Visualizing and Testing the Multivariate Linear Regression Model

Abstract

Recent results make the multivariate linear regression model much easier to use. This model has $m \geq 2$ response variables. Results by Kakizawa (2009) and Su and Cook (2012) can be used to explain the large sample theory of the least squares estimator and of the widely used Wilks' Λ, Pillai's trace, and Hotelling Lawley trace test statistics. Kakizawa (2009) shows that these statistics have the same limiting distribution. This paper reviews these results and gives two theorems to show that the Hotelling Lawley test generalizes the usual partial F test for $m = 1$ response variable to $m \geq 1$ response variables. Plots for visualizing the model are also given, and can be used to check goodness and lack of fit, to check for outliers and influential cases, and to check whether the error distribution is multivariate normal or from some other elliptically contoured distribution.

Keywords: DD plot, Hotelling Lawley trace test, Pillai's trace test, response plot, Wilks' Λ test

1. Introduction

The *multivariate linear regression model* is $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$. This paper suggests some plots and reviews some recent results by Kakizawa (2009) and Su and Cook (2012) that make this model easier to use.

The model 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 $x_{i1} = 1$. 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 $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma_\epsilon} = (\sigma_{ij})$ for $k = 1, ..., n$. Then the $p \times m$ coefficient matrix $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \ldots & \boldsymbol{\beta}_m \end{bmatrix}$ and the $m \times m$ covariance matrix $\boldsymbol{\Sigma_\epsilon}$ are to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$. Multiple linear regression corresponds to $m = 1$ response variable, and is written in matrix form as $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$. Subscripts are needed for the m multiple linear regression models $\boldsymbol{Y}_j = \boldsymbol{X\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where $E(\boldsymbol{e}_j) = \boldsymbol{0}$. For the multivariate linear regression model, $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\ \boldsymbol{I}_n$ for $i, j = 1, ..., m$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

The $n \times m$ matrix of response variables and $n \times m$ matrix of errors are

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \ldots & \boldsymbol{Y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix} \text{ and } \boldsymbol{E} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \ldots & \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix},$$

while the $n \times p$ design matrix of predictor variables is $\boldsymbol{X}$.

Least squares is the classical method for fitting the multivariate linear model. The *least squares estimators* are $\hat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 & \hat{\boldsymbol{\beta}}_2 & \ldots & \hat{\boldsymbol{\beta}}_m \end{bmatrix}$. The matrix of *predicted values* or *fitted values* $\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \begin{bmatrix} \hat{\boldsymbol{Y}}_1 & \hat{\boldsymbol{Y}}_2 & \ldots & \hat{\boldsymbol{Y}}_m \end{bmatrix}$. The matrix of *residuals* $\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} = \begin{bmatrix} \boldsymbol{r}_1 & \boldsymbol{r}_2 & \ldots & \boldsymbol{r}_m \end{bmatrix}$. These quantities can be found from the m multiple linear regressions of Y_j on the predictors: $\hat{\boldsymbol{\beta}}_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_j$, $\hat{\boldsymbol{Y}}_j = \boldsymbol{X}\hat{\boldsymbol{\beta}}_j$ and $\boldsymbol{r}_j = \boldsymbol{Y}_j - \hat{\boldsymbol{Y}}_j$ 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,

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

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

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

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.

Plots for checking the model are given in Section 2.1. Kakizawa (2009) examines testing for the multivariate linear regression model, showing that the Wilks, Pillai, and Hotelling Lawley test statistics perform well asymptotically for a large class of zero mean error distributions. Section 2.2 reviews these results and shows that the Hotelling Lawley test statistic is closely related to the partial F statistic for multiple linear regression. Section 3 gives an example and some simulations.

2. Method

2.1 Plots for the multivariate linear regression model

This subsection suggests using residual plots, response plots, and the DD plot to examine the multivariate linear regression model. These plots will be described below since the response plot and DD plots are not as well known as the residual plot. The residual plots are often used to check for lack of fit of the model. The response plots are used to check linearity and to detect influential cases and outliers. The response and residual plots are used exactly as in the $m = 1$ case corresponding to multiple linear regression. See Olive and Hawkins (2005) and Cook and Weisberg (1999a, p. 432; 1999b).

