Robust Multivariate Linear Regression

David J. Olive*
Southern Illinois University

May 3, 2013

Abstract

A robust multivariate linear regression estimator can be obtained by replacing the least squares estimator with the robust `hbreg` estimator. Then the robust multivariate linear regression estimator is asymptotically equivalent to the classical multivariate linear regression estimator since the probability that the robust estimator is equal to the classical estimator goes to one in probability as the sample size $n \to \infty$ for a large class of iid zero mean error distributions. This paper discusses the robust estimator and some tests and prediction regions using the robust estimator that are asymptotically equivalent to those using the classical estimator. A second robust estimator that is useful for outlier detection is derived.

KEY WORDS: outliers; prediction regions.

*David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. E-mail address: dolive@siu.edu.

1 INTRODUCTION

Olive (2013b), using results from Su and Cook (2012) and Kakizawa (2009), derived a useful prediction region for the classical multivariate linear regression model, and gave F approximations to the widely used Wilk, Pillai, and Hotelling Lawley test statistics. This paper will show that these large sample tests and prediction regions also work for the robust multivariate linear regression estimator that replaces least squares with the `hbreg` estimator. This section reviews the multivariate linear regression model and the results from Olive (2013b). Section 2 reviews the `hbreg` estimator and derives the robust estimator and section 3 gives some examples and simulations. Section 4 derives a robust estimator that is useul for outlier detection.

1.1 The Multivariate Linear Regression Model

The *multivariate linear regression model* $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $x_1, x_2, ..., x_p$. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$ where the constant $x_{i1} = 1$ could be omitted from the case. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$ where the matrices are defined below. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma_\epsilon} = (\sigma_{ij})$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\mathrm{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $\boldsymbol{e}_i$ is defined below. Then the $p \times m$ coefficient matrix $\boldsymbol{B} = \left[\begin{array}{cccc} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \ldots & \boldsymbol{\beta}_m \end{array} \right]$ and the $m \times m$ covariance matrix $\boldsymbol{\Sigma_\epsilon}$ are to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$.

The $n \times m$ matrix of response variables

$$\boldsymbol{Z} = \left[\begin{array}{cccc} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \ldots & \boldsymbol{Y}_m \end{array} \right] = \left[\begin{array}{c} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{array} \right].$$

The $n \times p$ design matrix of predictor variables

$$\boldsymbol{X} = \left[\begin{array}{cccc} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \ldots & \boldsymbol{v}_p \end{array} \right] = \left[\begin{array}{c} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{array} \right]$$

where $\boldsymbol{v}_1 = \boldsymbol{1}$.

The $n \times m$ matrix of errors

$$\boldsymbol{E} = \left[\begin{array}{cccc} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \ldots & \boldsymbol{e}_m \end{array} \right] = \left[\begin{array}{c} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{array} \right].$$

Least squares is the classical method for fitting the multivariate linear model. The *least squares estimators* are $\hat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z} = \left[\begin{array}{cccc} \hat{\boldsymbol{\beta}}_1 & \hat{\boldsymbol{\beta}}_2 & \ldots & \hat{\boldsymbol{\beta}}_m \end{array} \right]$. The *predicted*

values or *fitted values* $\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \left[\begin{array}{cccc} \hat{\boldsymbol{Y}}_1 & \hat{\boldsymbol{Y}}_2 & \ldots & \hat{\boldsymbol{Y}}_m \end{array} \right]$. The *residuals*

$$\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} = \left[\begin{array}{c} \hat{\boldsymbol{\epsilon}}_1^T \\ \hat{\boldsymbol{\epsilon}}_2^T \\ \vdots \\ \hat{\boldsymbol{\epsilon}}_n^T \end{array} \right] = \left[\begin{array}{cccc} \hat{\boldsymbol{r}}_1 & \hat{\boldsymbol{r}}_2 & \ldots & \hat{\boldsymbol{r}}_m \end{array} \right].$$

These quantities can be found from the m multiple linear regressions of Y_j on the predictors: $\widehat{\boldsymbol{\beta}}_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_j$, $\widehat{\boldsymbol{Y}}_j = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j$ and $\hat{\boldsymbol{r}}_j = \boldsymbol{Y}_j - \hat{\boldsymbol{Y}}_j$ for $j = 1, ..., m$. Hence $\hat{\epsilon}_{i,j} = Y_{i,j} - \hat{Y}_{i,j}$ where $\hat{\boldsymbol{Y}}_j = (\hat{Y}_{1,j}, ..., \hat{Y}_{n,j})^T$. Finally,

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} = \frac{(\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T(\boldsymbol{Z} - \hat{\boldsymbol{Z}})}{n-d} = \frac{(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})^T(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})}{n-d} = \frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-d} = \frac{1}{n-d}\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T.$$

The choices $d = 0$ and $d = p$ are common. If $d = 1$, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=1} = \boldsymbol{S}_r$, the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ since the sample mean of the $\hat{\boldsymbol{\epsilon}}_i$ is $\boldsymbol{0}$. Let $\hat{\boldsymbol{\Sigma}}_\epsilon = \hat{\boldsymbol{\Sigma}}_{\epsilon,p}$ be the unbiased estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$.

The $\boldsymbol{\epsilon}_i$ are assumed to be iid. Some important joint distributions for $\boldsymbol{\epsilon}$ are completely specified by an $m \times 1$ population *location* vector $\boldsymbol{\mu}$ and an $m \times m$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. An important model is the elliptically contoured $EC_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with probability density function

$$f(\boldsymbol{z}) = k_m|\boldsymbol{\Sigma}|^{-1/2}g[(\boldsymbol{z} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z} - \boldsymbol{\mu})]$$

where $k_m > 0$ is some constant and g is some known function. The multivariate normal (MVN) $N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution is a special case.

Some additional notation will be useful. Assume that $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_n$ are iid from a multivariate distribution. The classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ of multivariate location and dispersion is the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \quad \text{and} \quad \boldsymbol{S} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\boldsymbol{x}_\mathrm{i} - \overline{\boldsymbol{x}})(\boldsymbol{x}_\mathrm{i} - \overline{\boldsymbol{x}})^\mathrm{T}. \tag{1}$$

Let the $p \times 1$ column vector T be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T, \boldsymbol{C}) = (\boldsymbol{x}_i - T)^T\boldsymbol{C}^{-1}(\boldsymbol{x}_i - T) \tag{2}$$

for each observation $\boldsymbol{x}_i$. Notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center T is $D_i(T, \boldsymbol{I}_p)$. The classical Mahalanobis distance uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. Following Johnson (1987, pp. 107-108), the population squared Mahalanobis distance

$$U \equiv D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \tag{3}$$

and for elliptically contoured distributions, U has probability density function (pdf)

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)}k_p u^{p/2-1}g(u). \tag{4}$$

1.2 A Prediction Region

Following Olive (2013b), given n cases of training or past data $(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$ and a vector of predictors $\boldsymbol{x}_f$, suppose it is desired to predict a future vector $\boldsymbol{y}_f$. Then a large sample $(1-\delta)100\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{y}_f \in \mathcal{A}_n) \to 1-\delta$ as $n \to \infty$, and is asymptotically optimal if the volume of the region converges in probability to the volume of the population minimum volume covering region.

If the $\boldsymbol{\epsilon}_i$ are iid from an $EC_m(\boldsymbol{0}, \boldsymbol{\Sigma}, g)$ distribution with continuous decreasing g and nonsingular covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = c\boldsymbol{\Sigma}$ for some constant $c > 0$, then the population asymptotically optimal prediction region is $\{\boldsymbol{y} : D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}\}$ where $P(D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}) = 1-\delta$. For example, if the iid $\boldsymbol{\epsilon}_i \sim N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then $D_{1-\delta} = \sqrt{\chi^2_{m,1-\delta}}$. If the error distribution is not elliptically contoured, then the above region still has $100(1-\delta)\%$ coverage, but prediction regions with smaller volume may exist.

Theorem 1, see Olive (2013b). Suppose $\boldsymbol{y}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i = \hat{\boldsymbol{y}}_i + \hat{\boldsymbol{\epsilon}}_i$ where $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, and where the zero mean $\boldsymbol{\epsilon}_f$ and the $\boldsymbol{\epsilon}_i$ are iid for $i = 1, ..., n$. Given $\boldsymbol{x}_f$, suppose the fitted model produces $\hat{\boldsymbol{y}}_f$ and nonsingular $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Let $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ and

$$D_i^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$$

for $i = 1, ..., n$. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + m/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta m/n), \quad \text{otherwise.}$$

If $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let $0 < \delta < 1$ and $h = D_{(U_n)}$ where $D_{(U_n)}$ is the q_nth sample quantile of the D_i. Let the nominal $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$ be given by $\{\boldsymbol{z} : (\boldsymbol{z} - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{z} - \hat{\boldsymbol{y}}_f) \leq D^2_{(U_n)}\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D^2_{(U_n)}\} =$

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}\}. \tag{5}$$

a) Consider the n prediction regions for the data where $(\boldsymbol{y}_{f,i}, \boldsymbol{x}_{f,i}) = (\boldsymbol{y}_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. If the order statistic $D_{(U_n)}$ is unique, then U_n of the n prediction regions contain $\boldsymbol{y}_i$ where $U_n/n \to 1 - \delta$ as $n \to \infty$.

b) If $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then (5) is a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$.

c) If $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and the $\boldsymbol{\epsilon}_i$ come from an elliptically contoured distribution such that the highest density region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}\}$, then the prediction region (5) is asymptotically optimal.

