

This article was originally published in a journal published by Elsevier, and the attached copy is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues that you know, and providing a copy to your institution's administrator.

All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:

<http://www.elsevier.com/locate/permissionusematerial>



ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Statistics & Probability Letters 77 (2007) 1201–1213

STATISTICS &
PROBABILITY
LETTERS

www.elsevier.com/locate/stapro

Thinking outside the box: Statistical inference based on Kullback–Leibler empirical projections

Kjell Doksum¹, Akichika Ozeki*, Jihoon Kim, Elias Chaibub Neto

Department of Statistics, University of Wisconsin, Madison, WI 53706, USA

Available online 16 March 2007

Abstract

Suppose that X is a random vector with probability distribution P and suppose that \mathcal{P} denotes a proposed model that involves interesting parameters and relationship between variables. We consider statistical inference procedures for the case where $P \notin \mathcal{P}$ constructed as follows: let $\theta(P)$ denote the parameter of the distribution $Q \in \mathcal{P}$ that minimizes a Kullback–Leibler (K–L)-type discrepancy $K(Q, P)$ between Q and P . We take $\theta(P)$ to be the parameter of interest. The estimate of $\theta(P)$, when it exists, is defined by $\hat{\theta} = \theta(\hat{P})$ where \hat{P} is the empirical probability. We call $\theta(\hat{P})$ a Kullback–Leibler empirical projection (KLEP). When $\theta(\hat{P})$ does not exist, we extend the concept of a K–L discrepancy to limits of empirical likelihoods to obtain KLEP procedures. Properties of inference procedures based on $\hat{\theta}$ are considered when $P \notin \mathcal{P}$. In particular we compare the naive procedure that uses the standard error applicable when $P \in \mathcal{P}$, the sandwich formula standard error, and the bootstrap standard error using asymptotic methods and Monte Carlo simulation. For regression experiments with a model based on transforming both response and covariates, we use results of Hernandez and Johnson [1980. The large-sample behavior of transformations to normality. *J. Amer. Statist. Assoc.* 75, 855–861] to derive KLEP procedures.

© 2007 Elsevier B.V. All rights reserved.

Keywords: KLEP; Box-Cox transformation; Outside the box; K-L divergence; Sandwich formula; Bootstrap; Classification; Covariate transformations

1. Introduction

Box (1979) captured the spirit of much of the work on statistical modeling when he said “Models of course, are never true but fortunately it is only necessary that they be useful”. We consider, as have many others, the effect on statistical inference of the true distribution being outside the working model. The probability distribution P generating the data is outside the model class \mathcal{P} being used in the statistical analysis.

A common example is linear regression models \mathcal{P} where the response Y is modeled to depend linearly on covariates X_1, \dots, X_d . In this case, if the true distribution P generating (X_1, \dots, X_d, Y) falls outside \mathcal{P} , the statistical least squares analysis applies to the coefficients in the linear model “closest” to P .

*Corresponding author.

E-mail addresses: doksum@stat.wisc.edu (K. Doksum), ozeki@cs.wisc.edu (A. Ozeki).

¹Supported in part by NSF Grant DMS-0505651, NIH R01GM076274-01, and CNPq Brazil.

In general, the Q in \mathcal{P} closest to P is determined by minimizing a measure of how much Q differs from P . We consider the Kullback–Leibler (K–L) divergence

$$K(Q, P) = -E_P(\log[q(\mathbf{Z})/p(\mathbf{Z})]),$$

where \mathbf{Z} denotes a random vector and p and q are the densities of P and Q with respect to some common measure μ . Let

$$J(Q, P) = E_P(\log q(\mathbf{Z}))$$

be the entropy of Q under P and note that

$$\arg \inf\{K(Q, P) : Q \in \mathcal{P}\} = \arg \sup\{J(Q, P) : Q \in \mathcal{P}\}. \quad (1.1)$$

Because P is unknown, we consider replacing it with the empirical distribution \hat{P} which assigns probability $\hat{p}_i = n^{-1}$ to each observed \mathbf{z}_i in the realization of an i.i.d. sample of size n from P .

If $K(Q, \hat{P})$ does not exist in R , using (1.1), we can instead use

$$\hat{Q} = \arg \sup\{J(Q, \hat{P}) : Q \in \mathcal{P}\} = \arg \sup\left\{\sum_{i=1}^n \log q(\mathbf{z}_i) : Q \in \mathcal{P}\right\} \quad (1.2)$$

which is the method of maximum likelihood when \mathcal{P} is a parametric model.

When \mathcal{P} is non- or semiparametric, then \hat{Q} as defined in (1.2) may not exist. In this case we can use a variation of an empirical likelihood (Owen, 1988, 2001; Shao, 2003; Bickel and Doksum, 2008) instead of $J(Q, \hat{P})$. Here (Bickel and Doksum, 2008), we let $\tilde{\mathcal{P}}$ be the closure in weak convergence of the union of \mathcal{P} and a class \mathcal{P}_D of discrete distributions Q that are consistent with the model \mathcal{P} and assign positive probability $q_i = Q(\{\mathbf{z}_i\}) > 0$ to each \mathbf{z}_i , $i = 1, \dots, n$, $\sum_{i=1}^n q_i = 1$. Then we define the maximum empirical likelihood estimate (MELE)

$$\hat{Q}_E = \arg \sup\left\{\sum_{i=1}^n \log q_i : Q \in \tilde{\mathcal{P}}\right\}. \quad (1.3)$$

Let $\hat{q}_E(\mathbf{z}_i)$ denote the q_i that maximizes (1.3). The empirical K–L divergence corresponding to \hat{Q}_E is

$$K_E(\hat{Q}_E, \hat{P}) = -E_{\hat{P}}(\log[\hat{q}_E(\mathbf{Z})/\hat{p}(\mathbf{Z})]) = \log \frac{1}{n} - \frac{1}{n} \sum_{i=1}^n \log \hat{q}_E(\mathbf{z}_i),$$

which is non-negative and equals zero when \mathcal{P} contains all distributions. Note that \hat{P} is uniform and maximizes the entropy among all distributions on $\{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ (e.g. Cover and Thomas, 2003), while \hat{Q}_E maximizes the entropy among these distributions over $Q \in \tilde{\mathcal{P}}$.

One possible definition of \mathcal{P}_D is

$$\mathcal{P}_D = \left\{ Q_D : Q_D(\mathbf{z}_i) = q(\mathbf{z}_i) / \sum_{j=1}^n q(\mathbf{z}_j) : Q \in \mathcal{P} \right\}.$$

In this case, if we set $\bar{q} = E_P(q(\mathbf{Z}))$ and $Q_D \in \mathcal{P}_D$, then

$$K(Q_D, \hat{P}) \rightarrow_P -E_P(\log[q(\mathbf{Z})/\bar{q}]).$$

Thus, we can measure the “distance” between P and the model \mathcal{P} by

$$I(\mathcal{P}, P) \equiv \inf\{-E_P(\log[q(\mathbf{Z})/\bar{q}]) : Q \in \mathcal{P}\}.$$

Often \mathcal{P} is parametrized as $\{P_\theta : \theta \in \Theta\}$ where $\theta = (\boldsymbol{\beta}, \eta)$ consists of a Euclidean parameter $\boldsymbol{\beta}$ and a function η . We take the parameter of interest to be the $\theta(P) \in \Theta$ that is the parameter of the distribution Q in \mathcal{P} that is closest in a K–L sense (as defined above) to the probability P that generates the data in an experiment. The estimate of $\theta(P)$ is the Kullback–Leibler empirical projection (KLEP) $\hat{\theta} = \theta(\hat{P})$. This estimate provides the best of two worlds: if $P \in \mathcal{P}$, then $\hat{\theta}$ is asymptotically optimal by maximum (empirical) likelihood theory. If $P \notin \mathcal{P}$, then $\hat{\theta}$ is nonparametrically asymptotically optimal by the theory of efficient influence functions (e.g. Bickel et al., 1993, 1998; van der Vaart, 1998). We investigate the sandwich and bootstrap approach to statistical

inference for the situation where $P \notin \mathcal{P}$ for linear, Box–Cox, and logistic models and find that, for the linear model, both the sandwich and bootstrap approaches give good results while for the Box–Cox model, the bootstrap is the best choice.

