# A note on least squares sensitivity in single-index model estimation and the benefits of response transformations

## Alexandra L. Garnham and Luke A. Prendergast

*Department of Mathematics and Statistics*
*La Trobe University*
*Bundoora, VIC 3086, Australia*
*e-mail:* algarnham@students.latrobe.edu.au*;* luke.prendergast@latrobe.edu.au

**Abstract:** Ordinary Least Squares (OLS) is recognised as being useful in the context of multiple linear regression but can also be effective under the more general framework of the single-index model. In cases where it is ineffective, transformations to the response can improve performance while still allowing for interpretation on the original scale. In this paper we introduce an influence diagnostic for OLS that can be used to assess its effectiveness in the general setting and which can also be used following response transformations. These findings are further emphasized and verified via some simulation studies.

## Contents

## 1. Introduction

It is well known that Ordinary Least Squares (OLS) is an efficient and effective estimator of coefficients from a multiple linear regression (MLR) model. Additionally it is also applicable to a much wider framework of models. Consider a regressor vector $\mathbf{x} \in \mathbb{R}^p$, random response variable $Y \in \mathbb{R}$ and random error term $\epsilon$ assumed to be independent of $\mathbf{x}$ with mean 0. In extending results from Brillinger (1977, 1983), Li & Duan (1989) showed that, for mild conditions on $\mathbf{x}$, the OLS slope vector is equal to $c\boldsymbol{\beta}$ for a $c \in \mathbb{R}$ under the assumed model $Y = f(\boldsymbol{\beta}^\top \mathbf{x}, \epsilon)$. Here the form of $f$ is unknown. However, given that the OLS slope vector, which we will denote as $\mathbf{b}$, is a scalar multiple of $\boldsymbol{\beta}$, a plot of $Y$ versus $\mathbf{b}^\top \mathbf{x}$ may be used to visually explore possibilities for $f$.

Depending on the form of $f$, OLS is not always a good estimator of the direction of $\boldsymbol{\beta}$ in practice. Nevertheless as the error term is not necessarily additive, response transformations are possible that may result in significant improvements in estimation. The purpose of this paper is two-fold. Firstly, we introduce an influence diagnostic for OLS in dimension reduction based on recent results from Prendergast & Smith (2010). This influence diagnostic can be used to assess the effectiveness of OLS as a dimension reduction method under various assumed models. The second purpose is to discuss the suitability of response transformations when using OLS. To do so, we use our influence diagnostic to show that simple response transformations can lead to very notable improvements in estimation.

We discuss OLS in Section 2 and show by example how poorly it can perform. We also explain how response transformations are supported by the theory. In Section 3 we introduce the influence diagnostic for OLS before establishing the associated diagnostic following a response transformation in Section 4. This section affords general insights as to how and when simple transformation functions can improve estimation. Finally, the main findings of the paper are summarized in Section 5 and further research is discussed.

## 2. Sufficient summary plots using ordinary least squares

Consider a random response variable denoted $Y \in \mathbb{R}$ and a vector of predictors $\mathbf{x} \in \mathbb{R}^p$. The MLR model assumes $Y = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x} + \epsilon$ where $\beta_0$ is the unknown intercept coefficient and $\boldsymbol{\beta}$ is the unknown slope vector consisting of regression coefficients for the predictors. Here $\epsilon$ is the random error assumed independent of $\mathbf{x}$ with $E(\epsilon) = 0$. Let $\{Y_i, \mathbf{x}_i\}_{i=1}^n$ denote a random sample of $n$ observations drawn from the MLR model. Then, based on this random sample, OLS is an efficient estimator of both $\beta_0$ and $\boldsymbol{\beta}$. The population OLS slope vector is denoted

$$\mathbf{b} = \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{xy} \tag{1}$$

where $\boldsymbol{\Sigma} = \text{var}(\mathbf{x})$ and $\boldsymbol{\Sigma}_{xy} = \text{cov}(\mathbf{x}, Y)$. Under the MLR conditions $\mathbf{b} = \boldsymbol{\beta}$. The OLS estimator for the slope vector is then $\widehat{\mathbf{b}} = \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\Sigma}}_{xy}$ where $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Sigma}}_{xy}$ are the usual sample-based estimators of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{xy}$ respectively.

### *2.1. The single-index model*

An important development for OLS was provided by Brillinger ([1977](#), [1983](#)) which allowed OLS to be applied to a wider variety of models. Li & Duan ([1989](#)) provided further generalizations and extensions, the result of which is a very diverse range of suitable applications for OLS. Continuing the previous definitions for $Y$, $\boldsymbol{\beta}$, $\mathbf{x}$ and $\epsilon$, the single-index model is of the form

$$Y = f(\boldsymbol{\beta}^\top \mathbf{x}, \epsilon) \tag{2}$$

where $f$ is an unknown *link function* on which there are no conditions (see for example Theorem 2.1 of Li & Duan, [1989](#)). In seeking information regarding $f$, $\mathbf{x} \in \mathbb{R}^p$ can be replaced by $\boldsymbol{\beta}^\top \mathbf{x} \in \mathbb{R}$ without loss of information. Li ([1991](#)) and Carroll & Li ([1992](#)) refer to this as dimension reduction in a visualisation sense as $Y$ is plotted against the lower-dimensional $\boldsymbol{\beta}^\top \mathbf{x}$ enabling inference regarding the form of $f$.

Li & Duan ([1989](#)) gave the following condition, generally known as the Linear Design Condition (LDC):

**Condition 1.** *For any* $\mathbf{d} \in \mathbb{R}^p$, $E(\mathbf{d}^\top \mathbf{x} | \boldsymbol{\beta}^\top \mathbf{x})$ *is linear in* $\boldsymbol{\beta}^\top \mathbf{x}$.

This condition holds when $\mathbf{x}$ is elliptically symmetrically distributed (e.g. see Cook & Weisberg, [1991](#)). Moreover, Hall & Li ([1993](#)) show that Condition [1](#) often holds approximately when $p$ is large. Under Condition [1](#) and when the model in ([2](#)) holds, Li & Duan ([1989](#)) show that $\mathbf{b} = c\boldsymbol{\beta}$ for a $c \in \mathbb{R}$. Consequently, when $c \neq 0$, OLS is expected to find the direction of $\boldsymbol{\beta}$. Since $f$ in ([2](#)) is unknown, it is only the direction of $\boldsymbol{\beta}$ that is important. For instance, for any $\boldsymbol{\gamma}$ which is a non-zero scalar multiple of $\boldsymbol{\beta}$, the model can be re-specified as $Y = g(\boldsymbol{\gamma}^\top \mathbf{x}, \epsilon)$. A plot of $Y$ versus $\boldsymbol{\gamma}^\top \mathbf{x}$ is called a Sufficient Summary Plot (SSP, e.g. see Cook, [1998](#)). The sample-based equivalent is referred to as an Estimated SSP (ESSP) for which the $Y_i$'s are plotted against the $\widehat{\boldsymbol{\gamma}}^\top \mathbf{x}_i$'s such that $\widehat{\boldsymbol{\gamma}}$ is an estimate to the direction of $\boldsymbol{\beta}$. As a consequence, $\widehat{\mathbf{b}}$ can be used to obtain an ESSP in an effort to seek an appropriate link function when Condition [1](#) holds.

### *2.2. When OLS is ineffective as a dimension reduction estimator*

OLS can often work very well under the framework of ([2](#)), yet in some cases it can fail spectacularly even when Condition [1](#) is satisfied. As noted in Section [2.1](#), the OLS slope vector under the required conditions is equal to $\mathbf{b} = c\boldsymbol{\beta}$ for $c \in \mathbb{R}$. To successfully find the direction of $\boldsymbol{\beta}$ it is naturally required that $c \neq 0$. A cause of $c = 0$ is when $f$ is symmetric around $\boldsymbol{\beta}^\top E(\mathbf{x})$ for an elliptically symmetric $\mathbf{x}$. For example, suppose $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_p)$ and $Y = (\boldsymbol{\beta}^\top \mathbf{x})^2 + \epsilon$ so that $\text{cov}(\mathbf{x}, Y) = \mathbf{0}$. OLS is expected to fail since $\mathbf{b} = \mathbf{0}$. However, there are other situations for which OLS will fail that are not so obvious. To show this we will consider the following example.
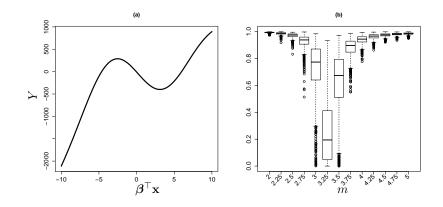
FIG 1. *For Example 2.1.(a) Plot of the model in (3) for which $m = 3.25$. (b) Boxplots of $cor(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\widehat{\mathbf{b}})^2$ plotted as a function of m clearly showing the estimate of the direction of $\boldsymbol{\beta}$ produced by OLS deteriorating in the neighborhood of 3.25. Each simulation is run 1000 times on a sample of 1000 observations. Please refer to main text for interpretation.*

**Example 2.1.** For $\mathbf{x} = [X_1, \ldots, X_{10}]^\top \sim N(\mathbf{0}, \mathbf{I}_{10})$, suppose that

$$Y = (\boldsymbol{\beta}^\top \mathbf{x} - 2)^3 - 400 \sin\left(\frac{1}{2}\boldsymbol{\beta}^\top \mathbf{x}\right) \tag{3}$$

where $\boldsymbol{\beta} = [1, m, 0, \ldots, 0]^\top$ with $m \in \mathbb{R}$.