Notice that for the data $(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$, if $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}$ exists, then $100q_n\%$ of the n cases are in their corresponding prediction region, and $q_n \to 1 - \delta$ even if $(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is not a good estimator. Hence the coverage q_n of the training data is robust to model assumptions. Of course the volume of the prediction region could be large if a poor estimator $(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is used or if the $\boldsymbol{\epsilon}_i$ do not come from an elliptically contoured distribution. The nonparametric region uses $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ from the classical estimator in (5).

1.3 Testing

Following Olive (2013b), next consider testing a linear hypothesis $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{LB} \neq \boldsymbol{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix. Let $\boldsymbol{H} = \hat{\boldsymbol{B}}^T \boldsymbol{L}^T [\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}$. Let the error or residual sum of squares and cross products matrix be

$$\boldsymbol{W}_e = \hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}} = (\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T(\boldsymbol{Z} - \hat{\boldsymbol{Z}}) = \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{Z}^T[\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Z}.$$

Then $\boldsymbol{W}_e/(n-p) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the ordered eigenvalues of $\boldsymbol{W}_e^{-1}\boldsymbol{H}$. Then there are four commonly used test statistics.

The Roy's maximum root statistic is $\lambda_{max}(\boldsymbol{L}) = \lambda_1$.

The Wilk's Λ statistic is $\Lambda(\boldsymbol{L}) = |(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{W}_e| = |\boldsymbol{W}_e^{-1}\boldsymbol{H} + \boldsymbol{I}|^{-1} = \prod_{i=1}^{m}(1+\lambda_i)^{-1}$.

The Pillai's trace statistic is $V(\boldsymbol{L}) = tr[(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{H}] = \sum_{i=1}^{m} \frac{\lambda_i}{1+\lambda_i}$.

The Hotelling-Lawley trace statistic is $U(\boldsymbol{L}) = tr[\boldsymbol{W}_e^{-1}\boldsymbol{H}] = \sum_{i=1}^{m} \lambda_i$.

Theorem 2, Olive (2013b). The Hotelling-Lawley trace statistic

$$U(\boldsymbol{L}) = \frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]. \quad (6)$$

Some notation is useful to show (6) and to show that $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$ under mild conditions if H_0 is true. Following Henderson and Searle (1979), let matrix $\boldsymbol{A} = [\boldsymbol{a}_1 \ \ \boldsymbol{a}_2 \ \ \ldots \ \ \boldsymbol{a}_p]$. Then the vec operator stacks the columns of $\boldsymbol{A}$ on top of one another so

$$vec(\boldsymbol{A}) = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_p \end{pmatrix}.$$

Let $\boldsymbol{A} = (a_{ij})$ be an $m \times n$ matrix and $\boldsymbol{B}$ a $p \times q$ matrix. Then the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the $mp \times nq$ matrix

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2n}\boldsymbol{B} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix}.$$

An important fact is that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonsingular square matrices, then $[\boldsymbol{A} \otimes \boldsymbol{B}]^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$. The following assumption is important.

Assumption D1: Let h_i be the ith diagonal element of $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Assume $\max_{1 \leq i \leq n} h_i \to 0$ as $n \to \infty$, assume that the zero mean iid errors have finite fourth moments, and assume that $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \xrightarrow{P} \boldsymbol{W}^{-1}$.

Then for the least squares estimator, Su and Cook (2012) show that if assumption D1 holds, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ is $\sqrt{n}$ consistent and $\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$.

Theorem 3, Olive (2013b). If assumption D1 holds and if H_0 is true, then $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

Kakizawa (2009) shows, under stronger assumptions than Theorem 3, that for a large class of iid error distributions, the following test statistics have the same χ^2_{rm} limiting distribution when H_0 is true, and the same noncentral $\chi^2_{rm}(\omega^2)$ limiting distribution with noncentrality parameter ω^2 when H_0 is false under a local alternative. Hence the three tests are robust to the assumption of normality. The limiting null distribution is well known when the zero mean errors are iid from a multivariate normal distribution. See Khattree and Naik (1999, p. 68): $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, $(n-p)V(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, and $-[n-p-0.5(m-r+3)]\log(\Lambda(\boldsymbol{L})) \xrightarrow{D} \chi^2_{rm}$. Results from Kshirsagar (1972, p. 301) suggest that the chi-square approximation is very good if $n \geq 3(m^2+p^2)$ for multivariate normal errors.

Theorems 2 and 3 are useful for relating multivariate tests with the partial F test for multiple linear regression that tests whether a reduced model that omits some of the predictors can be used instead of the full model that uses all p predictors. The partial F test statistic is

$$F_R = \left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F)$$

where the residual sums of squares $SSE(F)$ and $SSE(R)$ and degrees of freedom df_F and df_r are for the full and reduced model while the mean square error $MSE(F)$ is for the full model. Let the null hypothesis for the partial F test be $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ where $\boldsymbol{L}$ sets the coefficients of the predictors in the full model but not in the reduced model to 0. Seber and Lee (2003, p. 100) shows that

$$F_R = \frac{[\boldsymbol{L}\hat{\boldsymbol{\beta}}]^T(\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}[\boldsymbol{L}\hat{\boldsymbol{\beta}}]}{r\hat{\sigma}^2}$$

is distributed as $F_{r,n-p}$ if H_0 is true and the errors are iid $N(0,\sigma^2)$. Note that for multiple linear regression with $m = 1$, $F_R = (n-p)U(\boldsymbol{L})/r$ since $\hat{\boldsymbol{\Sigma}}_\epsilon^{-1} = 1/\hat{\sigma}^2$. Hence the scaled Hotelling Lawley test statistic is the partial F test statistic extended to $m > 1$ predictor variables by Theorem 2.

By Theorem 3, for example, $rF_R \xrightarrow{D} \chi^2_r$ for a large class of nonnormal error distribution. If $Z_n \sim F_{k,d_n}$, then $Z_n \xrightarrow{D} \chi^2_k/k$ as $d_n \to \infty$. Hence using the $F_{r,n-p}$ approximation gives a large sample test with correct asymptotic level, and the partial F test is robust to nonnormality.

Similarly, using an $F_{rm,n-pm}$ approximation for the following test statistics gives large sample tests with correct asymptotic level by Kakizawa (2009) and similar power for large n. The large sample test will have correct asymptotic level as long as the denominator degrees of freedom $d_n \to \infty$ as $n \to \infty$, and $d_n = n - pm$ reduces to the partial F test if $m = 1$ and $U(\boldsymbol{L})$ is used. Then the three test statistics are

$$\frac{-[n-p-0.5(m-r+3)]}{rm} \log(\Lambda(\boldsymbol{L}), \quad \frac{n-p}{rm} V(\boldsymbol{L}), \text{ and } \frac{\mathrm{n-p}}{\mathrm{rm}} \mathrm{U}(\boldsymbol{L}).$$

Following Khattree and Naik (1999, p. 67) for the Roy's largest root test, if $h =$

$\max(r, m)$, use

$$\frac{n-p-h+r}{h}\lambda_{max}(\boldsymbol{L}) \approx F(h, n-p-h+r).$$

Simulations in Olive (2013b) suggest that this approximation is good for $r = 1$ but poor for $r > 1$. Anderson (1984, p. 333) states that Roy's largest root test has the greatest power if $r = 1$ but is an inferior test for $r > 1$.

Multivariate analogs of tests for multiple linear regression can be derived with appropriate choice of $\boldsymbol{L}$. Assume a constant $x_1 = 1$ is in the model. The analog of the ANOVA F test for multiple linear regression is the MANOVA F test that uses $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_{p-1}]$ to test whether the nontrivial predictors are needed in the model.

The F_j test of hypotheses uses $\boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$, where the 1 is in the jth position, to test whether the jth predictor is needed in the model given that the other $p-1$ predictors are in the model. This test is an analog of the t tests for multiple linear regression.

The MANOVA partial F test is used to test whether a reduced model is good where the reduced model deletes r of the variables from the full model. For this test, the ith row of $\boldsymbol{L}$ has a 1 in the position corresponding to the ith variable to be deleted. Omitting the jth variable corresponds to the F_j test while omitting variables $x_2, ..., x_p$ corresponds to the MANOVA F test. Using $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_k]$ tests whether the last k predictors are needed in the multivariate linear regression model given that the remaining predictors are in the model.

