Chapter 12

Multivariate Linear Regression

12.1 Introduction

Definition 12.1. The **response variables** are the variables that you want to predict. The **predictor variables** are the variables used to predict the response variables.

Notation. 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 $X_1, X_2, ..., X_p$ where $X_1 = 1$ is the trivial predictor. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (1, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$ where the 1 could be omitted.

In matrix form, the model is $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$, and the data matrix $\boldsymbol{W} = [\boldsymbol{X} \;\; \boldsymbol{Y}]$ except usually the first column $\boldsymbol{1}$ of $\boldsymbol{X}$ is omitted. The $n \times m$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,m} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n,1} & Y_{n,2} & \dots & Y_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \dots & \boldsymbol{Y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix}.$$

The $n \times p$ matrix

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

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

The $p \times m$ matrix

$$\boldsymbol{B} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \dots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \dots & \beta_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p,1} & \beta_{p,2} & \dots & \beta_{p,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \dots & \boldsymbol{\beta}_m \end{bmatrix}.$$

The $n \times m$ matrix

$$\boldsymbol{E} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \dots & \epsilon_{1,m} \\ \epsilon_{2,1} & \epsilon_{2,2} & \dots & \epsilon_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \dots & \epsilon_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \dots & \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix}.$$

Warning: The $\boldsymbol{e}_i$ are error vectors, not orthonormal eigenvectors.

Definition 12.2. In the *multiple linear regression model*,

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{12.1}$$

for $i = 1, \dots, n$. In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{12.2}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Equivalently,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{12.3}$$

The e_i are iid with zero mean and variance σ^2, and multiple linear regression is used to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2.

Each response variable in a multivariate linear regression model follows a multiple linear regression model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is

assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$. Hence the errors corresponding to the jth response are uncorrelated with variance $\sigma_j^2 = \sigma_{jj}$. Notice that the **same design matrix** $\boldsymbol{X}$ of predictors is used for each of the m models, but the jth response variable vector $\boldsymbol{Y}_j$, coefficient vector $\boldsymbol{\beta}_j$ and error vector $\boldsymbol{e}_j$ change and thus depend on j.

Now consider the ith case $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T)$ which corresponds to the ith row of $\boldsymbol{Z}$ and the ith row of $\boldsymbol{X}$. Then

$$\begin{bmatrix} Y_{i1} = \beta_{11}x_{i1} + \cdots + \beta_{p1}x_{ip} + \epsilon_{i1} = \boldsymbol{x}_i^T\boldsymbol{\beta}_1 + \epsilon_{i1} \\ Y_{i2} = \beta_{12}x_{i1} + \cdots + \beta_{p2}x_{ip} + \epsilon_{i2} = \boldsymbol{x}_i^T\boldsymbol{\beta}_2 + \epsilon_{i2} \\ \vdots \\ Y_{im} = \beta_{1m}x_{i1} + \cdots + \beta_{pm}x_{ip} + \epsilon_{im} = \boldsymbol{x}_i^T\boldsymbol{\beta}_m + \epsilon_{im} \end{bmatrix}$$

or $\boldsymbol{y}_i = \boldsymbol{\mu}_{\boldsymbol{x}_i} + \boldsymbol{\epsilon}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i$ where

$$E(\boldsymbol{y}_i) = \boldsymbol{\mu}_{\boldsymbol{x}_i} = \boldsymbol{B}^T\boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{x}_i^T\boldsymbol{\beta}_1 \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_m \end{bmatrix}.$$

The notation $\boldsymbol{y}_i|\boldsymbol{x}_i$ and $E(\boldsymbol{y}_i|\boldsymbol{x}_i)$ is more accurate, but usually the conditioning is suppressed. Taking $\boldsymbol{\mu}_{\boldsymbol{x}_i}$ to be a constant (or condition on $\boldsymbol{x}_i$ if the predictor variables are random variables), $\boldsymbol{y}_i$ and $\boldsymbol{\epsilon}_i$ have the same covariance matrix. In the multivariate regression model, this covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ does not depend on i. Observations from different cases are uncorrelated (often independent), but the m errors for the m different response variables for the *same case* are correlated. If $\boldsymbol{X}$ is a random matrix, then assume $\boldsymbol{X}$ and $\boldsymbol{E}$ are independent and that expectations are conditional on $\boldsymbol{X}$.

Definition 12.3. The **multivariate linear regression model** $\boldsymbol{y}_k = \boldsymbol{B}^T\boldsymbol{x}_k + \boldsymbol{\epsilon}_k$ for $k = 1, ..., n$ is written in matrix form as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = ((\sigma_{ij}))$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$. Then $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{X}\boldsymbol{B}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$. Considering the kth row of $\boldsymbol{Z}, \boldsymbol{X}$ and $\boldsymbol{E}$ shows that $\boldsymbol{y}_k^T = \boldsymbol{x}_k^T\boldsymbol{B} + \boldsymbol{\epsilon}_k^T$.

Example 12.1. Suppose it is desired to predict the response variables $Y_1 =$ *height* and $Y_2 =$ *height at shoulder* of a person from partial skeletal

remains. A model for prediction can be built from nearly complete skeletons or from living humans, depending on the population of interest (eg ancient Egyptians or modern US citizens). The predictor variables might be $x_1 \equiv 1$, $x_2 = \textit{femur length}$ and $x_3 = \textit{ulna length}$. The two heights of individuals with $x_2 = 200mm$ and $x_3 = 140mm$ should be shorter on average than the two heights of individuals with $x_2 = 500mm$ and $x_3 = 350mm$. In this example Y_1, Y_2, x_2 and x_3 are quantitative variables. If $x_4 = \textit{gender}$ is a predictor variable, then gender (coded as male = 1 and female = 0) is qualitative.

Definition 12.4. Least squares is the classical method for fitting multivariate linear regression. 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*

$$\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \left[\begin{array}{cccc} \hat{\boldsymbol{Y}}_1 & \hat{\boldsymbol{Y}}_2 & \ldots & \hat{\boldsymbol{Y}}_m \end{array} \right] = \left[\begin{array}{cccc} \hat{Y}_{1,1} & \hat{Y}_{1,2} & \ldots & \hat{Y}_{1,m} \\ \hat{Y}_{2,1} & \hat{Y}_{2,2} & \ldots & \hat{Y}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{Y}_{n,1} & \hat{Y}_{n,2} & \ldots & \hat{Y}_{n,m} \end{array} \right].$$

The *residuals* $\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} =$

$$\left[\begin{array}{c} \hat{\boldsymbol{\epsilon}}_1^T \\ \hat{\boldsymbol{\epsilon}}_2^T \\ \vdots \\ \hat{\boldsymbol{\epsilon}}_n^T \end{array} \right] = \left[\begin{array}{cccc} \hat{\boldsymbol{r}}_1 & \hat{\boldsymbol{r}}_2 & \ldots & \hat{\boldsymbol{r}}_m \end{array} \right] = \left[\begin{array}{cccc} \hat{\epsilon}_{1,1} & \hat{\epsilon}_{1,2} & \ldots & \hat{\epsilon}_{1,m} \\ \hat{\epsilon}_{2,1} & \hat{\epsilon}_{2,2} & \ldots & \hat{\epsilon}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\epsilon}_{n,1} & \hat{\epsilon}_{n,2} & \ldots & \hat{\epsilon}_{n,m} \end{array} \right].$$

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 $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Also,

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} = (n-d)^{-1}\boldsymbol{Z}^T[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}]\boldsymbol{Z},$$

and

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

Theorem 12.1, (Johnson and Wichern (1988, p. 304): Suppose $\boldsymbol{X}$ has full rank $p < n$ and the covariance structure of Definition 12.3 holds. Then $E(\hat{\boldsymbol{B}}) = \boldsymbol{B}$ so $E(\hat{\boldsymbol{\beta}}_j) = \boldsymbol{\beta}_j$, $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_k) = \sigma_{jk}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ for $j, k = 1, ..., p$. Also $\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{B}}$ are uncorrelated, $E(\hat{\boldsymbol{E}}) = \boldsymbol{0}$ and

$$E(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = E\left(\frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-p}\right) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}.$$

Theorem 12.2. $\boldsymbol{S}_r = \boldsymbol{\Sigma}_\epsilon + O_P(n^{-1/2})$ if $\boldsymbol{B} - \hat{\boldsymbol{B}} = O_P(n^{-1/2})$, $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{x}_i^T = O_P(1)$, $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T = O_P(n^{1/2})$ and $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + O_P(n^{-1/2})$.

Proof. Note that $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i = \hat{\boldsymbol{B}}^T\boldsymbol{x}_i + \hat{\boldsymbol{\epsilon}}_i$. Hence $\hat{\boldsymbol{\epsilon}}_i = (\boldsymbol{B} - \hat{\boldsymbol{B}})^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$. Thus

$$\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T = \sum_{i=1}^n (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i)(\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i)^T = \sum_{i=1}^n [\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T + \boldsymbol{\epsilon}_i(\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i)^T + (\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i)\hat{\boldsymbol{\epsilon}}_i^T] =$$

$$\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T + (\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{x}_i^T)(\boldsymbol{B} - \hat{\boldsymbol{B}}) + (\boldsymbol{B} - \hat{\boldsymbol{B}})^T(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{\epsilon}_i^T) + (\boldsymbol{B} - \hat{\boldsymbol{B}})^T(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T)(\boldsymbol{B} - \hat{\boldsymbol{B}}).$$

Thus $\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T +$

$$O_P(1)O_P(n^{-1/2}) + O_P(n^{-1/2})O_P(1) + O_P(n^{-1/2})O_P(n^{1/2})O_P(n^{-1/2}),$$

and the result follows since $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + O_P(n^{-1/2})$ and

$$\boldsymbol{S}_r = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T.$$

12.2 Checking the Model

12.2.1 Plots

Notation. Plots will be used to simplify regression analysis, and in this text a plot of W versus Z uses W on the horizontal axis and Z on the vertical axis.

Definition 12.5. A **response plot** for the jth response variable is a plot of the fitted values $\widehat{Y}_{ij}$ versus the response Y_{ij}. 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}.

Remark 12.1. Make the m response and residual plots for any multivariate linear regression. In a response plot, the vertical deviations from the identity line are the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij}$. If the model is appropriate, then the plotted points 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. If the model is good, then the each of the m residual plots should be ellipsoidal with no trend and should be centered about the $r = 0$ line. There should not be any pattern in the residual plot: as a narrow vertical strip is moved from left to right, the behavior of the residuals within the strip should show little change. Outliers and patterns such as curvature or a fan shaped plot are bad.

Notation. A *rule of thumb* is a rule that often but not always works well in practice.

Rule of thumb 12.1. Use multivariate linear regression if $n > 10 \max(p, m)$. The m response and residual plots should all look good. Make the DD plot of the $\hat{\boldsymbol{\epsilon}}_i$. If a residual plot would look good after several points have been deleted, and if these deleted points were not gross outliers (points far from the point cloud formed by the bulk of the data), then the residual plot is probably good. Beginners often find too many things wrong with a good model. For practice, use the computer to generate several multivariate linear regression data sets, and make the m response and residual plots for these data sets. This exercise will help show that the plots can have considerable variability even when the multivariate linear regression model is good.

Rule of thumb 12.2. If the plotted points in the residual plot look like a left or right opening megaphone, the first model violation to check is the assumption of nonconstant variance. (This is a rule of thumb because it is possible that such a residual plot results from another model violation such as nonlinearity, but nonconstant variance is much more common.)

Remark 12.2. Residual plots *magnify departures* from the model while the response plots emphasizes *how well the multivariate linear regression model fits the data.*

Definition 12.6. An **RR plot** is a scatterplot matrix of the m sets of residuals $\boldsymbol{r}_1, ..., \boldsymbol{r}_m$.

Definition 12.7. An **FF plot** is a scatterplot matrix of the m sets of fitted values of response variables $\hat{\boldsymbol{Y}}_1, ..., \hat{\boldsymbol{Y}}_m$. The m response variables $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_m$ can be added to the plot.

Remark 12.3. Multivariate linear regression makes the most sense if the m errors are linearly related, eg from an elliptically contoured distribution. Make the RR plot and a DD plot of the residuals $\hat{\boldsymbol{\epsilon}}_i$ to check that the errors are linearly related. Make a DD plot of the continuous predictor variables to check for x-outliers. Make a DD plot of $Y_1, ..., Y_m$ to check for outliers, especially if it is assumed that the response variables come from an elliptically contoured distribution.

Example 12.2. Tremearne (1911) presents a data set of about 17 measurements on 115 people of Hausa nationality. We deleted 3 cases (107, 108 and 109) because of missing values and used *height* as the response variable Y_1. Suppose Y_2 is the other response variable and that the response and residual plots for Y_2 are well behaved. Along with a constant $x_{i,1} \equiv 1$, the five additional predictor variables used were *height when sitting*, *height when kneeling*, *head length, nasal breadth*, and *span* (perhaps from left hand to right hand). Figure 12.1 presents the response and residual plots corresponding the response variable $Y_1 =$ *height* for this data set. These plots show that the model should be useful for the data since the plotted points in the response plot are linear and follow the identity line while the plotted points in the residual plot follow the $r = 0$ line with no other pattern (except for a possible outlier marked 44).

To use the response plot to visualize the conditional distribution of $Y_1|\boldsymbol{x}^T\boldsymbol{\beta}_1$, use the fact that the fitted values $\hat{Y}_1 = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}_1$. For example, suppose the height given fit = 1700 is of interest. Mentally examine the plot about a narrow vertical strip about fit = 1700, perhaps from 1675 to 1725. The cases in the narrow strip have a mean close to 1700 since they fall close to the identity line. Similarly, when the fit = w for w between 1500 and 1850, the cases have heights near w, on average.