Initially, we expected OLS to provide a reasonable estimate of the direction of $\boldsymbol{\beta}$ due to the lack of symmetry of the link function about $\boldsymbol{\beta}^\top E(\mathbf{x}) = 0$. For example, plot (a) of Figure 1 depicts $Y$ versus $\boldsymbol{\beta}^\top \mathbf{x}$ when $m = 3.25$ where it is clear symmetric dependency is not evident. However, simulation results show that OLS is typically a poor estimator for choices of $m$ close to 3.25 even for very large sample sizes. To assess OLS as an estimator we will report $cor(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\widehat{\mathbf{b}})^2$ where $\mathbf{X}$ is the matrix whose $i$th row contains the predictor for the $i$th observation. This is the squared correlation between the true and estimated dimension reduced predictors. Plot (b) of Figure 1 provides boxplots of $cor(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\widehat{\mathbf{b}})^2$ for varying $m$ run over 1000 simulations with $n = 1000$ observations. It is clear that as $m$ approaches 3.25, the ability of OLS to successfully find the direction of $\boldsymbol{\beta}$ deteriorates since $cor(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\widehat{\mathbf{b}})^2$ is typically small for $m$ in this vicinity.

To further explore the cause of OLS failing, we consider the population slope vector $\mathbf{b} = \text{cov}(\mathbf{x}, Y)$. For any $m$, it can be shown that

$$\mathbf{b} = \left[15 + 3m^2 - 200 \exp\left(-\frac{1}{8}(1 + m^2)\right)\right] \boldsymbol{\beta}. \tag{4}$$

As Li & Duan (1989) suggest, the OLS slope in (4) is a scalar multiple of $\boldsymbol{\beta}$. However, $\mathbf{b} = \mathbf{0}$ when $m = \pm 3.258$ and close to $\mathbf{0}$ for $m$ in this vicinity. Hence OLS estimates the zero vector for this $m$ resulting in poor performance.

## 2.3. Response transformations

The original results of Brillinger (1977, 1983) assumed that the error term was additive. However, the generalization by Li & Duan (1989) to allow for a non-additive error as in (2) is important when contemplating response transformations. Consider a function $t : \mathbb{R} \mapsto \mathbb{R}$. Under (1), we may write

$$t(Y) = g(\boldsymbol{\beta}^\top \mathbf{x}, \epsilon) \tag{5}$$

for a new and still unknown link function $g$. Now let

$$\mathbf{b}_t = \boldsymbol{\Sigma}^{-1} \text{cov}\left[\mathbf{x}, t(Y)\right] \tag{6}$$

denote the OLS slope vector with respect to the transformed $Y$. Since Condition 1 is unaffected by the transformation, $\mathbf{b}_t$ is a scalar multiple of $\boldsymbol{\beta}$ providing that Condition 1 holds. Two obvious options for SSP's are:

1. Plot $t(Y)$ versus $\mathbf{b}_t^\top \mathbf{x}$ to seek $g$; or
2. Plot $Y$ versus $\mathbf{b}_t^\top \mathbf{x}$ to seek $f$

with similarly defined ESSP's in practice. The second option is generally most appealing as it allows for examination of the predictors and response on the original scale. Consequently, it is important to realize that the transformation need only occur in the step involving the estimation of the direction of $\boldsymbol{\beta}$.

## 3. Influence analysis of OLS for ESSP's

Prendergast (2008) introduced an influence diagnostic that can be used in practice to detect influential observations when using OLS for ESSP's. Here we introduce an influence diagnostic that can be used to provide general insights with respect to various models. The diagnostic shares similarities with Hampel's influence function (Hampel, 1974) and extends on work by Prendergast & Smith (2010) who introduce diagnostics for certain dimension reduction methods.

Consider $(Y, \mathbf{x}) \sim F$ and let the contamination distribution be defined as

$$F_\epsilon = (1 - \epsilon)F + \epsilon \Delta_{(y_0, \mathbf{x}_0)} \tag{7}$$

where $\Delta_{(y_0, \mathbf{x}_0)}$ is the Dirac distribution that places all of its mass at the contamination point $(y_0, \mathbf{x}_0)$, and $\epsilon$ denotes the proportion of contamination.

Now let $T$ be a functional for an estimator defined in both $F$ and $F_\epsilon$. The Influence Function, IF, (Hampel, 1974) is defined by

$$\text{IF}(T, F; y_0, \mathbf{x}_0) = \lim_{\epsilon \downarrow 0} \frac{T(F_\epsilon) - T(F)}{\epsilon} = \left. \frac{\partial}{\partial \epsilon} T(F_\epsilon) \right|_{\epsilon = 0} \tag{8}$$

which can be used to quantify the relative influence of contamination on an estimator. For example, a power series expansion of $T(F_\epsilon)$ gives

$$T(F_\epsilon) = T(F) + \epsilon \, \text{IF}(T, F; y_0, \mathbf{x}_0) + \frac{1}{2}\epsilon^2 \left. \frac{\partial^2}{\partial \epsilon^2} T(F_\epsilon) \right|_{\epsilon = 0} + O(\epsilon^3). \tag{9}$$

When $\epsilon$ is small, $T(F_\epsilon) \approx T(F) + \epsilon \text{ IF}(T, F; y_0, \mathbf{x}_0)$ hence $\text{IF}(T, F; y_0, \mathbf{x}_0)$ can be used to quantify the approximate influence of the contamination. This type of sensitivity analysis has important connections with the empirical setting. For example, $\text{IF}(y_i, \mathbf{x}_i)$ is proportional to the difference in estimation with and without the $i$th observation.

A generalization of the model in (2), although not applicable to OLS, allows for numerous vectors of regressor coefficients. For such a model, an influence diagnostic for some dimension reduction estimators based on average canonical correlations was introduced by Prendergast & Smith (2010). Whilst their diagnostic is applicable to the model in (2), it is not directly relevant to OLS due to scaling differences. For OLS, the influence diagnostic given by Prendergast & Smith (2010) is based on the squared correlation between the $\mathbf{b}^\top \mathbf{x}$ and $\mathbf{b}_\epsilon^\top \mathbf{x}$ where $\mathbf{b}$ is the OLS slope vector at $F$ and $\mathbf{b}_\epsilon$ is the OLS slope vector at $F_\epsilon$. Recall the target for OLS is $c\boldsymbol{\beta}^\top \mathbf{x}$ for any non-zero $c \in \mathbb{R}$. Consequently, if contamination results in $\mathbf{b}_\epsilon = c_\epsilon \boldsymbol{\beta}$ where $c_\epsilon$ is nonzero, then an influence diagnostic based on the squared correlation between $\mathbf{b}^\top \mathbf{x}$ and $\mathbf{b}_\epsilon^\top \mathbf{x}$ will appropriately indicate zero influence. We now consider the following theorem, the proof of which can be found in Appendix A.

**Theorem 1.** *Let b denote the functional for the OLS slope vector estimator where $b(F) = \mathbf{b}$ and $b(F_\epsilon) = \mathbf{b}_\epsilon$. Then, for $\mathbf{x} \sim F$ with $E(\mathbf{x}) = \boldsymbol{\mu}$ and $Cov(\mathbf{x}) = \boldsymbol{\Sigma}$, the power series for $cor^2(\mathbf{b}^\top \mathbf{x}, \mathbf{b}_\epsilon^\top \mathbf{x})$ can be written as*

$$cor^2(\mathbf{b}^\top \mathbf{x}, \mathbf{b}_\epsilon^\top \mathbf{x}) = 1 - \epsilon^2 \rho_b(y_0, \mathbf{x}_0) + O(\epsilon^3)$$

*where, for $\mathbf{P}_b = (\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b})^{-1} \boldsymbol{\Sigma}^{1/2} \mathbf{b} \mathbf{b}^\top \boldsymbol{\Sigma}^{1/2}$ and $r_0 = y_0 - E(Y) - \mathbf{b}^\top (\mathbf{x}_0 - \boldsymbol{\mu})$*

$$\rho_b(y_0, \mathbf{x}_0) = \frac{r_0^2}{\left\| \boldsymbol{\Sigma}^{1/2} \mathbf{b} \right\|^2} \left\| (\mathbf{I}_p - \mathbf{P}_b) \boldsymbol{\Sigma}^{-1/2} (\mathbf{x}_0 - \boldsymbol{\mu}) \right\|^2.$$

**Remark 3.1.** The sensitivity analysis and resulting diagnostic that we consider within this paper is based on Hampels influence function (Hampel, 1974). However, further research could also explore other more general sensitivity analysis approaches that have been considered for least squares. One alternative could be to consider small perturbations of the design matrix $\mathbf{X}$ and vector of responses. Golub & Van Loan (1996) p. 242, for example, detail this approach allowing for the computation of error bounds. Other sensitivity approaches have also been discussed by Chatterjee & Hadi (1988). For example, Chatterjee & Hadi (1988) consider an asymptotic approach which examines the effect of measurement errors in $\mathbf{X}$ on estimates.

