Chapter 1
Introduction

This chapter provides a preview of the book, and some techniques useful for visualizing data in the background of the data are given in Section 1.2. Sections 1.3 and 1.7 review the multivariate normal distribution and multiple linear regression. Section 1.4 suggests methods for outlier detection. Some large sample theory is presented in Section 1.5, and Section 1.6 covers mixture distributions.

1.1 Overview

Statistical Learning could be defined as the statistical analysis of multivariate data. Machine learning, data mining, analytics, business analytics, data analytics, and predictive analytics are synonymous terms. The techniques are useful for Data Science and Statistics, the science of extracting information from data. The *R* software will be used. See R Core Team (2020).

Let $\boldsymbol{z} = (z_1, ..., z_k)^T$ where $z_1, ..., z_k$ are k random variables. Often $\boldsymbol{z} = (Y, \boldsymbol{x}^T)^T$ where $\boldsymbol{x}^T = (x_1, ..., x_p)$ is the vector of predictors and Y is the variable of interest, called a response variable. Predictor variables are also called independent variables, covariates, or features. The response variable is also called the dependent variable. Usually context will be used to decide whether $\boldsymbol{z}$ is a random vector or the observed random vector.

Definition 1.1. A **case** or **observation** consists of k random variables measured for one person or thing. The ith case $\boldsymbol{z}_i = (z_{i1}, ..., z_{ik})^T$. The **training data** consists of $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$. A statistical model or method is fit (trained) on the training data. The **test data** consists of $\boldsymbol{z}_{n+1}, ..., \boldsymbol{z}_{n+m}$, and the test data is often used to evaluate the quality of the fitted model.

Following James et al. (2013, p. 30), the previously unseen test data is not used to train the Statistical Learning method, but interest is in how well the

method performs on the test data. If the training data is $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_n, Y_n)$, and the previously unseen test data is $(\boldsymbol{x}_f, Y_f)$, then particular interest is in the accuracy of the estimator $\hat{Y}_f$ of Y_f obtained when the Statistical Learning method is applied to the predictor $\boldsymbol{x}_f$. The two Pelawa Watagoda and Olive (2021b) prediction intervals, developed in Section 2.2, will be tools for evaluating Statistical Learning methods for the additive error regression model $Y_i = m(\boldsymbol{x}_i) + e_i = E(Y_i|\boldsymbol{x}_i) + e_i$ for $i = 1, ..., n$ where $E(W)$ is the expected value of the random variable W. The multiple linear regression (MLR) model, $Y_i = \beta_1 + x_2\beta_2 + \cdots + x_p\beta_p + e = \boldsymbol{x}^T\boldsymbol{\beta} + e$, is an important special case. Olive, Rathnayake, and Haile (2022) give prediction intervals for parametric regression models such as generalized linear models (GLMs), generalized additive models (GAMs), and some survival regression models.

The estimator $\hat{Y}_f$ is a *prediction* if the response variable Y_f is continuous, as occurs in regression models. If Y_f is categorical, then $\hat{Y}_f$ is a *classification*. For example, if Y_f can be 0 or 1, then $\boldsymbol{x}_f$ is classified to belong to group i if $\hat{Y}_f = i$ for $i = 0$ or 1.

Following Marden (2006, pp. 5,6), the focus of *supervised learning* is predicting a future value of the response variable Y_f given $\boldsymbol{x}_f$ and the training data $(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_1)$. Hence the focus is not on hypothesis testing, confidence intervals, parameter estimation, or which model fits best, although these four inference topics can be useful for better prediction. The focus of *unsupervised learning* is to group $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ into clusters. *Data mining* is looking for relationships in large data sets.

Notation: Typically lower case boldface letters such as $\boldsymbol{x}$ denote column vectors, while upper case boldface letters such as $\boldsymbol{S}$ or $\boldsymbol{Y}$ are used for matrices or column vectors. If context is not enough to determine whether $\boldsymbol{y}$ is a random vector or an observed random vector, then $\boldsymbol{Y} = (Y_1, ..., Y_p)^T$ may be used for the random vector, and $\boldsymbol{y} = (y_1, ..., y_p)^T$ for the observed value of the random vector. An upper case letter such as Y will usually be a random variable. A lower case letter such as x_1 will also often be a random variable. An exception to this notation is the generic multivariate location and dispersion estimator $(T, \boldsymbol{C})$ where the location estimator T is a $p \times 1$ vector such as $T = \overline{\boldsymbol{x}}$. $\boldsymbol{C}$ is a $p \times p$ dispersion estimator and conforms to the above notation.