Cases 3, 44 and 63 are highlighted. The 3rd person was very tall while the 44th person was rather short. Beginners often label too many points as outliers. Mentally draw a box about the bulk of the data ignoring any outliers. Double the width of the box (about the identity line for the response

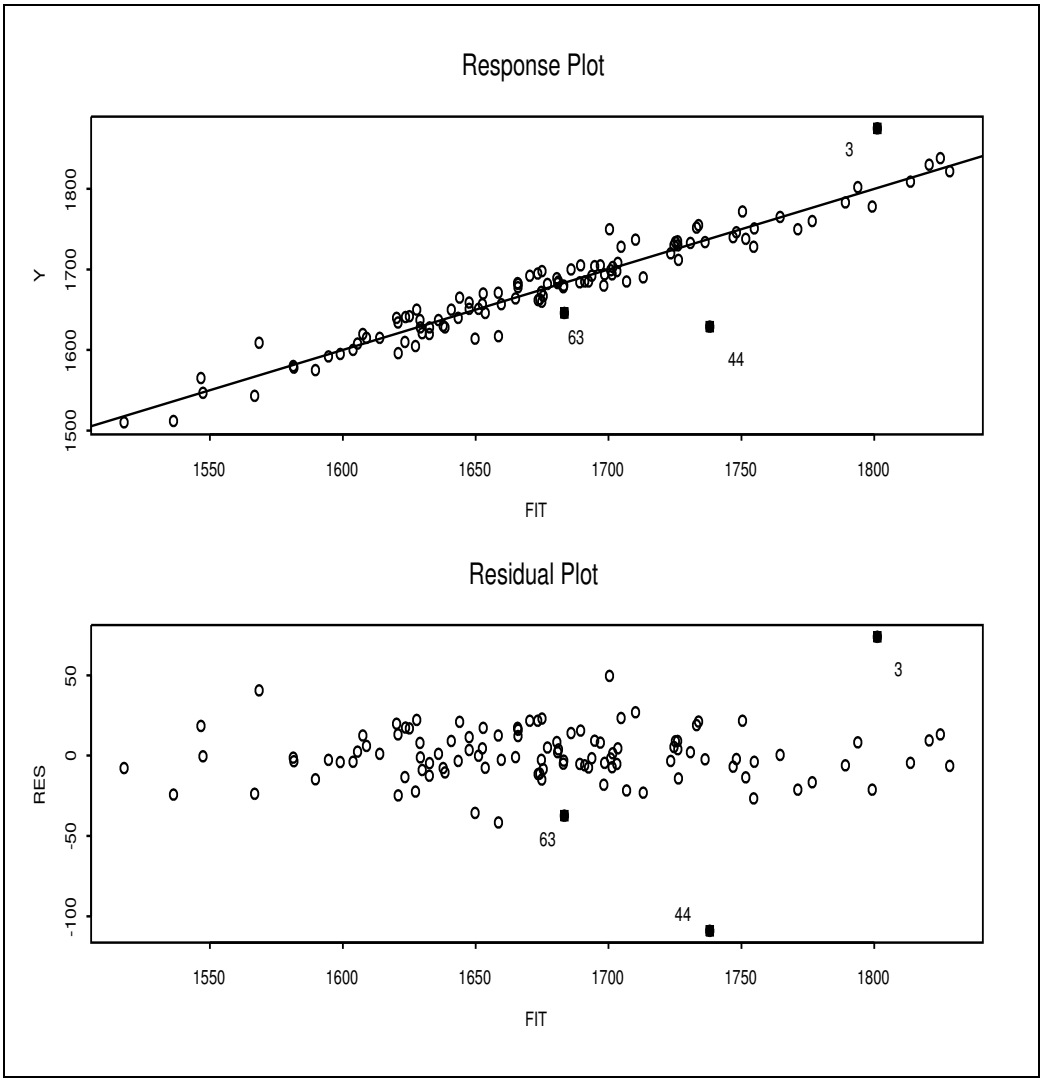

Figure 12.1: Residual and Response Plots for the Response Variable Height

plot and about the horizontal $r = 0$ line for the residual plot). Cases outside of this imaginary doubled box are potential outliers. Alternatively, visually estimate the standard deviation of the residuals in both plots. In the residual plot look for residuals that are more than 5 standard deviations from the $r = 0$ line. In Figure 12.1, the standard deviation of the residuals appears to be around 10. Hence cases 3 and 44 are certainly worth examining.

The plots corresponding to Y_1 can be made with the following commands. In general store $Y1, Y2, ..., Ym$ and make the *MLRplot(X, Y)* command m times for $Y = Y1, ..., Ym$.

```
source("G:/mpack.txt")
#assume the data is stored in R matrix major
X<-major[,-6]; Y1 <- major[,6]; MLRplot(X,Y1)
```

12.2.2 Predictor and Response Transformations

Predictor transformations for the continuous predictors can be made exactly as in Section 2.4.

Warning: The Rule of thumb 2.1 does not always work. For example, the log rule may fail. If the relationships in the scatterplot matrix are already linear or if taking the transformation does not increase the linearity, then no transformation may be better than taking a transformation. For the *Arc* data set `evaporat.lsp`, the log rule suggests transforming the response variable *Evap*, but no transformation works better.

Response transformations can also be made as in Section 2.4, but there is an alternative graphical method for response transformations once the predictors are fixed. Discussion will first be given for multiple linear regression with response variable Y. Then for multivariate regression, simply use the transformation plots for each of the m response variables $Y_1, ..., Y_m$.

An important class of *response transformation models* adds an additional unknown transformation parameter λ_o, such that

$$Y_i = t_{\lambda_o}(Z_i) \equiv Z_i^{(\lambda_o)} = E(Y_i|\boldsymbol{x}_i) + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i. \tag{12.4}$$

If λ_o was known, then $Y_i = t_{\lambda_o}(Z_i)$ would follow a multiple linear regression model with p predictors including the constant. Here, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients depending on λ_o, $\boldsymbol{x}$ is a $p \times 1$ vector of predictors

that are assumed to be measured with negligible error, and the errors e_i are assumed to be iid with zero mean.

Definition 12.8. Assume that **all** of the values of the "response" Z_i are **positive**. A *power transformation* has the form $Y = t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$ where

$$\lambda \in \Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\}.$$

Definition 12.9. Assume that **all** of the values of the "response" Z_i are **positive**. Then the *modified power transformation family*

$$t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda} \tag{12.5}$$

for $\lambda \neq 0$ and $Z_i^{(0)} = \log(Z_i)$. Often $Z_i^{(1)}$ is replaced by Z_i for $\lambda = 1$. Generally $\lambda \in \Lambda$ where Λ is some interval such as $[-1, 1]$ or a coarse subset such as Λ_L. This family is a special case of the response transformations considered by Tukey (1957).

A graphical method for response transformations refits the model using the same fitting method: changing only the "response" from Z to $t_\lambda(Z)$. Compute the "fitted values" $\hat{W}_i$ using $W_i = t_\lambda(Z_i)$ as the "response." Then a *transformation plot* of $\hat{W}_i$ versus W_i is made for each of the seven values of $\lambda \in \Lambda_L$ with the identity line added as a visual aid. Vertical deviations from the identity line are the "residuals" $r_i = W_i - \hat{W}_i$. Then a candidate response transformation $Y = t_{\lambda^*}(Z)$ is reasonable if the plotted points follow the identity line in a roughly evenly populated band. Curvature from the identity line suggests that the candidate response transformation is inappropriate.

Definition 12.10. A *transformation plot* is a plot of $\hat{W}$ versus W with the identity line added as a visual aid.

There are several reasons to use a coarse grid of powers. First, several of the powers correspond to simple transformations such as the log, square root, and cube root. These powers are easier to interpret than $\lambda = .28$, for example. According to Mosteller and Tukey (1977, p. 91), the **most commonly used power transformations** are the $\lambda = 0$ (log), $\lambda = 1/2$, $\lambda = -1$ and $\lambda = 1/3$ transformations in decreasing frequency of use. Secondly, if the estimator $\hat{\lambda}_n$ can only take values in Λ_L, then sometimes $\hat{\lambda}_n$ will

converge (eg in probability) to $\lambda^* \in \Lambda_L$. Thirdly, Tukey (1957) showed that neighboring power transformations are often very similar, so restricting the possible powers to a coarse grid is reasonable. Note that powers can always be added to the grid Λ_L. Useful powers are $\pm 1/4, \pm 2/3, \pm 2$, and ± 3. Powers from numerical methods can also be added.

Application 12.1. This graphical method for selecting a response transformation is very simple. Let $W_i = t_\lambda(Z_i)$. Then for each of the seven values of $\lambda \in \Lambda_L$, perform least squares (OLS) on $(W_i, \boldsymbol{x}_i)$ and make the transformation plot of $\hat{W}_i$ versus W_i. If the plotted points follow the identity line for λ^*, then take $\hat{\lambda}_o = \lambda^*$, that is, $Y = t_{\lambda^*}(Z)$ is the response transformation. (Note that this procedure can be modified to create a graphical diagnostic for a numerical estimator $\hat{\lambda}$ of λ_o by adding $\hat{\lambda}$ to Λ_L.) Note that for multivariate regression, use $W = Y_j$ for $j = 1, ..., m$. Hence $7m$ plots will be made.

If more than one value of $\lambda \in \Lambda_L$ gives a linear plot, take the simplest or most reasonable transformation or the transformation that makes the most sense to subject matter experts. Also check that the corresponding "residual plots" of $\hat{W}$ versus $W - \hat{W}$ look reasonable. The values of λ in decreasing order of importance are $1, 0, 1/2, -1$ and $1/3$. So the log transformation would be chosen over the cube root transformation if both transformation plots look equally good.

After selecting the transformations, the usual checks on the multivariate regression model should be made. In particular, make the m response and residual plots. In particular, the transformation plot for the selected transformation is the response plot, and a residual plot should also be made.

The following two examples illustrates the procedure for a single response variable $Y = Y_1$, and the plots show $t_\lambda(Z)$ on the vertical axis. The label "TZHAT" of the horizontal axis are the "fitted values" that result from using $t_\lambda(Z)$ as the "response" in the OLS software. In general for multivariate regression, the plots would be made for $Z_1, ..., Z_m$ resulting in response variables $Y_1 = t_1(Z_1), ..., Y_m = t_m(Z_m)$.

Example 12.3: Textile Data. In their pioneering paper on response transformations, Box and Cox (1964) analyze data from a 3^3 experiment on the behavior of worsted yarn under cycles of repeated loadings. The "response" Z is the *number of cycles to failure* and a constant is used along with the three predictors *length, amplitude* and *load.* Using the normal profile log likelihood for λ_o, Box and Cox determine $\hat{\lambda}_o = -0.06$ with approximate 95

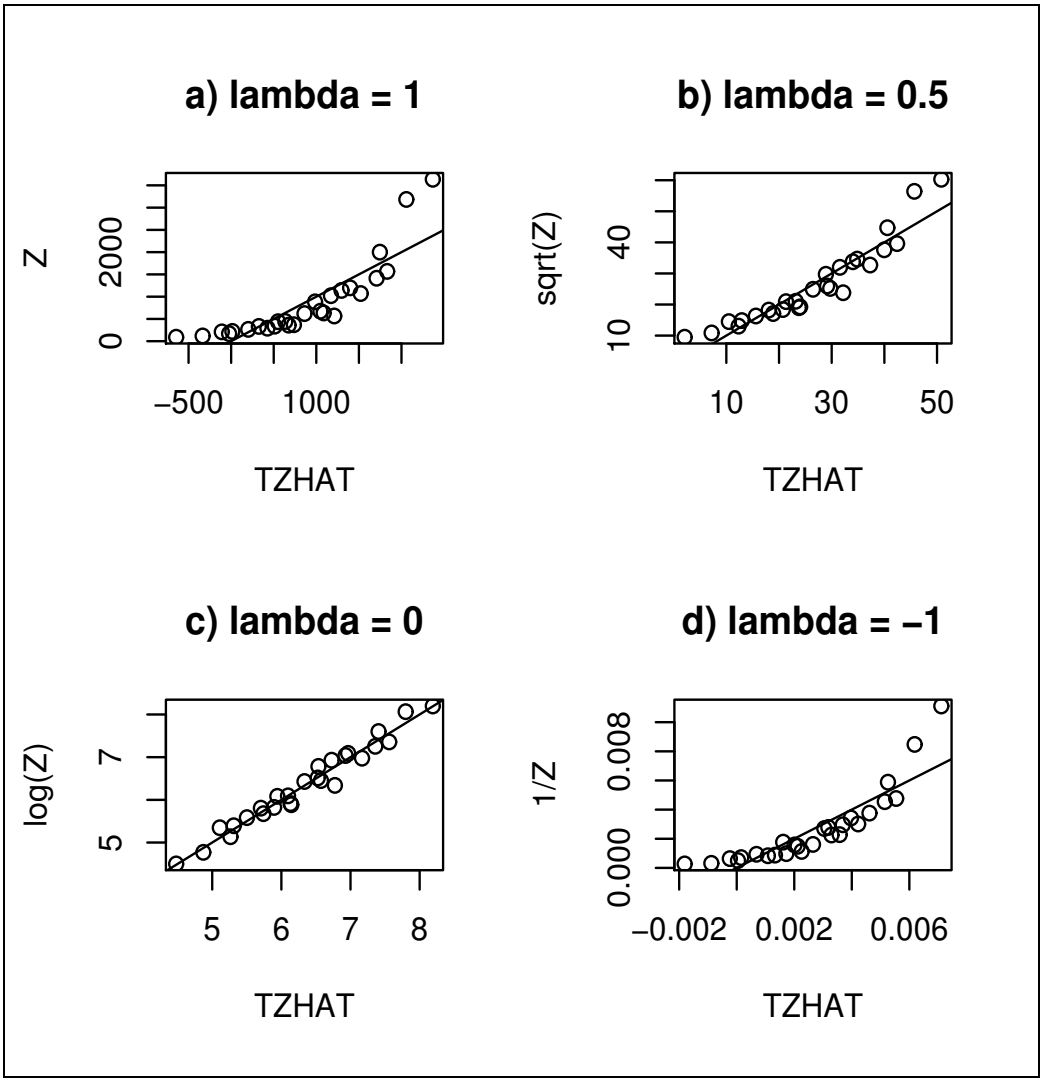

Figure 12.2: Four Transformation Plots for the Textile Data

percent confidence interval -0.18 to 0.06. These results give a strong indication that the log transformation may result in a relatively simple model, as argued by Box and Cox. Nevertheless, the numerical Box–Cox transformation method provides no direct way of judging the transformation against the data.

Shown in Figure 12.2 are transformation plots of $\hat{Z}$ versus Z^λ for four values of λ except $\log(Z)$ is used if $\lambda = 0$. The plots show how the transformations bend the data to achieve a homoscedastic linear trend. Perhaps more importantly, they indicate that the information on the transformation is spread throughout the data in the plot since changing λ causes all points along the curvilinear scatter in Figure 12.2a to form along a linear scatter in Figure 12.2c. Dynamic plotting using λ as a control seems quite effective for judging transformations against the data and the log response transformation does indeed seem reasonable.

Note the simplicity of the method: Figure 12.2a shows that a response transformation is needed since the plotted points follow a nonlinear curve while Figure 12.2c suggests that $Y = \log(Z)$ is the appropriate response transformation since the plotted points follow the identity line. If all 7 plots were made for $\lambda \in \Lambda_L$, then $\lambda = 0$ would be selected since this plot is linear. Also, Figure 12.2a suggests that the log rule is reasonable since $\max(Z)/\min(Z) > 10$.

The essential point of the next example is that observations that influence the choice of the usual Box–Cox numerical power transformation are often easily identified in the transformation plots. The transformation plots are especially useful if the bivariate relationships of the predictors, as seen in the scatterplot matrix of the predictors, are linear.