$\rho_b(y_0, \mathbf{x}_0)$ shares similarities with the sample-based leave-one-out diagnostic introduced by Prendergast (2008) and, from Theorem 1, we can use it to assess sensitivity of OLS as an estimator in the dimension reduction setting. The influence diagnostic can provide interesting general insights to OLS estimation in this setting. For example, even extreme predictor outliers can have very small influence regardless of the choice of accompanying response. We highlight this in the example below.
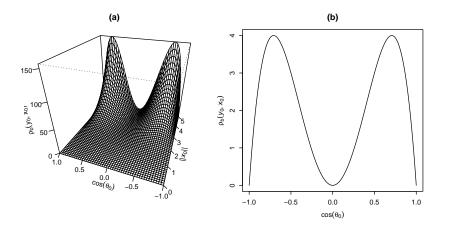
Fɪɢ 2. *Plots of $\rho_b(y_0, \mathbf{x}_0)$ for Example 3.2 for (a) varying $\cos(\theta_0)$ and $\|\mathbf{x}_0\|$ and (b) varying $\cos(\theta_0)$ with a fixed $\|\mathbf{x}_0\| = 2$.*

**Example 3.1.** For simplicity suppose that $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}_p$. Also suppose that $\mathbf{x}_0$ is in the same direction as $\boldsymbol{\beta}$ and therefore in the same direction as $\mathbf{b}$. From Theorem 1, $\rho_b(y_0, \mathbf{x}_0) = 0$ for all $y_0 \in \mathbb{R}$ since $(\mathbf{I}_p - \mathbf{P}_b)\mathbf{x}_0 = \mathbf{0}$. This also holds for $\mathbf{x}_0$ in the opposite direction to $\boldsymbol{\beta}$.

Zero influence resulted in Example 3.1 since the contaminant regressor $\mathbf{x}_0$ was in the same, or opposite, direction to $\boldsymbol{\beta}$. The corollary below is useful for assessing the role that the direction of $\mathbf{x}_0$ has in exertion of influence on OLS.

**Corollary 1.** *Let $\mathbf{z}_0 = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x}_0 - \boldsymbol{\mu})$ and $\theta_0$ denote the angle between $\mathbf{z}_0$ and $\boldsymbol{\Sigma}^{1/2}\mathbf{b}$. Then the influence diagnostic can be represented as*

$$\rho_b(y_0, \mathbf{x}_0) = \left[y_0 - \mu_y - \|\boldsymbol{\Sigma}^{1/2}\mathbf{b}\|.\|\mathbf{z}_0\|.\cos(\theta_0)\right]^2 [1 - \cos^2(\theta_0)].\frac{\|\mathbf{z}_0\|^2}{\|\boldsymbol{\Sigma}^{1/2}\mathbf{b}\|^2}.$$

*Proof.* The proof is straightforward by noting that, since $(\mathbf{I}_p - \mathbf{P}_b)$ is idempotent, $\|(\mathbf{I}_p - \mathbf{P}_b)\mathbf{z}_0\|^2 = \mathbf{z}_0^\top (\mathbf{I}_p - \mathbf{P}_b)\mathbf{z}_0$ and that $\cos(\theta_0) = \mathbf{b}^\top (\mathbf{x}_0 - \boldsymbol{\mu})/(\|\boldsymbol{\Sigma}^{1/2}\mathbf{b}\|.\|\mathbf{z}_0\|)$. □

From Corollary 1, the influence diagnostic is a function of $\|\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}_0 - \boldsymbol{\mu})\|$ and $\cos^2(\theta_0)$. We now consider another example to highlight the role of the angle between the predictor contaminant and $\mathbf{b}$.

**Example 3.2.** For simplicity suppose $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}_p$, and let $\|\mathbf{b}\| = 1$ and $y_0 = E(Y)$ so that $y_0$ is a typical response. Then, from Corollary 1, we have $\rho_b(y_0, \mathbf{x}_0) = \|\mathbf{x}_0\|^4 \cos^2(\theta_0)[1 - \cos^2(\theta_0)]$. In plot (a) of Figure 2 we show $\rho_b(y_0, \mathbf{x}_0)$ for varying $\cos(\theta_0)$ and $\|\mathbf{x}_0\|$. It is clear that the influence diagnostic is highly sensitive to the angle between $\mathbf{x}_0$ and $\mathbf{b}$. That is, for different $\mathbf{x}_0$'s with similar $\|\mathbf{x}_0\|$, there can be very different influence in estimation. This is further emphasized in plot (b) where we provide a cross-section of (a) with $\|\mathbf{x}_0\| = 2$.

These plots also show that even extreme predictor outliers (i.e. large $\|\mathbf{x}_0\|$) may not have a large influence. For example, when $\cos(\theta_0) = 0$ or $\pm 1$ (which occurs when $\mathbf{x}_0 = c\mathbf{b}$ for some $c \in \mathbb{R}$) then $\rho_b(y_0, \mathbf{x}_0) = 0$.

Prendergast & Smith (2010) also reported the expected value of the influence diagnostic for the contaminant equal to a random $(Y, \mathbf{x})$ from $F$. Noting that the influence function can be used to derive the asymptotic variance of an estimator as, for an estimator with functional $T$, the asymptotic variance, at $F$, is $\text{ASV}(T) = E[\text{IF}(T, F; Y, \mathbf{X})^2])$; then the $E[\rho_b(Y, \mathbf{x})]$ provides similar insight in the sense that it is the average sensitivity in estimation across all $(Y, \mathbf{x}) \sim F$. Therefore, $E[\rho_b(Y, \mathbf{x})]$ will provide a useful means for which to compare estimators and we report this in the following corollary to Theorem 1.

**Corollary 2.** *For $(Y, \mathbf{x}) \sim F$ where $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,*

$$E[\rho_b(Y, \mathbf{x})] = (p - 1) \left[ \frac{Var(Y)}{\|\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{b}\|^2} - 1 \right].$$

*Proof.* Using similar arguments to those by Prendergast & Smith (2010) in the proof of their Corollary 3, it can be shown that $E[\rho_b(Y, \mathbf{x})] = (p - 1)E[c(Y, \mathbf{x})]$ for $c(Y, \mathbf{x}) = [Y - \mu_y - \mathbf{b}^\top(\mathbf{x} - \boldsymbol{\mu})]^2 / \|\boldsymbol{\Sigma}^{1/2}\mathbf{b}\|^2$. For $\boldsymbol{\Sigma}_{xy} = \text{cov}(\mathbf{x}, Y)$,

$$\begin{aligned} E[c(Y, \mathbf{x})] &= E\left\{ [Y - \mu_Y - \mathbf{b}^\top(\mathbf{x} - \boldsymbol{\mu})]^2 \right\} \|\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{b}\|^{-2} \\ &= \left[ \text{Var}(Y) - 2\mathbf{b}^\top\boldsymbol{\Sigma}_{xy} + \mathbf{b}^\top\boldsymbol{\Sigma}\mathbf{b} \right] \|\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{b}\|^{-2}. \end{aligned}$$

Since $\mathbf{b} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{xy}$, $E[c(Y, \mathbf{x})] = [\text{Var}(Y) - \mathbf{b}^\top\boldsymbol{\Sigma}\mathbf{b}] / \|\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{b}\|^2$ which completes the proof. □

Although we do not discuss Corollary 2 further here, the results will be important in the considerations provided in the next section.

## 4. Influence following response transformations

As discussed in Section 2.3, response transformations may be employed without disrupting the theory. Here we consider such response transformations and use the influence diagnostic introduced previously to compare the estimator with and without transformation. In the following corollary we provide the influence diagnostic from Theorem 1 following response transformation. This result also holds for transformation functions that are affected by contamination, see Appendix A, when Condition 1 holds. Thus we have two cases:

   (i) $t$ is unaffected by contamination,
   (ii) $t$ is affected by contamination but Condition 1 holds.

**Corollary 3.** *For both cases consider a response transformation function $t$ defined at $F$ which may be different at $F_\epsilon$. The influence diagnostic presented in Theorem 1 is*

$$\rho_t(y_0, \mathbf{x}_0) = \frac{r_{t,0}^2}{\left\|\boldsymbol{\Sigma}^{1/2}\mathbf{b}_t\right\|^2} \left\|(\mathbf{I}_p - \mathbf{P}_t)\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}_0 - \boldsymbol{\mu})\right\|^2$$

where $(Y, \mathbf{x}) \sim F$, $\mathbf{b}_t$ is the OLS slope vector for the regression of $t(Y)$ on $\mathbf{x}$, $r_{t,0} = t(y_0) - E[t(Y)] - \mathbf{b}_t^\top(\mathbf{x}_0 - \boldsymbol{\mu})$ and $\mathbf{P}_t = (\mathbf{b}_t^\top \boldsymbol{\Sigma} \mathbf{b}_t)^{-1} \boldsymbol{\Sigma}^{1/2} \mathbf{b}_t \mathbf{b}_t^\top \boldsymbol{\Sigma}^{1/2}$. Additionally, from Corollary 2,

$$E\left[\rho_t(Y, \mathbf{x})\right] = (p-1)\left[\frac{Var\left[t(Y)\right]}{\|\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{b}_t\|^2} - 1\right].$$

Note that when the model in (2) and Condition 1 hold, $\mathbf{b}$ and $\mathbf{b}_t$ are scalar multiples of $\boldsymbol{\beta}$. Hence, provided they are not the zero vector, $\mathbf{P}_b$ and $\mathbf{P}_t$ are identical. Subsequently, a comparison of OLS with and without a response transformation can be carried out by comparing $r_{t,0}^2/\|\boldsymbol{\Sigma}^{1/2}\mathbf{b}_t\|^2$ and $r_0^2/\|\boldsymbol{\Sigma}^{1/2}\mathbf{b}\|^2$. Next we will consider some example transformations.