2 Robust Estimators

2.1 Resistant Regression Estimators for Multiple Linear Regression

Consider the multiple linear regression model, written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. This model is a special case of the multivariate linear regression model with $m = 1$.

Resistant estimators are useful for detecting certain types of outliers. Resistant estimators are often created by computing several trial fits $\boldsymbol{b}_i$ that are estimators of $\boldsymbol{\beta}$. Then a criterion is used to select the trail fit to be used in the resistant estimator. Suppose $c \approx n/2$. The LMS(c) criterion is $Q_{LMS}(\boldsymbol{b}) = r^2_{(c)}(\boldsymbol{b})$ where $r^2_{(1)} \leq \cdots \leq r^2_{(n)}$ are the ordered squared residuals, and the LTS(c) criterion is $Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c} r^2_{(i)}(\boldsymbol{b})$. The LTA($c$) criterion is $Q_{LTA}(\boldsymbol{b}) = \sum_{i=1}^{c} |r(\boldsymbol{b})|_{(i)}$ where $|r(\boldsymbol{b})|_{(i)}$ is the ith ordered absolute residual. Three impractical high breakdown robust estimators are the Hampel (1975) least median of squares (LMS) estimator, the Rousseeuw (1984) least trimmed sum of squares (LTS) estimator, and the Hössjer (1991) least trimmed sum of absolute deviations (LTA) estimator. These estimators correspond to the $\hat{\boldsymbol{\beta}}_L \in \mathcal{R}^p$ that minimizes the corresponding criterion.

A good resistant estimator is the Olive (2005) *median ball algorithm* (MBA or `mbareg`). The Euclidean distance of the ith vector of predictors $\boldsymbol{x}_i$ from the jth vector of predictors $\boldsymbol{x}_j$ is

$$D_i(\boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)}.$$

For a fixed $\boldsymbol{x}_j$ consider the ordered distances $D_{(1)}(\boldsymbol{x}_j), ..., D_{(n)}(\boldsymbol{x}_j)$. Next, let $\hat{\boldsymbol{\beta}}_j(\alpha)$ denote the OLS fit to the $\min(p+3+\lfloor \alpha n/100 \rfloor, n)$ cases with the smallest distances where the approximate percentage of cases used is $\alpha \in \{1, 2.5, 5, 10, 20, 33, 50\}$. (Here $\lfloor x \rfloor$ is the greatest integer function so $\lfloor 7.7 \rfloor = 7$. The extra $p+3$ cases are added so that OLS can be computed for small n and α.) This yields seven OLS fits corresponding to the cases with predictors closest to $\boldsymbol{x}_j$. A fixed number K of cases are selected at random without replacement to use as the $\boldsymbol{x}_j$. Hence $7K$ OLS fits are generated. We use $K = 7$ as the default. A robust criterion Q is used to evaluate the $7K$ fits and the OLS fit to all of the data. Hence $7K + 1$ OLS fits are generated and the MBA estimator is the fit that minimizes the criterion. The median squared residual is a good choice for Q.

Three ideas motivate this estimator. First, $\boldsymbol{x}$-outliers, which are outliers in the predictor space, tend to be much more destructive than Y-outliers which are outliers in the response variable. Suppose that the proportion of outliers is γ and that $\gamma < 0.5$. We would like the algorithm to have at least one "center" $\boldsymbol{x}_j$ that is not an outlier. The probability of drawing a center that is not an outlier is approximately $1 - \gamma^K > 0.99$ for $K \geq 7$ and this result is free of p. Secondly, by using the different percentages of coverages, for many data sets there will be a center and a coverage that contains no outliers. Thirdly, the MBA estimator is a $\sqrt{n}$ consistent estimator.

The Olive and Hawkins (2011) `hbreg` estimator is a robust estimator that is asymptotically equivalent to the least squares estimator for many error distributions. Assume that the multiple linear regression model contains an intercept and that the median absolute deviation (MAD) of the Y values from their median is finite. Make an OLS fit to the c_n cases whose Y values are closest to the median Y, and use this fit as the start for concentration: find the c_n cases with the smallest squared residuals and fit OLS to these cases. Use 10 concentration steps and let the attractor be the final estimator, denoted by $\hat{\boldsymbol{\beta}}_B$. It can be shown that $\hat{\boldsymbol{\beta}}_B$ is a high breakdown estimator.

With these preliminaries, we now define our high breakdown procedure. This is made up of three components.
1) The OLS estimator $\hat{\boldsymbol{\beta}}_C$ that is consistent for clean data.
2) The practical $\sqrt{n}$ consistent `mbareg` estimator $\hat{\boldsymbol{\beta}}_A$ that is effective for outlier identification.
3) The practical high-breakdown estimator $\hat{\boldsymbol{\beta}}_B$.

By selecting one of these three estimators according to the features each of them uncovers in the data, we may inherit the good properties of each of them.

The `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ is defined as follows. Pick a constant $a > 1$ and set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_C$. If $aQ_L(\hat{\boldsymbol{\beta}}_A) < Q_L(\hat{\boldsymbol{\beta}}_C)$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_A$. If $aQ_L(\hat{\boldsymbol{\beta}}_B) < \min[Q_L(\hat{\boldsymbol{\beta}}_C), aQ_L(\hat{\boldsymbol{\beta}}_A)]$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_B$.

That is, find the smallest of the three scaled criterion values $Q_L(\hat{\boldsymbol{\beta}}_C)$, $aQ_L(\hat{\boldsymbol{\beta}}_A)$, $aQ_L(\hat{\boldsymbol{\beta}}_B)$. According to which of the three estimators attains this minimum, set $\hat{\boldsymbol{\beta}}_H$ to $\hat{\boldsymbol{\beta}}_C, \hat{\boldsymbol{\beta}}_A$ or $\hat{\boldsymbol{\beta}}_B$ respectively.

Large sample theory for `hbreg` is simple and given in the following theorem. Let $\hat{\boldsymbol{\beta}}_L$ be the LMS, LTS or LTA estimator that minimizes the criterion Q_L. Note that the impractical estimator $\hat{\boldsymbol{\beta}}_L$ is never computed. The following theorem shows that $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C = \hat{\boldsymbol{\beta}}_{OLS}$. The clean data are in general position if any p

clean cases give a unique estimate of $\hat{\boldsymbol{\beta}}$. The LTA criterion will be used in the simulations.

Theorem 4, Olive and Hawkins (2011). Assume the clean data are in general position, and suppose that both $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_C$ are consistent estimators of $\boldsymbol{\beta}$ where the regression model contains a constant. Then the `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ is high breakdown and is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$ since the probability that $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_C$ goes to one as $n \to \infty$.

2.2 Robust Multivariate Linear Regression

The classical multivariate linear regression estimator is found from m least squares multiple linear regressions of Y_j on the predictors. The robust multivariate linear regression estimator is found from m `hbreg` multiple linear regressions of Y_j on the predictors. By Theorem 4, the probability that the robust estimator is equal to the classical estimator goes to one as $n \to \infty$ for a large class of error distributions.

Hence the large sample nonparametric prediction region and the large sample Wilk's test, Pillai's test and Hotelling Lawley test using the robust estimator are asymptotically equivalent to their analogs using the classical estimator for a large class of error distributions. The next section investigates whether reasonable sample sizes result in good results for the robust estimator.

3 Plots, Examples and Simulations

3.1 Plots

A *response plot* for the jth response variable is a plot of the fitted values $\hat{Y}_{ij}$ versus the response Y_{ij} where $i = 1, ..., n$. The identity line with slope one and zero intercept is added to the plot as a visual aid. A *residual plot* corresponding to the jth response variable is a plot of $\hat{Y}_{ij}$ versus r_{ij}.

Make the m response and residual plots for the multivariate linear regression model. In a response plot, the vertical deviations from the identity line are the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij}$. The plotted points in the response plot should cluster about the identity line in each of the m response plots. If outliers are present or if the plot is not linear, then the current model or data need to be changed or corrected. The response and residual plots are used just as for multiple linear regression where $m = 1$. See Olive and Hawkins (2005) and Cook and Weisberg (1999, p. 432).

The Rousseeuw and Van Driessen (1999) DD plot is a plot of classical Mahalanobis distances versus robust Mahalanobis distances. Results from Olive (2002) suggest the plotted points in the DD plot will cluster about the identity line if the $\boldsymbol{\epsilon}_i$ are iid from a multivariate normal $N_m(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ distribution and about some line through the origin with slope greater than one for a large class of elliptically contoured distributions. Make a DD plot of the residuals $\hat{\boldsymbol{\epsilon}}_i$ to check the error distribution. Make a DD plot of the continuous predictor variables to check for $\boldsymbol{x}$-outliers.