The main focus of the first three chapters is developing tools to analyze the multiple linear regression (MLR) model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$. Classical regression techniques use (ordinary) least squares (OLS) and assume $n >> p$, but Statistical Learning methods often give useful results if $p >> n$. OLS forward selection, lasso, ridge regression, marginal maximum likelihood (MMLE), one component partial least squares (OPLS), the elastic net, partial least squares (PLS), and principal component regression (PCR) will be some of the techniques examined. See Chapter 3.

Chapter 2 develops prediction regions and inference after variable selection. Prediction intervals are a special case of prediction regions, and applying the large sample nonparametric prediction region on the bootstrap sample results in a bootstrap confidence region. These tools will be useful for inference when n/p is large. Prediction intervals are developed that can be useful even if $p \geq n$.

For classical regression and multivariate analysis, we often want $n \geq 10p$, and a model with $n < 5p$ is overfitting: the model does not have enough data to estimate parameters accurately. Statistical Learning methods often use a model with a complexity measure d, where $n \geq Jd$ with $J \geq 5$ and preferably $J \geq 10$. For several regression models with lasso, d is the number of variables with nonzero lasso coefficients.

Acronyms are widely used in regression and Statistical Learning, and some of the more important acronyms appear in Table 1.1. Also see the text's index.

Remark 1.1. There are several important Statistical Learning principles.
1) There is more interest in prediction or classification, e.g. producing $\hat{Y}_f$, than in other types of inference such as parameter estimation, hypothesis testing, confidence intervals, or which model fits best.
2) Often the focus is on extracting useful information for *high dimensional statistics* where n/p is not large, e.g. $p > n$. If d is a complexity measure for the fitted model, we want n/d large. A *sparse model* has few nonzero coefficients. We can have sparse population models and sparse fitted models. Sometimes sparse fitted models are useful even if the population model is not sparse. Often the number of nonzero coefficients of a *sparse fitted model* $= d$. Sparse fitted models are often useful for prediction.
3) Interest is in how well the method performs on test data. Performance on training data is overly optimistic for estimating performance on test data.
4) Some methods are *flexible* while others are *unflexible*. For unflexible regression methods, the sufficient predictor is often a hyperplane $SP = \boldsymbol{x}^T\boldsymbol{\beta}$ (see Definition 1.2), and often the mean function $E(Y|\boldsymbol{x}) = M(\boldsymbol{x}^T\boldsymbol{\beta})$ where the function M is known but the $p \times 1$ vector of parameters $\boldsymbol{\beta}$ is unknown and must be estimated (GLMs). Flexible methods tend to be useful for more complicated regression methods where $E(Y|\boldsymbol{x}) = m(\boldsymbol{x})$ for an unknown function m or $SP \neq \boldsymbol{x}^T\boldsymbol{\beta}$ (GAMs). Flexibility tends to increase with d. See Chapter 4, Table 1.1, and Definition 1.2 for GLMs and GAMs.

1.2 Response Plots and Response Transformations

This section will consider tools for visualizing the regression model in the background of the data. The definitions in this section tend not to depend

Table 1.1 Acronyms

Acronym	Description
AER	additive error regression
AP	additive predictor = SP for a GAM
cdf	cumulative distribution function
cf	characteristic function
CI	confidence interval
CLT	central limit theorem
CV	cross validation
DA	discriminant analysis
EC	elliptically contoured
EAP	estimated additive predictor = ESP for a GAM
ESP	estimated sufficient predictor
ESSP	estimated sufficient summary plot = response plot
FDA	Fisher's discriminant analysis
GAM	generalized additive model
GLM	generalized linear model
iid	independent and identically distributed
KNN	K–nearest neighbors discriminant analysis
lasso	an MLR method
LDA	linear discriminant analysis
LR	logistic regression
MAD	the median absolute deviation
MCLT	multivariate central limit theorem
MED	the median
mgf	moment generating function
MLD	multivariate location and dispersion
MLR	multiple linear regression
MMLE	marginal maximum likelihood
MVN	multivariate normal
OLS	ordinary least squares
OPLS	one component partial least squares
PCA	principal component analysis
PCR	principal component(s) regression
PLS	partial least squares
pdf	probability density function
PI	prediction interval
pmf	probability mass function
QDA	quadratic discriminant analysis
SE	standard error
SP	sufficient predictor
SSP	sufficient summary plot
SVM	support vector machine

on whether n/p is large or small, but the estimator $\hat{h}$ tends to be better if n/p is large. In regression, the response variable is the variable of interest: the variable you want to predict. The predictors or features $x_1, ..., x_p$ are variables used to predict Y.