### 4.1. Log-transformation for exponential growth models

Consider an exponential growth model defined by

$$Y = a^{b_0 + b_1 \boldsymbol{\beta}^\top \mathbf{x} + \sigma \epsilon} \tag{10}$$

for an $a \in \mathbb{R}^+$, $b_0 \in \mathbb{R}$, $b_1 \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$. This model satisfies the model in (2) so that the OLS slope vector is equal to $c\boldsymbol{\beta}$ for a $c \in \mathbb{R}$. Ignoring trivial cases such as $a = 1$ or $b_1 = 0$, OLS is expected to find the direction of $\boldsymbol{\beta}$. Throughout suppose that $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\epsilon \sim N(0, 1)$ and define $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} = 1$. This is not a restriction since $b_1$ in (10) can be chosen to be any value.

To compare influence with and without the log transformation, we will derive $E[\rho_b(Y, \mathbf{x})]$ and $E[\rho_{\log}(Y, \mathbf{x})]$, the expected influence at $(Y, \mathbf{x}) \sim F$ for the original and log transformed OLS estimators respectively. These are provided below and the associated derivations can be found in Appendix B.

For ln denoting the natural log we have

$$(p-1)^{-1}E[\rho_b(Y, \mathbf{x})] = \frac{\exp\left\{[\ln(a)]^2\left(b_1^2 + \sigma^2\right)\right\} - 1 - [\ln(a)]^2 b_1^2}{[\ln(a)]^2 b_1^2}. \tag{11}$$

Following log transformation of $Y$, with any base,

$$(p-1)^{-1}E[\rho_{\log}(Y, \mathbf{x})] = \frac{\sigma^2}{b_1^2} \tag{12}$$

which does not depend on the base chosen or $a$. Interestingly, for a fixed error variance $\sigma^2$, the expected influence following a log transformation decreases with increasing $b_1$. This is not necessarily the case with $E[\rho_b(Y, \mathbf{x})]$ in (11) where a large value of $b_1^2$ can lead to very large $\exp\{[\ln(a)]^2(b_1^2 + \sigma^2)\}$ and consequently large influence despite the denominator $[\ln(a)]^2 b_1^2$.

**Example 4.1.** In Figure 3 we plot $E[\rho_b(Y, \mathbf{x})]$ and $E[\rho_{\log}(Y, \mathbf{x})]$ for the model in (10). For plots (a)-(c) $a = 2$, $\sigma$ is chosen to be 0.5, 1.0 and 2.0 respectively. For small $|b_1|$ relative to $\sigma$, it is clear both approaches are expected to perform
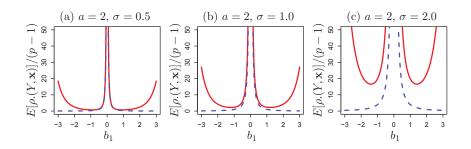
FIG 3. *Plots of $E[\rho_b(Y,\mathbf{x})]/(p-1)$ (red solid line) and $E[\rho_{\log}(Y,\mathbf{x})]$ (blue dashed line) for Example 4.1 for three choices of $\sigma$.*

TABLE 1

*Values of $E[\rho_b(Y,\mathbf{x})]$ (OLS with no transformation) and $E[\rho_t(Y,\mathbf{x})]$ (OLS following log transformation) for Model 1 with varying a and fixed $b_1 = 1$ including average $cor(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\hat{\mathbf{b}})^2$ (standard deviations, SD, in parentheses) for $1000$ trials of $n = 50$ randomly sampled observations*

| | Model: $Y = a^{1+\boldsymbol{\beta}^\top\mathbf{x}+\epsilon}$ | | | |
|---|---|---|---|---|
| $a$ | No transformation | | Log transformation | |
| | $E[\rho_b(Y,\mathbf{x})]$ | Mean $cor(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\mathbf{b})^2$ (SD) | $E[\rho_{\log}(Y,\mathbf{x})]$ | Mean $cor(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\mathbf{b})^2$ (SD) |
| 1.25 | 9.927 | 0.934 (0.030) | 9.0 | 0.964 (0.019) |
| 1.5 | 12.312 | 0.880 (0.060) | 9.0 | 0.964 (0.018) |
| 1.75 | 16.024 | 0.825 (0.085) | 9.0 | 0.965 (0.019) |
| 2 | 21.235 | 0.773 (0.103) | 9.0 | 0.965 (0.018) |
| 2.25 | 28.302 | 0.730 (0.115) | 9.0 | 0.965 (0.018) |
| 2.5 | 37.749 | 0.701 (0.126) | 9.0 | 0.964 (0.019) |
| 2.75 | 50.295 | 0.671 (0.131) | 9.0 | 0.964 (0.020) |
| 3 | 66.891 | 0.653 (0.139) | 9.0 | 0.965 (0.019) |
| 3.25 | 88.786 | 0.632 (0.134) | 9.0 | 0.965 (0.018) |
| 3.5 | 117.602 | 0.612 (0.138) | 9.0 | 0.965 (0.019) |
| 3.75 | 155.438 | 0.598 (0.142) | 9.0 | 0.965 (0.018) |
| 4 | 204.991 | 0.594 (0.139) | 9.0 | 0.965 (0.019) |

poorly due to the large expected influence. This is not surprising due to the small contribution of $b_1\boldsymbol{\beta}^\top\mathbf{x}$ to $Y$ relative to the error. As $|b_1|$ increases, we see a marked difference in the expected performance of both approaches. For the usual OLS approach performance quickly deteriorates even for moderate sized $b_1$. For the model, extremely large responses are possible when $b_1$ is not small which should highly influence OLS.

Conversely, OLS following a log transformation is expected to perform well with small expected influence and, as shown in (12), improve further for increasing $|b_1|$. Using the log transformation, sensitivity of OLS to very large response values is reduced. This can be seen in Corollary 3 where the diagnostic depends on the log of the response only. We saw similar results for other choices of $a$.

For verification, we now give our simulated results in Tables 1 and 2 for varying $a$ and $b_1$ respectively. Each simulation was run 1000 times with $n = 50$.

*Values of $E[\rho_b(Y, \mathbf{x})]$ (OLS with no transformation) and $E[\rho_t(Y, \mathbf{x})]$ (OLS following log transformation) for Model 1 with varying $b_1$ and fixed $a = 2$ including average $cor(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\widehat{\mathbf{b}})^2$ (standard deviations, SD, in parentheses) for $1000$ trials of $n = 50$ randomly sampled observations*

| | Model: $Y = 2^{1+b_1\boldsymbol{\beta}^\top\mathbf{x}+\epsilon}$ | | | |
| $b_1$ | No transformation | | Log transformation | |
| | $E[\rho_b(Y,\mathbf{x})]$ | Mean $cor(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\widehat{\mathbf{b}})^2$ (SD) | $E[\rho_{\log}(Y,\mathbf{x})]$ | Mean $cor(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\widehat{\mathbf{b}})^2$ (SD) |
|---|---|---|---|---|
| 1.25 | 20.075 | 0.743 (0.109) | 5.760 | 0.977 (0.012) |
| 1.5 | 22.352 | 0.702 (0.123) | 4.000 | 0.984 (0.008) |
| 1.75 | 27.955 | 0.677 (0.126) | 2.939 | 0.988 (0.006) |
| 2 | 38.056 | 0.650 (0.128) | 2.250 | 0.991 (0.005) |
| 2.25 | 55.411 | 0.618 (0.127) | 1.778 | 0.993 (0.005) |
| 2.5 | 85.610 | 0.602 (0.129) | 1.440 | 0.994 (0.003) |
| 2.75 | 140.075 | 0.594 (0.133) | 1.190 | 0.995 (0.003) |
| 3 | 242.975 | 0.574 (0.131) | 1.000 | 0.996 (0.002) |
| 3.25 | 447.827 | 0.558 (0.132) | 0.852 | 0.997 (0.002) |
| 3.5 | 879.031 | 0.553 (0.132) | 0.735 | 0.997 (0.002) |
| 3.75 | 1840.818 | 0.542 (0.131) | 0.640 | 0.997 (0.001) |
| 4 | 4117.064 | 0.532 (0.129) | 0.563 | 0.998 (0.001) |

Here $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_p)$, $\epsilon \sim N(0, 1)$, $\boldsymbol{\beta} = [1, -2, 0, 0.5, 0, \ldots, 0]^\top$ and $p = 10$. From Table 1, a significant improvement due to the log transformation is evident. Included is $E[\rho_b(Y, \mathbf{x})]$ and $E[\rho_{\log}(Y, \mathbf{x})]$, which show that without transformation, increasing $a$ is expected to diminish performance of OLS. This is emphasized by the mean $cor(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\widehat{\mathbf{b}})^2$ decreasing and associated standard deviation increasing with $a$. Conversely, the transformation gives small, approximately constant mean $cor(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\widehat{\mathbf{b}})^2$ with small standard deviation across all values of $a$.

From Table 2, increasing $b_1$ results in standard OLS deteriorating in performance, both in mean influence and the correlations. In contrast a vast improvement has been achieved using the log transformation which, as mentioned previously, has improved performance with increasing $b_1$.