The goal is to select \mathcal{P} to be a useful model. One of the most useful models is the multivariate normal model because of the intuitive interpretations of its parameters. On the other hand, the distance between P and \mathcal{P} should be small to avoid systematic distortions. Hernandez and Johnson’s (1980) results can be used to unite the goals of being useful and realistic by using a model based on transforming the original data to multivariate normality. We show in Section 5 how to use Hernandez and Johnson (1980) to derive KLEM estimates of the parameters in a regression model based on transforming both the response and the covariates using Box–Cox transformations (Box and Cox, 1964).

2. Regression

Let $(X_{i1}, \dots, X_{id}, Y_i), i = 1, \dots, n$, be i.i.d. observations on (X_1, \dots, X_d, Y) . Let \mathcal{P} be a class of linear models or a class of distributions of (X_1, \dots, X_d, Y) satisfying

$$Y = \sum_{j=0}^d \alpha_j X_j + \varepsilon, \quad X_0 = 1,$$

$$E(\varepsilon) = 0, \quad \text{Var}(\varepsilon) = \sigma^2,$$

where X_1, \dots, X_d are independent of ε . Let $\mathbf{X} = (X_0, X_1, \dots, X_d)^\top$, and $\Sigma_X = \text{Cov}(X_1, \dots, X_d)^\top$. The above model is semiparametric with the distribution \mathbf{X} and ε arbitrary satisfying the following condition where P denotes a probability distribution of (X_1, \dots, X_d, Y) :

$$(A.1) \quad 0 < E_P Y^2 < \infty, \quad 0 < \text{Var}_P(X_j) < \infty, \quad j \geq 1, \quad \Sigma_X \text{ invertible.}$$

Define

$$\boldsymbol{\beta} = \boldsymbol{\beta}(P) = \arg \min\{E_P[Y - \mathbf{a}^\top \mathbf{X}]^2 : \mathbf{a} \in \mathbb{R}^{d+1}\}.$$

Clearly,

$$\boldsymbol{\beta} = (E[\mathbf{X}\mathbf{X}^\top])^{-1} E(\mathbf{X}Y).$$

The vector $\boldsymbol{\beta}$ is the coefficient vector of the distribution P_0 in \mathcal{P} closest to the true P in the sense that $\boldsymbol{\beta}(P) = \boldsymbol{\beta}(P_0)$ with

$$P_0 = \arg \min\{K(Q, P) : Q \in \mathcal{P}_0\},$$

where \mathcal{P}_0 is the subset of \mathcal{P} where $\varepsilon \sim N(0, \sigma^2)$ and K is the K–L discrepancy.

Least squares analysis corresponds to defining $\hat{\boldsymbol{\beta}}$ to be the empirical plug-in estimate:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{P}) = \arg \min\{E_{\hat{P}}[Y - \mathbf{a}^\top \mathbf{X}]^2 : \mathbf{a} \in \mathbb{R}^{d+1}\},$$

where \hat{P} is the empirical probability which assigns probability n^{-1} to each observed data point $(x_{i1}, \dots, x_{id}, y_i), 1 \leq i \leq n$. Thus, $\hat{\boldsymbol{\beta}} = (\mathbf{X}_D^\top \mathbf{X}_D)^{-1} \mathbf{X}_D^\top \mathbf{Y}$ where \mathbf{X}_D is the $n \times (d+1)$ design matrix and $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$. The estimate $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal with adjusted variance.

Proposition 2.1. *If P satisfies (A.1), then $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow \mathcal{L} \mathcal{N}(\mathbf{0}, \Sigma)$, where*

$$\Sigma = \Sigma_{1X}^{-1} \Sigma_{Xe} \Sigma_{1X}^{-1}, \quad e = Y - \boldsymbol{\beta}^\top \mathbf{X},$$

$$\Sigma_{1X} = E(X_j X_k), \quad \Sigma_{Xe} = E(e^2 X_j X_k), \quad 0 \leq j \leq d, \quad 0 \leq k \leq d. \tag{2.1}$$

The usual (naive) asymptotic covariance matrix of the least squares estimate which is based on assuming that P is in the box is $\sigma_0^2 \Sigma_{1X}^{-1}$ where $\sigma_0^2 = \text{Var}_P(e)$. Thus, this naive variance is automatically inflated when P is outside the box. However, it is often incorrect. See Sections 2.2 and 3.3 for cases where it is correct.

In the context of heteroscedastic linear models, formula (2.1) is called a *sandwich formula*, e.g. Bickel and Doksum (2007, Example 6.6.4). Here (2.1) applies without any model assumptions except second moment conditions. Having the X 's spread out makes the Σ_{1X}^{-1} terms in (2.1) smaller, but the Σ_{Xe} term larger when X and e are correlated. An interpretation is that having the X 's spread out reduces variance but assuming linearity in X over a wide range of X 's is questionable and leads to an inflation of variance when P is outside the linear box \mathcal{P} . There is no bias problem because the target is linear, but (2.1) inflates the variance when there is a lack of model fit.

In a minimax sense, $\hat{\beta}$ is the most efficient estimate of β if P is in the box. To see this note that for the conditional distribution $\mathcal{L}_Q(Y|X_D)$ of Y given X_D , $\hat{\beta}$ has the variance $\sigma_e^2(X_D^T X_D)^{-1}$ for all $Q \in \mathcal{P}$. Let $Q_0 \in \mathcal{P}_0$ and let β^* be any estimate which is unbiased for $\mathcal{L}_{Q_0}(Y|X_D)$, then

$$\sup_{Q \in \mathcal{P}_1} \text{MSE}_Q(\beta^*|X_D) \geq \text{MSE}_{Q_0}(\beta^*|X_D) = \text{Var}_{Q_0}(\beta^*|X_D) \geq \text{Var}_{Q_0}(\hat{\beta}|X_D),$$

where \mathcal{P}_1 is the class of probabilities in \mathcal{P} with σ_X^2 and σ_e^2 fixed. The second inequality follows from the information inequality.

Proposition 2.2. *The least square estimate $\hat{\beta}$ minimizes*

$$\sup\{\text{MSE}_Q(\beta^*)|Q \in \mathcal{P}_1\},$$

over the class of estimates unbiased for $\mathcal{L}_Q(Y|X_D)$, $Q \in \mathcal{P}_0$.

Suppose P is not in \mathcal{P} . Because $\hat{\beta} = \beta(\hat{P})$ where \hat{P} is the empirical probability, there are a number of results that imply that in a class of regular estimates, $\hat{\beta}$ is nonparametrically asymptotically optimal. We refer to Bickel et al. (1993, 1998, Sections 1.3, 3.3, and 5.5) and van der Vaart (1998, Chapter 25). The influence function of the functional corresponding to $X_D^T Y$ is given by (2.6) in Doksum and Samarov (1995). This can be used to show that the estimate based on the efficient score function coincides with $\hat{\beta}$. See also Bickel and Ritov (2003).

2.1. Variance estimation, confidence regions, and variance adjustment

In order to find asymptotic confidence regions for the β coefficients, we need an estimate of Σ_{Xe} (our estimate of Σ_{1X} is $\hat{\Sigma}_{1X} = n^{-1}X_D^T X_D$). For this purpose, we introduce

$$X_{De} = (X_{ij}\hat{e}_i), \quad 1 \leq i \leq n, \quad 0 \leq j \leq d.$$

Then our estimate of Σ is the *sandwich estimate*

$$\hat{\Sigma} = n(X_D^T X_D)^{-1}(X_{De}^T X_{De})(X_D^T X_D)^{-1}.$$

When $d = 1$, the estimate of $\text{Var}(\hat{\beta}_1)$ is

$$\text{SE}^2(\hat{\beta}_1) \equiv \text{Var}(\hat{\beta}_1) = \frac{\sum (X_{i1} - \bar{X}_1)^2 \hat{e}_i^2}{[\sum (X_{i1} - \bar{X}_1)^2]^2}$$

and

$$n\text{SE}^2(\hat{\beta}_1) \xrightarrow{P} \frac{E_P[(X_1 - \mu_1)^2 e^2]}{[\text{Var}_P(X_1)]^2},$$

which reduces to the usual (naive) $\text{Var}_P(e)/\text{Var}_P(X_1)$ when X_1 and e are independent, that is, when P is the linear box \mathcal{P} . This $d = 1$ case provides the insight that:

Proposition 2.3. *Assume (A.1). When $d = 1$, the naive least square asymptotic variance of $\hat{\beta}_1$ is larger than the sandwich formula variance iff $(X_1 - \mu_1)^2$ and e^2 are negatively correlated.*

Let $\hat{\sigma}_e^2 = (n - d - 1)^{-1} \sum_{i=1}^n \hat{e}_i^2$ be the usual estimate of residual variance. We obtain confidence regions for β if we replace $\hat{\Sigma}_{1X}^{-1} \hat{\sigma}_e^2$ in the usual linear model confidence regions with the sandwich estimate $\hat{\Sigma}$. These

confidence regions have the correct asymptotic coverage probabilities for P satisfying our moment conditions (A.1).