Some notation is needed to describe the DD plot. 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

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

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 notation MD will be used to denote the classical Mahalanobis distances $MD_i = D_i(\overline{\boldsymbol{x}}, \boldsymbol{S})$, and RD will denote distances $RD_i = D_i(T, \boldsymbol{C})$ using the robust RMVN estimator $(T, \boldsymbol{C})$ described in Zhang, Olive, and Ye (2012). The classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ is a consistent estimator of the population mean and covariance matrix $(\boldsymbol{\mu_x}, \boldsymbol{\Sigma_x})$, and the RMVN estimator $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu_x}, c\boldsymbol{\Sigma_x})$ for a large class of elliptically contoured distributions where $c > 0$ depends on the distribution and $c = 1$ for the multivariate normal distribution.

The Rousseeuw and Van Driessen (1999) DD plot is a plot of classical Mahalanobis distances MD versus robust Mahalanobis distances RD, and is used to check the error distribution and to detect outliers. The DD plot suggests that the error distribution is elliptically contoured if the plotted points cluster tightly about a line through the origin as $n \to \infty$. The plot suggests that the error distribution is multivariate normal if the line is the identity line RD=MD with unit slope and zero intercept. If n is large and the plotted points do not cluster tightly about a line through the origin, then the error distribution may not be elliptically contoured. Make a DD plot of the continuous predictor variables to check for $\boldsymbol{x}$-outliers. These applications of the DD plot for iid multivariate data are discussed in Olive (2002, 2013).

Make the m response and residual plots for the multivariate linear model. A *response plot* for the jth response variable is a plot of the fitted values $\widehat{Y_{ij}}$ versus the response Y_{ij} where $i = 1, ..., n$. The identity line 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}. In a response plot, the vertical deviations from the identity line are the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij}$. Suppose the model is good, the error distribution is not highly skewed, and $n \geq 10p$. Then the plotted points should cluster about the identity line in each of the m response plots and about the $r = 0$ line in the m residual plots. If outliers are present or if the plots are not linear, then the current model or data needs to be transformed or corrected. See Example 1.

2.2 Testing hypotheses

This subsection reviews useful results from Kakizawa (2009) and Su and Cook (2012). These results will show that the Hotelling Lawley test statistic is an extension of the partial F test statistic.

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 Wilks' 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 $U(\boldsymbol{L}) = tr[\boldsymbol{W}_e^{-1}\boldsymbol{H}] = \sum_{i=1}^{m} \lambda_i$.

Before proving Theorem 1 and showing that $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$ under mild conditions if H_0 is true, we first introduce some necessary notations. Following Henderson and Searle (1979), let matrix $\boldsymbol{A} = [\boldsymbol{a}_1\ \ \boldsymbol{a}_2\ \ \dots\ \ \boldsymbol{a}_p]$. Then the vec operator stacks the columns of $\boldsymbol{A}$ on top of one another, and $\boldsymbol{A} \otimes \boldsymbol{B}$ is the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$. An important fact is that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonsingular square matrices, then $[\boldsymbol{A} \otimes \boldsymbol{B}]^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$.

Theorem 1 *The Hotelling-Lawley trace statistic*

$$U(\boldsymbol{L}) = \frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]. \tag{3}$$

Proof. Using the Searle (1982, p. 333) identity $tr(\boldsymbol{A}\boldsymbol{G}^T\boldsymbol{D}\boldsymbol{G}\boldsymbol{C}) = [vec(\boldsymbol{G})]^T[\boldsymbol{C}\boldsymbol{A} \otimes \boldsymbol{D}^T][vec(\boldsymbol{G})]$, it follows that $(n-p)U(\boldsymbol{L}) = tr[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}] = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] = T$ where $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}, \boldsymbol{G} = \boldsymbol{L}\hat{\boldsymbol{B}}, \boldsymbol{D} = [\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}$, and $\boldsymbol{C} = \boldsymbol{I}$. Hence (3) holds. □

Kakizawa (2009) gives a result that can be shown to be equivalent to (3) using a commutation matrix $\boldsymbol{K}_{mn}$ where $\boldsymbol{K}_{mn}vec(\boldsymbol{A}) = vec(\boldsymbol{A}^T)$, $\boldsymbol{K}_{mn}^{-1} = \boldsymbol{K}_{nm}$, $\boldsymbol{K}_{pm}(\boldsymbol{A} \otimes \boldsymbol{B})\boldsymbol{K}_{nq} = \boldsymbol{B} \otimes \boldsymbol{A}$ and $vec(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = (\boldsymbol{C}^T \otimes \boldsymbol{A})vec(\boldsymbol{B})$. The above proof avoids commutation matrix algebra, and equation (3) will be used to show that the Hotelling Lawley test generalizes the usual partial F test for $m = 1$ response variable to $m \geq 1$ response variables.