## 4.2. Response-discretization transformations

Let the discretization function be

$$t_F(Y) = \begin{cases} D_1, & Y \in S_1 \\ D_2, & Y \in S_2 \\ \vdots & \vdots \\ D_H, & Y \in S_H \end{cases} \tag{13}$$

where $D_h \in \mathbb{R}$ with $D_1 < D_2 < \cdots < D_H$ and $S_1, \ldots, S_H$ are non-overlapping subranges of the range of $Y$ such that $\bigcup_{h=1}^{H} S_h = \text{range}(Y)$. We denote $P(Y \in S_h)$ as $p_h$. The attractive feature of this transformation is that it retains some co-variability between the response and predictor whilst limiting the minimum and maximum transformed response values. This transformation shares some

similarities with SIR (Li, 1991) where the role of the response is to determine positioning of observations within slices. Consider the model of the form

$$Y = f(\boldsymbol{\beta}^\top \mathbf{x} + \sigma\epsilon) \tag{14}$$

where $f$ is a strictly increasing link function, $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\epsilon \sim N(0,1)$ and $\sigma \geq 0$. Before we move on, we will provide the form of $\mathbf{b}_t$ for the model in (14) under the assumption of a normal $\mathbf{x}$; the proof can be found in Appendix C.

**Corollary 4.** *For the model in* (14) *where $f$ is a monotonic increasing link function, $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\epsilon \sim N(0,1)$ and $\sigma \geq 0$, the OLS slope vector following the discretization described in* (13) *is*

$$\mathbf{b}_t = c_t \boldsymbol{\beta} \ \ where \ \ c_t = \frac{1}{\sqrt{\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} + \sigma^2}} \sum_{h=1}^{H} D_h \ [\phi(Z_{h-1}) - \phi(Z_h)]$$

*where $Z_j$ is the $(j/H) \times 100$th percentile from the $N(0,1)$ distribution such that $P(Z \leq Z_j) = j/H$ for $Z \sim N(0,1)$ and where $\phi$ denotes the standard normal probability density function.*

A simple choice for the $D_h$'s in (13) is to let $D_h = h$ and for the $S_h$'s to be chosen such that $P(Y \in S_h) = 1/H$. For now we will proceed with this approach and refer to this transformation as $H$-*fold equally weighted response discretization*. It should be noted for this transformation function, $t$ depends on $F$ since the $S_h$'s depend on $F$. As such, when applied to $F_\epsilon$, the contamination can affect a change in $t$ through changes to the $S_h$'s. However, Corollary 3 holds when Condition 1 holds.

**Corollary 5.** *Suppose that $(Y, \mathbf{x}) \sim F$ and where $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, for $t$ being the discretization function described in* (13) *with $D_h = h$ and $p_h = 1/H$ for all $h = 1, \ldots, H$ ($H$-fold equally weighted response discretization),*

$$E\left\{\rho_F[t(Y), \mathbf{x}]\right\} = (p-1)\left[\frac{(H^2 - 1)}{12\|\boldsymbol{\Sigma}^{1/2}\mathbf{b}_t\|^2} - 1\right]$$

*where $\mathbf{b}_t$ is provided in Corollary 4 with each $D_h$ replaced by $h$.*

*Proof.* The proof is straightforward from Corollary 3 when noting that $t(Y)$ is a discrete uniform random variable such that $\mathrm{Var}[t(Y)] = (H^2 - 1)/12$. □

We now consider the effects of $H$-fold equally weighted response discretization on a general model with monotone link function. From Corollary 4, $\mathbf{b}_t$ is a simple expression in terms of the standard normal density and specific standard normal percentiles that are dependent on $H$. When applying this result to Corollary 5, it is possible to calculate the expected sensitivity $E\{\rho_F[t(Y), \mathbf{x}]\}$.

In Table 3 we present values of $E\{\rho_F[t(Y), \mathbf{x}]\}/(p-1)$ for the model in (14) when $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$, $\epsilon \sim N(0,1)$ and $\|\boldsymbol{\beta}\| = 1$, for various choices of $H$ and $\sigma$. For the three choices of $\sigma$, $E\{\rho_F[t(Y), \mathbf{x}]\}/(p-1)$ decreases with increasing $H$. Many other choices of $H$ were also considered, each showing similar decreases although

TABLE 3

*Values of $E\{\rho_F[t(Y),\mathbf{x}]\}/(p-1)$ following $H$-fold equally weighted response discretization for the model in (14) with various $H$ and $\sigma$ and with $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$ and $\|\boldsymbol{\beta}\| = 1$*

| $\sigma = 0$ | $H$ | 2 | 5 | 10 | 20 | 50 | 100 | 500 | 1000 |
|---|---|---|---|---|---|---|---|---|---|
| | $\dfrac{E\left[\rho_F(t_F(Y),\mathbf{x})\right]}{(p-1)}$ | 0.57 | 0.13 | 0.07 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| $\sigma = 0.5$ | $H$ | 2 | 5 | 10 | 20 | 50 | 100 | 500 | 1000 |
| | $\dfrac{E\left[\rho_F(t_F(Y),\mathbf{x})\right]}{(p-1)}$ | 0.96 | 0.41 | 0.34 | 0.32 | 0.31 | 0.31 | 0.31 | 0.31 |
| $\sigma = 1$ | $H$ | 2 | 5 | 10 | 20 | 50 | 100 | 500 | 1000 |
| | $\dfrac{E\left[\rho_F(t_F(Y),\mathbf{x})\right]}{(p-1)}$ | 2.14 | 1.25 | 1.14 | 1.11 | 1.10 | 1.09 | 1.09 | 1.09 |

the difference became negligible for larger $H$. The limit of $E\{\rho_F[t(Y),\mathbf{x}]\}/(p-1)$ is unknown, however simulations suggest it should be very similar to that for $H = 1000$. These results imply a reduction in estimator variability for this model when $H$ is chosen to be as large as possible. In practice when there are a finite number of observed responses, $n$, this would suggest an $n$-fold discretization such that each observed response is replaced with its respective rank. At $F$, the rank transformation OLS slope is

$$\mathbf{b}_t = \boldsymbol{\Sigma}^{-1}\mathrm{cov}\left[\mathbf{x}, \mathrm{rank}(Y)\right]. \tag{15}$$

A benefit of the rank transformation least squares estimator in (15) is that $H$ does not need to be chosen. It should be noted that the rank transformation has been widely considered and discussed with respect to OLS in the multiple linear regression contex when response normality is violated. However, the rank transformation in OLS here is simpler as it does not involve additional complications to ensure that the unique $\boldsymbol{\beta}$ is estimated. For example, Cuzick (1988) considers models of the form given in (14), where $f^{-1}$ exists, with an emphasis on one predictor variable, $p = 1$, but where extensions to $p > 1$ are provided. The approach involves replacing the response with a score based on rank and also requires estimation of $f^{-1}$ to target the specific predictor coefficient. Other approaches, such as $R$-estimates, again under the multiple linear regression model (see, for e.g., Chapter 3 of Hettmansperger & McKean, 1998), involve minimization over an objective function based on residuals and scores of the ranks of residuals. Our approach in (15) is simple since (i) it requires no knowledge of estimation of $f$, and (ii) given the ranks of the observed responses, the estimate in practice exists in a simple closed form and does not require computational minimization over an objective function.

**Example 4.2.** Recall the model in (3). Earlier we discussed how poorly OLS estimated the direction of $\boldsymbol{\beta}$ when $m$ was close to 3.25. We showed that OLS without transformation was estimating the zero vector as opposed to the direction of $\boldsymbol{\beta}$ for this $m$, causing the $\mathrm{cor}(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\widehat{\mathbf{b}})^2$ to decrease significantly. We now demonstrate how discretization can result in a marked improvement to the results. Here we consider the rank transformation with a large sample size of $n = 1000$, thereby replacing all response values with their corresponding rank.
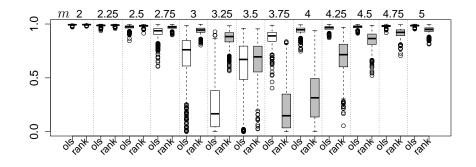
FIG 4. *Boxplots of cor$(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\widehat{\mathbf{b}})^2$ for OLS (white boxes labelled OLS) and OLS following rank transformation of the response (grey boxes labeled rank) for 1000 trials with $n = 1000$ for Model 2.*

Figure 4 compares the cor$(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\widehat{\mathbf{b}})^2$ for 1000 simulated trials where $\widehat{\mathbf{b}}$ has been estimated by OLS with and without the rank transformation. The white boxplots, like those in plot (b) of Figure 1, show the estimate of the direction $\boldsymbol{\beta}$ deteriorating as $m$ approaches 3.258 for standard OLS. However the grey boxplots show that good results can be obtained in this vicinity of $m$ using the rank transformation. This does not necessarily mean that the rank transformation works well for all $m$, since small correlation values now occur at approximately 3.75 whereas OLS without transformation performs typically well here. Therefore, if both methods were used in practice, it is likely that one if not both approaches would provide a good ESSP.

**Example 4.3.** For $\mathbf{x} = [X_1, \ldots, X_p]^\top \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\epsilon \sim N(0, 1)$ consider the Cubic, $Y = (1 + b_1\boldsymbol{\beta}^\top\mathbf{x} + \sigma\epsilon)^3$, and Exponential, $Y = \exp(1 + b_1\boldsymbol{\beta}^\top\mathbf{x} + \sigma\epsilon)$, models.

In Figure 5 we show plots of the expected influence for these models, with and without the $H$-fold equally weighted response transformation. It is clear the expected influence decreases with the transformation, particulary for larger values of $H$. The exponential model is notably affected where the expected influence for this model increases rapidly with increasing $b_1$. The large $H$ transformation significantly decreases the influence of both models, indicating the transformation is expected to improve the performance of OLS.