The Olive and Hawkins (2010) RMVN estimator $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ is an easily computed $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ under regularity conditions (E1) that include a large class of elliptically contoured distributions, and $c = 1$ for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. The RMVN estimator also gives useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data even when certain types of outliers are present, and will be the robust estimator used in the DD plots. Also see Zhang, Olive and Ye (2012).

Consider the DD plot applied to the $\hat{\boldsymbol{z}}_i$ based on the robust estimator. The nonparametric region based on the robust estimator uses the sample mean and sample covariance matrix applied to the $\hat{\boldsymbol{z}}_i$. The DD plot will have a vertical line at the cutoff $D_{(U_n)}$. Hence points to the left of the line correspond to cases that are in the nonparametric region. The RMVN estimator can be applied to the $\hat{\boldsymbol{z}}_i$. The region that uses $D_i(T_{RMVN}, \boldsymbol{C}_{RMVN}) \leq D_{(U_n)}(T_{RMVN}, \boldsymbol{C}_{RMVN})$ will be called the semiparametric region, while the parametric MVN region uses $D_i(T_{RMVN}, \boldsymbol{C}_{RMVN}) \leq \sqrt{\chi^2_{p,q_n}}$ where $P(W \leq \chi^2_{p,q_n}) = q_n$ if $W \sim \chi^2_p$. These two regions are only conjectured to be large sample prediction regions, but are added to the DD plot as visual aids. Cases below the horizontal line that crosses the identity line correspond to the semiparametric region while cases below the horizontal line that ends at the identity line correspond to the parametric MVN region. A vertical line dropped down from this point of intersection does correspond to a large sample prediction region for multivariate normal data. Note that $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$, and adding a constant $\hat{\boldsymbol{y}}_f$ to all of the residual vectors does not change the Mahalanobis distances, so the DD plot of the residual vectors can be used to display the prediction regions.

3.2 Examples and Simulations

Example 1. Cook and Weisberg (1999, p. 351, 433, 447) give a data set on 82 mussels sampled off the coast of New Zealand. Let $Y_1 = \log(S)$ and $Y_2 = \log(M)$ where S is the shell mass and M is the muscle mass. The predictors are $X_2 = L$, $X_3 = \log(W)$ and $X_4 = H$: the shell length, log(width) and height. Figures 1 and 2 give the response and residual plots for Y_1 and Y_2. For Y_1, case 79 sticks out while for Y_2, cases 8, 25 and 48 are not fit well. Figure 3 shows the DD plot of the residual vectors. The nonparametric 90% prediction region for the residuals consists of the points to the left of the vertical line $MD = 2.60$. Cases 8, 48 and 79 have especially large distances. For this data set, the classical and robust estimators were identical, and hence the Cook (1977) distances can be computed. Highlighted cases had Cook's distance $> \min(0.5, 2p/n)$. The response, residual and DD plots are effective for finding influential cases, for checking linearity and whether the error distribution is multivariate normal or some other elliptically contoured distribution, and for displaying the nonparametric prediction region. Note that cases to the right of the vertical line correspond to cases that are not in their prediction region. These are the cases corresponding to residual vectors with large Mahalanobis distances. Also adding a constant does not change the distance, so the DD plot for the residuals is the same as the DD plot for the $\hat{\boldsymbol{z}}_i$.

Example 2. Buxton (1920) gives various measurements of 88 men. *Head length* and person's *height* were the response variables while an intercept, *nasal height, bigonal*

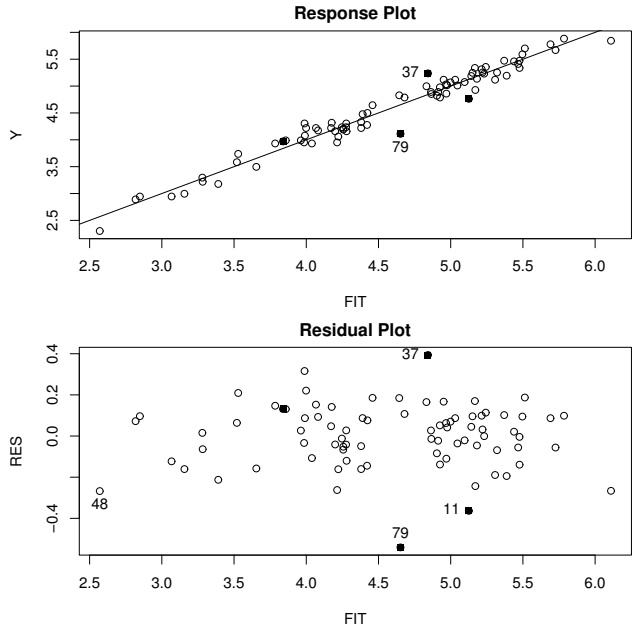

Figure 1: Plots for $Y_1 = \log(S)$.

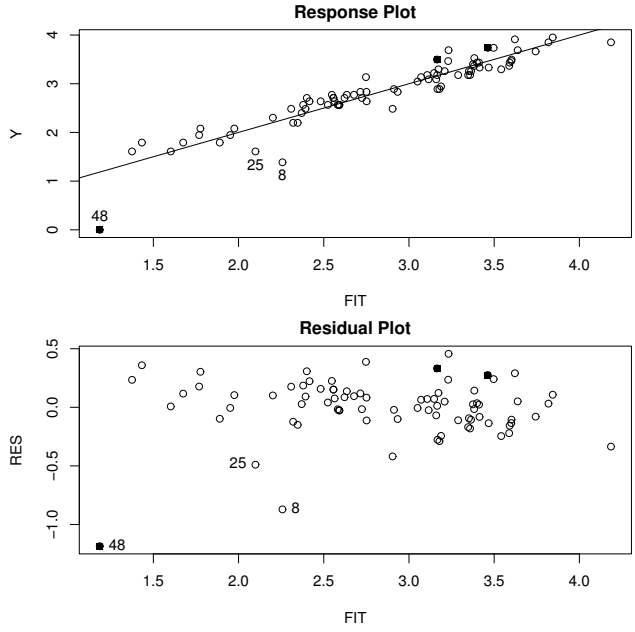

Figure 2: Plots for $Y_2 = \log(M)$.

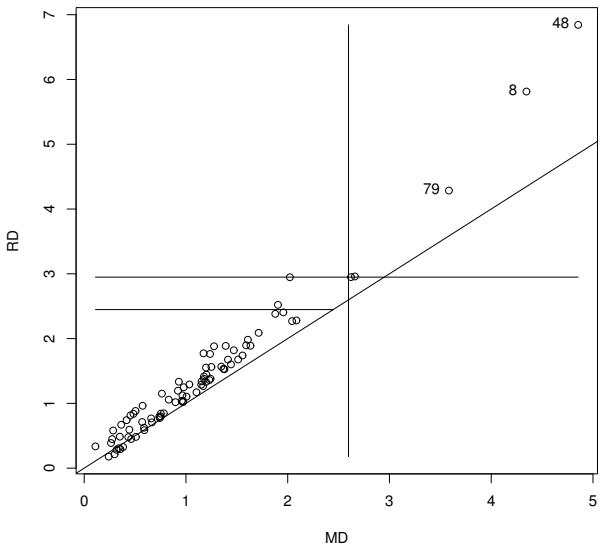

Figure 3: DD Plot of the Residual Vectors.

breadth, and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, numbers 62–66, were reported to be about 0.75 inches tall with head lengths well over five feet! Figure 4 shows the response and residual plots corresponding to Y_1 for the robust estimator. The response plot for the classical estimator, not shown, has the identity line tilted slightly above most of the plotted points in the lower part of the plot, while the plotted points in the lower part of the residual plot follow a line with negative slope instead of the $r = 0$ line. Figure 5 shows the response and residual plots corresponding to Y_2 for the robust estimator. The response plot for the classical estimator, not shown, has the identity line tilted slightly below most of the plotted points in the upper part of the plot, while the plotted points in the upper part of the residual plot follow a line with negative slope instead of the $r = 0$ line. Figure 6 shows the DD plot. The 90% semiparametric and nonparametric regions use the 95th percentile which is a linear combination of an outlying case with a nonoutlying case. The parametric MVN region contains cases below the RD = 2.448 line, which is obscured by the identity line. The tests of hypotheses for the robust estimator are not robust to outliers because all $n = 87$ residual vectors are used to make $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. As is typically the case, outliers can be detected with the plots using the classical or robust estimator.