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 error vectors have finite fourth moments, and assume that $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \xrightarrow{P} \boldsymbol{W}^{-1}$.

Su and Cook (2012) give a central limit type theorem for the multivariate linear regression model: if assumption D1 holds, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ *is a* $\sqrt{n}$ *consistent estimator of* $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, and $\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$.

Their theorem also shows that for multiple linear regression (m = 1), $\hat{\sigma}^2 = MSE$ is a $\sqrt{n}$ consistent estimator of σ^2. Note that it is not assumed that the error vectors have an elliptically contoured distribution.

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

Proof. By Su and Cook (2012), $\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$. Then under H_0, $\sqrt{n}\ \ vec(\boldsymbol{L}\hat{\boldsymbol{B}}) \xrightarrow{D} N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)$, and $n\ \ [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}$. This result also holds if $\boldsymbol{W}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are replaced by $\hat{\boldsymbol{W}} = n(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Hence under H_0 and using the proof of Theorem 1, $T = (n-p)U(\boldsymbol{L}) = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}$. □

Kakizawa (2009) shows, under stronger assumptions than Theorem 2 (such as eighth moments instead of fourth moments) 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 $-[n-p-0.5(m-r+3)]\log(\Lambda(\boldsymbol{L})) \xrightarrow{D} \chi^2_{rm}$. Also see Fujikoshi, Ulyanov, and Shimizu (2010, ch. 7). Results from Kshirsagar (1972, p. 301) suggest that the third chi-square approximation is very good if $n \geq 3(m+p)^2$ for multivariate normal errors.

Theorems 1 and 2 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 1.

By Theorem 2, 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

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

By Berndt and Savin (1977) and Anderson (1984, pp. 333, 371),

$$V(\boldsymbol{L}) \le -\log(\Lambda(\boldsymbol{L})) \le U(\boldsymbol{L}).$$

Hence the Hotelling Lawley test will have the most power and Pillai's test will have the least power.

Following Khattree and Naik (1999, pp. 67-68), there are several approximations used by the `SAS` software. For the Roy's largest root test, if $h = \max(r, m)$, use

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

The simulations in Section 3 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 $r > 1$. Let $g = n-p-(m-r+1)/2$, $u = (rm-2)/4$ and $t = \sqrt{r^2m^2-4}/\sqrt{m^2+r^2-5}$ for $m^2+r^2-5 > 0$ and $t = 1$, otherwise. Assume H_0 is true. Thus $U \xrightarrow{P} 0$, $V \xrightarrow{P} 0$, and $\Lambda \xrightarrow{P} 1$ as $n \to \infty$. Then

$$\frac{gt-2u}{rm} \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \approx F(rm, gt-2u) \text{ or } \mathrm{(n-p)t} \frac{1-\Lambda^{1/\mathrm{t}}}{\Lambda^{1/\mathrm{t}}} \approx \chi^2_{\mathrm{rm}}.$$

For large n and $t > 0$, $-\log(\Lambda) = -t\log(\Lambda^{1/t}) = -t\log(1+\Lambda^{1/t}-1) \approx t(1-\Lambda^{1/t}) \approx t(1-\Lambda^{1/t})/\Lambda^{1/t}$. If it can not be shown that $(n-p)[-\log(\Lambda) - t(1-\Lambda^{1/t})/\Lambda^{1/t}] \xrightarrow{P} 0$ as $n \to \infty$, then it is possible that the approximate χ^2_{rm} distribution may be the limiting distribution for only a small class of iid error distributions. When the $\boldsymbol{\epsilon}_i$ are iid $N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, there are some exact results. For $r = 1$,

$$\frac{n-p-m+1}{m} \frac{1-\Lambda}{\Lambda} \sim F(m, n-p-m+1).$$

For $r = 2$,

$$\frac{2(n-p-m+1)}{2m} \frac{1-\Lambda^{1/2}}{\Lambda^{1/2}} \sim F(2m, 2(n-p-m+1)).$$