Our simulated results, shown in Figures 6 and 7, complement these findings. It should be noted that for each simulation, $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_p)$ and $\boldsymbol{\beta} = [b_1, 0, \ldots, 0]^\top$, where $b_1 \in \mathbb{R}$ and $p = 5$. For simplicity we consider the rank transformation. These boxplots show the cor$(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\widehat{\boldsymbol{\beta}})^2$ for 1000 simulated trials with 100 observations given by OLS with and without transformation for increasing values of $b_1$. Here we have three choices for $\sigma$ to compare how these models are affected by small and large amounts of error. These results demonstrate how estimation has been improved, with rank transformation consistently giving better results compared to general OLS. It is also clear how these boxplots correspond with
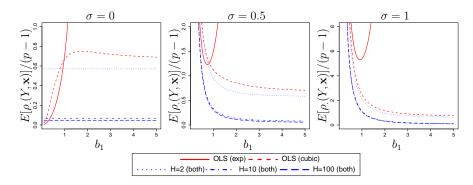
FIG 5. *Plots of $E[\rho_b(Y, \mathbf{x})]/(p-1)$ (red lines) and $E[\rho_F(t_F(Y), \mathbf{x})]$ (blue lines) when $Y = (1+b_1\boldsymbol{\beta}^\top\mathbf{x}+\sigma\epsilon)^3$ and $Y = \exp(1+b_1\boldsymbol{\beta}^\top\mathbf{x}+\sigma\epsilon)$, with the $E[\rho_F(t_F(Y), \mathbf{x})]/(p-1)$ at $H = 2, 10$ and 100.*

the expected influence, with our correlation results resembling the trends shown in Figure 5.

## 5. Discussion and further research

In this paper we have demonstrated how an influence diagnostic, similar to the influence function (Hampel, 1974), can be used to assess performance of OLS for single-index models. Importantly, we have shown how simple response transformations can greatly improve OLS. This was done with respect to average influence under some proposed models and further highlighted with some simulation studies. Although the log and response discretization transformations were the focus of this paper in an application sense, the results themselves are quite general and will be useful for further explorations. For example, the two extensions below are currently being investigated as part of the lead author's doctorate studies:

- Recently, Prendergast & Sheather (2013) carried out a sensitivity study of inverse response plot estimation and their simulation studies indicated that robust $M$-estimators could provide improved estimates compared to OLS even for normal data. The results of this paper can be used to theoretically compare inverse response plot estimators of the coefficient vector for various response transformations. For example, the rank transformation should be well suited to this type of estimation and other linearization transformations under assumptions such as normality are also being explored. This includes the performance of quantile transformations under normality assumptions for $\mathbf{x}$ and include a sensitivity analysis when $\mathbf{x}$ is non-normal.
- Li (1992) introduced a method call Principal Hessian Directions (pHd) that can be used in the multiple index setting allowing for more than one direction. As with OLS, very large or small observed values for the
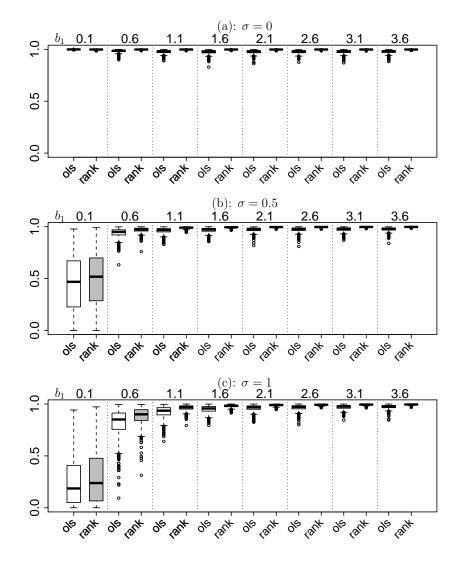
FIG 6. *Boxplots of cor($\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\widehat{\mathbf{b}}$)$^2$ where $Y = (1 + b_1\boldsymbol{\beta}^\top\mathbf{x} + \sigma\epsilon)^3$ for OLS (white boxes labeled OLS) and OLS following rank response transformation (grey boxes labeled rank), for 1000 trials with $n = 100$ and three choices of $\sigma$.*

response can have a large influence on estimation. In developing the work seen here, we are extending the theory to allow pHd to be used. Response transformations, such as those studied in this paper, are currently being considered as part of the first author's PhD dissertation.

To summarize, since response transformations still allow OLS to be used for single-index models with exploration sill on the original response scale, we recommend routinely considering response transformations in this setting. The
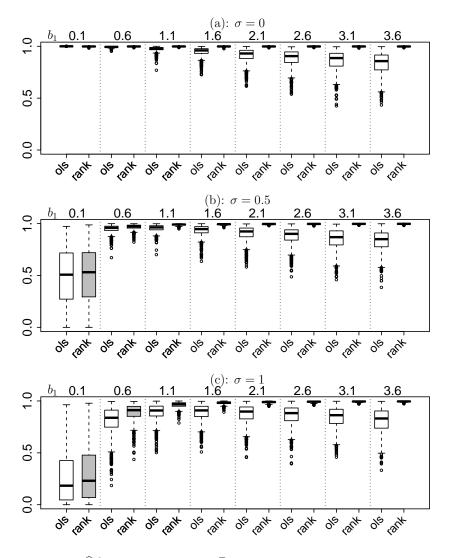
FIG 7. $cor(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\widehat{\mathbf{b}})^2$ where $Y = \exp(1 + b_1\boldsymbol{\beta}^\top\mathbf{x} + \sigma\epsilon)$ for OLS (white boxes labeled OLS) and OLS following rank response transformation (grey boxes labeled rank), for 1000 trials with $n = 100$ and three choices of $\sigma$.

response transformations we considered can be easily implemented in any statistical software package and can yield big improvements to estimation.

## Acknowledgements

## Appendix A

Although $F$ is a joint distribution function for $(Y, \mathbf{x})$, for simplicity throughout we take $Y \sim F$ to mean $Y$ distributed according to its marginal distribution from $F$. This simplicity is also carried out for $\mathbf{x}$ and similarly for $F_\epsilon$.

### *A.1. Proof of Theorem 1*

Recall $b(F) = \mathbf{b}$ and let $b(F_\epsilon) = \mathbf{b}_\epsilon$ be the perturbed equivalent at $F_\epsilon$. For simplicity let $c_\epsilon = \mathrm{cor}^2(\mathbf{b}^\top \mathbf{x}, \mathbf{b}_\epsilon^\top \mathbf{x})$ and let $c_\epsilon'$, $c_\epsilon''$ be the first and second derivatives of $c_\epsilon$ with respect to $\epsilon$. Furthermore, let $c_0'$, $c_0''$ be these derivatives at $\epsilon = 0$. The power series expansion for $c_\epsilon$, where $\mathbf{x} \sim F$, is

$$c_\epsilon = 1 + \epsilon c_0' + \epsilon^2 c_0''/2 + O(\epsilon^3). \tag{16}$$

For $\mathrm{cov}(\mathbf{x}) = \boldsymbol{\Sigma}$ and using the Product Rule,

$$c_\epsilon' = \frac{1}{\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}} \left[ \frac{\partial}{\partial \epsilon} \frac{\left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon\right)^2}{\mathbf{b}_\epsilon^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon} + \frac{1}{\mathbf{b}_\epsilon^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon} \frac{\partial}{\partial \epsilon} \left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon\right)^2 + \left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon\right)^2 \frac{\partial}{\partial \epsilon} \frac{1}{\mathbf{b}_\epsilon^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon} \right]. \tag{17}$$

Let the power series for $\mathbf{b}_\epsilon$ be $\mathbf{b} + \epsilon \mathbf{b}_1 + \epsilon^2 \mathbf{b}_2/2 + O(\epsilon^3)$ where $\mathbf{b}_1$ and $\mathbf{b}_2$ are the first and second order derivatives of $\mathbf{b}_\epsilon$ at $\epsilon = 0$. Then $\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon = \mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b} + \epsilon \mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_1 + \epsilon^2 \mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_2/2 + O(\epsilon^3)$ and $\mathbf{b}_\epsilon^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon = \mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b} + 2\epsilon \mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_1 + \epsilon^2 (\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_2 + \mathbf{b}_1^\top \boldsymbol{\Sigma} \mathbf{b}_1)$. Using the Chain Rule,

$$\left\{ \partial \left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon\right)^2 / \partial \epsilon \right\} \Big|_{\epsilon=0} = \left[ 2\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon \left\{ \partial \mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon / \partial \epsilon \right\} \right] \Big|_{\epsilon=0} = 2(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b})(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_1).$$

Similarly $\{\partial (\mathbf{b}_\epsilon^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon)^{-1}/(\partial \epsilon)\}|_{\epsilon=0} = -2\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_1/(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b})^2$ so that, from (17),

$$c_0' = \frac{1}{(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b})^2} \frac{\partial}{\partial \epsilon} \left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon\right)^2 \Big|_{\epsilon=0} + \left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}\right) \frac{\partial}{\partial \epsilon} \frac{1}{(\mathbf{b}_\epsilon^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon)} \Big|_{\epsilon=0} = 0. \tag{18}$$