Example 12.4: Mussel Data. Cook and Weisberg (1999a, p. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. Suppose the response Z is *muscle mass* M in grams, and the predictors are the *length* L and *height* H of the shell in mm, the logarithm $\log W$ of the *shell width* W, the logarithm $\log S$ of the *shell mass* S and a constant. With this starting point, we might expect a log transformation of M to be needed because M and S are both mass measurements and $\log S$ is being used as a predictor. Using $\log M$ would essentially reduce all measurements to the scale of length. The Box–Cox likelihood method gave $\hat{\lambda}_0 = 0.28$ with approximate 95 percent confidence interval 0.15 to 0.4. The log transformation

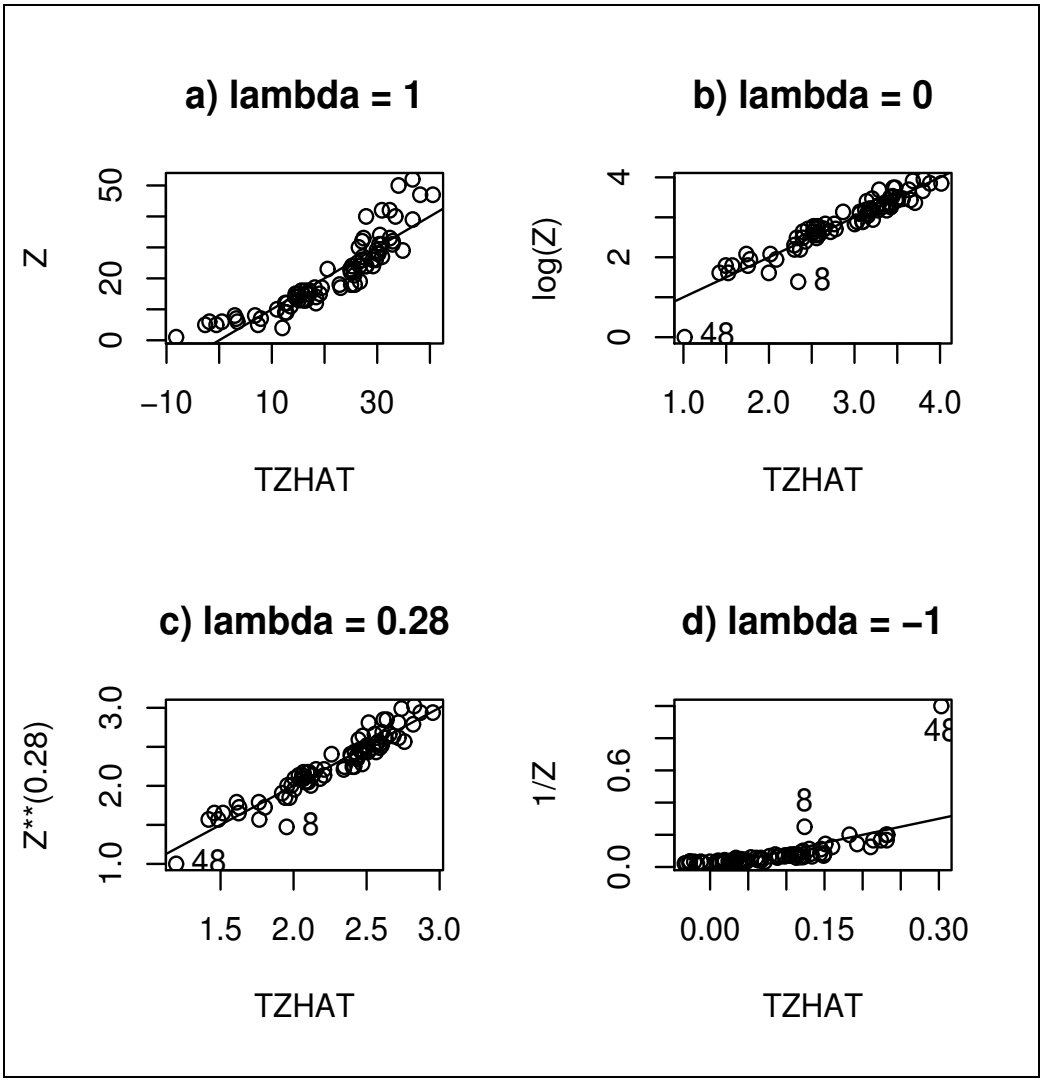

Figure 12.3: Transformation Plots for the Mussel Data

is excluded under this inference leading to the possibility of using different transformations of the two mass measurements.

Shown in Figure 12.3 are transformation plots for four values of λ. A striking feature of these plots is the two points that stand out in three of the four plots (cases 8 and 48). The Box–Cox estimate $\hat{\lambda} = 0.28$ is evidently influenced by the two outlying points and, judging deviations from the identity line in Figure 12.3c, the mean function for the remaining points is curved. In other words, the Box–Cox estimate is allowing some visually evident curvature in the bulk of the data so it can accommodate the two outlying points. Recomputing the estimate of λ_o without the highlighted points gives $\hat{\lambda}_o = -0.02$, which is in good agreement with the log transformation anticipated at the outset. Reconstruction of the transformation plots indicated that now the information for the transformation is consistent throughout the data on the horizontal axis of the plot.

Note that in addition to helping visualize $\hat{\lambda}$ against the data, the transformation plots can also be used to show the curvature and heteroscedasticity in the competing models indexed by $\lambda \in \Lambda_L$. Example 12.3 shows that the plot can also be used as a diagnostic to assess the success of numerical methods such as the Box–Cox procedure for estimating λ_o.

12.3 Variable Selection

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted without important loss of information. First we review variable selection for the multiple linear regression (MLR) model, and then adapt the techniques for multivariate linear regression.

12.3.1 Variable Selection for the MLR Model

This subsection follows Olive and Hawkins (2005) closely. A *model for variable selection* in multiple linear regression can be described by

$$Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = \boldsymbol{\beta}^T\boldsymbol{x} + e = \boldsymbol{\beta}_S^T\boldsymbol{x}_S + \boldsymbol{\beta}_E^T\boldsymbol{x}_E + e = \boldsymbol{\beta}_S^T\boldsymbol{x}_S + e \qquad (12.6)$$

where e is an error, Y is the response variable, $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is a $k_S \times 1$ vector and $\boldsymbol{x}_E$ is a $(p - k_S) \times 1$ vector.

Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \mathbf{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of k terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Then

$$Y = \boldsymbol{\beta}_I^T \boldsymbol{x}_I + \boldsymbol{\beta}_O^T \boldsymbol{x}_O + e. \tag{12.7}$$

Definition 12.11. The model $Y = \boldsymbol{\beta}^T \boldsymbol{x} + e$ that uses all of the predictors is called the *full model*. A model $Y = \boldsymbol{\beta}_I^T \boldsymbol{x}_I + e$ that only uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel*. The *sufficient predictor* (SP) is the linear combination of the predictor variables used in the model. Hence the full model has $SP = \boldsymbol{\beta}^T \boldsymbol{x}$ and the submodel has $SP = \boldsymbol{\beta}_I^T \boldsymbol{x}_I$.

Notice that the full model is a submodel. The estimated sufficient predictor (ESP) is $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ and the following remarks suggest that *a submodel I is worth considering if the correlation* $\text{corr}(ESP, ESP(I)) \geq 0.95$. Suppose that S is a subset of I and that model (12.6) holds. Then

$$SP = \boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}_S^T \boldsymbol{x}_S = \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T \boldsymbol{x}_{I/S} + \mathbf{0}^T \boldsymbol{x}_O = \boldsymbol{\beta}_I^T \boldsymbol{x}_I \tag{12.8}$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \mathbf{0}$ and the sample correlation $\text{corr}(\boldsymbol{\beta}^T \boldsymbol{x}_i, \boldsymbol{\beta}_I^T \boldsymbol{x}_{I,i}) = 1.0$ for the population model if $S \subseteq I$.

This subsection proposes a graphical method for evaluating candidate submodels. Let $\hat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$ obtained from the regression of Y on all of the terms $\boldsymbol{x}$. Denote the residuals and fitted values from the *full model* by $r_i = Y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i = Y_i - \hat{Y}_i$ and $\hat{Y}_i = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ respectively. Similarly, let $\hat{\boldsymbol{\beta}}_I$ be the estimate of $\boldsymbol{\beta}_I$ obtained from the regression of Y on $\boldsymbol{x}_I$ and denote the corresponding residuals and fitted values by $r_{I,i} = Y_i - \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ and $\hat{Y}_{I,i} = \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ where $i = 1, ..., n$. Two important summary statistics for a multiple linear regression model are R^2, the proportion of the variability of Y explained by the nontrivial predictors in the model, and the estimate $\hat{\sigma}$ of the error standard deviation σ.

Definition 12.12. The "fit–fit" or *FF plot* is a plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ while a "residual–residual" or *RR plot* is a plot $r_{I,i}$ versus r_i. A *response plot* is a plot of $\hat{Y}_{I,i}$ versus Y_i. A *residual plot* is a plot of $\hat{Y}_{I,i}$ versus $r_{I,i}$.

Many numerical methods such as forward selection, backward elimination, stepwise and all subset methods using the $C_p(I)$ criterion (Jones 1946, Mallows 1973), have been suggested for variable selection. We will use the FF plot, RR plot, the response plots from the full and submodel, and the residual plots (of the fitted values versus the residuals) from the full and submodel. These six plots will contain a great deal of information about the candidate subset provided that Equation (12.6) holds and that a good estimator for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_I$ is used.

For these plots to be useful, it is crucial to verify that a multiple linear regression (MLR) model is appropriate for the full model. **Both the response plot and the residual plot for the full model need to be used to check this assumption**. The plotted points in the response plot should cluster about the *identity line* (that passes through the origin with unit slope) while the plotted points in the residual plot should cluster about the line $r = 0$. Any nonlinear patterns or outliers in either plot suggests that an MLR relationship does not hold. Similarly, before accepting the candidate model, use the response plot and the residual plot from the candidate model to verify that an MLR relationship holds for the response Y and the predictors $\boldsymbol{x}_I$. If the submodel is good, then the residual and response plots of the submodel should be nearly identical to the corresponding plots of the full model. Assume that all submodels contain a constant.

Remark 12.4. To visualize whether a candidate submodel using predictors $\boldsymbol{x}_I$ is good, use the fitted values and residuals from the submodel and full model to make an RR plot of the $r_{I,i}$ versus the r_i and an FF plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$. Add the OLS line to the RR plot and identity line to both plots as visual aids. The subset I is good if the plotted points cluster tightly about the identity line in *both plots*. In particular, the OLS line and the identity line should "nearly coincide" so that it is difficult to tell that the two lines intersect at the origin in the RR plot.

The following notation will be useful. Suppose that all submodels include a constant and that $\boldsymbol{X}$ is the full rank $n \times p$ design matrix for the full model. Let the corresponding vectors of OLS fitted values and residuals be $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$ and $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, respectively. Suppose that $\boldsymbol{X}_I$ is the $n \times k$ design matrix for the candidate submodel and that the corresponding vectors of OLS fitted values and residuals are $\hat{\boldsymbol{Y}}_I = \boldsymbol{X}_I(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y} = \boldsymbol{H}_I\boldsymbol{Y}$ and $\boldsymbol{r}_I = (\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$, respectively. For

multiple linear regression, recall that if the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then the F_I statistic for testing whether the $p-k$ predictor variables in $\boldsymbol{x}_O$ can be deleted is

$$F_I = \frac{SSE(I) - SSE}{(n-k)-(n-p)} / \frac{SSE}{n-p} = \frac{n-p}{p-k}[\frac{SSE(I)}{SSE} - 1]$$

where SSE is the error sum of squares from the full model and SSE(I) is the error sum of squares from the candidate submodel. Then

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k$$

where MSE is the error mean square for the full model. Notice that $C_p(I) \leq 2k$ if and only if $F_I \leq p/(p-k)$. Remark 12.7 below suggests that for subsets I with k terms, submodels with $C_p(I) \leq \min(2k, p)$ are especially interesting.

Olive (2013, proposition 5.1) shows that

$$\text{corr}(r, r_I) = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}, \tag{12.9}$$

and that the plotted points in the FF, RR and response plots will cluster about the identity line. This proposition is a property of OLS and holds even if the data does not follow an MLR model.

Remark 12.5. Note that for large n, $C_p(I) < k$ or $F_I < 1$ will force corr(ESP,ESP(I)) to be high (≥ 0.95). Let d be a lower bound on $\text{corr}(r, r_I)$. If

$$C_p(I) \leq 2k + n\left[\frac{1}{d^2} - 1\right] - \frac{p}{d^2},$$

then $\text{corr}(r, r_I) \geq d$. The simple screen $C_p(I) \leq 2k$ corresponds to

$$d_n \equiv \sqrt{1 - \frac{p}{n}}.$$

To reduce the chance of overfitting, use the screen $C_p(I) \leq \min(2k, p)$.

A standard model selection procedure will often be needed to suggest models. For example, forward selection or backward elimination could be

used. If $p < 30$, Furnival and Wilson (1974) provide a technique for selecting a few candidate subsets after examining all possible subsets.

Rule of thumb 12.3 (assuming that the cost of each predictor is the same): a) After using a numerical method such as forward selection or backward elimination, let I_{min} correspond to the submodel with the smallest C_p. Find the submodel I_I with the fewest number of predictors such that $C_p(I_I) \leq C_p(I_{min}) + 1$. Then I_I is the initial submodel that should be examined. It is possible that $I_I = I_{min}$ or that I_I is the full model. Do not use more predictors than model I_I to avoid overfitting.

b) Models I with fewer predictors than I_I such that $C_p(I) \leq C_p(I_{min}) + 4$ are interesting and should also be examined.

c) Models I with k predictors, including a constant and with fewer predictors than I_I such that $C_p(I_{min}) + 4 < C_p(I) \leq \min(2k, p)$ should be checked but often underfit: important predictors are deleted from the model. Underfit is especially likely to occur if a predictor with one degree of freedom is deleted and the jump in C_p is large, greater than 4, say. (A factor has $c - 1$ degrees of freedom corresponding to the $c - 1$ indicator variables used to define the factor, and usually either all of the indicator variables are kept or deleted by variable selection software.)

d) If there are no models I with fewer predictors than I_I such that $C_p(I) \leq \min(2k, p)$, then model I_I is a good candidate for the best subset found by the numerical procedure.

Variable selection seeks a subset I of the variables to keep in the model. The submodel I will always contain a constant and will have $k - 1$ nontrivial predictors where $1 \leq k \leq p$.

Forward selection starts with a constant $= W_1 = X_1$. Step 1) $k = 2$: compute C_p for all models containing the constant and a single predictor X_i. Keep the predictor $W_2 = X_j$, say, that corresponds to the model with the smallest value of C_p.
Step 2) $k = 3$: Fit all models with $k = 3$ that contain W_1 and W_2. Keep the predictor W_3 that minimizes C_p. ...
Step j) $k = j + 1$: Fit all models with $k = j + 1$ that contains $W_1, W_2, ..., W_j$. Keep the predictor W_{j+1} that minimizes C_p. ...
Step $p - 1$): Fit the full model.

Backward elimination: All models contain a constant $= U_1 = X_1$.

Step 1) $k = p$: Start with the full model that contains $X_1, ..., X_p$. We will also say that the full model contains $U_1, ..., U_p$ where $U_1 = X_1$ but U_i need not equal X_i for $i > 1$.
Step 2) $k = p - 1$: fit each model with $p - 1$ predictors including a constant. Delete the predictor U_p, say, that corresponds to the model with the smallest C_p. Keep $U_1, ..., U_{p-1}$.
Step 3) $k = p-2$: fit each model with $p-2$ predictors and a constant. Delete the predictor U_{p-1} that corresponds to the smallest C_p. Keep $U_1, ..., U_{p-2}$. ...
Step j) $k = p - j + 1$: fit each model with $p - j + 1$ predictors and a constant. Delete the predictor U_{p-j+2} that corresponds to the smallest C_p. Keep $U_1, ..., U_{p-j+1}$. ...
Step $p - 1$) $k = 2$. The current model contains U_1, U_2 and U_3. Fit the model U_1, U_2 and the model U_1, U_3. Assume that model U_1, U_2 minimizes C_p. Then delete U_3 and keep U_1 and U_2.