For $m = 2$,

$$\frac{2(n-p)}{2r} \frac{1-\Lambda^{1/2}}{\Lambda^{1/2}} \sim F(2r, 2(n-p)).$$

Let $s = \min(r, m)$, $m_1 = (|r-m|-1)/2$ and $m_2 = (n-p-m-1)/2$. Note that $s(|r-m|+s) = \min(r, m)\max(r, m) = rm$. Then

$$\frac{n-p}{rm} \frac{V}{1-V/s} = \frac{n-p}{s(|r-m|+s)} \frac{V}{1-V/s} \approx \frac{2m_2+s+1}{2m_1+s+1} \frac{V}{s-V} \approx$$

$$F(s(2m_1+s+1), s(2m_2+s+1)) \approx F(s(|r-m|+s), s(n-p)) = F(rm, s(n-p)).$$

This approximation is asymptotically correct by Slutsky's theorem since $1 - V/s \xrightarrow{P} 1$. Finally,

$$\frac{n-p}{rm}U = \frac{n-p}{s(|r-m|+s)}U \approx \frac{2(sm_2+1)}{s^2(2m_1+s+1)}U \approx F(s(2m_1+s+1), 2(sm_2+1))$$

$$\approx F(s(|r-m|+s), s(n-p)) = F(rm, s(n-p)).$$

This approximation is asymptotically correct for a wide range of iid error distributions.

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 test for multiple linear regression. The statistic $F_j = \frac{1}{d_j}\hat{\boldsymbol{B}}_j^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \hat{\boldsymbol{B}}_j$ where $\hat{\boldsymbol{B}}_j^T$ is the jth row of $\hat{\boldsymbol{B}}$ and $d_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}$, the jth diagonal entry of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

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.

3. Results

Example 1 Cook and Weisberg (1999a, p. 351, 433, 447) gives 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. The response plots show strong linear relationships. For Y_1, case 79 sticks out while for Y_2, cases 8, 25 and 48 are not fit well. Highlighted cases had Cook's distance $> \min(0.5, 2p/n)$. See Cook (1977). Figure 3 shows the DD plot of the residual vectors. The plotted points are highly correlated but do not cover the identity line, suggesting an elliptically contoured error distribution that is not multivariate normal. The lines $MD = 2.60$, $RD = 2.80$ and $RD = 2.448$ in the DD plot correspond to the 95th percentiles of the MD_i, RD_i, and $\sqrt{\chi^2_{2,0.95}}$. Cases 8, 48 and 79 have especially large distances.

The response, residual, and DD plots are effective for finding influential cases, for checking linearity, and for checking whether the error distribution is multivariate normal or some other elliptically contoured distribution.

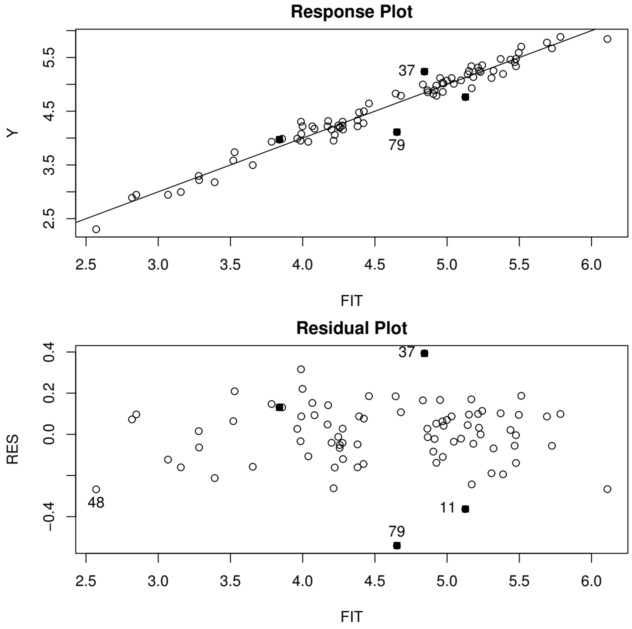

Figure 1: Plots for $Y_1 = \log(S)$.