How large is the variance adjustment when we replace the naive variance estimate with the sandwich estimate? We conduct a Monte Carlo (MC) study with $d = 1$ to compare the naive $(SE(\hat{\beta}_1)_{NAIVE})$, sandwich $(SE(\hat{\beta}_1)_{SA})$, bootstrap $(SE(\hat{\beta}_1)_{BOOT})$, and MC $(SE(\hat{\beta}_1)_{MC})$ estimates of $SE(\hat{\beta}_1)$ when the true P satisfies

$$Y = (1 - \gamma)(\alpha_0 + \alpha_1 X_1) + \gamma L(X_1) + \varepsilon, \tag{2.2}$$

where $L(t) = 5 + 2.5[1 + \exp(-10t)]^{-1}$, $\varepsilon \sim N(0, \sigma^2)$ and $X_1 \sim N(0, \sigma_0^2)$ are independent, $\sigma^2 \in \{0.1, 0.5, 0.75, 1.183\}$, $\sigma_0^2 = 0.1$, $\alpha_0 = 6.25$, $\alpha_1 = 0.5$, and $\gamma \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Here $\sigma^2 = 1.183$ is determined by solving power = 0.5 for the classical t -test of $H_0 : \beta_1 = 0$, $H_1 : \beta_1 > 0$, when $\gamma = 0$.

MC simulation is based on 1000 different seeds with $n = 128$, i.e. generate $(y_1, x_{1,1}), \dots, (y_{128}, x_{1,128})$ for each simulation. For the i th MC simulation, we compute $\hat{\beta}_{0,i}$, $\hat{\beta}_{1,i}$, $SE(\hat{\beta}_1)_{NAIVE,i}$, and $SE(\hat{\beta}_1)_{SA,i}$. Then we use the usual formulae to compute the MC averages of these quantities. They are denoted as $\hat{\beta}_0$, $\hat{\beta}_1$, $SE(\hat{\beta}_1)_{NAIVE}$, and $SE(\hat{\beta}_1)_{SA}$.

We use the MC estimate of $SE(\hat{\beta}_1)$ to represent the true $SE(\hat{\beta}_1) = \sqrt{E_P(\hat{\beta}_1 - \beta_1)^2}$ which can only be computed when P is known. The MC estimate is given by

$$SE(\hat{\beta}_1)_{MC} = \sqrt{\frac{\sum_{i=1}^{1000} (\hat{\beta}_{1,i} - \hat{\beta}_1)^2}{1000}}. \tag{2.3}$$

For the bootstrap, we first use (2.2) to generate 128 data pairs, $(y_1, x_{1,1})^{[1]}, \dots, (y_{128}, x_{1,128})^{[1]}$. From these 128 (x, y) pairs, we draw 100 bootstrap samples of size 128 and compute the bootstrap estimate of $SE(\hat{\beta}_1)$. This is repeated 200 times and the Monte Carlo estimate of the bootstrap estimate is the average of the 200 bootstrap estimates of $SE(\hat{\beta}_1)$.

The MC $\hat{\beta}_0$ and $\hat{\beta}_1$ are plotted against the nonlinearity parameter γ in Fig. 1. Note that $\hat{\beta}_0$ is nearly constant while $\hat{\beta}_1$ increases linearly with γ . A variety of σ values gives very similar results. When $\gamma = 0$, i.e. $P \in \mathcal{P}$, $\hat{\beta}_0$ and $\hat{\beta}_1$ are close to α_0 and α_1 as expected.

We compare the naive, sandwich (SA), and the bootstrap (BOOT) SE's to the "true" SE for model (2.2) in Fig. 2, where the "true" SE is represented by the MC SE (2.3). We see from Fig. 2 and other graphs not given here that when $\gamma = 0$, i.e. $P \in \mathcal{P}$, the four SE's are close. For σ^2 small, the naive estimate does poorly as γ increases, that is, as the true P moves away from the model \mathcal{P} , while for larger σ^2 , all three methods perform well. Both the sandwich and bootstrap approaches do well for all σ^2 and all γ . An interesting result is observed with $\sigma^2 = 1.183$. All four estimates are close together when the noise is large. In summary, $SE(\hat{\beta}_1)_{SA}$ and $SE(\hat{\beta}_1)_{BOOT}$ are always close to $SE(\hat{\beta}_1)_{MC}$. $SE(\hat{\beta}_1)_{NAIVE}$ is close to $SE(\hat{\beta}_1)_{MC}$ only when the true distribution P is in \mathcal{P} and when the noise is large relative to $SE(\hat{\beta}_1)$.

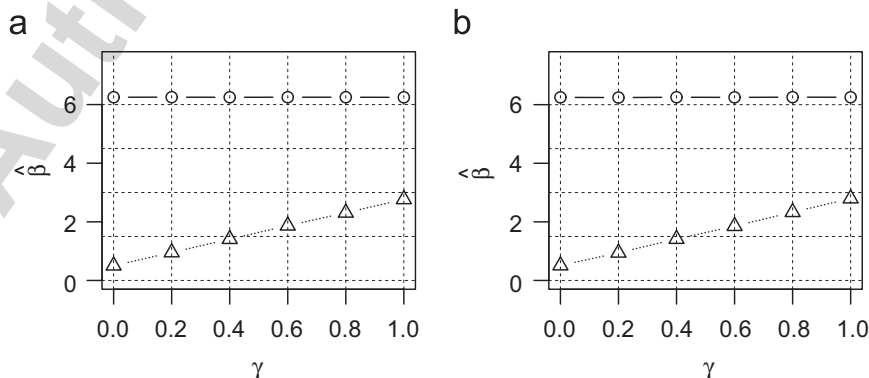


Fig. 1. Monte Carlo values of $\hat{\beta}_0$ (○) and $\hat{\beta}_1$ (△) plotted against the nonlinearity parameter γ . (a) $\sigma^2 = 0.1$ and (b) $\sigma^2 = 1.183$.

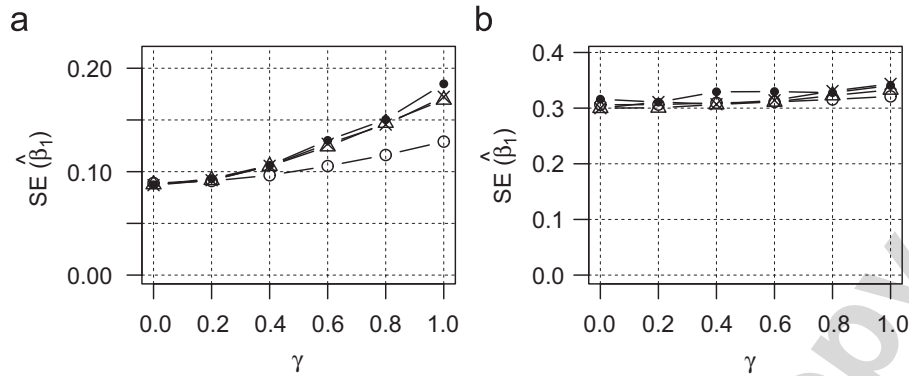


Fig. 2. The naive (○), sandwich (SA △), bootstrap (BOOT ×), and Monte Carlo (MC ●) $\hat{\beta}_1$ standard errors as functions of the nonlinearity parameter γ . (a) $\sigma^2 = 0.1$ and (b) $\sigma^2 = 1.183$.