Definition 1.2. *Regression* investigates how the response variable Y changes with the value of a $p \times 1$ vector $\boldsymbol{x}$ of predictors. Often this *conditional distribution* $Y|\boldsymbol{x}$ is described by a *1D regression model*, where Y is conditionally independent of $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$, written

$$Y \perp\!\!\!\perp \boldsymbol{x}|SP \text{ or } Y \perp\!\!\!\perp \boldsymbol{x}|h(\boldsymbol{x}), \tag{1.1}$$

where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. The *estimated sufficient predictor* $\text{ESP} = \hat{h}(\boldsymbol{x})$. An important special case is a model with a linear predictor $h(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ where $\text{ESP} = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ and often $\alpha = 0$. This class of models includes the *generalized linear model* (GLM). Another important special case is a *generalized additive model* (GAM), where Y is independent of $\boldsymbol{x} = (x_1, ..., x_p)^T$ given the *additive predictor* $AP = SP = \alpha + \sum_{j=1}^{p} S_j(x_j)$ for some (usually unknown) functions S_j. The *estimated additive predictor* $\text{EAP} = \text{ESP} = \hat{\alpha} + \sum_{j=1}^{p} \hat{S}_j(x_j)$.

Notation. Often the index i will be suppressed. For example, the *multiple linear regression model*

$$Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{1.2}$$

for $i = 1, ..., n$ where $\boldsymbol{\beta}$ is a $p \times 1$ unknown vector of parameters, and e_i is a random error. This model could be written $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$. More accurately, $Y|\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta} + e$, but the conditioning on $\boldsymbol{x}$ will often be suppressed. Often the errors $e_1, ..., e_n$ are **iid** (independent and identically distributed) from a distribution that is known except for a scale parameter. For example, the e_i's might be iid from a normal (Gaussian) distribution with *mean* 0 and unknown *standard deviation* σ. For this Gaussian model, estimation of $\boldsymbol{\beta}$ and σ is important for inference and for predicting a new future value of the response variable Y_f given a new vector of predictors $\boldsymbol{x}_f$.

1.2.1 Response and Residual Plots

Definition 1.3. An *estimated sufficient summary plot* (ESSP) or **response plot** is a plot of the ESP versus Y. A *residual plot* is a plot of the ESP versus the residuals.

Notation: In this text, a plot of x versus Y will have x on the horizontal axis, and Y on the vertical axis. For the *additive error regression* model $Y = m(\boldsymbol{x}) + e$, the ith residual is $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) = Y_i - \hat{Y}_i$ where $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i)$

is the ith fitted value. The additive error regression model is a 1D regression model with sufficient predictor $SP = h(\boldsymbol{x}) = m(\boldsymbol{x})$.

For the additive error regression model, the response plot is a plot of $\hat{Y}$ versus Y where the *identity line* with unit slope and zero intercept is added as a visual aid. The residual plot is a plot of $\hat{Y}$ versus r. Assume the errors e_i are iid from a unimodal distribution that is not highly skewed. Then the plotted points should scatter about the identity line and the $r = 0$ line (the horizontal axis) with no other pattern if the fitted model (that produces $\hat{m}(\boldsymbol{x})$) is good.