Assume that the full model has p predictors including a constant and that the submodel I has k predictors including a constant. Assume that the full model has good response and residual plots and that $n > 5p$. Then we would like following properties i) – xi) (roughly in order of decreasing importance) to hold. Often we can not find a submodel where i) – xi) all hold simultaneously. Do not use more predictors than model I_I to avoid overfitting.

Then the submodel I is good if
i) the response and residual plots for the submodel looks like the response and residual plots for the full model.
ii) corr(ESP,ESP(I)) = corr($\hat{\mathrm{Y}}, \hat{\mathrm{Y}}_{\mathrm{I}}$) ≥ 0.95.
iii) The plotted points in the FF plot cluster tightly about the identity line.
iv) Want the p-value ≥ 0.01 for the partial F test that uses I as the reduced model.
v) Want $k \leq n/10$.
vi) The plotted points in the RR plot cluster tightly about the identity line.
vii) Want $R^2(I) > 0.9R^2$ and $R^2(I) > R^2 - 0.07$ ($R^2(I) \leq R^2(full)$ since adding predictors to I does not decrease $R^2(I)$).
viii) Want $C_p(I_{min}) \leq C_p(I) \leq \min(2k, p)$ with no big jumps in C_p (the increase should be less than four) as variables are deleted.
ix) Want hardly any predictors with p-values > 0.05.
x) Want few predictors with p-values between 0.01 and 0.05.
xi) Want MSE(I) to be smaller than or not much larger than the MSE from

the full model.

Example 12.5. The FF and RR plots can be used as a diagnostic for whether a given numerical method is including too many variables. Gladstone (1905-1906) attempts to estimate the *weight* of the human brain (measured in grams after the death of the subject) using simple linear regression with a variety of predictors including *age* in years, *height* in inches, *head height* in mm, *head length* in mm, *head breadth* in mm, *head circumference* in mm, and *cephalic index*. The *sex* (coded as 0 for females and 1 for males) of each subject was also included. The variable *cause* was coded as 1 if the cause of death was acute, 3 if the cause of death was chronic, and coded as 2 otherwise. A variable *ageclass* was coded as 0 if the age was under 20, 1 if the age was between 20 and 45, and as 3 if the age was over 45. Head *size*, the product of the *head length*, *head breadth*, and *head height*, is a volume measurement, hence $(size)^{1/3}$ was also used as a predictor with the same physical dimensions as the other lengths. Thus there are 11 nontrivial predictors and one response, and all models will also contain a constant. Nine cases were deleted because of missing values, leaving 267 cases.

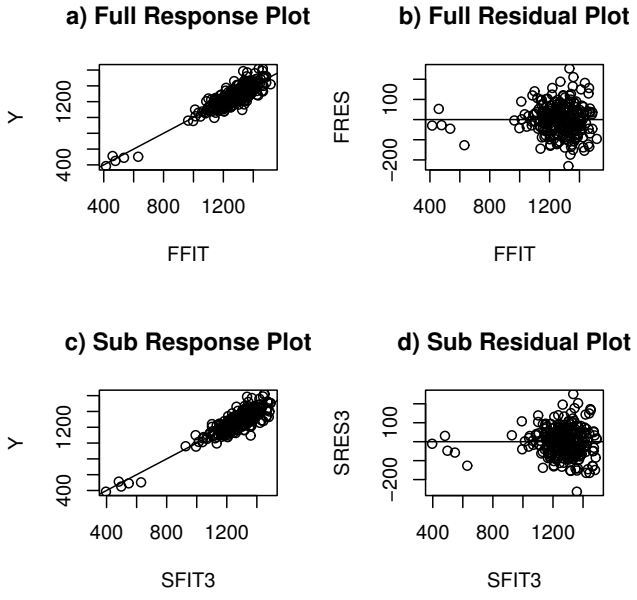

Figure 12.4: Gladstone data: comparison of the full model and the submodel.

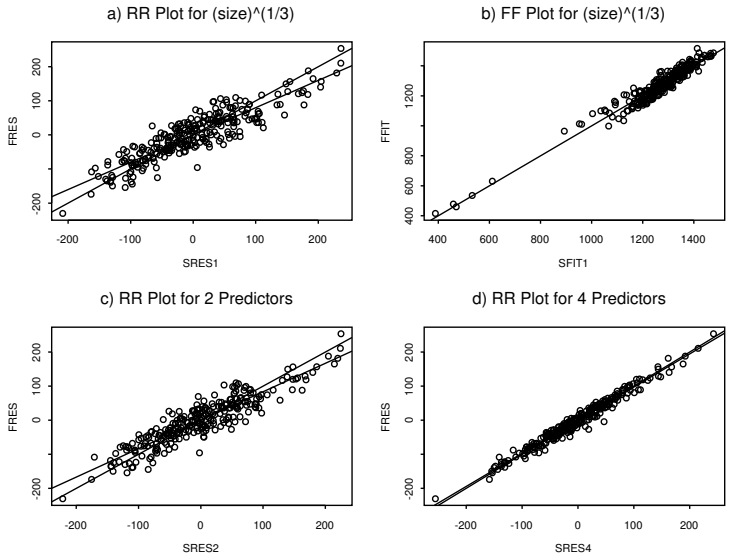

Figure 12.5: Gladstone data: submodels added $(size)^{1/3}$, sex, age and finally $breadth$.

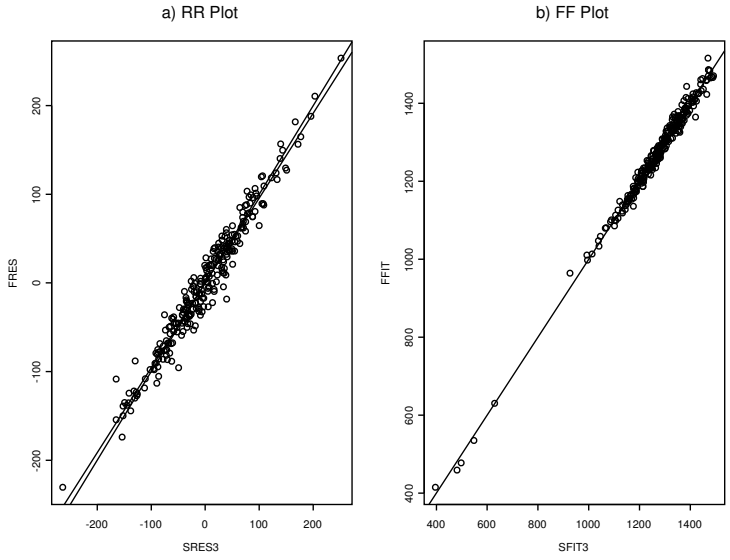

Figure 12.6: Gladstone data with Predictors $(size)^{1/3}$, sex, and age

Figure 12.4 shows the response plots and residual plots for the full model and the final submodel that used a constant, $size^{1/3}$, *age* and *sex*. The five cases separated from the bulk of the data in each of the four plots correspond to five infants. These may be outliers, but the visual separation reflects the small number of infants and toddlers in the data. A purely numerical variable selection procedure would miss this interesting feature of the data. We will first perform variable selection with the entire data set, and then examine the effect of deleting the five cases. Using forward selection and the C_p statistic on the Gladstone data suggests the subset I_5 containing a constant, $(size)^{1/3}$, *age*, *sex*, *breadth*, and *cause* with $C_p(I_5) = 3.199$. The p–values for breadth and cause were 0.03 and 0.04, respectively. The subset I_4 that deletes *cause* has $C_p(I_4) = 5.374$ and the p–value for *breadth* was 0.05. Figure 12.5d shows the RR plot for the subset I_4. Note that the correlation of the plotted points is very high and that the OLS and identity lines nearly coincide.

A scatterplot matrix of the predictors and response suggests that $(size)^{1/3}$ might be the best single predictor. First we regressed $Y =$ *brain weight* on the eleven predictors described above (plus a constant) and obtained the residuals r_i and fitted values $\hat{Y}_i$. Next, we regressed Y on the subset I containing $(size)^{1/3}$ and a constant and obtained the residuals $r_{I,i}$ and the fitted values $\hat{Y}_{I,i}$. Then the RR plot of $r_{I,i}$ versus r_i, and the FF plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ were constructed.

For this model, the correlation in the FF plot (Figure 12.5b) was very high, but in the RR plot the OLS line did not coincide with the identity line (Figure 12.5a). Next *sex* was added to I, but again the OLS and identity lines did not coincide in the RR plot (Figure 12.5c). Hence *age* was added to I. Figure 12.6a shows the RR plot with the OLS and identity lines added. These two lines now nearly coincide, suggesting that a constant plus $(size)^{1/3}$, *sex*, and *age* contains the relevant predictor information. This subset has $C_p(I) = 7.372$, $R_I^2 = 0.80$, and $\hat{\sigma}_I = 74.05$. The full model which used 11 predictors and a constant has $R^2 = 0.81$ and $\hat{\sigma} = 73.58$. Since the C_p criterion suggests adding *breadth* and *cause*, the C_p criterion may be leading to an overfit.

Figure 12.6b shows the FF plot. The five cases in the southwest corner correspond to five infants. Deleting them leads to almost the same conclusions, although the full model now has $R^2 = 0.66$ and $\hat{\sigma} = 73.48$ while the submodel has $R_I^2 = 0.64$ and $\hat{\sigma}_I = 73.89$.

12.3.2 Variable Selection for Multivariate Linear Regression

We still have the full model $\boldsymbol{x} = (\boldsymbol{x}_I^T, \boldsymbol{x}_O)^T$ where $\boldsymbol{x}_I$ is a candidate submodel. It is crucial to verify that a multivariate regression model is appropriate for the full model. **For each of the m response variables, use the response plot and the residual plot for the full model to check this assumption**.

To obtain the candidate subset for multivariate regression, do numerical variable selection such as forward selection or backward elimination for multiple linear regression for each response variable Y_j. Very often predictor variables are highly correlated and often similar sets of predictor variables will be used by each of the m multiple linear regressions. See if there is a pattern to the most important and least important predictors. Try to get rid of predictors that are not needed in any of the m multiple linear regressions. It is better to keep too many predictors than to possible delete a predictor that is needed in at least one of the m multiple linear regression, but want $n > 10p$.

Check the submodel $\boldsymbol{x}_I$ for multivariate linear regression with the FF, RR plots and the response and residual plots for the full model and for the candidate model for each of the m response variables $Y_1, ..., Y_m$. The submodels use Y_{Ij} for $j = 1, ..., m$.

12.4 Prediction

12.4.1 Prediction Intervals for Multiple Linear Regression

This subsection gives estimators for predicting a future or new value Y_f of the vector of response variables given the predictors $\boldsymbol{x}_f$. The following subsection will extend the results to multivariate regression.

Warning: All too often the MLR model seems to fit the data

$$(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_n, Y_n)$$

well, but when new data is collected, a very different MLR model is needed to fit the new data well. In particular, the MLR model seems to fit the data $(\boldsymbol{x}_i, Y_i)$ well for $i = 1, ..., n$, but when the researcher tries to predict Y_f for a

new vector of predictors $\boldsymbol{x}_f$, the prediction is very poor in that $\hat{Y}_f$ is not close to the Y_f actually observed. **Wait until after the MLR model has been shown to make good predictions before claiming that the model gives good predictions!**

There are several reasons why the MLR model may not fit new data well. i) The model building process is usually iterative. Data Z, $w_1, ..., w_k$ is collected. If the model is not linear, then functions of Z are used as a potential response and functions of the w_i as potential predictors. After trial and error, the functions are chosen, resulting in a final MLR model using Y and $x_1, ..., x_p$. Since the same data set was used during the model building process, biases are introduced and the MLR model fits the "training data" better than it fits new data. Suppose that Y, $x_1, ..., x_p$ are specified before collecting data and that the residual and response plots from the resulting MLR model look good. Then predictions from the prespecified model will often be better for predicting new data than a model built from an iterative process.

ii) If $(\boldsymbol{x}_f, Y_f,)$ come from a different population than the population of $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_n, Y_n)$, then prediction for Y_f can be arbitrarily bad.

iii) Even a good MLR model may not provide good predictions for an $\boldsymbol{x}_f$ that is far from the $\boldsymbol{x}_i$ (extrapolation).

iv) The MLR model may be missing important predictors (underfitting).

v) The MLR model may contain unnecessary predictors (overfitting).

Two remedies for i) are a) use previously published studies to select an MLR model before gathering data. b) Do a trial study. Collect some data, build an MLR model using the iterative process. Then use this model as the prespecified model and collect data for the main part of the study. Better yet, do a trial study, specify a model, collect more trial data, improve the specified model and repeat until the latest specified model works well. Unfortunately, trial studies are often too expensive or not possible because the data is difficult to collect. Also, often the population from a published study is quite different from the population of the data collected by the researcher. Then the MLR model from the published study is not adequate.

Definition 12.13. Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ and the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Let $h_i = h_{ii}$ be the ith diagonal element of $\boldsymbol{H}$ for $i = 1, ..., n$. Then h_i is called the ith **leverage** and $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$. Suppose new data is to be collected with predictor vector $\boldsymbol{x}_f$. Then the

leverage of $\boldsymbol{x}_f$ is $h_f = \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f$. **Extrapolation** occurs if $\boldsymbol{x}_f$ is far from the $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$.

Rule of thumb 12.4. Predictions based on extrapolation are not reliable. A rule of thumb is that extrapolation occurs if $h_f > \max(h_1, ..., h_n)$. This rule works best if the predictors are linearly related in that a plot of x_i versus x_j should not have any strong nonlinearities. If there are strong nonlinearities among the predictors, then $\boldsymbol{x}_f$ could be far from the $\boldsymbol{x}_i$ but still have $h_f < \max(h_1, ..., h_n)$.

Example 12.6. Consider predicting $Y = \textit{weight}$ from $x = \textit{height}$ and a constant from data collected on men between 18 and 24 where the minimum height was 57 and the maximum height was 79 inches. The OLS equation was $\hat{Y} = -167 + 4.7x$. If $x = 70$ then $\hat{Y} = -167 + 4.7(70) = 162$ pounds. If $x = 1$ inch, then $\hat{Y} = -167 + 4.7(1) = -162.3$ pounds. It is impossible to have negative weight, but it is also impossible to find a 1 inch man. This MLR model should not be used for x far from the interval $(57, 79)$.

The following theorem is analogous to the central limit theorem and the theory for the t–interval for μ based on $\overline{Y}$ and the sample standard deviation (SD) S_Y. If the data $Y_1, ..., Y_n$ are iid with mean 0 and variance σ^2, then $\overline{Y}$ is asymptotically normal and the t–interval will perform well if the sample size is large enough. The result below suggests that the OLS estimators $\hat{Y}_i$ and $\hat{\boldsymbol{\beta}}$ are good if the sample size is large enough. The condition $\max h_i \to 0$ in probability usually holds if the researcher picked the design matrix $\boldsymbol{X}$ or if the $\boldsymbol{x}_i$ are iid random vectors from a well behaved population. Outliers can cause the condition to fail.

