E^{\tilde{\beta}\tilde{\mathbf{x}}}\tilde{\mathbf{x}}' &= E^{\tilde{\beta}\tilde{\mathbf{x}}}(\mathbf{u} + \tilde{\beta}\tilde{\mathbf{x}} \cdot \|\tilde{\beta}\|^{-2} \cdot \tilde{\beta}) \\ &= \tilde{\beta}\tilde{\mathbf{x}} \cdot \|\tilde{\beta}\|^{-2} \cdot \tilde{\beta},\end{aligned}$$

proving (A.2) with $c_1 = E\tilde{\psi}(\tilde{\beta}\tilde{\mathbf{x}}) \cdot \tilde{\beta}\tilde{\mathbf{x}} \cdot \|\tilde{\beta}\|^{-2}$. Similarly,

$$E^{\tilde{\beta}\tilde{\mathbf{x}}}\tilde{\mathbf{x}}\tilde{\mathbf{x}}' = (\tilde{\beta}\tilde{\mathbf{x}})^2 \cdot \|\tilde{\beta}\|^{-2} \cdot \tilde{\beta}'\tilde{\beta} + E^{\tilde{\beta}\tilde{\mathbf{x}}}\mathbf{u}'\mathbf{u}$$

and

$$E^{\tilde{\beta}\tilde{x}}\mathbf{u}'\mathbf{u} = c'(I - \|\tilde{\beta}\|^{-2}\tilde{\beta}\tilde{\beta}'),$$

where $c' = E^{\tilde{\beta}\tilde{x}}\{\|\tilde{x}\|^2 - (\tilde{\beta}\tilde{x})^2\|\tilde{\beta}\|^{-2}\}/(p-1)$. Therefore (A.3) holds with

$$\begin{aligned} c_2 &= E\tilde{\psi}(\tilde{\beta}\tilde{x})\left\{\|\tilde{x}\|^2 - (\tilde{\beta}\tilde{x})^2\|\tilde{\beta}\|^{-2}\right\}/(p-1) \\ &= (p-1)^{-1}E\psi(\beta x)\left\{(x-\mu)'V^{-1}(x-\mu)\right. \\ &\qquad \left. -(\beta(x-\mu))^2(\beta V\beta')^{-1}\right\}. \end{aligned} \tag{A.4}$$

□

REMARK. (A.2) can be proved directly from assumption (A.2), but (A.3) depends on the elliptic symmetry assumption (A.2‴).

PROOF OF THEOREM 5.3.1. Without loss of generality, assume that $\mu = 0$. Then by Lemma 5.1, we see that C and Λ take the block form

$$C = \begin{pmatrix} * & c_1\beta V \\ c_1 V\beta' & c_2 V + c_3 V\beta'\beta V \end{pmatrix},$$

$$\Lambda = \begin{pmatrix} * & c_4\beta V \\ c_4 V\beta' & c_5 V + c_6 V\beta'\beta V \end{pmatrix},$$

where c_1–c_6 are scalars, with $c_5 = \eta^{-1}$ and $c_2 = \lambda\eta^{-1}$ [see (A.4)]. Now noticing that Λ takes the same form as the information matrix I in Section 7.2, we may use the same argument to find

$$\Lambda^{-1} = \begin{pmatrix} * & c_7\beta \\ c_7\beta' & c_8 V^{-1} + c_9\beta'\beta \end{pmatrix},$$

where $c_8 = c_5^{-1} = \eta$. Now the result (5.3.1) follows immediately from matrix multiplication.

To see that (5.3.5) holds, we simply observe that in the proof of Lemma 5.1, the scalar $c' = 1$ under the normality assumption. □

PROOF OF THEOREM 5.4.1. First, it is clear that without loss of generality we may assume $\mu = 0$ and $V = I$. Similar to the usual parametric model case, we approximate $l_n(a, b)$ locally by a quadratic function

$$l_n(a, c) \approx l_n + (a - \alpha^*, b - \beta^*)s_n + \tfrac{1}{2}(a - \alpha^*, b - \beta^*)i_n(a - \alpha^*, b - \beta^*)',$$

where l_n, s_n, i_n are defined in Section 5.2 with the argument (α^*, β^*) being omitted for simplicity. Now minimizing the right-hand side of this expression over $(a, b) \in R^{p+1}$ and over $(a, b) = (a, \mathbf{v}A)$ for $\mathbf{v} \in R^h$, respectively, we see that

$$\begin{aligned} Q &\approx \min_{\mathbf{v}\in R^h,\, a\in R} \left\| i_n^{-1/2}s_n - i_n^{1/2}(a, \mathbf{v}A)' \right\|^2 \\ &= \left\| H_n\left(i_n^{-1/2}s_n\right) \right\|^2, \end{aligned}$$

where H_n is the projection matrix from R^{p+1} to the orthogonal complement of the linear space $\{(n^{-1}i_n)^{1/2}(a, \mathbf{v}A)\colon a \in R, \mathbf{v} \in R^h\}$. Now asymptotically, $i^{-1/2}s_n$ is normal, with mean 0 and covariance

$$\Lambda^{-1/2}C\Lambda^{-1/2} = \begin{pmatrix} * & c_1\beta^* \\ c_1\beta^{*\prime} & \lambda I + c_3\beta^{*\prime}\beta^* \end{pmatrix}.$$

On the other hand, H_n converges to a projection matrix H with the property

$$0 = H\Lambda^{1/2}\begin{pmatrix} a \\ A'\mathbf{v}' \end{pmatrix}$$

$$= H\begin{pmatrix} * & c_4\beta^* \\ c_4\beta^{*\prime} & I + c_5\beta^{*\prime}\beta^* \end{pmatrix}\begin{pmatrix} a \\ A'\mathbf{v}' \end{pmatrix}$$

for any $a \in R$, $\mathbf{v} \in R^h$. Since β^* is of the form $\mathbf{v}A$, we see that the asymptotic covariance for $H_n(i_n^{-1/2}s_n)$ is an idempotent matrix with rank $p - h$ after being rescaled by λ^{-1}, proving the desired result. □

Acknowledgments. We appreciate useful discussions with Peter Bickel, David Brillinger, Art Goldberger and Chris Skinner. Peter Bickel suggested to us the connection of this work with adaptive estimation.