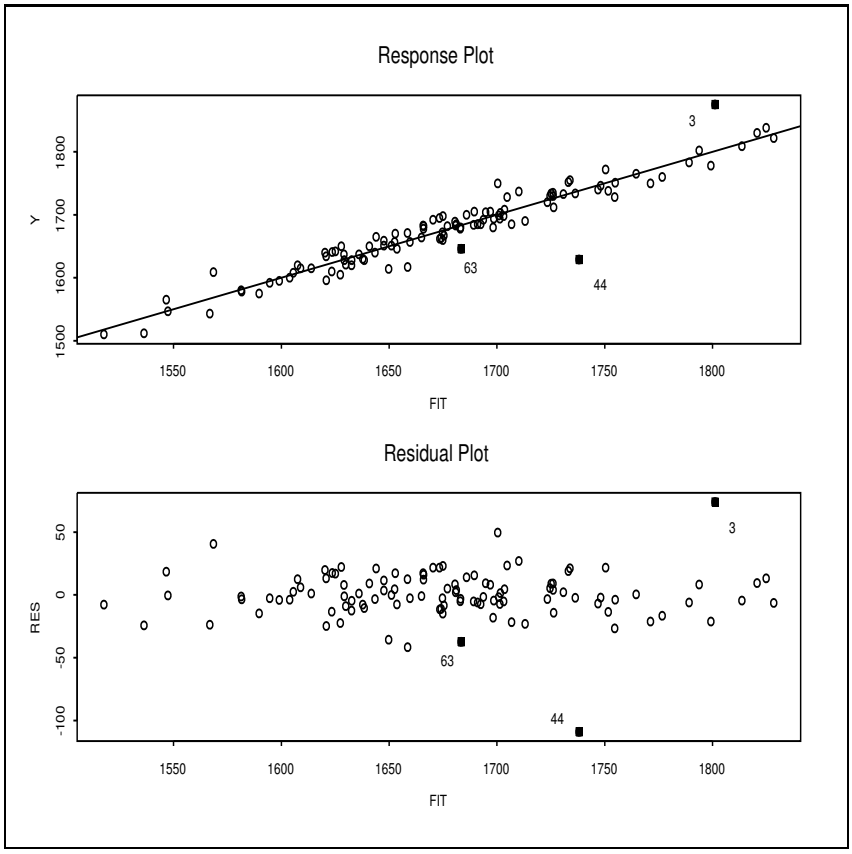

Fig. 1.1 Residual and Response Plots for the Tremearne Data

Example 1.1. Tremearne (1911) presents a data set of about 17 measurements on 115 people of Hausa nationality. We deleted 3 cases because of missing values and used *height* as the response variable Y. 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 1.1 presents the (ordinary) least squares (OLS) response and residual plots for this data set. These plots show that an MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ should be a useful model 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). Note that many important acronyms, such as OLS and MLR, appear in Table 1.1.

To use the response plot to visualize the conditional distribution of $Y|\boldsymbol{x}^T\boldsymbol{\beta}$, use the fact that the fitted values $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. 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 1685 to 1715. 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*: cases that lie far away from the bulk of the data. 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 plot and about the horizontal 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 1.1, the standard deviation of the residuals appears to be around 10. Hence cases 3 and 44 are certainly worth examining.

The identity line can also pass through or near an outlier or a cluster of outliers. Then the outliers will be in the upper right or lower left of the response plot, and there will be a large gap between the cluster of outliers and the bulk of the data. Figure 1.1 was made with the following *R* commands, using *slpack* function `MLRplot` and the *major.lsp* data set from the text's webpage.

```
major <- matrix(scan(),nrow=112,ncol=7,byrow=T)
#copy and paste the data set, then press enter
major <- major[,-1]
X<-major[,-6]
Y <- major[,6]
MLRplot(X,Y) #left click the 3 highlighted cases,
#then right click Stop for each of the two plots
```

A problem with response and residual plots is that there can be a lot of black in the plot if the sample size n is large (more than a few thousand). A variant of the response plot for the additive error regression model would plot the identity line, the two lines parallel to the identity line corresponding to the Section 2.2 large sample $100(1 - \delta)\%$ prediction intervals for Y_f that

depends on $\hat{Y}_f$. Then plot points corresponding to training data cases that do not lie in their $100(1-\delta)\%$ PI. Use $\delta = 0.01$ or 0.05. Try the following commands that used $\delta = 0.2$ since n is small. The commands use the *slpack* function `AERplot`. See Problem 1.10.

```
out<-lsfit(X,Y)
res<-out$res
yhat<-Y-res
AERplot(yhat,Y,res=res,d=2,alph=1) #usual response plot
AERplot(yhat,Y,res=res,d=2,alph=0.2)
#plots data outside the 80% pointwise PIs

n<-100000; q<-7
b <- 0 * 1:q + 1
x <- matrix(rnorm(n * q), nrow = n, ncol = q)
y <- 1 + x %*% b + rnorm(n)
out<-lsfit(x,y)
res<-out$res
yhat<-y-res
dd<-length(out$coef)
AERplot(yhat,y,res=res,d=dd,alph=1) #usual response plot
AERplot(yhat,y,res=res,d=dd,alph=0.01)
#plots data outside the 99% pointwise PIs
AERplot2(yhat,y,res=res,d=2)
#response plot with 90% pointwise prediction bands
```