Theorem 12.3: Huber (1981, p. 157-160). Consider the MLR model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ and assume that the errors are independent with zero mean and the same variance: $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$. Also assume that $\max_i(h_1, ..., h_n) \to 0$ in probability as $n \to \infty$. Then

a) $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} \to E(Y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$ in probability for $i = 1, ..., n$ as $n \to \infty$.

b) All of the least squares estimators $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ are asymptotically normal where $\boldsymbol{a}$ is any fixed constant $p \times 1$ vector.

Theorem 12.4. The least squares estimator satisfies $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = o_P(1)$ if

$$\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{X}^T\boldsymbol{e}}{n}\right) = o_P(1).$$

Proof:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\frac{\boldsymbol{X}^T \boldsymbol{X}}{n} \right)^{-1} \left(\frac{\boldsymbol{X}^T \boldsymbol{e}}{n} \right).$$

Definition 12.14. A large sample $100(1-\delta)\%$ prediction interval (PI) has the form $(\hat{L}_n, \hat{U}_n)$ where $P(\hat{L}_n < Y_f < \hat{U}_n) \xrightarrow{P} 1-\delta$ as the sample size $n \to \infty$.

The interpretation of a 100 $(1-\delta)\%$ PI for a random variable Y_f is similar to that of a confidence interval (CI). Collect data, then form the PI, and repeat for a total of k times where k trials are independent from the same population. If Y_{fi} is the ith random variable and PI_i is the ith PI, then the probability that $Y_{fi} \in PI_i$ for m of the PIs follows a binomial$(k, \rho = 1-\delta)$ distribution. Hence if 100 95% PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times.

The length of the CI goes to 0 as the sample size n goes to ∞ while the length of the PI converges to some nonzero number L, say. To see this, consider $\boldsymbol{x}_f$ such that the heights Y of women between 18 and 24 is normal with a mean of 66 inches and an SD of 3 inches. A 95% CI for $E(Y|\boldsymbol{x}_f)$ should be centered at about 66 and the length should go to zero as n gets large. But a 95% PI needs to contain about 95% of the heights so the PI should converge to the interval $66 \pm 1.96(3)$. This result follows because if $Y \sim N(66, 9)$ then $P(Y < 66 - 1.96(3)) = P(Y > 66 + 1.96(3)) = 0.025$. In other words, the endpoints of the PI estimate the 97.5 and 2.5 percentiles of the normal distribution. However, the percentiles of a parametric error distribution depend heavily on the parametric distribution and the parametric formulas are violated if the assumed error distribution is incorrect.

Let ξ_δ be the δ percentile of the error e, ie, $P(e \leq \xi_\delta) = \delta$. Let $\hat{\xi}_\delta$ be the sample δ percentile of the residuals. The percentiles of the residuals are consistent estimators, $\hat{\xi}_\delta \xrightarrow{P} \xi_\delta$, under "mild" regularity conditions, and this consistency is the basis for using QQ plots. For multiple linear regression with iid errors with constant variance σ^2, sufficient conditions are $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ and the $\boldsymbol{x}_i$ are bounded in probability. See Olive (2011), Olive and Hawkins (2003), Welsh (1986) and Rousseeuw and Leroy (1987, p. 128).

For many error distributions,

$$E(MSE) = E\left(\sum_{i=1}^{n} \frac{r_i^2}{n-p}\right) = \sigma^2 = E\left(\sum_{i=1}^{n} \frac{e_i^2}{n}\right).$$

This result suggests that

$$\sqrt{\frac{n}{n-p}} r_i \approx e_i.$$

Let

$$a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}}\sqrt{(1+h_f)}. \tag{12.10}$$

Following Olive (2007), a PI is asymptotically optimal if it has the shortest asymptotic length that gives the desired asymptotic coverage. If the error distribution is unimodal, an asymptotically optimal PI can be created by applying the shorth(c) estimator to the residuals where $c = \lceil n(1-\delta) \rceil$ and $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. That is, let $r_{(1)}, ..., r_{(n)}$ be the order statistics of the residuals. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, ..., r_{(n)} - r_{(n-c+1)}$. Let $(r_{(d)}, r_{(d+c-1)}) = (\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2})$ correspond to the interval with the smallest distance. Then the large sample 100 $(1-\delta)$% PI for Y_f is

$$(\hat{Y}_f + a_n \tilde{\xi}_{\delta_1}, \hat{Y}_f + a_n \tilde{\xi}_{1-\delta_2}). \tag{12.11}$$

12.4.2 Prediction Intervals for Multivariate linear Regression

For multivariate linear regression, want to predict a future or new value $\boldsymbol{Y}_f = (Y_{1f}, ..., Y_{mf})^T$ of the vector of m response variables given the vector of predictors $\boldsymbol{x}_f$.

The collection of m prediction intervals $(L_{1n}, U_{1n}), ..., (L_{mn}, U_{mn})$ are *large sample simultaneous conservative* $100(1-\delta)$% *prediction intervals* for Y_{jf} if the m prediction intervals all hold simultaneously, that is all m PIs (L_{jn}, U_{jn}) contain Y_{jf}, with probability $1 - \gamma_n$ where $1 - \gamma_n \to 1 - \gamma \geq 1 - \delta$ as $n \to \infty$.

The *Bonferroni* simultaneous PIs are made by increasing the coverage of a single PI from $1-\delta$ to $(1-\delta/m)$. Hence 90% large sample simultaneous PIs will use coverage 0.95 if $m = 2$ and coverage 0.99 if $m = 10$. Let E_j be an event with $P(E_j) = 1 - \delta_j$. Let $\overline{E}_j$ be the compliment of E_j so $P(\overline{E}_j) = \delta_j$.

Then Bonferroni's inequality is

$$P(\cap_{j=1}^{m} E_j) = 1 - P(\overline{\cap_{j=1}^{m} E_j}) = 1 - P(\cup_{j=1}^{m} \overline{E_j}) \geq 1 - \sum_{j=1}^{m} P(\overline{E}_j) =$$

$= 1 - \sum_{j=1}^{m} \delta_j = 1 - \delta$ if $\delta_j = \delta/m$. To use this inequality for simultaneous intervals, let E_j be the event that the jth PI contains Y_{jf}. Then $P(\cap_{j=1}^{m} E_j)$ is the probability that all m PIs contain Y_{jf} for $j = 1, ..., m$.

Let $\tau = \delta/m$. Then the m large sample simultaneous conservative $100(1-\delta)\%$ PIs are

$$(\hat{Y}_{jf} + a_n \tilde{\xi}_{\tau_1}, \hat{Y}_{jf} + a_n \tilde{\xi}_{1-\tau_2}) \tag{12.12}$$

for $j = 1, ..., m$ using Equation (12.11) and residuals $r_{1,j}, ..., r_{n,j}$. That is, make the $100(1-\tau)\%$ PI (12.11) for Y_{jf} for $j = 1, ..., m$ corresponding to the multiple linear regression of the jth response variable Y_j on $\boldsymbol{X}$.

These PIs make no use of the fact that $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, but no parametric distribution for the $\boldsymbol{\epsilon}_i$ is needed. The classical simultaneous prediction region for $\boldsymbol{y}_f$ assumes that the $\boldsymbol{\epsilon}_i$ are iid $N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ and tend to have large undercoverage (are too liberal) when the normality assumption is violated, which is usually the case.

12.4.3 Prediction Regions

Suppose a prediction region for $\boldsymbol{y}_f$ given a vector of predictors $\boldsymbol{x}_f$ is desired. If we had many cases $\boldsymbol{z}_i = \boldsymbol{B}^T \boldsymbol{x}_f + \boldsymbol{\epsilon}_i$, then we could make a prediction region for $\boldsymbol{z}_i$ using Section 5.2. Instead, use $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{B}}^T \boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$. This takes the data cloud of the n residual vectors $\hat{\boldsymbol{\epsilon}}_i$ and centers the cloud at $\hat{\boldsymbol{y}}_f$. Note that $\hat{\boldsymbol{z}}_i = (\boldsymbol{B} - \boldsymbol{B} + \hat{\boldsymbol{B}})^T \boldsymbol{x}_f + (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i) = \boldsymbol{z}_i + (\hat{\boldsymbol{B}} - \boldsymbol{B})^T \boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i = \boldsymbol{z}_i + O_P(n^{-1/2})$. Hence the distances based on the $\boldsymbol{z}_i$ and the distances based on the $\hat{\boldsymbol{z}}_i$ should have the same quantiles, asymptotically.

Theorem 12.5. 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 $\boldsymbol{\epsilon}_f$ and the $\boldsymbol{\epsilon}_i$ are iid for $i = 1, ..., n$. 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 $q_n = \min(1 - \alpha + 0.05, 1 - \alpha + m/n)$ for $\alpha > 0.1$ and

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

If $q_n < 1 - \alpha + 0.001$, set $q_n = 1 - \alpha$. Let $0 < \alpha < 1$ and $h = D_{(U_n)}$ where $D_{(U_n)}$ is the q_nth sample quantile of the D_i. Consider the nominal $100(1 - \alpha)\%$ prediction region for $\boldsymbol{y}_f$

$$\{\boldsymbol{z} : (\boldsymbol{z} - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{z} - \hat{\boldsymbol{y}}_f) \leq D^2_{(U_n)}\} =$$

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D^2_{(U_n)}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}\}. \tag{12.13}$$

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 $U_n/n \to 1 - \alpha$ 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 (12.13) is a large sample $100(1 - \alpha)\%$ 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 $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\alpha}\}$, then the prediction region (12.13) is asymptotically optimal.

Proof. a) Suppose $(\boldsymbol{x}_f, \boldsymbol{y}_f) = (\boldsymbol{x}_i, \boldsymbol{y}_i)$. Then

$$D^2_{\boldsymbol{y}_i}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i) = \hat{\boldsymbol{\epsilon}}_i^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \hat{\boldsymbol{\epsilon}}_i = D^2_{\hat{\boldsymbol{\epsilon}}_i}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}).$$

Hence $\boldsymbol{y}_i$ is in the ith prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ iff $\hat{\boldsymbol{\epsilon}}_i$ is in prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$, but exactly U_n of the $\hat{\boldsymbol{\epsilon}}_i$ are in the latter region by construction, if $D_{(U_n)}$ is unique. Since $D_{(U_n)}$ is the $(1 - \alpha)$ percentile of the D_i asymptotically, $U_n/n \to 1 - \alpha$.

b) Let $P[D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\alpha}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})] = 1 - \alpha$. Since $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, Proposition 5.1 shows that if $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} (E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ then $D(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Hence the percentiles of the distances also converge in probability, and the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{1-\alpha}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ converges to $1 - \alpha =$ the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\alpha}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$.

c) The asymptotically optimal prediction region is the region with the smallest volume (hence highest density) such that the coverage is $1 - \alpha$, as $n \to \infty$. This region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\alpha}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$ if the

asymptotically optimal region for the $\boldsymbol{\epsilon}_i$ is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\alpha}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$. Hence the result follows by b). $\square$

Multivariate linear regression satisfies Theorem 12.5, and applying a prediction region from Section 5.2 on the $\hat{\boldsymbol{z}}_i$ results in a large sample $100(1-\alpha)\%$ prediction region for $\boldsymbol{y}_f$ given the vector of predictors $\boldsymbol{x}_f$. The prediction region is asymptotically optimal if the $\boldsymbol{\epsilon}_i$ are iid from an $EC_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$ distribution for a large class of elliptically contoured distributions.

To see the above claim, note that if the $\boldsymbol{\epsilon}_i$ are iid from an elliptically contoured distribution with nonsingular covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, then the population asymptotically optimal prediction region is $\{\boldsymbol{y} : D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) < D_{1-\alpha}\}$ where $P(D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) < D_{1-\alpha}) = 1 - \alpha$. For example, if the iid $\boldsymbol{\epsilon}_i \sim N_m(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then $D_{1-\alpha} = \sqrt{\chi^2_{m,1-\alpha}}$. If the error distribution is not elliptically contoured, then the above region still has $100(1-\alpha)\%$ coverage, but prediction regions with smaller volume may exist. In general these quantities need to be estimated. If many errors $\boldsymbol{\epsilon}_i$ were available and $\boldsymbol{B}$ was known, could estimate $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ with $\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T/n$, compute $\boldsymbol{z}_i = \boldsymbol{B}^T\boldsymbol{x}_F + \boldsymbol{\epsilon}_i$ and estimate $D_{1-\alpha}$ with $D_{(\lceil n(1-\alpha)\rceil)}$, the sample $(1-\alpha)$ percentile of the $D_{\boldsymbol{z}_i}$. These quantities are unavailable, but the plug in estimators are $\hat{\boldsymbol{y}}_f = \hat{\boldsymbol{B}}^T\boldsymbol{x}_f$, $\boldsymbol{S}_r = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = (n-1)^{-1}\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T$, $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ and $\hat{D}_{1-\alpha}$, the sample $(1-\alpha)$ percentile of the $D_{\hat{\boldsymbol{z}}_i}$.

Following Section 5.2, suppose $(T, \boldsymbol{C})$ is the sample mean and scaled sample covariance matrix applied to the $\hat{\boldsymbol{z}}_i$ where the multivariate linear regression used least squares. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{y} : (\boldsymbol{y}-T)^T\boldsymbol{C}^{-1}(\boldsymbol{y}-T) \leq h^2\} = \{\boldsymbol{y} : D^2_{\boldsymbol{y}} \leq h^2\} = \{\boldsymbol{y} : D_{\boldsymbol{y}} \leq h\}. \quad (12.14)$$

A future observation (random vector) $\boldsymbol{y}_f$ is in the region (12.14) if $D_{\boldsymbol{y}_f} \leq h$. Set up the prediction region (12.14) using $h = D_{(U_n)}$ as described in Theorem 2.5. Following Section 5.2, this prediction region (12.14) will be called the nonparametric prediction region.

The nonparametic prediction region has some interesting properties. Let $\boldsymbol{S}_r$ be the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$. The sample mean of the residual vectors is $\mathbf{0}$ since least squares was used. Hence the $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ have sample covariance matrix $\boldsymbol{S}_r$, and sample mean $\hat{\boldsymbol{y}}_f$. Hence $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$, and the $D_i(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ are used to compute $D_{(U_n)}$. So if there are 100 different values $(\boldsymbol{x}_{jf}, \boldsymbol{y}_{jf})$ to be predicted, only need to update $\hat{\boldsymbol{y}}_{jf}$

for $j = 1, ..., 100$, do not need to update the covariance matrix $\boldsymbol{S}_r$.

The geometry of the nonparametric region is simple. Let R_r be the nonparametric prediction region applied to the residuals $\hat{\boldsymbol{\epsilon}}_i$, and let (12.14) be the nonparametric prediction region using $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ when the multivariate regression is fit by least squares. Then R_r is a hyperellipsoid with center $\mathbf{0}$, and the nonparametric prediction region (12.14) is the hyperellipsoid R_r translated to have center $\hat{\boldsymbol{y}}_f$.