REFERENCES

BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671.

BICKEL, P. J. and DOKSUM, K. A. (1981). An analysis of transformations revisited. *J. Amer. Statist. Assoc.* **76** 293–311.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. B* **26** 211–252.

BOX, G. and DRAPER, N. (1959). A basis for the selection of a response surface design. *J. Amer. Statist. Assoc.* **54** 622–654.

BRILLINGER, D. R. (1977). The identification of a particular nonlinear time series system. *Biometrika* **64** 509–515.

BRILLINGER, D. R. (1983). A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr., eds.) 97–114. Wadsworth, Belmont, Calif.

CHENG, C.-S. and LI, K.-C. (1984). The strong consistency of *M*-estimators in linear models. *J. Multivariate Anal.* **15** 91–98.

CHUNG, C. F. and GOLDBERGER, A. S. (1984). Proportional projections in limited dependent variable models. *Econometrica* **52** 531–534.

COX, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press, Princeton, N.J.

DIACONIS, P. and FREEDMAN, D. A. (1982). On inconsistent *M*-estimators. *Ann. Statist.* **10** 454–461.

DUAN, N. (1986). Scale sensitivity for prediction procedures in the Box–Cox transformation family. Unpublished.

EFRON, B. (1982). Transformation theory: How normal is a family of distributions? *Ann. Statist.* **10** 323–339.

EFRON, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.* **81** 709–721.

ENGLE, R. F., GRANGER, C. W. J., RICE, J. and WEISS, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81** 310–320.

EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Dekker, New York.

FISHER, R. A. (1936). The use of multiple measurements in taxomonic problems. *Ann. Eugenics* **7** 179–188.

FRIEDMAN, J. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

GALIL, Z. and KIEFER, J. (1977). Comparison of Box–Draper and *D*-optimum designs for experiments with mixtures. *Technometrics* **19** 441–444.

GOLDBERGER, A. S. (1981). Linear regression after selection. *J. Econometrics* **15** 357–366.

GREENE, W. (1981). On the asymptotic bias of ordinary least squares estimates of the Tobit model. *Econometrica* **49** 505–514.

GREENE, W. (1983). Estimation of limited dependent models by ordinary least squares and the method of moments. *J. Econometrics* **21** 195–212.

HAGGSTROM, G. W. (1983). Logistic regression and discriminant analysis by ordinary least squares. *J. Bus. Econ. Statist.* **1** 229–238.

HALTON, J. H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numer. Math.* **2** 84–90.

HINKLEY, D. V. and RUNGER, G. (1984). The analysis of transformed data (with discussion). *J. Amer. Statist. Assoc.* **79** 302–320.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–233. Univ. California Press.

HUBER, P. J. (1981). *Robust Statistics.* Wiley, New York.

HUBER, P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.* **13** 435–525.

KIEFER, J. (1973). Optimal designs for fitting biased multiresponse surfaces. In *Multivariate Analysis—III. Proc. Third International Symposium on Multivariate Analysis* (P. R. Krishnaiah, ed.) 287–297. Academic, New York.

LE CAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *Univ. California Publ. Statist.* **1** 277–330.

LEHMANN, E. L. (1983). *Theory of Point Estimation.* Wiley, New York.

LI, K.-C. (1982). Minimaxity of the method of regularization on stochastic processes. *Ann. Statist.* **10** 937–942.

LI, K.-C. (1984). Robust regression designs when the design space consists of finitely many points. *Ann. Statist.* **12** 269–282.

MARCUS, M. B. and SACKS, J. (1977). Robust designs for regression problems. In *Statistical Decision Theory and Related Topics II* (S. S. Gupta and D. S. Moore, eds.) 245–268. Academic, New York.

MCCULLAGH, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11** 59–67.

MOURIER, E. (1953). Elements aléatoires dans un espace de Banache. *Ann. Inst. H. Poincaré* **13** 159–244.

NELDER, J. A. and PREGIBON, D. (1986). Quasi-likelihood and generalized linear models. Unpublished.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

PORTNOY, S. (1985). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. II: Normal approximation. *Ann. Statist.* **13** 1403–1417.

RUUD, P. (1983). Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. *Econometrica* **51** 225–228.

SACKS, J. and YLVISAKER, D. (1984). Some model robust designs in regression. *Ann. Statist.* **12** 1324–1348.

SPECKMAN, P. (1979). Minimax estimates of linear functionals in a Hilbert space. Unpublished.

TSIATIS, A. A. (1981). A large sample study of Cox's regression model. *Ann. Statist.* **9** 93–108.
WAHBA, G. (1984). Partial spline models for the semi-parametric estimation of functions of several variables. In *Statistical Analysis of Time Series* 319–329. Inst. Statist. Math., Tokyo.
WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61** 439–447.
WHITE, H. (1981). Consequences and detection of misspecified nonlinear regression models. *J. Amer. Statist. Assoc.* **76** 419–433.
YOHAI, V. J. and MARRONA, R. A. (1979). Asymptotic behavior of M-estimators for the linear model. *Ann. Statist.* **7** 258–268.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
LOS ANGELES, CALIFORNIA 90024

THE RAND CORPORATION
1700 MAIN STREET
SANTA MONICA, CALIFORNIA 90406-2138