2.2. Testing

Suppose we want to test whether X_1 and Y are related while controlling for X_2, \dots, X_d . Then, because X_1 and the other X 's may be dependent (confounded), we test $H_0 : X_1$ is independent of X_2, \dots, X_d, Y . Let P_0 denote this null distribution. If we use the linear model \mathcal{P} , H_0 is translated into $H_0(\mathcal{P}) : \beta_1 = 0$, where, as we have seen,

$$\beta_1 = \sum_{k=1}^d \sigma^{1k} \text{Cov}(X_k, Y)$$

with $(\sigma^{jk})_{d \times d} \equiv \Sigma_X^{-1}$. Thus, H_0 implies $H_0(\mathcal{P})$, but not vice versa. The hypothesis $H_0(\mathcal{P})$ states that in the linear model closest to the true P , we cannot detect any relationship between X_1 and Y after controlling for the other X 's.

Under H_0 , the sandwich formula simplifies. In particular, as $n \rightarrow \infty$,

$$\text{Var}_{H_0}(\hat{\beta}_1) \asymp \frac{\sigma_e^2}{n\sigma_1^2},$$

where $\sigma_1^2 = \text{Var}(X_1)$ and σ_e^2 is computed under H_0 . The implication of this is that:

Proposition 2.4. *For testing H_0 , the usual t -test based on $t_1 = \hat{\beta}_1 / \text{SE}(\hat{\beta}_1)$ with $\text{SE}(\hat{\beta}_1) = \hat{\sigma}_e / n^{1/2} \hat{\sigma}_1$ has an asymptotic standard normal distribution provided only that P_0 satisfies (A.1). That is, even when the null distribution P_0 is outside the linear model box \mathcal{P} .*

2.3. Prediction

The advantage of parsimonious parametric models is that they give simple interpretable formulas that connect a response Y to covariates \mathbf{X} . In particular, the model provides formulae for predicting the response Y_0 of a case with covariate vector \mathbf{X}_0 . In our framework, we introduce $e_0 = Y_0 - \boldsymbol{\beta}^T \mathbf{X}_0$ and can then write

$$Y_0 = \boldsymbol{\beta}^T \mathbf{X}_0 + e_0,$$

where the distribution of e_0 given $\mathbf{X}_0 = \mathbf{x}_0$ depends on \mathbf{x}_0 whenever $P \notin \mathcal{P}$. However, for all P satisfying (A.1), when \mathbf{X}_0 is random, $E_P(e_0) = 0$. Thus, a population predictor for Y_0 is $\boldsymbol{\beta}^T \mathbf{X}_0$, with corresponding empirical predictor

$$\hat{Y}_0 = \hat{\boldsymbol{\beta}}^T \mathbf{X}_0.$$

The accuracy of the predictor can be judged by cross-validation, bootstrapping, or using 50% (say) of the sample (a training sample) to estimate $\boldsymbol{\beta}$, and the remaining 50% (a test sample) to compute residuals $\hat{Y}_i - \hat{\boldsymbol{\beta}}^T \mathbf{X}_i, i = [n/2], \dots, n$, that can be used to determine prediction accuracy and prediction intervals. These automatically incorporate lack of model fit, that is, when P is outside the box.

3. Transformation and single index models. Box–Cox revisited and rebutted, revisited

Again we observe $(X_{i1}, \dots, X_{id}, Y_i)$, $i = 1, \dots, n$, i.i.d. as $(\mathbf{X}^T, Y) = (1, X_1, \dots, X_d, Y)$ and assume that (\mathbf{X}^T, Y) has an unknown distribution P , except now the parameter $\boldsymbol{\beta}$ is defined as the coefficient vector of the closest linear transformation model to P . Thus let $Y^{(\lambda)} = g(Y, \lambda)$ be an increasing differentiable transformation of Y with derivative $g'(y, \lambda)$ with respect to y and set

$$\boldsymbol{\beta}(\lambda), \sigma^2(\lambda) \equiv \arg \min \{ \tau^{-2} E_P [Y^{(\lambda)} - \mathbf{a}^T \mathbf{X}]^2 + \log \tau^2 : \mathbf{a} \in \mathbb{R}^{d+1}, \tau^2 > 0 \}, \tag{3.1}$$

$$\lambda_0 \equiv \lambda_0(P) = \arg \min \{ \frac{1}{2} \log \sigma^2(\lambda) - E_P [\log g'(Y, \lambda)] : \lambda \in \mathcal{A} \}, \tag{3.2}$$

where $\log g'(Y, \lambda)$ is the Jacobian term in transformation model and \mathcal{A} is the range of the transformation parameter λ . The coefficient parameter is

$$\boldsymbol{\beta}_0 = (E[\mathbf{X}\mathbf{X}^T])^{-1} E(\mathbf{X}Y^{(\lambda_0)}).$$

The variance of the residual $e = Y^{(\lambda_0)} - \boldsymbol{\beta}_0^T \mathbf{X}$ is

$$\sigma_0^2 \equiv \text{Var}_P(e) = E_P [Y^{(\lambda_0)} - \boldsymbol{\beta}_0^T \mathbf{X}]^2.$$

The box \mathcal{P} is now the class of linear transformation models, that is, the class of probability distributions $P_{\boldsymbol{\alpha}, \lambda, \sigma^2}$ of (X_1, \dots, X_d, Y) where for some $\lambda \in \mathcal{A}$, $\boldsymbol{\alpha} \in \mathbb{R}^{d+1}$, and ε independent of X_1, \dots, X_d ,

$$g(Y, \lambda) = \sum_{j=0}^d \alpha_j X_j + \varepsilon.$$

The true distribution P of (X_1, \dots, X_d, Y) is unknown and satisfies (A.1) as in Section 2. Our parameter vector $\boldsymbol{\beta}_0$ is the coefficient vector of the linear transformation model $P_{\boldsymbol{\beta}_0, \lambda_0, \sigma_0^2}$ closest to the true probability distribution P . Here “closest” is in the sense of minimizing the expressions in (3.1) and (3.2). These are intuitive distances if our goal is to predict $Y^{(\lambda)}$ or estimate $E(Y^{(\lambda)} | \mathbf{X} = \mathbf{x})$. If $Y^{(\lambda)} - \boldsymbol{\beta}_0^T(\lambda)\mathbf{X}$ is independent of \mathbf{X} and has a normal distribution, then (3.1) and (3.2) are proportional to expected log likelihoods, that is, to K–L divergences.

More generally let $h_\lambda(\mathbf{y})$ denote the density of $\mathbf{Y}^{(\lambda)} = (Y_1^{(\lambda)}, \dots, Y_n^{(\lambda)})^T$ given $\mathbf{X} = \mathbf{x}$ and let $f(\mathbf{y}; \boldsymbol{\beta}, \sigma)$ be a multivariate normal distribution with mean $\boldsymbol{\beta}^T \mathbf{x}$ and covariance matrix $\sigma^2 I_n$ where I_n is the $(n \times n)$ identity matrix. Define the K–L discrepancy

$$I(f, h) = \int h_\lambda(\mathbf{y}) \{ \log h_\lambda(\mathbf{y}) - \log f(\mathbf{y}, \boldsymbol{\beta}, \sigma) \} d\mathbf{y}.$$

Let $g(y, \lambda)$ be the Box–Cox transformation. If we select $\lambda_*, \boldsymbol{\beta}_*$, and σ_* to minimize this discrepancy, then Hernandez and Johnson (1980) show that $\boldsymbol{\beta}_* = (X_D^T X_D)^{-1} X_D^T E(Y^{(\lambda_*)} | \mathbf{X})$. Moreover, the arguments of Hernandez and Johnson show that $\lambda_* = \lambda_*(\mathbf{x})$ is the minimizer of (3.2) when P is replaced by the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$. See also Yeo and Johnson (2001).