It is common practice to examine how well the prediction regions work on the data. That is, for $i = 1, ..., n$, set $\boldsymbol{x}_f = \boldsymbol{x}_i$ and see if $\boldsymbol{y}_i$ is in the region with probability near to $1 - \alpha$ with a simulation study. Note that $\hat{\boldsymbol{y}}_f = \hat{\boldsymbol{y}}_i$ if $\boldsymbol{x}_f = \boldsymbol{x}_i$. Simulation is not needed for the nonparametric prediction region (12.14) for the data since the prediction region (12.14) centered at $\hat{\boldsymbol{y}}_i$ contains $\boldsymbol{y}_i$ iff R_r, the prediction region centered at $\mathbf{0}$, contains $\hat{\boldsymbol{\epsilon}}_i$ since $\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i = \hat{\boldsymbol{\epsilon}}_i$. Thus $100q_n\%$ of prediction regions corresponding to the data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ contain $\boldsymbol{y}_i$, and $100q_n\% \rightarrow 100(1 - \alpha)\%$. Hence the prediction regions work well on the data and should work well on $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ similar to the data. Of course simulation should be done for $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ that are not equal to data cases.

This result holds provided that the multivariate linear regression using least squares is such that the sample covariance matrix $\boldsymbol{S}_r$ of the residual vectors is nonsingular, **the multivariate regression model need not be correct**. Hence the coverage at the n data cases $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is very robust to model misspecification. Of course, the prediction regions may be very large if the model is severely misspecified, but severity of misspecification can be checked with the response and residual plots. Coverage can also be arbitrarily bad if there is extrapolation or if $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ comes from a different population than that of the data.

Example 12.5. Consider the Mussel data described in Example 2.2 with response variables $Y_1 = \log(S)$ and $Y_2 = \log(M)$ with predictors $X_2 = L, X_3 = \log(W)$, and $X_4 = height$. Figure 12.7 shows a scatterplot matrix of the data and Figure 12.8 shows a DD plot of the data with multivariate prediction regions added. These plots suggest that the data may come from an elliptically contoured distribution that is not multivariate normal. The semiparametric and nonparametric 90% prediction regions of Section 5.2 consist of the cases below the $RD = 5.86$ line and to the left of the $MD = 4.12$ line. These two lines intersect on a line through the origin that is followed by the plotted points. The parametric MVN prediction region is given by the points below the $RD = 3.33$ line and does not contain enough cases.

Figures 12.9 and 12.10 give the response and residual plots for Y_1 and Y_2.

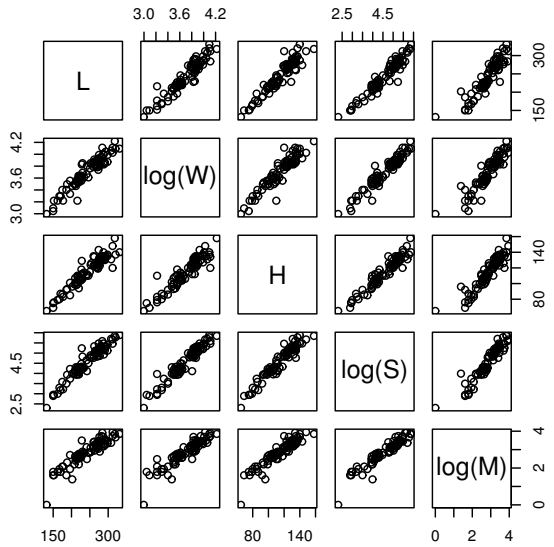

Figure 12.7: Scatterplot Matrix of the Mussels Data.

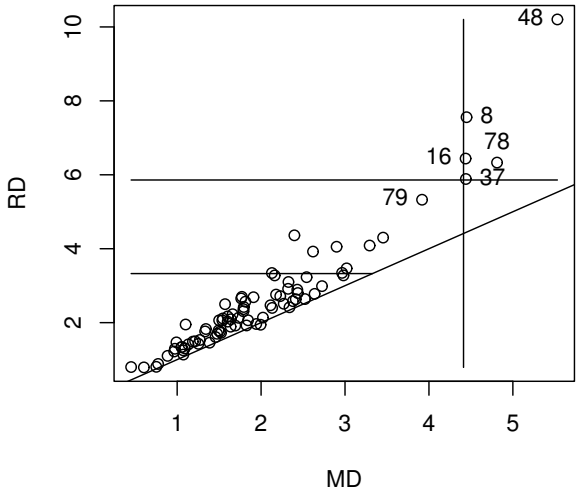

Figure 12.8: DD Plot of the Mussels Data.

For Y_2, cases 8, 25 and 48 are not fit well. A residual vector $\boldsymbol{r} = (\boldsymbol{r} - \boldsymbol{e}) + \boldsymbol{e}$ is a combination of $\boldsymbol{e}$ and a discrepancy $\boldsymbol{r} - \boldsymbol{e}$ that tends to have an approximate multivariate normal distribution. The $\boldsymbol{r} - \boldsymbol{e}$ term can dominate for small to moderate n when $\boldsymbol{e}$ is not multivariate normal, incorrectly suggesting that the distribution of the error $\boldsymbol{e}$ is closer to a multivariate normal distribution than is actually the case. Figure 12.11 shows the DD plot of the residual vectors. The nonparametric prediction region for the residuals consists of the points to the left of the vertical line $MD = 2.27$. Comparing Figure 12.8 and 12.11, the residual distribution is closer to a multivariate normal distribution. Cases 8, 48 and 79 have especially large distances. R code for producing the five figures is shown below.

```
y <- log(mussels)[,4:5]
x <- mussels[,1:3]
x[,2] <- log(x[,2])
z<-cbind(x,y)
pairs(z, labels=c("L","log(W)","H","log(S)","log(M)"))
ddplot4(z)
out <- mltreg(x,y)
ddplot4(out$res)
```

12.5 Testing Hypotheses

This section follows Khattree and Naik (1999, p. 66-67) closely.

Definition 12.15. Assume rank$(\boldsymbol{X}) = p$. The total corrected (for the mean) sum of squares and cross products matrix is

$$\boldsymbol{T} = \boldsymbol{R} + \boldsymbol{W} = \boldsymbol{Z}^T(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T)\boldsymbol{Z}.$$

Note that $\boldsymbol{T}/(n-1)$ is the usual sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}}$ if all n of the $\boldsymbol{y}_i$ are iid so that $\boldsymbol{B} = \boldsymbol{0}$. The regression sum of squares and cross products matrix is

$$\boldsymbol{R} = \boldsymbol{Z}^T[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T]\boldsymbol{Z} = \boldsymbol{Z}^T\boldsymbol{X}\hat{\boldsymbol{B}} - \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{1}\boldsymbol{1}^t\boldsymbol{Z}.$$

The error or residual sum of squares and cross products matrix is

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

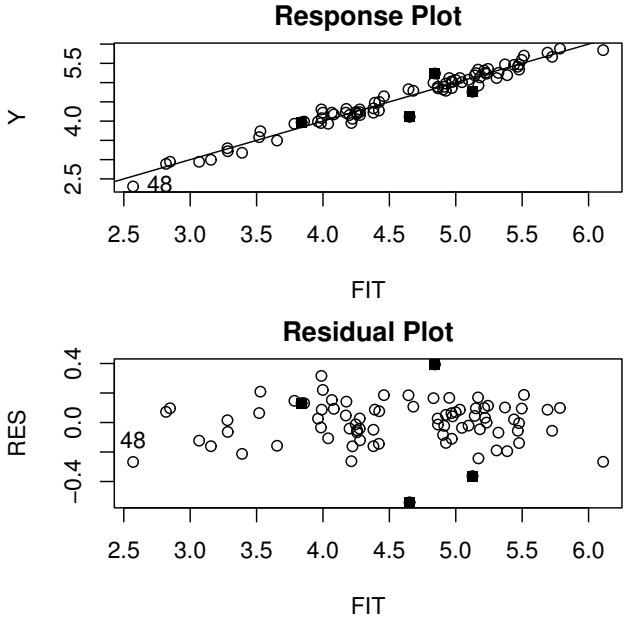

Figure 12.9: Plots for $Y_1 = \log(W)$.

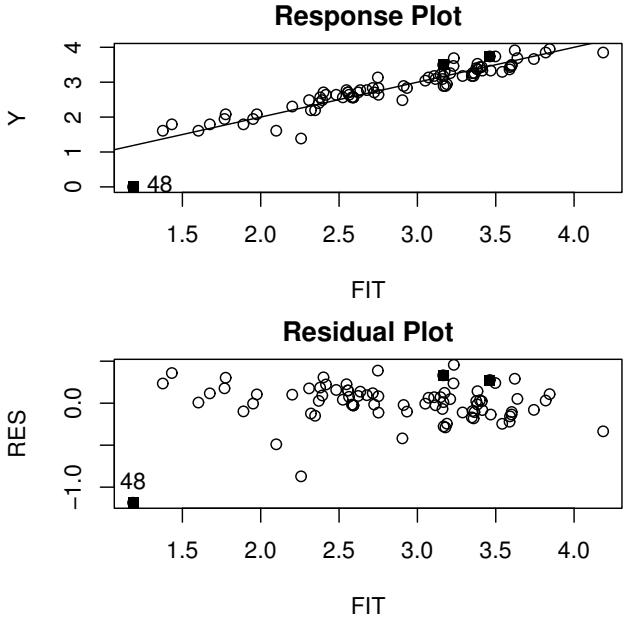

Figure 12.10: Plots for $Y_2 = \log(M)$.

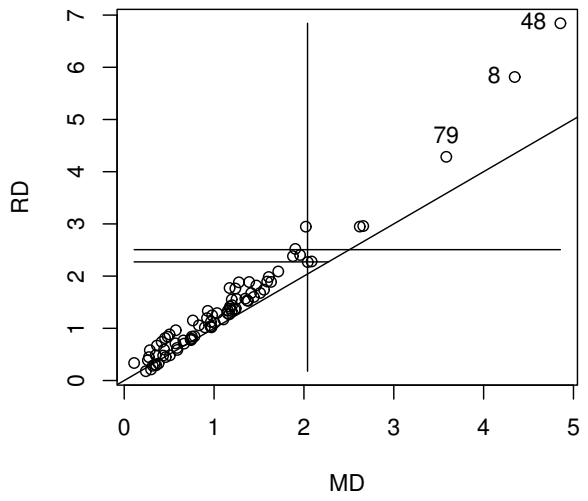

Figure 12.11: DD Plot of the Residual Vectors.

Note that $\boldsymbol{W}_e = \hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}$ and $\boldsymbol{W}_e/(n-p) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$.

Warning: *SAS* output uses $\boldsymbol{E}$ instead of $\boldsymbol{W}_e$.

The MANOVA table is shown below.

Summary MANOVA Table

Source	matrix	df
Regression or Treatment	$\boldsymbol{R}$	$p-1$
Error or Residual	$\boldsymbol{W}_e$	$n-p$
Total (corrected)	$\boldsymbol{T}$	$n-1$

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

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

Typically some function of one of the four above statistics is used to get pval, the estimated pvalue. Output often gives the pvals for all four test statistics. Be cautious about inference if the four test statistics do not lead to the same conclusions. Pillai's trace statistic is thought to be the most robust against nonnormality of the $\boldsymbol{\epsilon}_i$.

The four steps of the MANOVA test of linear hypotheses follow.
i) State the hypotheses $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ and $H_1 : \boldsymbol{LB} \neq \boldsymbol{0}$.
ii) Get test statistic from output.
iii) Get pval from output.
iv) State whether you reject H_0 or fail to reject H_0. If pval $\leq \alpha$, reject H_0 and conclude that $\boldsymbol{LB} \neq \boldsymbol{0}$. If pval $> \alpha$, fail to reject H_0 and conclude that $\boldsymbol{LB} = \boldsymbol{0}$ or that there is not enough evidence to conclude that $\boldsymbol{LB} \neq \boldsymbol{0}$. As a textbook convention, use $\alpha = 0.05$ if α is not given.

The MANOVA test of $H_0 : \boldsymbol{B} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{B} \neq \boldsymbol{0}$ is the special case corresponding to $\boldsymbol{L} = \boldsymbol{I}$ and $\boldsymbol{H} = \hat{\boldsymbol{B}}^T \boldsymbol{X}^T \boldsymbol{X} \hat{\boldsymbol{B}} = \hat{\boldsymbol{Z}}^T \hat{\boldsymbol{Z}}$.

12.6 Justification of the Hotelling Lawley Test

Some notation is needed. Following Henderson and Searle (1979), let matrix $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_p \end{bmatrix}$. 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 $[\boldsymbol{A} \otimes \boldsymbol{B}]^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$.

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. For now assume the error distribution is multivariate normal $N_p(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon})$. Then

$$vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m \end{pmatrix} \sim N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

where

$$\boldsymbol{C} = \boldsymbol{\Sigma_\epsilon} \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} \sigma_{11}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{12}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{1p}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \sigma_{21}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{22}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{2p}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{p1}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{p2}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{pp}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{bmatrix}.$$

Now let $\boldsymbol{A}$ be a $rm \times pm$ block diagonal matrix: $\boldsymbol{A} = diag(\boldsymbol{L}, ..., \boldsymbol{L})$. Then $\boldsymbol{A}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = vec(\boldsymbol{L}(\hat{\boldsymbol{B}} - \boldsymbol{B})) =$

$$\begin{pmatrix} \boldsymbol{L}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \\ \boldsymbol{L}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \\ \vdots \\ \boldsymbol{L}(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m) \end{pmatrix} \sim N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)$$

where $\boldsymbol{D} = \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T = \boldsymbol{ACA}^T =$

$$\begin{bmatrix} \sigma_{11}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{12}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{1p}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \\ \sigma_{21}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{22}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{2p}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{p1}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{p2}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{pp}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \end{bmatrix}.$$

Under H_0, $vec(\boldsymbol{LB}) = \boldsymbol{A}$ $vec(\boldsymbol{B}) = \boldsymbol{0}$, and

$$vec(\boldsymbol{L}\hat{\boldsymbol{B}}) = \begin{pmatrix} \boldsymbol{L}\hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_m \end{pmatrix} \sim N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T).$$

Hence under H_0,

$$[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \sim \chi^2_{rm},$$

and

$$T = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}. \quad (12.15)$$

A large sample level δ test will reject H_0 if $pval < \delta$ where

$$pval = P(\frac{T}{rm} < F_{rm,n-mp}). \quad (12.16)$$

Since least squares estimators are asymptotically normal, for a large class of distributions,

$$\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = \sqrt{n}\ \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m \end{pmatrix} \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{W})$$

where

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{W}^{-1}.$$

Then under H_0,

$$\sqrt{n}\ \ vec(\boldsymbol{L}\hat{\boldsymbol{B}}) = \sqrt{n}\ \begin{pmatrix} \boldsymbol{L}\hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_m \end{pmatrix} \xrightarrow{D} N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{LWL}^T),$$

and

$$n\ \ [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{LWL}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}.$$

Hence (12.15) holds, and (12.16) gives a large sample level δ test if the least squares estimators are asymptotically normal.