From (17), using the Product Rule,

$$c_0'' = \left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}\right) \frac{\partial^2}{\partial \epsilon^2} \frac{1}{(\mathbf{b}_\epsilon^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon)} \Big|_{\epsilon=0} + \frac{1}{(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b})^2} \frac{\partial^2}{\partial \epsilon^2} \left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon\right)^2 \Big|_{\epsilon=0}$$
$$+ \frac{2}{(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b})} \left[ \frac{\partial}{\partial \epsilon} \frac{1}{(\mathbf{b}_\epsilon^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon)} \Big|_{\epsilon=0} \right] \frac{\partial}{\partial \epsilon} \left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon\right)^2 \Big|_{\epsilon=0}. \tag{19}$$

From the expansion for $\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon$ prior to (18), we have $(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon)^2 = (\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b})^2 + 2\epsilon (\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b})(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_1) + \epsilon^2 [(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_1)^2 + (\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b})(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_2)] + O(\epsilon^3)$. Hence,

$$\left\{ \partial^2 \left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon\right)^2 / (\partial \epsilon^2) \right\} \Big|_{\epsilon=0} = 2 \left[ \left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_1\right)^2 + \left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}\right) \left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_2\right) \right].$$

Similarly, the Chain Rule and Product Rule give

$$\left.\frac{\partial^2}{\partial \epsilon^2} \frac{1}{\mathbf{b}_\epsilon^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon}\right|_{\epsilon=0} = \left\{ \frac{2}{(\mathbf{b}_\epsilon^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon)^3} \left[\frac{\partial}{\partial \epsilon} \mathbf{b}_\epsilon^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon\right]^2 - \frac{1}{(\mathbf{b}_\epsilon^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon)^2} \frac{\partial^2}{\partial \epsilon^2} \left(\mathbf{b}_\epsilon^\top \boldsymbol{\Sigma} \mathbf{b}_\epsilon\right) \right\}\Bigg|_{\epsilon=0}$$

$$= 8\frac{\left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_1\right)^2}{\left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}\right)^3} - 2\frac{\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_2}{\left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}\right)^2} - 2\frac{\mathbf{b}_1^\top \boldsymbol{\Sigma} \mathbf{b}_1}{\left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}\right)^2}.$$

Using these results (19) simplifies to, noting that $\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_1 = \mathbf{b}_1^\top \boldsymbol{\Sigma} \mathbf{b}$,

$$c_0'' = \frac{2}{(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b})} \left[\frac{\left(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}_1\right)^2}{(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b})} - \left(\mathbf{b}_1^\top \boldsymbol{\Sigma} \mathbf{b}_1\right)\right] = \frac{2}{(\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b})} \mathbf{b}_1^\top \boldsymbol{\Sigma}^{1/2} \left(\mathbf{P} - \mathbf{I}_p\right) \boldsymbol{\Sigma}^{1/2} \mathbf{b}_1 \tag{20}$$

where $\mathbf{P} = \boldsymbol{\Sigma}^{1/2} \mathbf{b} \mathbf{b}^\top \boldsymbol{\Sigma}^{1/2} / (\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b})$. The proof is complete by noting that $\mathbf{b}_1 = \text{IF}(b, F; y_0, \mathbf{x}_0)$ where (see, for e.g., page 243 of Staudte & Sheather, 1990), $\text{IF}(b, F; y_0, \mathbf{x}_0) = r_0 \boldsymbol{\Sigma}^{-1}(\mathbf{x}_0 - \boldsymbol{\mu})$ and by noting that, since $\mathbf{I}_p - \mathbf{P}$ is idempotent,

$$(\mathbf{x}_0 - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1/2} \left(\mathbf{P} - \mathbf{I}_p\right) \boldsymbol{\Sigma}^{-1/2}(\mathbf{x}_0 - \boldsymbol{\mu}) = -\|\left(\mathbf{I}_p - \mathbf{P}\right) \boldsymbol{\Sigma}^{-1/2}(\mathbf{x}_0 - \boldsymbol{\mu})\|^2.$$

### A.2. Proof of Corollary 3

Let $t$ and $t_\epsilon$ denote the response transformation functions for $Y \sim F$ and $Y \sim F_\epsilon$ respectively. Furthermore, let $T$ denote the predictor vector mean functional where, at an arbitrary $G$, $E_G(\mathbf{x}) = T(G) = \int \mathbf{x} dG$. It is straightforward to show that $T(F_\epsilon) = (1 - \epsilon)\boldsymbol{\mu} + \epsilon \mathbf{x}_0$. Consequently,

$$\text{cov}_{F_\epsilon} \left[t_\epsilon(Y), \mathbf{x}\right] = \int t_\epsilon(Y) \left[(\mathbf{x} - \boldsymbol{\mu}) - \epsilon(\mathbf{x}_0 - \boldsymbol{\mu})\right] dF_\epsilon$$

$$= (1 - \epsilon)\text{cov}_F \left[t_\epsilon(Y), \mathbf{x}\right] - \epsilon(\mathbf{x}_0 - \boldsymbol{\mu}) \int t_\epsilon(Y) dF$$

$$+ \epsilon \int t_\epsilon(Y)(\mathbf{x} - \boldsymbol{\mu}) d\Delta_{(y_0, \mathbf{x}_0)} + O(\epsilon^2). \tag{21}$$

Note $\text{cov}_F[t_\epsilon(Y), \mathbf{x}] = d_\epsilon \boldsymbol{\Sigma} \boldsymbol{\beta}$ for a $d_\epsilon \in \mathbb{R}$ when the model in (2) and Condition 1 hold. Under these conditions, (21) may be rewritten as

$$\text{cov}_{F_\epsilon} \left[t_\epsilon(Y), \mathbf{x}\right] = d_\epsilon \boldsymbol{\Sigma} \boldsymbol{\beta} - \epsilon(\mathbf{x}_0 - \boldsymbol{\mu}) \int t_\epsilon(Y) dF + \epsilon \int t_\epsilon(Y)(\mathbf{x} - \boldsymbol{\mu}) d\Delta_{(y_0, \mathbf{x}_0)} + O(\epsilon^2). \tag{22}$$

Let $b_t$ be the OLS slope functional with respect to a transformed response ($t(Y)$ at $F$ or $t_\epsilon(Y)$ at $F_\epsilon$). Also, let $C$ denote the functional for the usual estimator of the covariance of $\mathbf{x}$. Then $b_t(F_\epsilon) = [C(F_\epsilon)]^{-1} \text{cov}_{F_\epsilon}[t_\epsilon(Y), \mathbf{x}]$ so that, by using the Product Rule,

$$\text{IF}(b_t, F; y_0, \mathbf{x}_0) = \left.\frac{\partial}{\partial \epsilon} b_t(F_\epsilon)\right|_{\epsilon=0} = \left\{\left.\frac{\partial}{\partial \epsilon} [C(F_\epsilon)]^{-1}\right|_{\epsilon=0}\right\} \text{Cov}_F \left[t(Y), \mathbf{x}\right]$$

$$+ \boldsymbol{\Sigma}^{-1} \left.\frac{\partial}{\partial \epsilon} \text{cov}_{F_\epsilon} \left[t_\epsilon(Y), \mathbf{x}\right]\right|_{\epsilon=0}. \tag{23}$$

Using the Chain Rule, it can be shown the partial derivative involving $C(F_\epsilon)$ in the first term of (23), evaluated at $\epsilon = 0$, is equal to $-\boldsymbol{\Sigma}^{-1}\text{IF}(C, F; \mathbf{x}_0)\boldsymbol{\Sigma}^{-1}$. Also let $d_0'$ be the derivative of $d_\epsilon$ evaluated at $\epsilon = 0$. Using this, the fact that $\mathbf{b}_t = \boldsymbol{\Sigma}^{-1}\text{cov}_F[t(Y), \mathbf{x}]$ and (22), then (23) reduces to

$$\text{IF}(b_t, F; y_0, \mathbf{x}_0) = d_0'\boldsymbol{\beta} - \boldsymbol{\Sigma}^{-1}\text{IF}(C, F; \mathbf{x}_0)\mathbf{b}_t + \left[t(y_0) - \mu_{t(Y)}\right]\boldsymbol{\Sigma}^{-1}(\mathbf{x}_0 - \boldsymbol{\mu}).$$

Now, $\text{IF}(C, F; \mathbf{x}_0) = (\mathbf{x}_0 - \boldsymbol{\mu})(\mathbf{x}_0 - \boldsymbol{\mu})^\top - \boldsymbol{\Sigma}$ (see, for e.g., Critchley, 1985) which gives

$$\text{IF}(b_t, F; y_0, \mathbf{x}_0) = d_0'\boldsymbol{\beta} - \mathbf{b}_t + \left[t(y_0) - \mu_{t(Y)} - \mathbf{b}_t^\top(\mathbf{x}_0 - \boldsymbol{\mu})\right]\boldsymbol{\Sigma}^{-1}(\mathbf{x}_0 - \boldsymbol{\mu}). \quad (24)$$

From this point the proof is identical as to that for Theorem 1 but with $\mathbf{b}_1 = \text{IF}(b_t, F; y_0, \mathbf{x}_0)$ instead of $\text{IF}(b, F; y_0, \mathbf{x}_0)$ when the model in (2) and Condition 1 hold due to the following. In (20) of the proof for Theorem 1, when the aforementioned conditions hold, $\mathbf{P}$ is a projection matrix onto the space spanned by $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}$. Subsequently, $\mathbf{I}_p - \mathbf{P}$ projects onto the orthogonal complement of this space so that, from (24),

$$(\mathbf{I}_p - \mathbf{P})\boldsymbol{\Sigma}^{1/2}\mathbf{b}_1 = (\mathbf{I}_p - \mathbf{P})\boldsymbol{\Sigma}^{1/2}\text{IF}(b_t, F; y_0, \mathbf{x}_0) = r_{t,0}(\mathbf{I}_p - \mathbf{P})\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}_0 - \boldsymbol{\mu})$$

for $r_{t,0} = [t(y_0) - \mu_{t(Y)} - \mathbf{b}_t^\top(\mathbf{x}_0 - \boldsymbol{\mu})]$. Hence, form this point the result is in the same form as that for usual OLS but under a transformed response with the transformation function chosen at $F$. This completes the proof.