3.1. Identifiability, interpretability, and stability

Box and Cox (1982) in the rebuttal of Bickel and Doksum (1981) have pointed out that if λ_0 is unknown, then $\boldsymbol{\beta}(\lambda_0)$ is a vector of regression coefficients on an unknown scale, and therefore could be difficult to interpret. However, Brillinger (1983) in the case of normally distributed X 's and Stoker (1986), generally, have shown that independent of λ_0 , $\beta_j(\lambda_0)$ has an interpretation as the average relative change in $E(Y | \mathbf{X} = \mathbf{x})$ as x_j is perturbed. Here is a summary (see also Johnson and Doksum, 2002): assume that there exists a function h and vector $\boldsymbol{\alpha}$ such that

$$E_P(Y | \mathbf{X}) = h(\boldsymbol{\alpha}^T \mathbf{X}). \tag{3.3}$$

Then, if ∇ denotes the gradient with respect to \mathbf{X} and $\|\cdot\|$ denotes the Euclidean norm:

$$\boldsymbol{\alpha} = \frac{E[\nabla E(Y|\mathbf{X})]}{\|E[\nabla E(Y|\mathbf{X})]\|}.$$

Thus, in model (3.3) which includes the Box–Cox model, $\boldsymbol{\alpha}$ is identifiable, and it is interpretable as a coefficient vector that gives the average directional change in $E_P(Y|\mathbf{X})$ as \mathbf{X} is perturbed. Model (3.3) has generated a lot of work in statistics and econometrics. It is called a *single index model* (Stoker, 1986) and its analysis is related to *projection pursuit* (Friedman and Stuetzle, 1981). In our framework, (3.3) defines the box \mathcal{P} and our analysis is aimed at the parameter $\boldsymbol{\alpha} = \boldsymbol{\beta}(P)$ of the distribution in \mathcal{P} closest to the true P .

This discussion is related to the idea of a *stable parameter*. This is a parameter that has a stable value and interpretation for a wide variety of models and values of the nuisance parameters in these models. Some references are Brillinger (1983), Stoker (1986), Cox and Reid (1987), Chen et al. (2002), and Johnson and Doksum (2002).

3.2. Asymptotics. Confidence regions

For the Box–Cox model \mathcal{P} , the sandwich formula for $P \notin \mathcal{P}$ was derived by Hernandez and Johnson (1980) and Cho et al. (2001). A general treatment was given by Huber (1967). Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a parametric model with vector parameter θ for the random vector \mathbf{Z} . Let $l(\mathbf{z}; \theta)$ denote the log likelihood for the i.i.d. sample Z_1, \dots, Z_n from P_θ . In the Box–Cox model, $\mathbf{Z} = (\mathbf{X}, Y)$, $\theta = (\boldsymbol{\beta}, \sigma, \lambda)$, and

$$l(\theta; \mathbf{Z}) = -\frac{1}{2}n \log(2\pi) - n \log \sigma - \frac{1}{2}\sigma^{-2} \sum_{i=1}^n [Y_i^{(\lambda)} - \boldsymbol{\beta}^T \mathbf{X}_i]^2 + (\lambda - 1) \sum \log Y_i. \quad (3.4)$$

Now θ and $\hat{\theta}$ are the minimizers of $E(l(\theta; \mathbf{Z}))$ and $l(\theta; \mathbf{Z})$ where l is given in (3.4). It turns out that $\sqrt{n}(\hat{\beta}_0 - \beta_0)$ is difficult. We obtain a simple argument and result by considering $\boldsymbol{\beta}_d = (\beta_1, \dots, \beta_d)^T \equiv (\beta_1(\lambda_0), \dots, \beta_d(\lambda_0))^T$. The asymptotic distribution of $\hat{\boldsymbol{\beta}}_d = (\hat{\beta}_1, \dots, \hat{\beta}_d)^T$ is obtained by writing $\hat{\boldsymbol{\beta}}_d = \hat{\Sigma}_X^{-1} \hat{\Sigma}_{X\hat{Y}}$ where $\hat{\Sigma}_X$ is the $d \times d$ covariance matrix $n^{-1} \mathbf{X}_C^T \mathbf{X}_C$, $\mathbf{X}_C = (\mathbf{X}_{ij} - \bar{X}_j)_{n \times d}$, and $\hat{\Sigma}_{X\hat{Y}}$ is the $d \times 1$ covariance vector $n^{-1} \sum_{i=1}^n (\mathbf{X}_{ij} - \bar{X}_j) Y_i^{(\lambda)}$, $1 \leq j \leq d$. By the δ method, if we set

$$Y_i^{(\lambda)} = Y_i^{(\lambda_0)} + W_i^{(\lambda_0)}(\hat{\lambda} - \lambda_0) + o_p(n^{-1/2}),$$

where $W_i^{(\lambda)} = (\partial/\partial\lambda)Y_i^{(\lambda)}$, and use Slutsky's theorem, the LLN and the CLT, then we find, under regularity conditions,

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_d) &= \Sigma_X^{-1} \{n^{-1/2} \mathbf{X}_C^T [\mathbf{Y}^{(\lambda_0)} - \mathbf{X}_C \boldsymbol{\beta}_d]\} + \sqrt{n}(\hat{\lambda} - \lambda_0) \Sigma_X^{-1} \{n^{-1} \mathbf{X}_C^T \mathbf{W}^{(\lambda_0)}\} \\ &+ o_p(1) \rightarrow \mathcal{L} V_1 + \mathbf{c} V_2, \end{aligned} \quad (3.5)$$

where V_1 is $\mathcal{N}(0, \Sigma_d)$, $\Sigma_d = \Sigma_X^{-1} \Sigma_{Xe_0} \Sigma_X^{-1}$, $\Sigma_X = (\text{Cov}(X_i, X_j))_{d \times d}$, $\Sigma_{Xe_0} = E(e_0^2 (X_j - \mu_j)(X_h - \mu_h))_{d \times d}$, $e_0 = Y^{(\lambda_0)} - \boldsymbol{\beta}_d^T(\lambda_0) \mathbf{X}_d$, $\mathbf{X}_d = (\mathbf{X}_1 - \mu_1, \dots, \mathbf{X}_1 - \mu_d)^T$, $\mu_j = E(X_j)$, V_2 is the normal weak limit of $\sqrt{n}(\hat{\lambda} - \lambda_0)$, and

$$\mathbf{c} = \boldsymbol{\beta}_d^*(\lambda_0)$$

with $\boldsymbol{\beta}_d^*(\lambda_0) = \Sigma_X^{-1} \Sigma_{XW_0}$, $W_0 = W^{(\lambda_0)}$.

Also note that by applying Slutsky's theorem for matrices and thereby replacing $\hat{\Sigma}_X^{-1}$ by Σ_X^{-1} up front, and by ignoring the estimation of β_0 , we have an $o_p(1)$ remainder term that is relatively easy to control leading to simpler regularity conditions than other approaches e.g. Hernandez and Johnson (1980) and Chen et al. (2002).

In particular, we find the following conditions under which there is no inflation in the variance of the regression coefficients $\hat{\beta}_j$, $j = 1, \dots, d$, because λ_0 is unknown and has to be estimated.

Proposition 3.1. (a) Assume that P satisfies (A.1) with $Y^{(\lambda_0)}$ in place of Y , and assume that \mathbf{X} and Y are independent under P . Also assume that under P , Σ_{XW_0} exists, $\hat{\lambda} \rightarrow_p \lambda_0$, and that $W^{(\lambda)}$ exists and is continuous for λ

in a neighborhood of λ_0 . Then, under P ,

$$\sqrt{n}\hat{\boldsymbol{\beta}}_d \rightarrow \mathcal{L} \mathcal{N}(\mathbf{0}, \Sigma_d).$$

(b) Let P satisfy the conditions of (a). Suppose P_n is a sequence of probabilities contiguous to P with β_j tending to zero at the rate $n^{-1/2}$, $j = 1, \dots, d$. Then under P_n , $\sqrt{n}(\hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_d) \rightarrow \mathcal{L} \mathcal{N}(\mathbf{b}, \Sigma_d)$, where \mathbf{b} is the limit of $\sqrt{n}\boldsymbol{\beta}_d$.

Outline of proof. (a) holds because by mean value theorem $Y_i^{(\hat{\lambda})} = Y_i^{(\lambda_0)} + W_i^{(\lambda^*)}(\hat{\lambda} - \lambda_0)$ with $|\lambda^* - \lambda_0| \leq |\hat{\lambda} - \lambda_0|$, and $n^{-1}X_C^T W^{(\lambda_0)} = O_p(n^{-1/2})$ when X and Y are independent. (b) holds because if the remainder term in (a) tends to zero under independence, then it tends to zero for P_n contiguous to independence. For contiguity, see e.g. van der Vaart (1998) and Bickel and Doksum (2008).

Note that the simplification in the proof occurs because $\mathbf{c} = \mathbf{0}$ when X and Y are independent. No such simplification occurs if we include the intercept β_0 in the parameter vector.