Multivariate analogs of tests for multiple linear regression can be derived with appropriate choice of $\boldsymbol{L}$. Using $\boldsymbol{L} = [\boldsymbol{0} \ \boldsymbol{I}_{p-1}]$ tests whether the nontrivial predictors are needed in the multivariate linear regression model, an analog of the Anova 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 prredictors are in the model, an analog of the partial F test. Using $\boldsymbol{L} = (0, ..., 0, 1, 0, ..., 0)$, a row vector with a 1 in the jth position, tests whether the jth variable is needed in the multivariate linear regression model given that the other $p-1$ predictors are in the model, an analog to the t tests for multiple linear regression. This statistic has the form

$$T_j = \frac{1}{d_j}(\hat{\beta}_{j1}, \hat{\beta}_{j2}, ..., \hat{\beta}_{jm})\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \begin{pmatrix} \hat{\beta}_{j1} \\ \hat{\beta}_{j2} \\ \vdots \\ \hat{\beta}_{jm} \end{pmatrix}$$

where $d_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}$, the jth diagonal entry of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. The statistic T_j could be used for forward selection and backward elimination in variable selection.

12.7 Seemingly Unrelated Regressions

Each response variable in a multivariate linear regression model follows a multiple linear regression model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$. Hence the errors corresponding to the jth response are uncorrelated with variance $\sigma_j^2 = \sigma_{jj}$. Notice that **the same design matrix $\boldsymbol{X}$** of predictors is used for each of the m models, but the response variable vector $\boldsymbol{Y}_j$, coefficient vector $\boldsymbol{\beta}_j$ and error vector $\boldsymbol{e}_j$ change and thus depend on j.

The seemingly related regressions (SUR) model differs from the multivariate linear regression model in that each response model follows a multiple linear regression model $\boldsymbol{Y}_j = \boldsymbol{X}_j\boldsymbol{\beta}_j + \boldsymbol{e}_j$ with a different design matrix $\boldsymbol{X}_j$ and the $\boldsymbol{\beta}_j$ are $k_j \times 1$ vectors. Let $\boldsymbol{x}_{i,j} = (1, x_{2,j}, ..., x_{k_j,j})^T$. Then the ith case in the SUR model is $(Y_{i,1}, ..., Y_{i,m}, x_{2,1}, ..., x_{k_1,1}, x_{2,2}, ..., x_{k_2,2}, ..., x_{2,m}, ..., x_{k_m,m})$. That is, string $\boldsymbol{y}_i$ and the $\boldsymbol{x}_{i,j}$ into a vector, omitting the m ones.

The multivariate linear regression model can be regarded as the special case of the SUR model where all of the design matrices are equal $\boldsymbol{X}_j \equiv \boldsymbol{X}$ for $j = 1, ..., m$, and the SUR model can be regarded as a special case of the multivariate linear regression model where the design matrix $\boldsymbol{X}$ has columns corresponding to the constant $1, x_{2,1}, ..., x_{k_m,m}$. Hence if $k = \sum_{i=1}^{m} k_i$, then $\boldsymbol{X}$ is an $n \times (k - m + 1)$ matrix. Then the $(k - m + 1) \times 1$ vector $\boldsymbol{\beta}_j^* = (\beta_{1,j}, 0, ..., 0, \beta_{2,j}, ..., \beta_{k_j,j}, 0, ..., 0)^T$. Here $\boldsymbol{\beta}_j^*$ is the jth column of $\boldsymbol{B}$, and only k_j of the entries of $\boldsymbol{\beta}_j^*$ are nonzero. Hence most of the entries in $\boldsymbol{B}$ are zeroes.

A competitor of the SUR model would be the multivariate linear regression model where there are no restrictions on $\boldsymbol{B}$, so the columns $\boldsymbol{\beta}_j$ of $\boldsymbol{B}$ are estimated using least squares and $\boldsymbol{X}$. The SUR model says that the $Y_{i,1}, ..., Y_{i,m}$ are correlated, but only $\boldsymbol{x}_{i,j}$ is needed in the model for predicting the $Y_{i,j}$ when $\boldsymbol{x}_{i,1}, ..., \boldsymbol{x}_{i,m}$ are possible vectors of predictors. If this assumption is wrong, then the SUR model could be throwing away a lot of information from relevant predictors.

Definition 12.15. In the *seemingly unrelated regressions model*,

$$\boldsymbol{y}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i = \begin{pmatrix} \boldsymbol{x}_{i,1}^T \boldsymbol{\beta}_1 \\ \boldsymbol{x}_{i,2}^T \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{x}_{i,m}^T \boldsymbol{\beta}_m \end{pmatrix} + \begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \\ \vdots \\ \epsilon_{i,m} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_{i,1}^T \hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{x}_{i,2}^T \hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \boldsymbol{x}_{i,m}^T \hat{\boldsymbol{\beta}}_m \end{pmatrix} + \begin{pmatrix} \hat{\epsilon}_{i,1} \\ \hat{\epsilon}_{i,2} \\ \vdots \\ \hat{\epsilon}_{i,m} \end{pmatrix}$$

$= \hat{\boldsymbol{y}}_i + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$, where $\text{Cov}(\boldsymbol{\epsilon}_i) \equiv \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ is $m \times m$ and $E(\boldsymbol{\epsilon}_i) \equiv \boldsymbol{0}$. Here $\boldsymbol{x}_{i,j}$, $\boldsymbol{\beta}_j$ and $\hat{\boldsymbol{\beta}}_j$ are $k_j \times 1$ vectors where $\sum_{j=1}^{m} k_j = k$, and $\boldsymbol{y}_i = (y_{i1}, ..., y_{im})^T$.

There are several ways to estimate the $\hat{\boldsymbol{\beta}}_j$. First, estimate $\hat{\boldsymbol{\beta}}_j$ using least squares on the m multiple linear regression models $\boldsymbol{Y}_j = \boldsymbol{X}_j \boldsymbol{\beta}_j + \boldsymbol{e}_j$. This method should be equivalent to using the multivariate regression model where the $\boldsymbol{\beta}_j^*$ are the columns of $\boldsymbol{B}$ and the nonzero entries of $\hat{\boldsymbol{\beta}}_j^*$ are collected into the $k_j \times 1$ vectors $\hat{\boldsymbol{\beta}}_j$. Another method uses the seemingly unrelated regressions estimator (SURE) which uses the multivariate linear regression estimator as an initial estimator, and then uses generalized least squares. See Press (2005, § 8.5). In the discussion that follows, $\hat{\boldsymbol{\beta}}$ will be the SUR estimator which is thought to be more efficient than the alternatives. See White (1984, p. 166-171) for large sample theory of the SUR estimator.

Model checking and prediction for the SUR model is very similar to that for the multivariate regression model, but use the fitted values and residuals from the SUR model.

1) Make the m response and residual plots, and make the DD plot of the $\hat{\boldsymbol{\epsilon}}_i$.

2) Transformation plots and variable selection can be done using least squares on each of the m multiple linear regression models $\boldsymbol{Y}_j = \boldsymbol{X}_j = \boldsymbol{e}_j$ for $j = 1, ..., m$.

3) Simultaneous prediction intervals using (12.11) and (12.12) can be made using either least squares fits for each of the m models or using the fitted values and residuals from the SUR model.

4) A prediction region for $\boldsymbol{y}_f$ is made as in Section 12.4.3 using $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$ where $\hat{\boldsymbol{y}}_f = (\boldsymbol{x}_{f,1}^T \hat{\boldsymbol{\beta}}_1, ..., \boldsymbol{x}_{f,m}^T \hat{\boldsymbol{\beta}}_m)^T$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ and the $\hat{\boldsymbol{\beta}}_j$ are the SUR estimators.

```
mltreg(x,y,indices=c(3,4))

$partial
      partialF      Pval
[1,] 0.2001622 0.9349877

$Ftable
                Fj      pvals
[1,]   4.35326807 0.02870083
[2,] 600.57002201 0.00000000
[3,]   0.08819810 0.91597268
[4,]   0.06531531 0.93699302

$MANOVA
     MANOVAF         pval
[1,] 295.071 1.110223e-16
```

Example 12.2. The above output is for the Hebbler (1847) data from the the 1843 Prussia census. Sometimes if the wife or husband was not at the household, then s/he would not be counted. Y_1 = number of married civilian men in the district, Y_2 = number of women married to civilians in the district, x_2 = population of the district in 1843, x_3 = number of married

military men in the district, x_4 = number of women married to military men in the district. The reduced model deletes x_3 and x_4.

a) Do the MANOVA F test.

b) Do the F_2 test.

c) Do the F_4 test.

d) Do an appropriate 4 step test for the reduced model that deletes x_3 and x_4.

e) The output for the reduced model that deletes x_1 and x_2 is shown below. Do an appropriate 4 step test.

```
$partial
     partialF Pval
[1,] 569.6429    0
```

12.8 Summary

1) The multivariate linear regression model is a special case of the multivariate linear model where at least one predictor variable X_j is continuous. The MANOVA model is a multivariate linear model where all of the predictors are categorical variables so the X_j are coded and are often indicator variables.

2) 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 $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$. The constant $x_{i1} = 1$ is in the model, and is often omitted from the case and the data matrix. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = ((\sigma_{ij}))$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$. Then $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$.

3) Each response variable in a multivariate linear regression model follows a univariate linear regression model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$.

4) For each variable Y_k make a response plot of $\hat{Y}_{ik}$ versus Y_{ik} and a residual plot of $\hat{Y}_{ik}$ versus $r_{ik} = Y_{ik} - \hat{Y}_{ik}$. If the multivariate linear regression

model is appropriate, then the plotted points 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. If the model is good, then the each of the m residual plots should be ellipsoidal with no trend and should be centered about the $r = 0$ line. There should not be any pattern in the residual plot: as a narrow vertical strip is moved from left to right, the behavior of the residuals within the strip should show little change. Outliers and patterns such as curvature or a fan shaped plot are bad.

5) Make a scatterplot matrix of $Y_1, ..., Y_m$ and of the continuous predictors. Use power transformations to remove strong nonlinearities.

6) Consider testing $\boldsymbol{LB} = \boldsymbol{0}$ where $\boldsymbol{L}$ is a $r \times p$ full rank matrix. Let $\boldsymbol{W}_e = \hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}$ and $\boldsymbol{W}_e/(n-p) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. 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 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the ordered eigenvalues of $\boldsymbol{W}_e^{-1}\boldsymbol{H}$. Then there are four commonly used test statistics.

The 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 $U(\boldsymbol{L}) = tr[\boldsymbol{W}_e^{-1}\boldsymbol{H}] = \sum_{i=1}^{m}\lambda_i = \frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$.

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

7) Under regularity conditions, $-[n-p+1-0.5(m-r+3)]\log(\Lambda(\boldsymbol{L})) \xrightarrow{D} \chi^2_{rm}$,
$(n-p)V(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, $\quad (n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, and if $h = \max(r, m)$,

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

The Hotelling Lawley statistic is robust against nonnormality.

8) For the Wilk's Lambda test,

$$pval = P\left(\frac{-[n-p+1-0.5(m-r+3)]}{rm} \quad \log(\Lambda(\boldsymbol{L})) < F_{rm,n-rm}\right).$$

For the Pillai's trace test, $pval = P\left(\frac{n-p}{rm} \ \ V(\boldsymbol{L}) < F_{rm,n-rm}\right)$.

For the Hotelling Lawley trace test, $pval = P\left(\frac{n-p}{rm} \ \ U(\boldsymbol{L}) < F_{rm,n-rm}\right)$.

The above three tests are large sample tests, P(reject $H_0|H_0$ is true) $\rightarrow \alpha$ as $n \rightarrow \infty$, under regularity conditions.

For the Roy's largest root test, use
$pval = P\left(\frac{n-p-h+r}{h} \ \lambda_{max}(\boldsymbol{L}) < F_{h,n-p-h+r}\right)$.

The F statistic is an upper bound on the F statistic that provides a lower bound on the nominal level of significance, α, under regularity conditions.

9) The 4 step MANOVA F test of hypotheses uses $\boldsymbol{L} = [\boldsymbol{0} \ \ \boldsymbol{I}_{p-1}]$:
i) State the hypotheses H_0: the nontrivial predictors are not needed in the mreg model H_1: at least one of the nontrivial predictors is needed
ii) Find the test statistic F_o from output.
iii) Find the pval from output.
iv) If pval $< \alpha$, reject H_0. If pval $\geq \alpha$, fail to reject H_0. If H_0 is rejected, conclude that there is a mreg relationship between the response variables $Y_1, ..., Y_m$ and the predictors X_2, ..., X_p. If you fail to reject H_0, conclude that there is a not a mreg relationship between $Y_1, ..., Y_m$ and the predictors X_2, ..., X_p. (Get the variable names from the story problem.)

10) The 4 step F_j test of hypotheses uses $\boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$ where the 1 is in the jth position. Let $\boldsymbol{b}_j^T$ be the jth row of $\boldsymbol{B}$. i) State the hypotheses H_0 :
$b\boldsymbol{b}_j^T = \boldsymbol{0} \quad H_1 : \quad \boldsymbol{b}_j^T \neq \boldsymbol{0}$
ii) Find the test statistic F_j from output.
iii) Find pval from output.
iv) If pval $< \alpha$, reject H_0. If pval $\geq \alpha$, fail to reject H_0. Give a nontechnical sentence restating your conclusion in terms of the story problem. If H_0 is rejected, then conclude that X_j is needed in the mreg model for $Y_1, ..., Y_m$ given that the other predictors are in the model. If you fail to reject H_0, then conclude that X_j is not needed in the mreg model for $Y_1, ..., Y_m$ given that the other predictors are in the model. (Get the variable names from the story problem.)

11) The 4 step **MANOVA partial F test** of hypotheses has a full model using all of the variables and a reduced model where r of the variables are deleted. 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.
i) State the hypotheses H_0: the reduced model is good H_1: use the full model.
ii) Find the test statistic F_R from output.
iii) Find the pval from output.
iv) If pval $< \alpha$, reject H_0 and conclude that the full model should be used. If pval $\geq \alpha$, fail to reject H_0 and conclude that the reduced model is good.

12) The 4 step MANOVA F test should reject H_0 if the response and residual plots look good, n is large enough and at least one response plot does not look like the corresponding residual plot. A response plot for Y_j will look like a residual plot if the identity line appears almost horizontal, hence the range of $\hat{Y}_j$ is small.

13) The *mpack* function `mltreg` produces the m response and residual plots, gives $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$, the MANOVA partial F test statistic and pval corresponding to the reduced model that leaves out the variables given by indices (so X_2 and X_4 in the output below with $F = 0.77$ and pval $= 0.614$), F_j and the pval for the F_j test for variables 1, 2, ..., p (where $p = 4$ in the output below so $F_2 = 1.51$ with pval $= 0.284$) and F_0 and pval for the MANOVA F test (in the output below $F_0 = 3.15$ and pval$= 0.06$). The command `out <- mltreg(x,y,indices=c(2))` would produce a MANOVA partial F test corresponding to the F_2 test while the command `out <- mltreg(x,y,indices=c(2,3,4))` would produce a MANOVA partial F test corresponding to the MANOVA F test for a data set with $p = 4$ predictor variables. The Hotelling Lawley trace statistic is used in the tests.

```
out <- mltreg(x,y,indices=c(2,4))

$Bhat
             [,1]         [,2]         [,3]
[1,] 47.96841291 623.2817463 179.8867890
[2,]  0.07884384   0.7276600  -0.5378649
[3,] -1.45584256 -17.3872206   0.2337900
[4,] -0.01895002   0.1393189  -0.3885967

$Covhat
          [,1]      [,2]     [,3]
[1,]  21.91591  123.2557  132.339
[2,] 123.25566 2619.4996 2145.780
```

```
[3,] 132.33902 2145.7797 2954.082

$partial
      partialF      Pval
[1,] 0.7703294 0.6141573

$Ftable
             Fj      pvals
[1,] 6.30355375 0.01677169
[2,] 1.51013090 0.28449166
[3,] 5.61329324 0.02279833
[4,] 0.06482555 0.97701447

$MANOVA
      MANOVAF       pval
[1,] 3.150118 0.06038742
```

14) Given $\hat{\boldsymbol{B}} = [\hat{\boldsymbol{\beta}}_1 \;\; \hat{\boldsymbol{\beta}}_2 \;\; \cdots \;\; \hat{\boldsymbol{\beta}}_m]$ and $\boldsymbol{x}_f$, find $\hat{\boldsymbol{y}}_f = (\hat{y}_1, ..., \hat{y}_m)^T$ where $\hat{y}_i = \hat{\boldsymbol{\beta}}_i^T \boldsymbol{x}_f$.