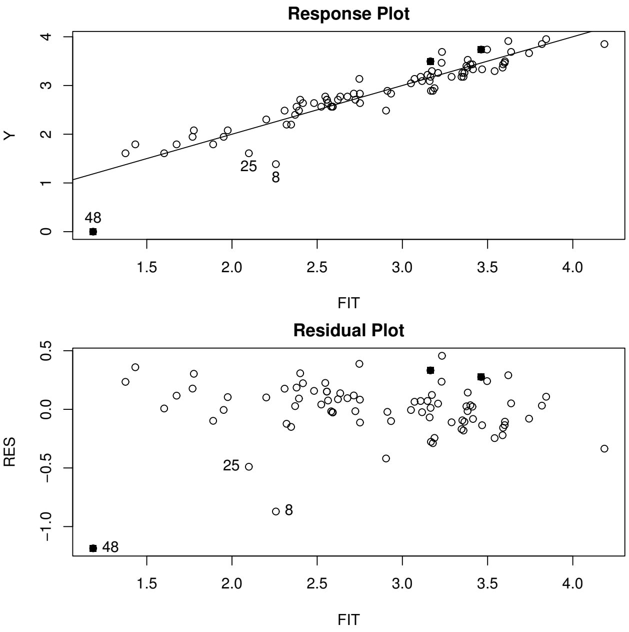

Figure 2: Plots for $Y_2 = \log(M)$.

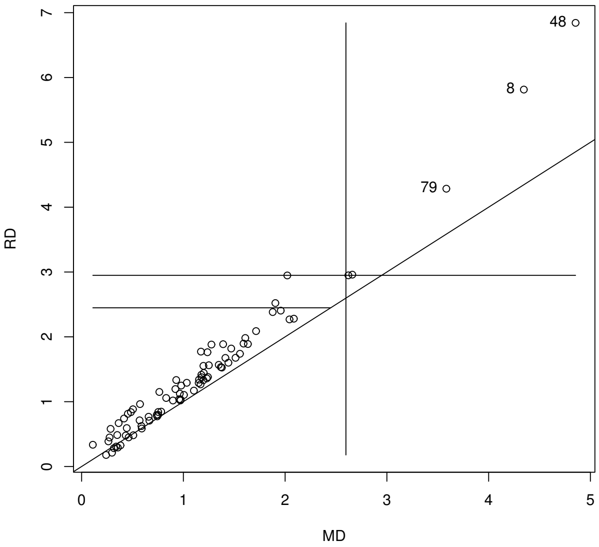

Figure 3: DD Plot of the Residual Vectors for the Mussels Data.

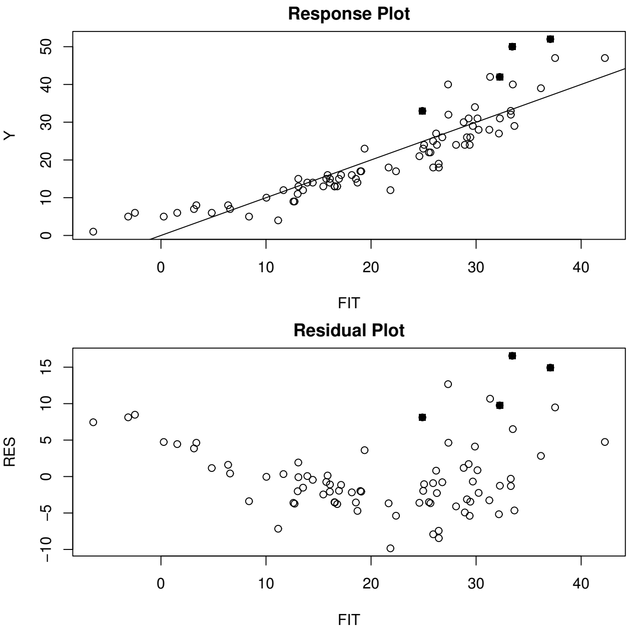

Figure 4: Response Plot for $Y_2 = M$.

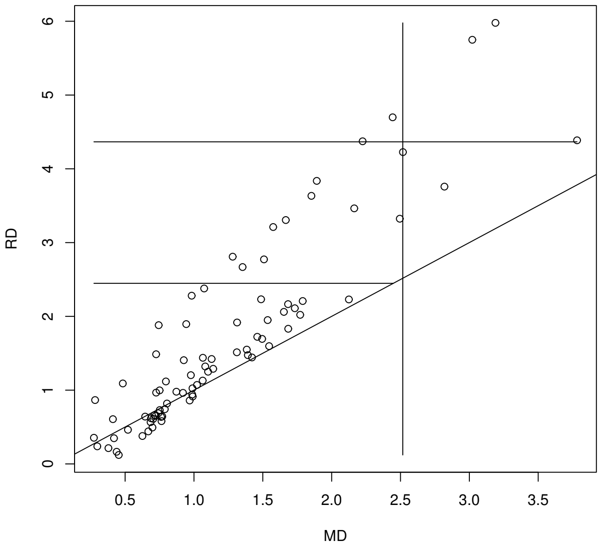

Figure 5: DD Plot When $Y_2 = M$.