A small simulation was used to study the prediction region and the Wilk's Lambda test, the Pillai's trace test, the Hotelling Lawley trace test, and the Roy's largest root test for the F_j tests and the MANOVA F test for multivariate linear regression. These test statistics were computed with the robust estimator $\hat{\boldsymbol{B}}$ instead of the classical estimator. The first row of $\boldsymbol{B}$ was always $\mathbf{1}^T$ and the last row of $\boldsymbol{B}$ was always $\mathbf{0}^T$. When the null hypothesis for the MANOVA F test is true, all but the first row corresponding to the constant are equal to $\mathbf{0}^T$. When $p \geq 3$ and the null hypothesis for the MANOVA F test

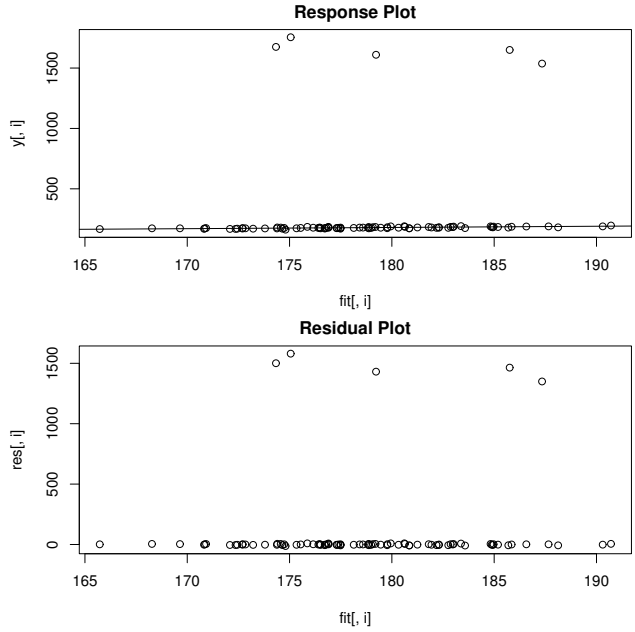

Figure 4: Plots for Y_1 = head length.

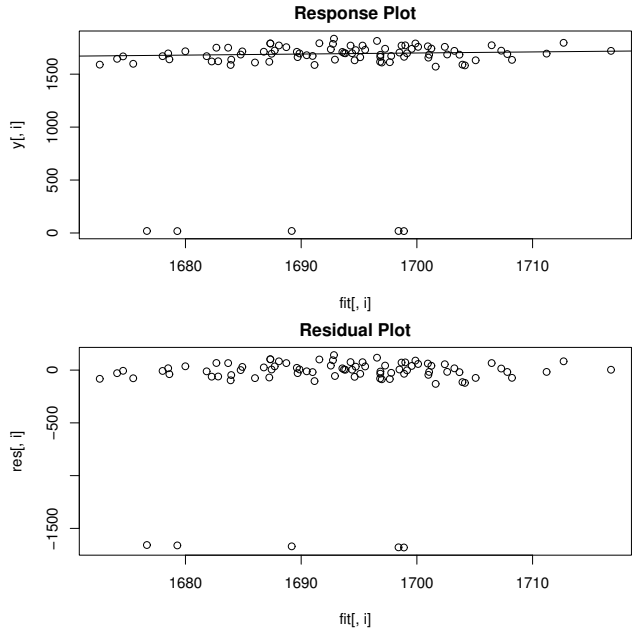

Figure 5: Plots for Y_2 = height.

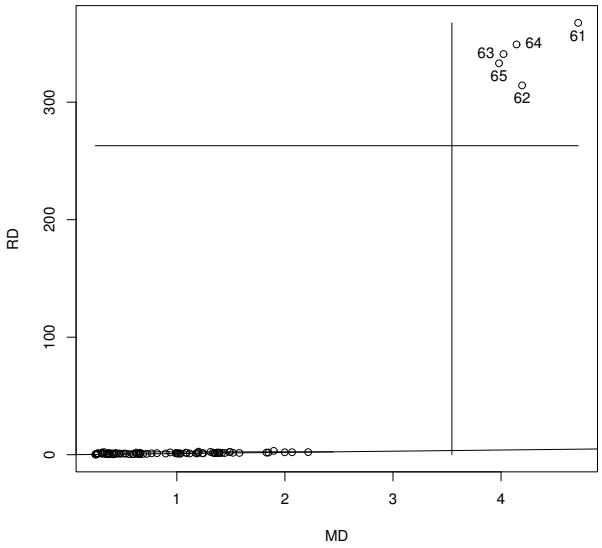

Figure 6: DD Plot of the Residual Vectors for the Buxton Data.

is false, then the second to last row of $\boldsymbol{B}$ is (1, 0, ..., 0), the third to last row is (1, 1, 0, ..., 0) etcetera as long as the first row is not changed from $\mathbf{1}^T$. First m iid errors $\boldsymbol{w}_i$ are generated such that the m errors are iid with variance σ^2. Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = \sigma^2[1 + (m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = \sigma^2[2\psi + (m-2)\psi^2]$ where $\psi = 0.10$. Hence the correlations $(2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$. As ψ gets close to 1, the data clusters about the line in the direction of $(1, ..., 1)^T$. Used $\boldsymbol{w}_i \sim N_m(\mathbf{0}, \boldsymbol{I}), \boldsymbol{w}_i \sim (1-\tau)N_m(\mathbf{0}, \boldsymbol{I}) + \tau N_m(\mathbf{0}, 25\boldsymbol{I})$ with $0 < \tau < 1$ and $\tau = 0.25$ in the simulation, $\boldsymbol{w}_i \sim$ multivariate t_d with $d = 7$ degrees of freedom, or $\boldsymbol{w}_i \sim$ lognormal - E(lognormal): where the m components of $\boldsymbol{w}_i$ were iid with distribution $e^z - E(e^z)$ where $z \sim N(0, 1)$. Only the lognormal distribution is not elliptically contoured.

The simulation used 5000 runs, and H_0 was rejected if the F statistic was greater than $F_{d_1,d_2}(0.95)$ where $P(F_{d_1,d_2} < F_{d_1,d_2}(0.95)) = 0.95$ with $d_1 = rm$ and $d_2 = n - mp$ for the test statistics

$$\frac{-[n-p-0.5(m-r+3)]}{rm} \log(\Lambda(\boldsymbol{L}), \quad \frac{n-p}{rm} V(\boldsymbol{L}), \text{ and } \frac{\mathrm{n-p}}{\mathrm{rm}} \mathrm{U}(\boldsymbol{L})$$

while $d_1 = h = \max(r, m)$ and $d_2 = n - p - h + r$ for the test statistic

$$\frac{n-p-h+r}{h}\lambda_{max}(\boldsymbol{L}).$$

Denote these statistics by W, P, HL and R. Let the coverage be the proportion of times that H_0 is rejected. Want coverage near 0.05 when H_0 is true and coverage close to 1 for good power when H_0 is false. With 5000 runs, coverage outside of (0.04,0.06) suggests

that the true coverage is not 0.05. Coverages are tabled for the F_1, F_2, F_{p-1}, and F_p test and for the MANOVA F test denoted by F_M. The null hypothesis H_0 was always true for the F_p test and always false for the F_1 test. When the MANOVA F test was true, H_0 was true for the F_j tests with $j \neq 1$. When the MANOVA F test was false, H_0 was false for the F_j tests with $j \neq p$, but the F_{p-1} test should be hardest to reject for $j \neq p$ by construction of $\boldsymbol{B}$ and the error vectors.

The `hbreg` estimator is asymptotically equivalent to OLS provided that OLS is consistent and the probability that `hbreg` selects $\hat{\boldsymbol{\beta}}_B$ goes to 0 as $n \to \infty$. For the simulated data with symmetric error distributions, $\hat{\boldsymbol{\beta}}_B$ appeared to give biased estimates of the slopes. However, for the simulated data with right skewed error distributions, $\hat{\boldsymbol{\beta}}_B$ appeared to give good estimates of the slopes but not the constant, and the probability that the `hbreg` estimator selected $\hat{\boldsymbol{\beta}}_B$ appeared to go to one. Removing $\hat{\boldsymbol{\beta}}_B$ from the `hbreg` estimator results in a $\sqrt{n}$ consistent estimator when OLS is $\sqrt{n}$ consistent, but massive sample sizes were still needed to get good estimates of the constants for highly skewed error distributions. Although the `mbareg` estimator is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$, if $m = 1$ and OLS needed $n = 1000$ to estimate the constant well, `mbareg` might need $n >$ one million. Getting all m `mbareg` estimators to estimate the constant well needs even larger sample sizes.

This paragraph will explain why `mbareg` needs large samples to give a good estimate of the constant for highly skewed error distributions when $m = 1$. Note that the LMS, LTA and LMS criteria use half sets. For simplicity, consider the LMS criterion that minimizes the median squared residual. Heuristically, for highly right skewed data, let the "left tail half set" shift the constant of the OLS hyperplane down so that the half set of cases closest to the plane are the half set with the smallest OLS residual values. These cases will have negative residuals and residuals close to zero, which are roughly the cases corresponding to the half set of errors in the left tail of the error distribution. Let the "OLS half set" correspond to the half set of cases with the smallest absolute OLS residuals, so the cases closest to the OLS hyperplane. Since the distribution is highly right skewed, the "OLS half set" has much more variability than the "left tail half set." (For the location model, OLS is the sample mean which is greater than the sample median for right skewed data. The "left tail half set" shifts the mean down to the midpoint c of the minimum value and the median value, and often $c \approx$ median/2 if the support of the highly right skewed distribution is $(0, \infty)$.) A trial fit that uses the same OLS slope estimates but which shifts the intercept down to use the "left tail half set" will have a smaller median squared residual than the median squared residual using OLS. The trial fits for `mbareg` are $\sqrt{n}$ consistent, so estimate the OLS intercept eventually. However, the trial fit that uses 1% of the data has less than 1% efficiency since the $\boldsymbol{x}$ values are close in distance rather than spread out. Unless the sample size is large, the `mbareg` estimator tends to produce some trial fits that shift the intercept down, and one of these trial fits is selected to be the final `mbareg` estimator since it has a smaller median squared residual than the other trial fits.