## Appendix B: Proofs for results in Section 4.1

Using various integration laws and Stein's Lemma (Stein, 1981), it is simple to derive the two influence measures shown in (11) and (12). Let $\phi$ denote the density function for the standard normal and $Z = \boldsymbol{\beta}^\top\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ such that $Z \sim N(0, 1)$. Then for some function $g$ and noting $Z$ and $\epsilon$ are independent,

$$E[g(Z)] = \int\int g(Z)\phi(Z)\phi(\epsilon)dZd\epsilon.$$

For $Y$ as in (10), $\text{Var}(Y) = a^{2b_0}\exp[\ln(a)^2(b_1^2 + \sigma^2)]\{\exp[\ln(a)^2(b_1^2 + \sigma^2)] - 1\}$. Using Stein's Lemma, $\text{Cov}(Y, Z) = b_1\ln(a)a^{b_0}\exp\{\frac{1}{2}[\ln(a)]^2(b_1^2 + \sigma^2)\}$. Since $\|\boldsymbol{\Sigma}^{1/2}\mathbf{b}\|^2 = \text{Cov}(Y, Z)^2$, using Corollary 2 gives $E[\rho_b(Y, \mathbf{x})]$ shown in (11).

For the log transformation, $E[t_F(Y)] = b_0\log(a)$, $E[t_F(Y^2)] = [\log(a)]^2(b_0^2 + b_1^2 + \sigma^2)$ and $\text{Var}[t_F(Y)] = [\log(a)]^2(b_1^2 + \sigma^2)$. Stein's lemma gives $\text{Cov}(t(Y), Z) = b_1\log(a)$. Using these in Corollary 3 gives $E[\rho_{\log}t_F(Y)]$.

## Appendix C: Proof of Corollary 4

For $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\epsilon \sim N(0, 1)$ with $\epsilon$ independent of $\mathbf{x}$, and $\boldsymbol{\beta}^\top\mathbf{x} + \sigma\epsilon \sim N(\boldsymbol{\beta}^\top\boldsymbol{\mu}, \boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma^2)$. Let $S_h^* = [q_{h-1}^*, q_h^*)$ where $q_j^*$ is the $\sum_{h=1}^j p_h$th quartile

of $N(\boldsymbol{\beta}^\top\boldsymbol{\mu}, \boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma^2)$ and let $Z_j$ denote the $\sum_{h=1}^{j} p_h$th quartile for $N(0,1)$. Then

$$E(\boldsymbol{\beta}^\top\mathbf{x} + \sigma\epsilon \mid \boldsymbol{\beta}^\top\mathbf{x} + \sigma\epsilon \in S_h^*) = \sqrt{\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma^2}\,E(Z \mid Z \in [Z_{h-1}, Z_h)) + \boldsymbol{\beta}^\top\boldsymbol{\mu}. \tag{25}$$

Since $\boldsymbol{\beta}^\top\boldsymbol{\mu}$ and $\sigma\epsilon$ are independent normal random variables, it follows that

$$E(\boldsymbol{\beta}^\top\mathbf{x} \mid \boldsymbol{\beta}^\top\mathbf{x} + \sigma\epsilon) = \boldsymbol{\beta}^\top\boldsymbol{\mu} + \frac{\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}}{\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma^2}(\boldsymbol{\beta}^\top\mathbf{x} + \sigma\epsilon - \boldsymbol{\beta}^\top\boldsymbol{\mu}). \tag{26}$$

See page 290 of the Handbook of the Normal Distribution (Patel & Read, 1982).

Recall the response discretization transformation in (13). Now since $Y = f(\boldsymbol{\beta}^\top\mathbf{x} + \sigma\epsilon)$ where $f$ is strictly increasing, then $t(Y)$ can be re-expressed as

$$t(Y) = \begin{cases} D_1, & \boldsymbol{\beta}^\top\mathbf{x} + \sigma\epsilon \in S_1^*; \\ D_2, & \boldsymbol{\beta}^\top\mathbf{x} + \sigma\epsilon \in S_2^*; \\ \vdots & \vdots \\ D_H, & \boldsymbol{\beta}^\top\mathbf{x} + \sigma\epsilon \in S_H^*. \end{cases} \tag{27}$$

Using (25) and (26), where $\phi$ is the density function for the standard normal

$$\boldsymbol{\beta}^\top\mathrm{Cov}[\mathbf{x}, t(Y)] = \sum_{h=1}^{H} p_h D_h\{E[E(\boldsymbol{\beta}^\top\mathbf{x} \mid \boldsymbol{\beta}^\top\mathbf{x} + \sigma\epsilon) \mid \boldsymbol{\beta}^\top\mathbf{x} + \sigma\epsilon \in S_h^*] - \boldsymbol{\beta}^\top\boldsymbol{\mu}\}$$

$$= \frac{\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma^2}} \sum_{h=1}^{H} D_h\{\phi(Z_{h-1}) - \phi(Z_h)\}. \tag{28}$$

Note that $\mathbf{b}_t = c_t\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\mathrm{Cov}[\mathbf{x}, t(Y)]$ so that pre-multiplying both sides by $\boldsymbol{\beta}^\top\boldsymbol{\Sigma}$ gives $\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\mathbf{b}_t = c_t\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} = \boldsymbol{\beta}^\top\mathrm{Cov}[\mathbf{x}, t(Y)]$. Therefore from (28) we find

$$c_t = \frac{1}{\sqrt{\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma^2}} \sum_{h=1}^{H} D_h\{\phi(Z_{h-1}) - \phi(Z_h)\}. \tag{29}$$

## References

BRILLINGER, D. R. 1977. The identification of a particular nonlinear time series system. *Biometrika*, **64**, 509–515. MR0483236

BRILLINGER, D. R. 1983. A generalized linear model with "Gaussian" regressor variables. *Pages 97–114 of: A Festschrift for Erich L. Lehmann.* Wadsworth Statist./Probab. Ser. Belmont, CA: Wadsworth. MR0689741

CARROLL, R. J., & LI, K.-C. 1992. Measurement error regression with unknown link: dimension reduction and data visualization. *J. Amer. Statist. Assoc.*, **87**(420), 1040–1050.

CHATTERJEE, S., & HADI, A. S. 1988. *Sensitivity analysis in linear regression.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons Inc.

COOK, R. D. 1998. *Regression graphics.* New York: John Wiley & Sons Inc. Ideas for studying regressions through graphics. MR1645673

COOK, R. D., & WEISBERG, S. 1991. Discussion of "Sliced inverse regression for dimension reduction." *J. Amer. Statist. Assoc.*, **86**, 328–332.

CRITCHLEY, F. 1985. Influence in principal components analysis. *Biometrika*, **72**(3), 627–636. MR0817577

CUZICK, J. 1988. Rank regression. *Ann. Statist.*, **16**, 1369–1389. MR0964929

GOLUB, G. H., & VAN LOAN, C. F. 1996. *Matrix computations.* Third edn. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press. MR1417720

HALL, P., & LI, K.-C. 1993. On almost linearity of low-dimensional projections from high-dimensional data. *Ann. Statist.*, **21**, 867–889.

HAMPEL, F. R. 1974. The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, **69**, 383–393. MR0362657

HETTMANSPERGER, T. P., & MCKEAN, J. W. 1998. *Robust nonparametric statistical methods.* London: Arnold.

LI, K.-C. 1991. Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, **86**, 316–327. MR1137117

LI, K.-C. 1992. On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Amer. Statist. Assoc.*, **87**, 1025–1039. MR1209564

LI, K.-C., & DUAN, N. 1989. Regression analysis under link violation. *Ann. Statist.*, **17**, 1009–1052.

PATEL, J. K., & READ, C. B. 1982. *Handbook of the normal distribution.* Statistics: Textbooks and Monographs, vol. 40. New York: Marcel Dekker Inc. MR0664762

PRENDERGAST, L. A. 2008. Trimming influential observations for improved single-index model estimated sufficient summary plots. *Comput. Statist. Data Anal.*, **52**, 5319–5327. MR2526597

PRENDERGAST, L. A., & SMITH, J. A. 2010. Influence functions for dimension reduction methods: an example influence study of principal Hessian direction analysis. *Scand. J. Stat.*, **37**, 588–611.

PRENDERGAST, L. A., & SHEATHER, S. J. 2013. On sensitivity of inverse response plot estimation and the benefits of a robust estimation approach. *Scand. J. Stat.*, **40**(2), 219–237.

STAUDTE, R. G., & SHEATHER, S. J. 1990. *Robust estimation and testing.* New York: John Wiley & Sons Inc.

STEIN, C. M. 1981. Estimation of the mean of a multivariate normal distribution. *Ann. statist.*, **9**(6), 1135–1151. MR0630098