Suppose the same model is used except $Y_2 = M$. Then the response and residual plots for Y_1 remain the same, but the plots shown in Figure 4 show curvature about the identity and $r = 0$ lines. Hence the linearity condition is violated. Figure 5 shows that the plotted points in the DD plot have correlation well less than one, suggesting that the error distribution is no longer elliptically contoured. Note that the plots can be used to quickly assess whether power transformations have resulted in a linear model.

The R function `mltreg` produces output for testing and makes the response and residual plots, and the function `ddplot4` makes the DD plot. The R commands for making the plots and output are shown below, assuming the data is stored in `mussels`. The output is very similar to the output for multiple linear regression. Bhat shows $\hat{\boldsymbol{B}}$, Ftable shows the F_j statistics and pvalues, while MANOVA shows the MANOVA F statistic and pvalue. The four Hotelling Lawley F_j statistics were greater than 5.77 with pvalues less than 0.005, and the MANOVA F statistic was 337.8 with pvalue ≈ 0.

```
y <- log(mussels)[,4:5]
x <- mussels[,1:3]
x[,2] <- log(x[,2])
out <- mltreg(x,y) #right click Stop 4 times
ddplot4(out$res) #right click Stop
y[,2] <- mussels[,5]
tem <- mltreg(x,y) #right click Stop 4 times
ddplot4(tem$res) #right click Stop
out
$Bhat       [,1]         [,2]
[1,] -2.322420435 -2.736457260
[2,]  0.004779329  0.002423747
[3,]  1.125434525  0.850428304
[4,]  0.013694060  0.016220043
$partial
     partialF         Pval
[1,] 19.12908 7.724099e-11
$Ftable
            Fj        pvals
[1,] 12.297891 2.448022e-05
[2,]  5.776937 4.663431e-03
[3,] 13.607901 9.276534e-06
[4,] 12.287095 2.467939e-05
$MANOVA
      MANOVAF pval
[1,] 337.7885    0
```

A small simulation was used to study the Wilks' Λ 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. The first row of $\boldsymbol{B}$ was always $\boldsymbol{1}^T$ and the last row of $\boldsymbol{B}$ was always $\boldsymbol{0}^T$. When the null hypothesis for the MANOVA F test is true, all but the first row corresponding to the constant are equal to $\boldsymbol{0}^T$. When $p \geq 3$ and the null hypothesis for the MANOVA F test 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) et cetera as long as the first row is not changed from $\boldsymbol{1}^T$. First $m \times 1$ error vectors $\boldsymbol{w}_i$ were 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 $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ 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 are $(2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$. As ψ gets close to 1, the error vectors cluster about the line in the direction of $(1, ..., 1)^T$. Used $\boldsymbol{w}_i \sim N_m(\boldsymbol{0}, \boldsymbol{I}), \boldsymbol{w}_i \sim (1-\tau)N_m(\boldsymbol{0}, \boldsymbol{I}) + \tau N_m(\boldsymbol{0}, 25\boldsymbol{I})$ 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

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

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	300	W	1	0.043	0.042	0.041	0.018
MVN	300	P	1	0.040	0.038	0.038	0.007
MVN	300	HL	1	0.059	0.058	0.057	0.045
MVN	300	R	1	0.051	0.049	0.048	0.993
MVN	600	W	1	0.048	0.043	0.043	0.034
MVN	600	P	1	0.046	0.042	0.041	0.026
MVN	600	HL	1	0.055	0.052	0.050	0.052
MVN	600	R	1	0.052	0.048	0.047	0.994
MIX	300	W	1	0.042	0.043	0.044	0.017
MIX	300	P	1	0.039	0.040	0.042	0.008
MIX	300	HL	1	0.057	0.059	0.058	0.039
MIX	300	R	1	0.050	0.050	0.051	0.993
MVT(7)	300	W	1	0.048	0.036	0.045	0.020
MVT(7)	300	P	1	0.046	0.032	0.042	0.011
MVT(7)	300	HL	1	0.064	0.049	0.058	0.045
MVT(7)	300	R	1	0.055	0.043	0.051	0.993
LN	300	W	1	0.043	0.047	0.040	0.020
LN	300	P	1	0.039	0.045	0.037	0.009
LN	300	HL	1	0.057	0.061	0.058	0.041
LN	300	R	1	0.049	0.055	0.050	0.994