In the simulations, `hbreg` estimated the slopes well for the highly skewed lognormal data, but not the constant. This results in incorrect residual vectors and test statistics. The `hbreg` tests can be used as diagnostics if the plotted points in the DD plot cluster tightly about a line through the origin. Also compare $\hat{\boldsymbol{B}}$ for the robust and classical

Table 1: Test Coverages: MANOVA F H_0 is True.

$\boldsymbol{w}$ dist	n	test	F_1	F_2	F_{p-1}	F_p	F_M
MVN	50	W	1	0.051	0.051	0.054	0.029
MVN	50	P	1	0.036	0.033	0.038	0.006
MVN	50	HL	1	0.104	0.110	0.119	0.130
MVN	50	R	1	0.098	0.100	0.110	0.727
MVN	200	W	1	0.044	0.046	0.044	0.044
MVN	200	P	1	0.042	0.042	0.042	0.035
MVN	200	HL	1	0.056	0.056	0.054	0.060
MVN	200	R	1	0.050	0.051	0.050	0.520
MIX	200	W	1	0.043	0.043	0.044	0.034
MIX	200	P	1	0.041	0.040	0.040	0.027
MIX	200	HL	1	0.054	0.053	0.053	0.048
MIX	200	R	1	0.050	0.049	0.048	0.518
MVT(7)	200	W	1	0.040	0.042	0.043	0.036
MVT(7)	200	P	1	0.038	0.040	0.040	0.028
MVT(7)	200	HL	1	0.049	0.049	0.053	0.051
MVT(7)	200	R	1	0.046	0.046	0.049	0.524

estimator. If the slopes are similar but not the intercepts, there may be highly skewed data or y-outliers.

Hence the simulation for tests of hypotheses used the symmetric elliptically contoured distributions. In Olive (2013b) for the classical estimator when the null hypothesis H_0 was true, simulated values started to get close to nominal levels for $n \geq 0.75(m+p)^2$, and were fairly good for $n \geq 1.5(m+p)^2$. The exception was Roy's test which rejects H_0 far too often if $r > 0$. Roy's test was very good for the F_j tests but very poor for the MANOVA F test.

The robust estimator needed larger values of n, and results are shown in Table 1 for $m = p = 5$. Want values for the F_1 test to be close to 1 since H_0 is false for the F_1 test and want values close to 0.05, otherwise. Results from Berndt and Savin (1977) suggest that Pillai's test will reject H_0 less often than Wilk's test which will reject less often than the Hotelling Lawley test. For Table 2 want values near 1 except for the F_p column.

The same type of data and 5000 runs were used to simulate the prediction regions for $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ for multivariate regression. With n=100, m=2, and p=4, the nominal coverage of the prediction region is 90%, and 92% of the training data is covered. Following Olive (2013a), consider the prediction region $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$. Then the ratio of the prediction region volumes

$$\frac{h_i^m \sqrt{det(\boldsymbol{C}_i)}}{h_2^m \sqrt{det(\boldsymbol{C}_2)}}$$

was recorded where $i = 1$ was the nonparametric region, $i = 2$ was the semiparametric

Table 2: Test Coverages: MANOVA F H_0 is False.

n	$\boldsymbol{w}$ dist	$m = p$	test	F_1	F_2	F_{p-1}	F_p	F_M
30	MVN	5	W	0.807	0.522	0.152	0.000	0.043
30	MVN	5	P	0.000	0.000	0.000	0.000	0.000
30	MVN	5	HL	0.995	0.889	0.570	0.009	0.878
30	MVN	5	R	1	0.956	0.756	0.076	0.999
50	MVN	5	W	1	0.999	0.921	0.057	1
50	MVN	5	P	1	0.998	0.911	0.042	0.997
50	MVN	5	HL	1	0.999	0.943	0.116	1
50	MVN	5	R	1	0.999	0.941	0.107	1
200	MVN	5	W	1	1	1	0.051	1
200	MVN	5	P	1	1	1	0.051	1
200	MVN	5	HL	1	1	1	0.060	1
200	MVN	5	R	1	1	1	0.056	1
50	MIX	5	W	0.803	0.569	0.278	0.010	0.502
50	MIX	5	P	0.747	0.518	0.236	0.006	0.230
50	MIX	5	HL	0.897	0.689	0.392	0.025	0.723
50	MIX	5	R	0.888	0.675	0.377	0.022	0.970
50	MVT	5	W	1	0.997	0.892	0.040	0.998
50	MVT	5	P	1	0.996	0.871	0.025	0.989
50	MVT	5	HL	1	0.999	0.925	0.087	1
50	MVT	5	R	1	0.998	0.922	0.080	1
450	MVN	20	W	1	1	1	0.015	1
450	MVN	20	P	1	1	1	0.013	1
450	MVN	20	HL	1	1	1	0.030	1
450	MVN	20	R	1	1	1	0.053	1

Table 3: Coverages for 90% Prediction Regions.

$\boldsymbol{w}$ dist	n	$m = p$	ncov	scov	mcov	nvol	mvol
MVN	300	5	0.903	0.899	0.902	1.006	1.017
MIX	300	5	0.897	0.907	0.688	0.885	0.001
MVT(7)	300	5	0.901	0.910	0.775	0.913	0.291
LN	300	5	0.915	0.916	0.592	0.688	0.008

region, and $i = 3$ was the parametric MVN region. Here h_1 and h_2 were the cutoff $D_{(U_n)}(T_i, \boldsymbol{C}_i)$ for $i = 1, 2$, and $h_3 = \sqrt{\chi^2_{m,q_n}}$.

If, as conjectured, the RMVN estimator is a consistent estimator when applied to the residual vectors instead of iid data, then the volume ratios converge in probability to 1 if the iid zero mean errors $\sim N_m(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon})$, and the volume ratio converges to 1 for $i = 1$ for a large class of elliptically contoured distributions. These volume ratios were denoted by voln and volm for the nonparametric and parametric MVN regions. The coverage was the proportion of times the prediction region contained $\boldsymbol{y}_f$ where ncov, scov and mcov are for the nonparametric, semiparametric and parametric MVN regions.

In the simulations, took $n = 3(m+p)^2 = 300$ and $m = p = 5$. Table 3 shows that the coverage of the nonparametric region was close to 0.9 in all cases. The volume ratio voln was fairly close to 1 for the three elliptically contoured distributions. Since the volume of the prediction region is proportional to h^m, the volume can be very small if h is too small and m is large. Parametric prediction regions usually give poor estimates of h when the parametric distribution is mispecified. Hence the parametric MVN region only performed well for multivariate normal data. The results using the robust estimator were nearly the same as those using the classical estimator. See Table 3 in Olive (2013b).

4 Another Robust Estimator

First we will review some results for multiple linear regression. Let $\boldsymbol{x} = (1, \boldsymbol{w}^T)^T$ and let

$$\mathrm{Cov}(\boldsymbol{w}) = \mathrm{E}[(\boldsymbol{w} - \mathrm{E}(\boldsymbol{w}))(\boldsymbol{w} - \mathrm{E}(\boldsymbol{w}))^{\mathrm{T}}] = \boldsymbol{\Sigma_w}$$

and $\mathrm{Cov}(\boldsymbol{w}, Y) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(Y - E(Y))] = \boldsymbol{\Sigma_{wY}}$. Let $\boldsymbol{\beta} = (\alpha, \boldsymbol{\eta}^T)^T$ be the population OLS coefficients from the regression of Y on $\boldsymbol{x}$ ($\boldsymbol{w}$ and a constant), where α is the constant and $\boldsymbol{\eta}$ is the vector of slopes. Let the OLS estimator be $\hat{\boldsymbol{\beta}} = (\hat{\alpha}, \hat{\boldsymbol{\eta}}^T)^T$. Then the population coefficients from an OLS regression of Y on $\boldsymbol{x}$ are

$$\alpha = E(Y) - \boldsymbol{\eta}^T E(\boldsymbol{w}) \ \text{ and } \ \boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{w}\mathrm{Y}}. \tag{7}$$

Then the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. The sample covariance matrix of $\boldsymbol{w}$ is

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T \ \text{ where the sample mean } \ \overline{\boldsymbol{w}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \boldsymbol{w}_{\mathrm{i}}.$$