To obtain approximately valid confidence regions, we need good estimates of the covariance matrix of $\hat{\boldsymbol{\beta}}_d$. We conduct an MC study of three candidates with $d = 1$ where the true distribution P satisfies

$$Y^{(\lambda)} = (1 - \gamma)(\alpha_0 + \alpha_1 X) + \gamma[L(X_1)] + \varepsilon, \tag{3.6}$$

where all parameter settings are the same as in (2.2) except $Y^{(\lambda)} = \lambda^{-1}(Y^\lambda - 1)$, $\lambda = 1$. Using the Box–Cox model, we compute MC $\hat{\beta}_0(\hat{\lambda})$, $\hat{\beta}_1(\hat{\lambda})$, $SE(\hat{\beta}_1(\hat{\lambda}))_{MC}$, $SE(\hat{\beta}_1(\hat{\lambda}))_{NAIVE}$, $SE(\hat{\beta}_1(\hat{\lambda}))_{NSA}$, and $SE(\hat{\beta}_1(\hat{\lambda}))_{BOOT}$. As in Section 2.1, our sandwich estimate is $\hat{\Sigma}_d = \hat{\Sigma}_X^{-1} \hat{\Sigma}_{X\varepsilon_0} \hat{\Sigma}_X^{-1}$ and thus is a naive sandwich (NSA) estimate because it leaves out the cV_2 term of the asymptotic expression (3.5) and is only valid under the conditions of Proposition 3.1(b).

The MC $\hat{\beta}_0(\hat{\lambda})$ and $\hat{\beta}_1(\hat{\lambda})$ are plotted against the nonlinearity parameter γ in Fig. 3. In this case these coefficients are unstable for small σ (as in Bickel and Doksum, 1981) and stable for large σ .

In Fig. 4 we compare the naive SE of $\hat{\beta}_1(\hat{\lambda})$ which is based on assuming that λ_0 is known and that P is in the box, the NSA SE which is based on assuming that λ_0 is known, and the nonparametric bootstrap to the “true” SE which is represented by the MC SE. We have omitted the actual sandwich SE because it involves unwieldy expressions and this is a situation where the bootstrap might be a good solution. Fig. 4 shows that the bootstrap does well, but that it is a little conservative.

3.3. Testing

As in Section 2.2, if we are interested in testing whether X_1 and Y are related while controlling for X_2, \dots, X_d we test $H_0 : X_1$ is independent of X_2, \dots, X_d, Y . In this case, Σ_X and $\Sigma_{X\varepsilon_0}$ have zeroes under H_0 wherever X_1 appears. Thus, in formula (3.5), $c_1 = 0$ and

$$\sqrt{n}[\hat{\boldsymbol{\beta}}_1(\hat{\lambda}) - \boldsymbol{\beta}_1(\lambda_0)] \rightarrow \mathcal{L} \mathcal{N}(0, \sigma_0^2)$$

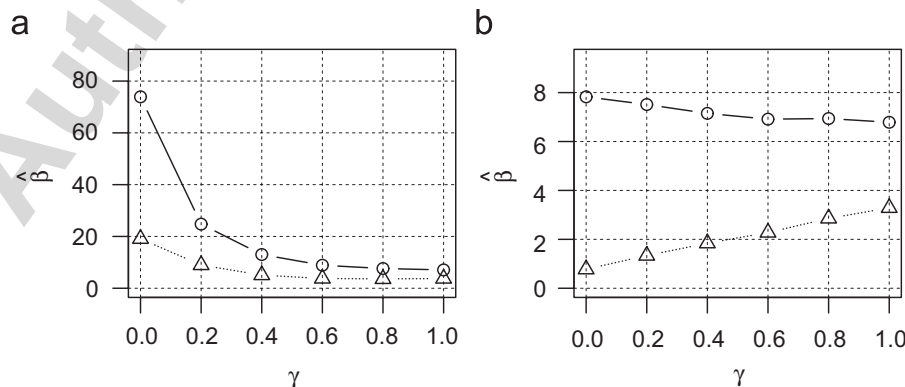


Fig. 3. Monte Carlo values of $\hat{\beta}_0(\hat{\lambda})$ (○) and $\hat{\beta}_1(\hat{\lambda})$ (△) plotted against the nonlinearity parameter γ . (a) $\sigma^2 = 0.1$ and (b) $\sigma^2 = 1.183$.

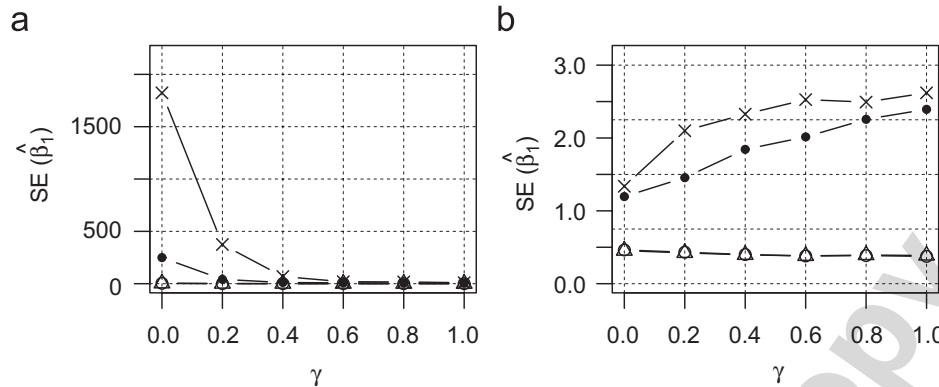


Fig. 4. The naive (○), naive sandwich (NSA △), bootstrap (BOOT ×) and Monte Carlo (MC ●) $\hat{\beta}_1(\hat{\lambda})$ standard errors plotted against the nonlinearity parameter γ for the Box–Cox model \mathcal{P} . Naive and NSA coincide. Here $P \notin \mathcal{P}$ for $\gamma > 0$. (a) $\sigma^2 = 0.1$ and (b) $\sigma^2 = 1.183$.

with $\sigma_0^2 = \sigma_e^2 / \text{Var}(X_1)$, $e = Y^{(\lambda_0)} - \beta^T(\lambda_0)X$. Thus, for testing H_0 , let

$$n\text{SE}^2(\hat{\beta}_1) = \left[\frac{1}{n-d-1} \sum_{i=1}^n (Y_i^{(\hat{\lambda})} - \hat{\beta}^T X_i)^2 \right] / \left[\frac{1}{n-1} \sum_{i=1}^n (X_{i1} - \bar{X}_1)^2 \right],$$

then the usual t -statistic $t_1 = \hat{\beta}_1(\hat{\lambda}) / \text{SE}(\hat{\beta}_1(\hat{\lambda}))$ has an asymptotic standard normal distribution under H_0 . We can ignore both the fact that $P \notin \mathcal{P}$ and that λ_0 needs to be estimated. The inflation in the variance is automatically corrected for by the t -test. See Doksum and Wong (1983) for similar results.

4. Binary regression

We observe $(X_{i1}, \dots, X_{id}, Y_i)$, $i = 1, \dots, n$, i.i.d. as (X_1, \dots, X_d, Y) where $Y \in \{0, 1\}$. Let \mathcal{P} be the class of logistic regression models for (X_1, \dots, X_d, Y) , that is, the class of probabilities $Q = Q_\alpha$, $\alpha \in R^{d+1}$, with

$$Q_\alpha(Y = 1 | X = x) = g(\alpha^T x),$$

where $g(t) = [1 + \exp\{-t\}]^{-1}$, $X = (X_0, X_1, \dots, X_d)^T$, $X_0 = 1$, $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_d)^T$, and we assume the usual identifiability conditions for Q_α . For any probability P for (X_1, \dots, X_d, Y) satisfying (A.1), define

$$\beta = \beta(P) = \arg \min \{K(Q_\alpha, P) : Q_\alpha \in \mathcal{P}, \alpha \in R^{d+1}\},$$

where K is the K–L discrepancy. Note that

$$\log L(X, Y; Q_\alpha) = Y \log g(\alpha^T X) + (1 - Y) \log [1 - g(\alpha^T X)]$$

and

$$\beta(P) = \arg \max_{\alpha} E_P[\log L(X, Y; Q_\alpha)].$$

Our estimate $\hat{\beta} = \beta(\hat{P})$ is the usual logistic regression analysis estimate

$$\hat{\beta} = \arg \max_{\alpha} n^{-1} \sum_{i=1}^n \log L(X_i, Y_i; Q_\alpha).$$