1.2.2 Response Transformations

A response transformation $Y = t_\lambda(Z)$ can make the MLR model or additive error regression model hold if the variable of interest Z is measured on the wrong scale. For MLR, $Y = t_\lambda(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e$, while for additive error regression, $Y = t_\lambda(Z) = m(\boldsymbol{x}) + e$. Predictor transformations are used to remove gross nonlinearities in the predictors, and this technique is often very useful. However, if there are hundreds or more predictors, graphical methods for predictor transformations take too long. Olive (2017a, Section 3.1) describes graphical methods for predictor transformations.

Power transformations are particularly effective, and a power transformation has the form $x = t_\lambda(w) = w^\lambda$ for $\lambda \neq 0$ and $x = t_0(w) = \log(w)$ for $\lambda = 0$. Often $\lambda \in \Lambda_L$ where

$$\Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\} \tag{1.3}$$

is called the *ladder of powers*. Often when a power transformation is needed, a transformation that goes "down the ladder," e.g. from $\lambda = 1$ to $\lambda = 0$ will

be useful. If the transformation goes too far down the ladder, e.g. if $\lambda = 0$ is selected when $\lambda = 1/2$ is needed, then it will be necessary to go back "up the ladder." Additional powers such as ± 2 and ± 3 can always be added. The following rules are useful for both response transformations and predictor transformations.

a) The **log rule** states that a positive variable that has the ratio between the largest and smallest values greater than ten should be transformed to logs. So $W > 0$ and $\max(W)/\min(W) > 10$ suggests using $\log(W)$.

b) The **ladder rule** appears in Cook and Weisberg (1999a, p. 86), and is used for a plot of two variables, such as ESP versus Y for response transformations or x_1 versus x_2 for predictor transformations.
Ladder rule: To spread *small* values of a variable, make λ *smaller*.
To spread *large* values of a variable, make λ *larger*.

Consider the ladder of powers. Often no transformation ($\lambda = 1$) is best, then the log transformation, then the square root transformation, then the reciprocal transformation.

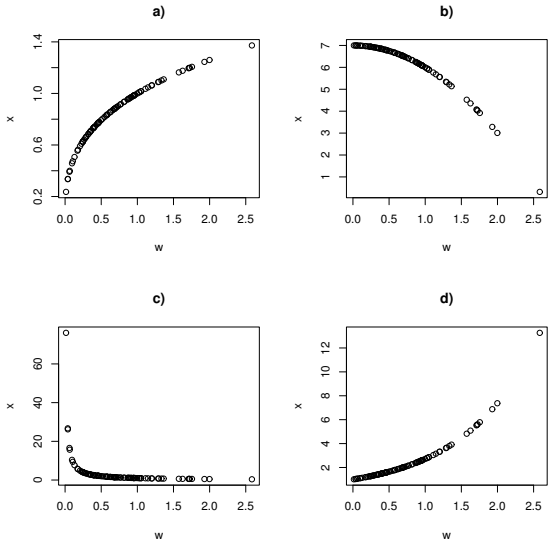

Fig. 1.2 Plots to Illustrate the Ladder Rule

Example 1.2. Examine Figure 1.2. Since w is on the horizontal axis, mentally add a narrow vertical slice to the plot. If a large amount of data falls in the slice at the left of the plot, then small values need spreading. Similarly, if a large amount of data falls in the slice at the right of the plot (compared to the middle and left of the plot), then large values need spreading. For the variable on the vertical axis, make a narrow horizontal slice. If the plot

looks roughly like the northwest corner of a square then small values of the horizontal and large values of the vertical variable need spreading. Hence in Figure 1.2a, small values of w need spreading. If the plot looks roughly like the northeast corner of a square, then large values of both variables need spreading. Hence in Figure 1.2b, large values of x need spreading. If the plot looks roughly like the southwest corner of a square, as in Figure 1.2c, then small values of both variables need spreading. If the plot looks roughly like the southeast corner of a square, then large values of the horizontal and small values of the vertical variable need spreading. Hence in Figure 1.2d, small values of x need spreading.