Similarly, define the sample covariance vector of $\boldsymbol{w}$ and Y to be

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(Y_i - \overline{Y}).$$

Suppose that $(Y_i, \boldsymbol{w}_i^T)^T$ are iid random vectors such that $\boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{w}Y}$ exist. Then

$$\hat{\alpha} = \overline{Y} - \hat{\boldsymbol{\eta}}^T\overline{\boldsymbol{w}} \xrightarrow{D} \alpha$$

and

$$\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y} \xrightarrow{D} \boldsymbol{\eta} \text{ as } \mathrm{n} \to \infty.$$

Now for multivariate linear regression, $\hat{\boldsymbol{\beta}}_j = (\hat{\alpha}_j, \hat{\boldsymbol{\eta}}_j^T)^T$ where $\hat{\alpha}_j = \overline{Y}_j - \hat{\boldsymbol{\eta}}_j^T\overline{\boldsymbol{w}}$ and $\hat{\boldsymbol{\eta}}_j = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y_j}$. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}\boldsymbol{y}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^T$ which has jth column $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y_j}$ for $j = 1, ..., m$. Let

$$\boldsymbol{u} = \begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{y} \end{pmatrix}, \quad E(\boldsymbol{u}) = \boldsymbol{\mu}_{\boldsymbol{u}} = \begin{pmatrix} E(\boldsymbol{w}) \\ E(\boldsymbol{y}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{w}} \\ \boldsymbol{\mu}_{\boldsymbol{y}} \end{pmatrix}, \text{ and } \mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} =$$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{w}\boldsymbol{w}} & \boldsymbol{\Sigma}_{\boldsymbol{w}\boldsymbol{y}} \\ \boldsymbol{\Sigma}_{\boldsymbol{y}\boldsymbol{w}} & \boldsymbol{\Sigma}_{\boldsymbol{y}\boldsymbol{y}} \end{pmatrix}.$$

Let the vector of constants be $\boldsymbol{\alpha}^T = (\alpha_1, ..., \alpha_m)$ and the matrix of slope vectors $\boldsymbol{B}_S = \begin{bmatrix} \boldsymbol{\eta}_1 & \boldsymbol{\eta}_2 & \dots & \boldsymbol{\eta}_m \end{bmatrix}$. Then the population least squares coefficient matrix is

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{\alpha}^T \\ \boldsymbol{B}_S \end{pmatrix}$$

where $\boldsymbol{\alpha} = \boldsymbol{\mu}_{\boldsymbol{y}} - \boldsymbol{B}_S^T\boldsymbol{\mu}_{\boldsymbol{w}}$ and $\boldsymbol{B}_S = \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{w}\boldsymbol{y}}$ where $\boldsymbol{\Sigma}_{\boldsymbol{w}} = \boldsymbol{\Sigma}_{\boldsymbol{w}\boldsymbol{w}}$.

If the $\boldsymbol{u}_i$ are iid with nonsingular covariance matrix $\mathrm{Cov}(\boldsymbol{u})$, the least squares estimator

$$\hat{\boldsymbol{B}} = \begin{pmatrix} \hat{\boldsymbol{\alpha}}^T \\ \hat{\boldsymbol{B}}_S \end{pmatrix}$$

where $\hat{\boldsymbol{\alpha}} = \overline{\boldsymbol{y}} - \hat{\boldsymbol{B}}_S^T\overline{\boldsymbol{w}}$ and $\hat{\boldsymbol{B}}_S = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}\boldsymbol{y}}$. The least squares multivariate linear regression estimator can be calculated by computing the classical estimator $(\overline{\boldsymbol{u}}, \boldsymbol{S}_{\boldsymbol{u}}) = (\overline{\boldsymbol{u}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}})$ of multivariate location and dispersion on the $\boldsymbol{u}_i$, and then plug in the results into the formulas for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{B}}_S$.

Let $(T, \boldsymbol{C}) = (\tilde{\boldsymbol{\mu}}_{\boldsymbol{u}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}})$ be a robust estimator of multivariate location and dispersion. If $\tilde{\boldsymbol{\mu}}_{\boldsymbol{u}}$ is a consistent estimator of $\boldsymbol{\mu}_{\boldsymbol{u}}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}$ is a consistent estimator of $c\,\boldsymbol{\Sigma}_{\boldsymbol{u}}$ for some constant $c > 0$, then a robust estimator of multivariate linear regression is the plug in estimator $\tilde{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}} - \tilde{\boldsymbol{B}}_S^T\tilde{\boldsymbol{\mu}}_{\boldsymbol{w}}$ and $\tilde{\boldsymbol{B}}_S = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}\boldsymbol{y}}$.

If $(T, \boldsymbol{C})$ is the RMVN estimator applied to the $\boldsymbol{u}_i$, then $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}_{\boldsymbol{u}}, c\,\boldsymbol{\Sigma}_{\boldsymbol{u}})$ if $\boldsymbol{u}$ is from a large class of $EC_d(\boldsymbol{\mu}_{\boldsymbol{u}}, \boldsymbol{\Sigma}_{\boldsymbol{u}}, g)$ distributions where $d = m + p - 1$. Thus the classical and robust estimators of multivariate linear regression are both $\sqrt{n}$ consistent estimator of $\boldsymbol{B}$ if the $\boldsymbol{u}_i$ are iid from a large class of elliptically

contoured distributions. This assumption is quite strong, but the robust estimator is useful for detecting outliers. When there are categorical predictors or the joint distribution of $\boldsymbol{u}$ is not elliptically contoured, it is possible that the robust estimator is bad and very different from the good classical least squares estimator.

Now the RMVN estimator computes the classical estimator of multivariate location and dispersion on the RMVN set of cases in highly concentrated ellipsoidal region, and then multiplies the dispersion estimator by a constant c. Hence the plug in robust multivariate linear regression estimator using RMVN is equivalent to the least squares multivariate linear regression estimator applied to the cases in the RMVN set. Call this estimator the `rmreg2` estimator.

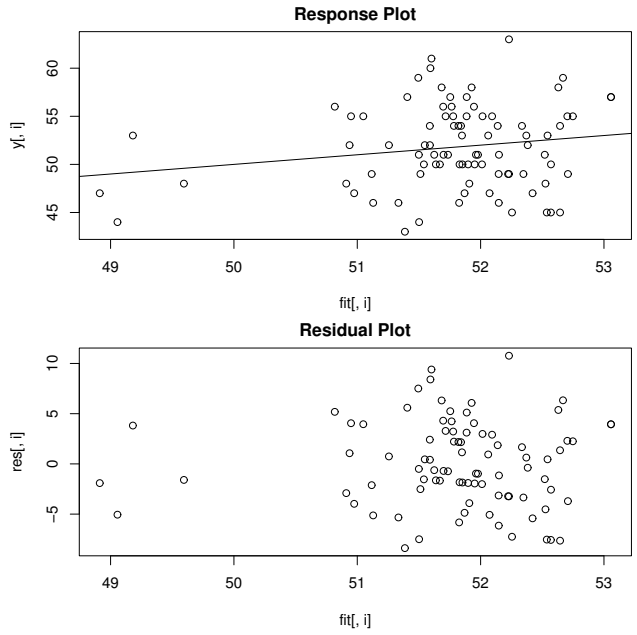

Figure 7: Plots for Y_1 = nasal height using `hbreg`.

Example 2, continued. The plots for the `rmreg2` estimator were very similar to Figures 5 and 6. Now let Y_1 = *nasal height* and Y_2 = *height* with x_2 = *head length*, x_3 = *bigonal breadth* and x_4 = *cephalic index*. Then Y_1 and x_2 have massive outliers. Then the response and residual plots for the classical estimator and the robust estimator using `hbreg` were nearly identical. Figures 7 and 8 show that the fit using `hbreg` went right through the outliers. Figures 9 and 10 show that the response and residual plots corresponding to `rmreg2` do not have fits that pass through the outliers.

5 Conclusions

Multivariate linear regression is a semiparametric method that is nearly as easy to use as multiple linear regression if m is small. The m response and residual plots should be made as well as the DD plot. For the classical estimator, response and residual plots can look good for $n \geq 10p$, but for testing and prediction regions, may need $n \geq k(m+p)^2$

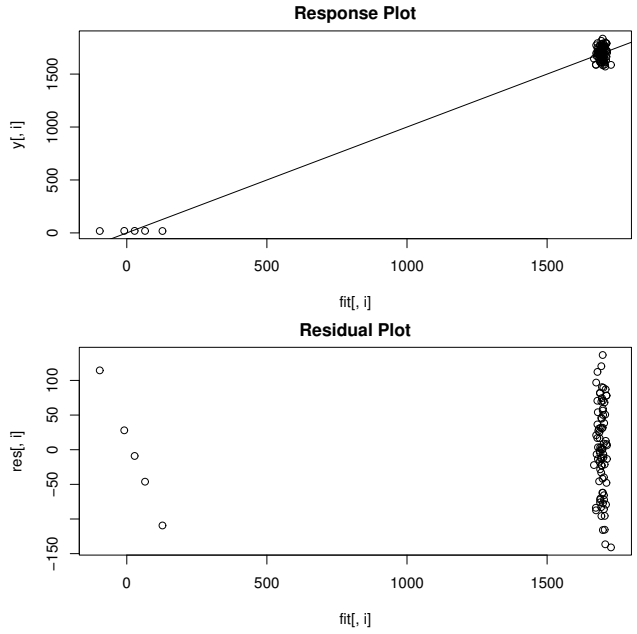

Figure 8: Plots for Y_2 = height using `hbreg`.