15) $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \dfrac{\hat{\boldsymbol{E}}^T \hat{\boldsymbol{E}}}{n-p} = \dfrac{1}{n-p} \sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^T$ while the sample covariance matrix of the residuals is $\boldsymbol{S}_r = \dfrac{n-p}{n-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \dfrac{\hat{\boldsymbol{E}}^T \hat{\boldsymbol{E}}}{n-1}$. Both $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ and $\boldsymbol{S}_r$ are $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ for a large class of error distributions for $\boldsymbol{\epsilon}_i$.

16) The $100(1-\alpha)\%$ nonparametric prediction region for $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ is the nonparametric prediction region from $\oint$ 5.2 applied to $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i = \hat{\boldsymbol{B}}^T \boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$. This takes the data cloud of the n residual vectors $\hat{\boldsymbol{\epsilon}}_i$ and centers the cloud at $\hat{\boldsymbol{y}}_f$. Let

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

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

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

If $q_n < 1 - \alpha + 0.001$, set $q_n = 1 - \alpha$. Let $0 < \alpha < 1$ and $h = D_{(U_n)}$ where $D_{(U_n)}$ is the q_nth sample quantile of the D_i. The $100(1-\alpha)\%$ nonparametric

prediction region for $\boldsymbol{y}_f$ is

$$\{\boldsymbol{z} : (\boldsymbol{z} - \hat{\boldsymbol{y}}_f)^T \boldsymbol{S}_r^{-1} (\boldsymbol{z} - \hat{\boldsymbol{y}}_f) \leq D^2_{(U_n)}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}\}.$$

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 $U_n/n \to 1 - \alpha$ as $n \to \infty$.

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

c) If $(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ 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 $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\alpha}\}$, then the nonparametric prediction region is asymptotically optimal.

17) On the DD plot for the residuals, the cases to the left of the vertical line correspond to cases that would have $\boldsymbol{y}_f = \boldsymbol{y}_i$ in the nonparametric prediction region if $\boldsymbol{x}_f = \boldsymbol{x}_i$ while the cases to the right of the line would not have $\boldsymbol{y}_f = \boldsymbol{y}_i$ in the nonparametric prediction region.

18) The DD plot for the residuals is interpreted almost exactly as a DD plot for iid multivariate data is interpreted. Plotted points clustering about the identity line suggests that the $\boldsymbol{\epsilon}_i$ may be iid from a multivariate normal distribution while plotted points that lie above the identity line but cluster about a line through the origin with slope greater than 1 suggests that the $\boldsymbol{\epsilon}_i$ may be iid from an elliptically contoured distribution that is not MVN. The semiparametric and parametric MVN prediction regions correspond to horizontal lines on the DD plot. Robust distances have not been shown to be consistent estimators of the population distances, but are useful for a graphical diagnostic.

19) A robust multivariate linear regression method replaces least squares with the hbreg estimator. The probability that the robust estimator equals the least squares estimator goes to 1 as $n \to \infty$ for a large class of error distributions. Hence the hypothesis tests and nonparametric prediction regions for the classical method can be applied to the robust method. The entries of $\hat{\boldsymbol{B}}$ are hard to drive to $\pm\infty$ for the robust estimator, and the residuals corresponding to outliers are often large. Since the residuals are used to compute $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$, the tests of hypothesis based on the robust estimator are not robust to the presence of outliers. But the robust estimator and classical estimator

tend to give different response and residual plots and test statistics when outliers are present.

12.9 Complements

The least squares estimator $\hat{\boldsymbol{\beta}}$ is a good estimator of $\boldsymbol{\beta}$ under very mild conditions by Theorem 12.3; however, Theorem 12.3 assumes that the model is known before gathering data. If variable selection and response transformation are performed to build a model, then the estimators are biased and results for inference fail to hold in that pvalues and coverage of confidence and prediction intervals will be wrong. See, for example, Berk (1978), Copas (1983), Miller (1984) and Rencher and Pun (1980). Hence it is a good idea to do a pilot study to suggest which transformations and variables to use. Then do a larger study without using variable selection and response transformations.

Cook and Olive (2001) and Olive (2004b, 2013) discuss response plots and transformation plots. Cook and Setodji (2003) use the FF plot while Wilcox (2009) has a robust method for multivariate regression. Su and Cook (2012) give an interesting alternative to least squares. Prediction regions for this method could be made following Section 12.4.3.

Khattree and Naik (1999, p. 91-98) discuss testing $H_0 : \boldsymbol{LBM} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{LBM} \neq \boldsymbol{0}$ where $\boldsymbol{M} = \boldsymbol{I}$ gives a linear test of hypotheses.

12.10 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

12.1*. Refer to the alternative form of the Hotelling Lawley test statistic. Let

$$T(\boldsymbol{W}) = n \ \ [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{LWL}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})].$$

Let

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} = \hat{\boldsymbol{W}}^{-1}.$$

Show $T(\hat{\boldsymbol{W}}) = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$.

12.2. Refer to the alternative form of the Hotelling Lawley test statistic. Let $T = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$. Let $\boldsymbol{L} = \boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$ have a 1 in the jth position. Let $\hat{\boldsymbol{b}}_j^T = \boldsymbol{L}\hat{\boldsymbol{B}}$ be the jth row of $\hat{\boldsymbol{B}}$. Let $d_j = \boldsymbol{L}_j(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}_j^T = (\boldsymbol{X}^T\boldsymbol{X})_{jj}^{-1}$, the jth diagonal entry of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Then $T_j = \frac{1}{d_j}\hat{\boldsymbol{b}}_j^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{b}}_j$. The Hotelling Lawley statistic $U = tr([(n-p)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}]^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}])$. Hence if $\boldsymbol{L} = \boldsymbol{L}_j$, then $U_j = \frac{1}{d_j(n-p)}tr(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{b}}_j\hat{\boldsymbol{b}}_j^T)$.

Using $tr(\boldsymbol{ABC}) = tr(\boldsymbol{CAB})$ and $tr(a) = a$ for scalar a, show the $(n-p)U_j = T_j$.

12.3. Refer to the alternative form of the Hotelling Lawley test statistic. Using the Searle (1982, p. 333) identity
$tr(\boldsymbol{AG}^T\boldsymbol{DGC}) = [vec(\boldsymbol{G})]^T[\boldsymbol{CA} \otimes \boldsymbol{D}^T][vec(\boldsymbol{G})]$, show
$(n-p)U(\boldsymbol{L}) = tr[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}]$
$= [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$ by identifying $\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{D}$, and $\boldsymbol{C}$.

```
$Ftable
            Fj        pvals
[1,] 82.147221 0.000000e+00
[2,] 58.448961 0.000000e+00
[3,] 15.700326 4.258563e-09
[4,]  9.072358 1.281220e-05
[5,] 45.364862 0.000000e+00

$MANOVA
      MANOVAF pval
[1,] 67.80145    0
```

12.4. The above output is for the R Seatbelts data set where $Y_1 = drivers$ = number of drivers killed or seriously injured, $Y_2 = front$ = number of front seat passengers killed or seriously injured, and $Y_3 = back$ = number of back seat passengers killed or seriously injured. The predictors were $x_2 = kms$ = distance driven, $x_3 = price$ = petrol price, $x_4 = van$ = number of van drivers killed, and $x_5 = law = 0$ if the law was in effect that month and 1 otherwise. The data consists of 192 monthly totals in Great Britain from

January 1969 to December 1984, and the compulsory wearing of seat belts law was introduced in February 1983.

a) Do the MANOVA F test.

b) Do the F_4 test.

12.5. a) Sketch a DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ for the multivariate linear regression model if the error vectors $\boldsymbol{\epsilon}_i$ are iid from a multivariate normal distribution. b) Does the DD plot change if the one way MANOVA model is used instead of the multivariate linear regression model?

```
y<-USJudgeRatings[,c(9,10,12)]
x<-USJudgeRatings[,-c(9,10,12)]
mltreg(x,y,indices=c(2,5,6,7,8))
$partial
     partialF      Pval
[1,] 1.649415 0.1855314

$MANOVA
      MANOVAF         pval
[1,] 340.1018 1.121325e-14
```

12.6. The above output is for the R judge ratings data set consisting of lawyer ratings for $n = 43$ judges. $Y_1 = oral$ = sound oral rulings, $Y_2 = writ$ = sound written rulings, and $Y_3 = rten$ = worthy of retention. The predictors were $x_2 = cont$ = number of contacts of lawyer with judge, $x_3 = intg$ = judicial integrity, $x_4 = dmnr$ = demeanor, $x_5 = dilg$ = diligence, $x_6 = cfmg$ = case flow managing, $x_7 = deci$ = prompt decisions, $x_8 = prep$ = preparation for trial, $x_9 = fami$ = familiarity with law, and $x_{10} = phys$ = physical ability.

a) Do the MANOVA F test.

b) Do the MANOVA partial F test for the reduced model that deletes x_2, x_5, x_6, x_7 and x_8.

12.7. Let $\boldsymbol{\beta}_i$ be $p \times 1$ and suppose

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \end{pmatrix} \sim N_{2p}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} \sigma_{11}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{12}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \sigma_{21}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{22}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{bmatrix} \right).$$

Find the distribution of

$$[\boldsymbol{L} \;\; \mathbf{0}] \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \end{pmatrix} = \boldsymbol{L}\hat{\boldsymbol{\beta}}_1$$

where $\boldsymbol{L}\boldsymbol{\beta}_1 = \mathbf{0}$ and $\boldsymbol{L}$ is $r \times p$ with $r \leq p$. Simplify.

R/Splus Problems

Warning: Use the command *source("G:/mpack.txt")* **to download the programs. See Preface or Section 15.2.** Typing the name of the `mpack` function, eg *ddplot*, will display the code for the function. Use the `args` command, eg *args(ddplot)*, to display the needed arguments for the function.

12.8. This problem examines multivariate linear regression on the Cook and Weisberg (1999a) mussels data with $Y_1 = \log(S)$ 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.

a) The R command for this part make the response and residual plots for each of the three variables. Click the rightmost mouse button and highlight *Stop* to advance the plot. When you have the response and residual plots for one variable on the screen, copy and paste the two plots into *Word*. Do this two times, once for each response variable. The plotted points fall in roughly evenly populated bands about the identity or $r = 0$ line.

b) Copy and paste the output produced from the R command for this part from \$partial on. This gives the output needed to do the MANOVA F test, MANOVA partial F test and the F_j tests.

c) The R command for this plot makes a DD plot of the residuals and adds the lines corresponding to the three prediction regions of Section 5.2. The robust cutoff is larger than the semiparametric cutoff. Place the plot in *Word*. Do the residuals appear to follow a multivariate normal distribution?

d) Do the MANOVA partial F test where the reduced model deletes X_3 and X_4.

e) Do the F_2 test.

f) Do the MANOVA F test.

12.9. This problem examines multivariate linear regression on SAS Institute (1985, p. 146) Fitness Club Data data with $Y_1 = chinups$, $Y_2 = situps$ and $Y_3 = jumps$. The predictors are $X_2 = weight$, $X_3 = waist$ and $X_4 = pulse$.

a) The R command for this part make the response and residual plots for each of the three variables. Click the rightmost mouse button and highlight *Stop* to advance the plot. When you have the response and residual plots for one variable on the screen, copy and paste the three plots into *Word*. Do this three times, once for each response variable. Are there any outliers?

b) The R command for this plot makes a DD plot of the residuals and adds the lines corresponding to the three prediction regions of Section 5.2. The robust cutoff is larger than the semiparametric cutoff. Place the plot in *Word*. Are there any outliers?

12.6. This problem uses the *mpack* function `mregsim` to simulate the Wilk's Lambda test, Pillai's trace test, Hotelling Lawley trace test, and Roy's largest root test for the F_j tests and the MANOVA F test for multivariate linear regression. When `mnull = T` the first row of $\boldsymbol{B}$ is $\mathbf{1}^T$ while the remaining rows are equal to $\mathbf{0}$. Hence the null hypothesis for the MANOVA F test is true. When `mnull = F` the null hypothesis is true for $p = 2$, but false for $p > 2$. Now the first row of $\boldsymbol{B}$ is $\mathbf{1}^T$ and the last row of $\boldsymbol{B}$ is $\mathbf{0}$. If $p > 2$, 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 $\boldsymbol{z}_i$ are generated such that the m errors are iid with variance σ^2. Then $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{z}_i$ so that $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T = ((\sigma_{ij}))$ where the diagonal entries $\sigma_{ii} = \sigma^2[1 + (m-1)\rho^2]$ and the off diagonal entries $\sigma_{ij} = \sigma^2[2\rho + (m-2)\rho^2]$ where $\rho = 0.10$. Terms like *Wilkcov* give the percentage of times the Wilk's test rejected the $F_1, F_2, ..., F_p$ tests. The \$mancv wcv pcv hlcv rcv fcv output gives the percentage of times the 4 test statistics reject the MANOVA F test. Here hlcov and fcov both correspond to the Hotelling Lawley test using the formulas in problem A).

5000 runs will be used so the simulation will take several minutes. Sample sizes $n = 10\min(m, p), n = 10\max(m, p)$ and $n = 10mp$ were interesting. Want coverage near 0.05 when H_0 is true and coverage close to 1 for good power when H_0 is false. Multivariate normal errors were used in a) and b) below.

a) Copy the coverage parts of the output produced by the R commands for this part. Used $n = 20, m = 2, p = 4$. Here H_0 is true except for the F_1 test. Wilk's and Pillai's tests had low coverage < 0.05 when H_0 was false. Roy's test was good for the F_j tests but why was Roy's test bad for the MANOVA F test?

b) Copy the coverage parts of the output produced by the R commands

for this part. Used $n = 20, m = 2, p = 4$. Here H_0 is false except for the F_4 test. Which two tests seem to be the best for this part?

12.11 This problem uses the *mpack* function `mpredsim` to simulate the prediction regions for $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ for multivariate regression. With 5000 runs this simulation takes several minutes. The R command for this problem generate iid lognormal errors then subtract the mean producing $\boldsymbol{z}_i$. Then the $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{z}_i$ are generated as in problem D). Used n=100, m=2, and p=4. The nominal coverage of the prediction region is 90%, and 92% of the training data is covered. The ncvr output gives the coverage of the nonparametric region. What was ncvr?