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

n	$m = p$	test	F_1	F_2	F_{p-1}	F_p	F_M
30	5	W	0.012	0.222	0.058	0.000	0.006
30	5	P	0.000	0.000	0.000	0.000	0.000
30	5	HL	0.382	0.694	0.322	0.007	0.579
30	5	R	0.799	0.871	0.549	0.047	0.997
50	5	W	0.984	0.955	0.644	0.017	0.963
50	5	P	0.971	0.940	0.598	0.012	0.871
50	5	HL	0.997	0.979	0.756	0.053	0.991
50	5	R	0.996	0.978	0.744	0.049	1
105	10	W	0.650	0.970	0.191	0.000	0.633
105	10	P	0.109	0.812	0.050	0.000	0.000
105	10	HL	0.964	0.997	0.428	0.000	1
105	10	R	1	1	0.892	0.052	1
150	10	W	1	1	0.948	0.032	1
150	10	P	1	1	0.941	0.025	1
150	10	HL	1	1	0.966	0.060	1
150	10	R	1	1	0.965	0.057	1
450	20	W	1	1	0.999	0.020	1
450	20	P	1	1	0.999	0.016	1
450	20	HL	1	1	0.999	0.035	1
450	20	R	1	1	0.999	0.056	1

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.

When the null hypothesis H_0 was true, simulated values started to get close to nominal levels for $n \geq 0.8(m + p)^2$, 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 $r > 1$. See Table 1 where 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. Roy's test was very good for the F_j tests but very poor for the MANOVA F test. Results are shown for $m = p = 10$. As expected from Berndt and Savin (1977), Pillai's test rejected H_0 less often than Wilks' test which rejected H_0 less often than the Hotelling Lawley test.

In Table 2, H_0 is only true for the F_p test where $p = m$, and want values in the F_p column near 0.05. Want values near 1 for high power otherwise. If H_0 is false, often H_0 will be rejected for small n. For example, if $n \geq 10p$, then the m residual plots should start to look good, and the MANOVA F test should be rejected. For the simulated data, had fair power for n not much larger than mp. Results are shown for the lognormal distribution.

4. Discussion

Multivariate linear regression is nearly as easy to use as multiple linear regression if m is small. The plots speed up the model building process for multivariate linear models since the success of power transformations achieving linearity can be quickly assessed and influential cases can often be quickly detected. The plots can also be used for competing methods such as the envelopes estimators of Su and Cook (2012). Variable selection for multivariate linear regression is discussed in Fujikoshi, Ulyanov, and Shimizu (2010). Often observations $(x_2, ..., x_p, Y_1, ..., Y_m)$ are collected on the same person or thing and hence are correlated. If transformations can be found such that the m response plots and residual plots look good, and n is large enough, then multivariate linear regression can be used to efficiently analyze the data. Examining m multiple linear regressions is an incorrect method for analyzing the data. From simulations, response and residual plots start to be informative for $n \geq 10p$. Cramér (1946, pp. 414-415) shows that when the e_i are iid $N(0, \sigma^2)$ and none of the $p - 1$ nontrivial predictors are needed in the multiple linear regression model, then $E(R^2) = (p - 1)/(n - 1)$ where R^2 is the coefficient of multiple determination.

For testing the multivariate linear regression model, we recommend $n \geq \max((m + p)^2, mp + 30)$ provided that the m response and residual plots look good. When $m = 1$ the model degrees of freedom $= n - p$. It is not clear what the model degrees of freedom is for $m > 1$. We used $n - mp$ which is likely too small (conservative), but using $k(n - p)$ for small integer $k > 1$ is likely too large. Based on a much larger simulation study Pelawa Watagoda (2013, pp. 111-112), using the four types of error distributions and $m = p$, the tests had approximately correct level if $n > 0.83(m + p)^2$ for the Hotelling Lawley test, if $n > 2.80(m + p)^2$ for the Wilks' test (agreeing with Kshirsagar (1972): $n \geq 3(m + p)^2$ for multivariate normal data), and if $n > 4.2(m + p)^2$ for Pillai's test.