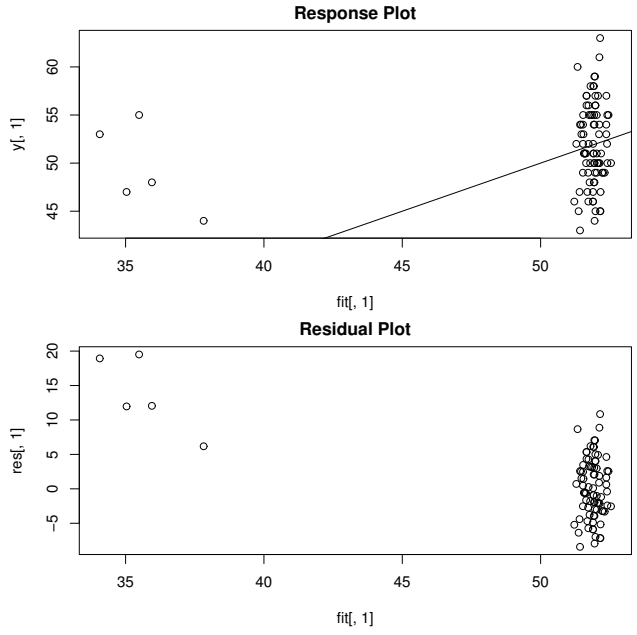

Figure 9: Plots for Y_1 = nasal height using `rmreg2`.

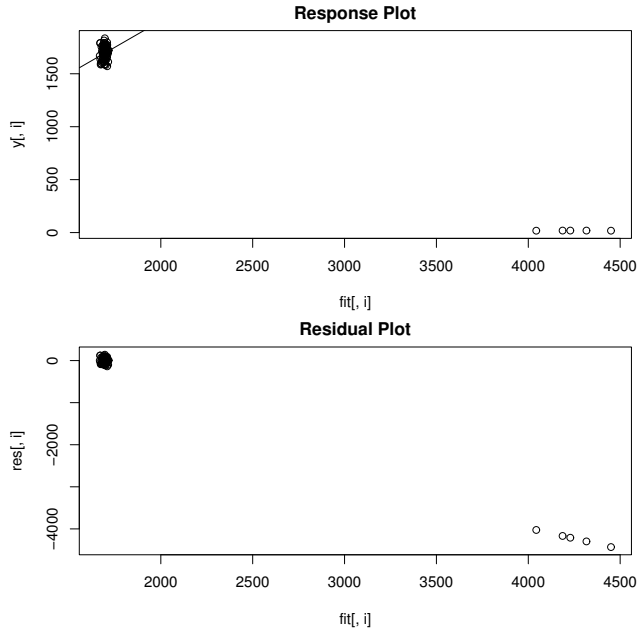

Figure 10: Plots for Y_2 = height using `rmreg2`.

where $0.5 \leq k \leq 3$ even for well behave elliptically contoured error distributions. For the robust estimator, larger sample sizes are needed, and for highly skewed data, the robust tests may fail. If the plotted points in the DD plot cluster tightly about a line through the origin, then an elliptically contoured error distribution may be reasonable, and then the first row of $\hat{\boldsymbol{B}}$ corresponding to the intercepts should be similar for both the robust and classical estimators.

The R software was used to make plots and software. See R Development Core Team (2011). The programs in the collection of functions *mpack.txt* are available at (www.math.siu.edu/olive/mpack.txt). The function `rmpredsim` was used to simulate the prediction regions, `rmregsim` was used to simulate the tests of hypotheses, and `rmregddsim` simulated the DD plots for various distributions. The function `rmltreg` makes the response and residual plots and computes the F_j, MANOVA F and MANOVA partial F test pvalues while the function `ddplot4` makes the DD plots. Similar functions for the classical estimator delete the initial "r." The highly outlier resistant multivariate linear regression estimator based on RMVN can be computed using the function `rmreg2`. When $m = 1$, this estimator is a highly outlier resistant multiple linear regression estimator.

The two methods for robust multivariate regression were to plug in a robust multiple linear regression estimator like `hbreg` in place of OLS, and to use a robust estimator of multivariate location and dispersion as a plug in estimator. Neither idea is new, but using practical highly outlier resistant estimators backed by theory, like `hbreg` and RMVN, is new. Rousseeuw, Van Aelst, Van Driessen and Agulló (2004) use FLTS and FMCD. Agulló, Croux and Van Aelst (2008) use the F-MLTS estimator. Kudraszow and Maronna (2011) use F-MM-type estimators. Also see Maronna and Morgenthaler (1986) for using a robust estimator of multivariate location and dispersion estimator to make a

robust regression estimator when $m = 1$.

References

Agulló, J., Croux, C. and Van Aelst, S. (2008), "The Multivariate Least-Trimmed Squares Estimator," *Journal of Multivariate Analysis*, 99, 311-338.

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis,* 2nd ed., Wiley, New York, NY.

Berndt, E.R. and Savin, N.E. (1977), "Conflict Among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model," *Econometrika*, 45, 1263-1277.

Buxton, L.H.D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

Cook, R.D., and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics,* Wiley, New York, NY.

Hampel, F.R. (1975), "Beyond Location Parameters: Robust Concepts and Methods," *Bulletin of the International Statistical Institute,* 46, 375-382.

Henderson, H.V., and Searle, S.R. (1977), "Vec and Vech Operators for Matrices, with Some Uses in Jacobians and Multivariate Statistics," *The Canadian Journal of Statistics,* 7, 65-81.

Hössjer, O. (1991), *Rank-Based Estimates in the Linear Model with High Breakdown Point*, Ph.D. Thesis, Report 1991:5, Department of Mathematics, Uppsala University, Uppsala, Sweden.

Johnson, M.E. (1987), *Multivariate Statistical Simulation,* Wiley, New York, NY.

Kakizawa, Y. (2009), "Third-order Power Comparisons for a Class of Tests for Multivariate Linear Hypothesis Under General Distributions," *Journal of Multivariate Analysis*, 100, 473-496.

Khattree, R., and Naik, D.N. (1999), *Applied Multivariate Statistics with SAS Software,* 2nd ed., SAS Institute, Cary, NC.

Kudraszow, N.L., and Maronna, R.A. (2011), "Estimates of MM Type for the Multivariate Linear Model," *Journal of Multivariate Analysis*, 102, 1280-1292.

Kshirsagar, A.M. (1972), *Multivariate Analysis*, Marcel Dekker, New York, NY.

Maronna, R.A., and Morgenthaler, S. (1986), "Robust Regression Through Robust Covariances," *Communications in Statistics: Theory and Methods*, 15, 1347-1365.

Olive, D.J. (2002), "Applications of Robust Distances for Regression," *Technometrics,* 44, 64-71.

Olive, D.J. (2005), "Two Simple Resistant Regression Estimators," *Computational Statistics and Data Analysis*, 49, 809-819.

Olive, D.J. (2013a), "Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data," *International Journal of Statistics and Probability*, 2, 90-100.

Olive, D.J. (2013b), "Plots, Prediction and Testing for the Multivariate Linear Model," preprint, see (www.math.siu.edu/olive/ppmultreg.pdf).

Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Olive, D.J., and Hawkins, D.M. (2010), "Robust Multivariate Location and Dispersion," Preprint, see (www.math.siu.edu/olive/preprints.htm).

Olive, D.J., and Hawkins, D.M. (2011), "Practical High Breakdown Regression," Preprint, see (www.math.siu.edu/olive/preprints.htm).

R Development Core Team (2011), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871-880.

Rousseeuw, P.J., Van Aelst, S., Van Driessen, K., and Agulló, J. (2004), "Robust Multivariate Regression," *Technometrics*, 46, 293-305.

Rousseeuw, P.J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis,* 2nd ed., Wiley, New York, NY.

Su, Z., and Cook, R.D. (2012), "Inner Envelopes: Efficient Estimation in Multivariate Linear Regression," *Biometrika*, 99, 687-702.

Zhang, J., Olive, D.J., and Ye, P. (2012), "Robust Covariance Matrix Estimation With Canonical Correlation Analysis," *International Journal of Statistics and Probability*, 1, 119-136.