It is known (e.g. Bickel and Doksum (2007, Section 6.4.3)) that with $m = (g(\alpha^T X_1), \dots, g(\alpha^T X_n))^T$, $\hat{\beta}$ solves $X_D^T(Y - m) = 0$, or equivalently $\hat{\beta}$ solves

$$\frac{1}{n} \sum_{i=1}^n \psi(X_i, Y_i; \alpha) = \mathbf{0},$$

where $\psi(X_i, Y_i; \alpha) = (X_{ij}e_i)_{0 \leq j \leq d}$ and $e_i = Y_i - m_i$. It follows that β is the solution to

$$E_P\{X_j[Y - g(\alpha^T X)]\} = 0, \quad j = 0, \dots, d.$$

Asymptotic theory (e.g. Huber, 1967; Shao, 2003, Section 5.4.3; Bickel and Doksum, 2007, Section 6.2.1) gives

$$n^{1/2}(\hat{\beta} - \beta) \rightarrow \mathcal{L} \mathcal{N}(0, \Sigma_P), \quad \Sigma_P = \Sigma_{IX}^{-1} \Sigma_{Xe} \Sigma_{IX}^{-1},$$

$$\Sigma_{IX} = E_P(l(\beta^T X) X_j X_k), \quad \Sigma_{Xe} = E_P(e^2 X_j X_k), \quad 0 \leq j \leq d, \quad 0 \leq k \leq d,$$

where $l(t) = \exp\{-t\}[1 + \exp\{-t\}]^{-2}$ is the logistic density. For $P = Q \in \mathcal{P}$, $\Sigma_Q = \Sigma_{Xe}^{-1}$ with $e = Y - g(\beta^T X)$ where

$$E_Q(e) = E_Q(E_Q(Y - g(\beta^T X)|X)) = E_Q[g(\beta^T X) - g(\beta^T X)] = 0.$$

We estimate Σ_P for general P using the sandwich estimate

$$\hat{\Sigma}_P = \hat{\Sigma}_{IX}^{-1} \hat{\Sigma}_{Xe} \hat{\Sigma}_{IX}^{-1},$$

where $\hat{\Sigma}_{Xe}$ is $n^{-1} Z_D^T Z_D$ with $Z_D = (X_{ij} \hat{e}_i)_{1 \leq i \leq n, 0 \leq j \leq d}$, $\hat{e}_i = Y_i - g(\hat{\beta}^T X_i)$ and $\hat{\Sigma}_{IX} = n^{-1} W_D^T W_D$ with $W_D = (X_{ij} l^{1/2}(\hat{\beta}^T X_i))_{n \times (d+1)}$. The estimate $\hat{\Sigma}_P$ can be used to construct confidence intervals for β_j and to test $H_0 : \beta_j = 0$.

We conducted an MC study to access the variance adjustment given by the sandwich estimate when the true P satisfies

$$(1 - \gamma) \prod_{i=1}^n \frac{[F(\alpha^T X_i)]^{Y_i}}{[1 - F(\alpha^T X_i)]^{Y_i-1}} + \gamma \prod_{i=1}^n \frac{[G(\alpha^T X_i)]^{Y_i}}{[1 - G(\alpha^T X_i)]^{Y_i-1}}, \tag{4.1}$$

where

$$F(\alpha^T X_i) = (1 + \exp\{-(\alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2})\})^{-1},$$

$$G(\alpha^T X_i) = \begin{cases} 0.1 & \text{if } (\alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2}) < 0, \\ 0.9 & \text{if } (\alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2}) \geq 0, \end{cases}$$

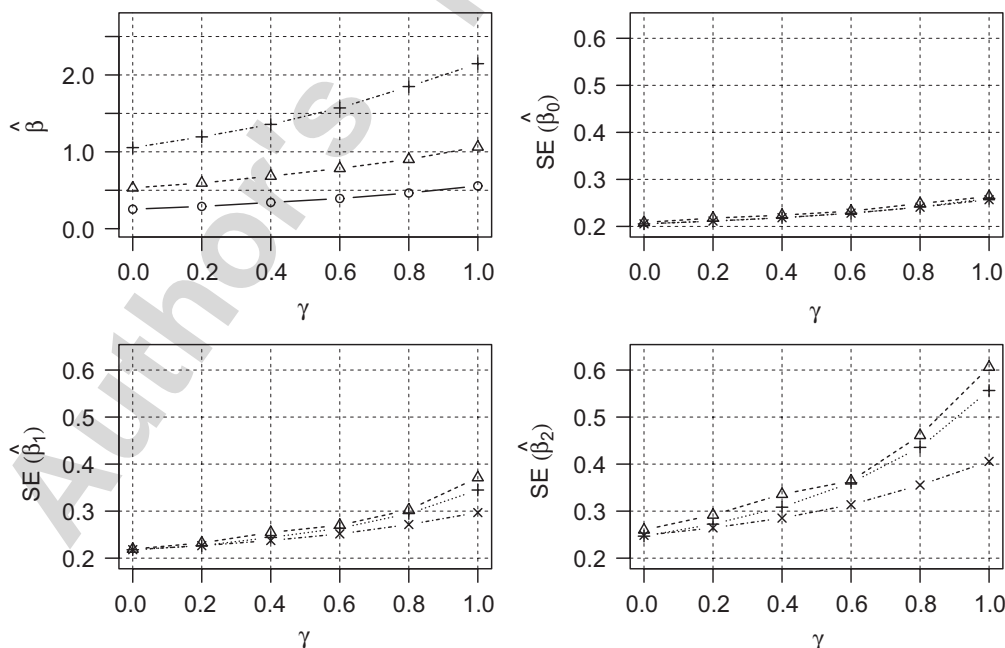


Fig. 5. Upper left panel presents Monte Carlo values of $\hat{\beta}_0$ (\circ), $\hat{\beta}_1$ (Δ), $\hat{\beta}_2$ ($+$) plotted against the mixture parameter γ . Upper right, lower left, and lower right panels represent the naive (NAIVE \times), sandwich (SA $+$), and Monte Carlo (MC Δ) standard errors as functions of the mixture parameter γ .

X_{i1} and X_{i2} are standard normal random variables, $n = 128$, $\alpha_0 = 0.25$, $\alpha_1 = 0.5$, and $\alpha_2 = 1$. We simulated from the above model 1000 times and compared the naive and sandwich (SA) SE's to the “true” SE for model (4.1) in Fig. 5, where the “true” SE is represented by the Monte Carlo (MC) SE.

The inspection of the upper left panel of Fig. 5 shows that the MC estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ diverge from the working model parameter values as γ increases. The general trend observed in the lower left and lower right panels is that when $\gamma = 0$ the MC estimates are similar, and as γ increases $SE(\hat{\beta}_j)_{SA}$ inflate in a similar way to $SE(\hat{\beta}_j)_{MC}$ whereas $SE(\hat{\beta}_j)_{NAIVE}$ show smaller inflation. The fact that all estimates in the upper right panel are close suggests that for small signal (recall that $\beta_0 = 0.25$) all estimates perform in a similar way.

4.1. Classification

Suppose (X_0, Y_0) is a random pair distributed as (X, Y) but only $X_0 = \mathbf{x}_0$ is observed. On the basis of observing \mathbf{x}_0 we are to classify Y_0 as zero or one. If $P \in \mathcal{P}$, $\boldsymbol{\beta}$ is known and the prior probability that $Y = 0$ is π , the optimal Bayes decision rule for 0–1 loss (e.g. Bickel and Doksum, 2007, Section 3.2) is “decide $Y_0 = 1$ iff $g(\boldsymbol{\beta}^T \mathbf{x}_0) > \pi$ ”. This suggests the following rule “classify Y_0 as 1 iff $g(\hat{\boldsymbol{\beta}}^T \mathbf{x}_0) > \pi$ ”. Note that for $P \in \mathcal{P}$, $g(\hat{\boldsymbol{\beta}}^T \mathbf{x}_0)$ is the MLE of $g(\boldsymbol{\beta}^T \mathbf{x}_0)$ and in this case is an asymptotically efficient estimate of $P(Y_0 = 1 | X_0 = \mathbf{x}_0) = g(\boldsymbol{\beta}^T \mathbf{x}_0)$.