The tests are also large sample tests for a robust estimator that is asymptotically equivalent to least squares on a large class of elliptically contoured distributions. See Rupasinghe Arachchige Don (2013). For the robust estimator the elliptically contoured assumption is important since the robust estimator and least squares give different estimators of the constant when the assumption is violated.

The *R* software was used in the simulation. See R Development Core Team (2011). Programs are in the collection of *R* functions *lregpack.txt* available from (http://lagrange.math.siu.edu/Olive/lregpack.txt). The mussels data set can be obtained from (http://lagrange.math.siu.edu/Olive/lregdata.txt). The function `mregsim` was used to simulate the tests of hypotheses. The function `mltreg` makes the residual and response plots, and computes the F_j, MANOVA F, and MANOVA partial F test pvalues. The function `mregddsim` simulates DD plots of residuals for the multivariate linear regression model. The function `MLRsim` simulates response and residual plots for various error distributions. The function `ddplot4` makes the DD plot of the residuals. See Example 1.

Acknowledgements

The authors thank the Editors and referees for their comments that improved this article.

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York, NY: Wiley.

Berndt, E. R., & Savin, N. E. (1977). Conflict among criteria for testing hypotheses in the multivariate linear regression model. *Econometrika*, 45, 1263-1277. http://dx.doi.org/10.2307/1914072

Cook, R. D. (1977). Deletion of influential observations in linear regression. *Technometrics,* 19, 15-18. http://dx.doi.org/10.2307/2286747

Cook, R. D., & Weisberg, S. (1999a). *Applied regression including computing and graphics.* New York, NY: Wiley.

Cook, R. D., & Weisberg, S. (1999b). Graphs in statistical analysis: Is the medium the message? *The American Statistician,* 53, 29-37. http://dx.doi.org/10.2307/2685649

Cramér, H. (1946). *Mathematical Methods of Statistics,* Princeton, NJ: Princeton University Press.

Fujikoshi, Y., Ulyanov, V. V., & Shimizu, R. (2010). *Multivariate statistics: high-dimensional and large-sample approximations*. Hoboken, NJ: Wiley.

Henderson, H. V., & 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. http://dx.doi.org/10.2307/3315017

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. http://dx.doi.org/10.1016/j.jmva.2008.06.002

Khattree, R., & Naik, D. N. (1999). *Applied multivariate statistics with SAS software* (2nd ed.). Cary, NC: SAS Institute.

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

Olive, D. J. (2002). Applications of robust distances for regression. *Technometrics,* 44, 64-71. http://dx.doi.org/10.1198/004017002753398335

Olive, D. J. (2013). Asymptotically optimal regression prediction intervals and prediction regions for multivariate data. *International Journal of Statistics and Probability*, 2, 90-100. http://dx.doi.org/10.5539/ijsp.v2n1p90

Olive, D. J., & Hawkins, D. M. (2005). Variable selection for 1D regression models. *Technometrics*, 47, 43-50. http://dx.doi.org/10.1198/004017004000000590

Pelawa Watagoda, L. C. R. (2013). Plots and testing for multivariate linear regression. Master's Paper, Southern Illinois University. http://lagrange.math.siu.edu/Olive/slasanthi.pdf

R Development Core Team (2011). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org.

Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics,* 41, 212-223. http://dx.doi.org/10.1080/00401706.1999.10485670

Rupasinghe Arachchige Don, H. S. (2013). Robust multivariate linear regression. Master's Paper, Southern Illinois University. http://lagrange.math.siu.edu/Olive/shasthika.pdf

Searle, S. R. (1982). *Matrix algebra useful for statistics*. New York, NY: Wiley.

Seber, G. A. F., & Lee, A. J. (2003). *Linear regression analysis* (2nd ed.). New York, NY: Wiley.

Su, Z., & Cook, R. D. (2012). Inner envelopes: Efficient estimation in multivariate linear regression. *Biometrika*, 99, 687-702. http://dx.doi.org/10.1093/biomet/ass024

Zhang, J., Olive, D. J., & Ye, P. (2012). Robust covariance matrix estimation with canonical correlation analysis. *International Journal of Statistics and Probability, 1*, 119-136. http://dx.doi.org/10.5539/ijsp.v1n2p119