5. Models with both response and covariate transformations

Consider the model \mathcal{P} for a regression experiment where after transformations of both Y and the X 's, a normal linear model holds. This leads to models of the form

$$h^{(\lambda)}(Y) = \alpha_0 + \sum_{j=1}^n \alpha_j h_j^{(\lambda_j)}(X_j) + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$. A problem with this model is that when $\alpha_j = 0$, then λ_j is not identifiable (see e.g. Fan et al., 2007). Instead of this approach we will use results of Hernandez and Johnson (1980) in combination with the empirical K–L idea of Section 1 to develop a transformation to normality approach for regression. Write $\mathbf{Z} \equiv (Z_1, \dots, Z_{d+1})^T \equiv (X_1, \dots, X_d, Y)^T$ where Y is the response and X_1, \dots, X_d are covariates. Now let $\mathbf{Z}^{(\lambda)} = (Z_1^{(\lambda_1)}, \dots, Z_{d+1}^{(\lambda_{d+1})})^T$ where $Z_j^{(\lambda_j)}$ is a 1–1 transformation of Z_j . Further let $P^{(\lambda)}$ denote the distribution of $\mathbf{Z}^{(\lambda)}$ and let $\Phi_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}$ be a $d + 1$ variate normal distribution with mean vector $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$. Now \mathcal{P} is the class for which there exist $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ such that $\mathbf{Z}^{(\lambda)} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For $(X, Y) \sim P$, we find $\boldsymbol{\lambda}^* \equiv \boldsymbol{\lambda}(P)$, $\boldsymbol{\mu}^* \equiv \boldsymbol{\mu}(P)$, and $\boldsymbol{\Sigma}^* \equiv \boldsymbol{\Sigma}(P)$ that make $P^{(\lambda)}$ and $\Phi_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}$ as close as possible in the K–L sense. The choice of the normal distribution is natural because the point of regression modeling is to introduce interesting interpretable parameters that indicate the strength of relationships between variables. Because the K–L discrepancy is equivariant under 1–1 transformations (e.g. Bickel and Doksum, 2007, p. 169), $\boldsymbol{\lambda}^*$, $\boldsymbol{\mu}^*$, and $\boldsymbol{\Sigma}^*$ also minimize the K–L discrepancy between P and \mathcal{P} . If $P \in \mathcal{P}$, then $\mathbf{Z}^{(\lambda^*)}$ is exactly multivariate normal, and the regression $E(Z_{d+1}^{(\lambda_{d+1}^*)} | Z_1^{(\lambda_1^*)}, \dots, Z_d^{(\lambda_d^*)})$ of $Z_{d+1}^{(\lambda_{d+1}^*)}$ on $Z_1^{(\lambda_1^*)}, \dots, Z_d^{(\lambda_d^*)}$ is linear. Thus, the parameter of interest is the coefficient vector $\boldsymbol{\beta}(P) = (\beta_0(P), \dots, \beta_d(P))^T$ in this linear regression as given in Section 2 with (X_1, \dots, X_d, Y) replaced by the transformed variables.

When $Z_j^{(\lambda_j)}$, $j = 1, \dots, d + 1$, are Box–Cox transformations, we find, using Hernandez and Johnson (1980), the following profile K–L solutions: first fix $\boldsymbol{\lambda}$ and minimize the K–L discrepancy over $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$:

$$\boldsymbol{\mu}^{(\lambda)}, \boldsymbol{\Sigma}^{(\lambda)} \equiv \arg \inf_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \{K(\Phi_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}, P^{(\lambda)})\} = E_P(\mathbf{Z}^{(\lambda)}), \text{Cov}_P(\mathbf{Z}^{(\lambda)}). \quad (5.1)$$

Next, according to Hernandez and Johnson (1980), we can find $\boldsymbol{\lambda}(P)$ as

$$\boldsymbol{\lambda}(P) = \arg \inf_{\boldsymbol{\lambda}} \{(\mathbf{1} - \boldsymbol{\lambda})^T E_P[\log(\mathbf{Z})] + \frac{1}{2} \log(\det[\text{Cov}_P(\mathbf{Z}^{(\lambda)})])\}. \quad (5.2)$$

The solution to (5.2) is substituted in (5.1) to give $\mu(P)$, $\Sigma(P)$, and $\beta(P)$. Now, because P is unknown we replace it with the empirical distribution \hat{P} which yields estimates that coincide with the classical linear model estimates with the original data replaced with the transformed data. Inference can be carried out using the nonparametric bootstrap.

References

- Bickel, P.J., Doksum, K.A., 2007. *Mathematical Statistics: Basic Ideas and Selected Topics*, second ed., vol. I. Pearson Prentice-Hall, Upper Saddle River, NJ (updated Printing).
- Bickel, P.J., Doksum, K.A., 2008. *Mathematical Statistics: Basic Ideas and Selected Topics*, vol. II. Pearson Prentice-Hall, Upper Saddle River, NJ.
- Bickel, P.J., Doksum, K.A., 1981. An analysis of transformations revisited. *J. Amer. Statist. Assoc.* 76, 296–311.
- Bickel, P.J., Ritov, Y., 2003. Non-parametric estimators which can be ‘plugged-in’. *Ann. Statist.* 31, 1033–1053.
- Bickel, P.J., Klaassen, C.A.J., Ritov, Y., Wellner, J.A., 1993, 1998. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York, NY.
- Box, G.E.P., 1979. Sampling and Bayes inference in scientific modeling and robustness (with discussion). *J. Roy. Statist. Soc.* 30, 383–430.
- Box, G.E.P., Cox, D.R., 1964. An analysis of transformations (with discussion). *J. Roy. Statist. Soc. Ser. B* 26, 211–252.
- Box, G.E.P., Cox, D.R., 1982. An analysis of transformations revisited, rebutted. *J. Amer. Statist. Assoc.* 77, 209–210.
- Brillinger, D.R., 1983. A generalized linear model with “Gaussian” regressor variables. In: Bickel, P.J., Doksum, K.A., Hodges, J.L., Jr. (Eds.), *Festschrift for Erich L. Lehmann*. Wadsworth, Belmont, CA, pp. 97–114.
- Chen, G., Lockhart, R.A., Stephens, M.A., 2002. Box–Cox transformations in linear models: large sample theory and test of normality. *Canad. J. Statist.* 30, 177–209.
- Cho, K., Yeo, I., Johnson, R.A., Loh, W.-Y., 2001. Asymptotic theory for Box–Cox transformations in linear models. *Statist. Probab. Lett.* 51, 337–343.
- Cover, T.M., Thomas, J.A., 2003. *Elements of Information Theory*. Wiley-Interscience, New York, NY.
- Cox, D.R., Reid, N.M., 1987. Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B* 49, 1–39.
- Doksum, K.A., Samarov, A., 1995. Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *Ann. Statist.* 23 (5), 1443–1473.
- Doksum, K.A., Wong, C.W., 1983. Statistical tests based on transformed data. *J. Amer. Statist. Assoc.* 78, 411–417.
- Fan, C., Fine, J., Jeong, J.-H., 2007. Cox regression model with covariate transformations. Technical Report, Statistics Department, University of Wisconsin Madison, WI.
- Friedman, J.H., Stuetzle, W., 1981. Projection pursuit regression. *J. Amer. Statist. Assoc.* 76, 817–823.
- Hernandez, F., Johnson, R.A., 1980. The large-sample behavior of transformations to normality. *J. Amer. Statist. Assoc.* 75, 855–861.
- Huber, P., 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability I*. University of California Press, Berkeley, CA, pp. 221–233.
- Johnson, R.A., Doksum, K.A., 2002. Comments on Box–Cox transformations in linear models: large sample theory and test of normality. *Canad. J. Statist.* 30, 215–219.
- Owen, A.B., 1988. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75 (2), 237–249.
- Owen, A.B., 2001. *Empirical likelihood*. Chapman & Hall, London.
- Shao, J., 2003. *Mathematical Statistics*, second ed. Springer, New York, NY.
- Stoker, T., 1986. Consistent estimation of scaled coefficients. *Econometrica* 54, 1461–1481.
- van der Vaart, A.W., 1998. *Asymptotic Statistics*. Cambridge University Press, Cambridge.
- Yeo, I., Johnson, R.A., 2001. A new family of power transformations to improve normality or symmetry. *Biometrika* 87, 954–959.