Consider the additive error regression model $Y = m(\boldsymbol{x}) + e$. Then the response transformation model is $Y = t_\lambda(Z) = m_\lambda(\boldsymbol{x}) + e$, and the graphical method for selecting the response transformation is to plot $\hat{m}_{\lambda_i}(\boldsymbol{x})$ versus $t_{\lambda_i}(Z)$ for several values of λ_i, choosing the value of $\lambda = \lambda_0$ where the plotted points follow the identity line with unit slope and zero intercept. For the multiple linear regression model, $\hat{m}_{\lambda_i}(\boldsymbol{x}) = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}_{\lambda_i}$ where $\hat{\boldsymbol{\beta}}_{\lambda_i}$ can be found using the desired fitting method, e.g. OLS or lasso.

Definition 1.4. 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 1.5. Assume that **all** of the values of the "response" Z_i are **positive**. Then the *modified power transformation family*

$$t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda} \tag{1.4}$$

for $\lambda \neq 0$ and $Z_i^{(0)} = \log(Z_i)$. 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 if the MLR or additive error regression model is reasonable for $Y = W$ and $\boldsymbol{x}$. Curvature from the identity line suggests that the candidate response transformation is inappropriate.

Notice that the graphical method is equivalent to making "response plots" for the seven values of $W = t_\lambda(Z)$, and choosing the "best response plot" where the MLR model seems "most reasonable." The seven "response plots" are called transformation plots below. Our convention is that a plot of X versus Y means that X is on the horizontal axis and Y is on the vertical axis.

Definition 1.6. A *transformation plot* is a plot of $\hat{W}$ versus W with the identity line added as a visual aid.

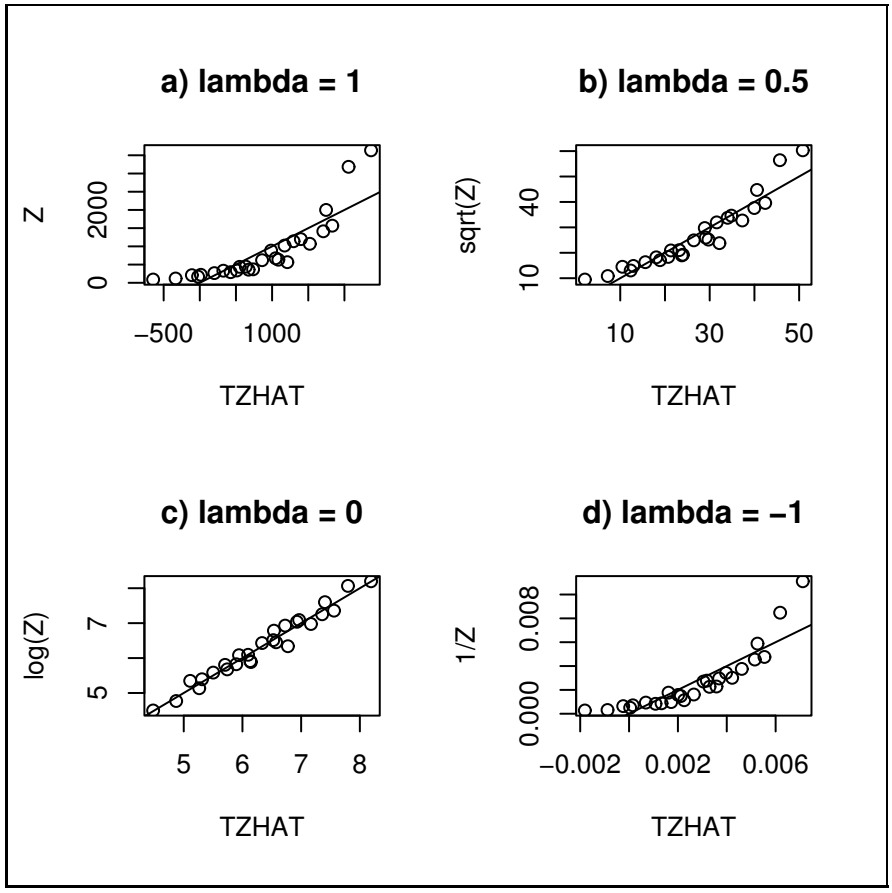

Fig. 1.3 Four Transformation Plots for the Textile Data

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 $\lambda = 0.28$, 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 (e.g. 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 1.1. 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 the regression fitting method, such as OLS or lasso, 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.

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 transformation, the usual checks should be made. In particular, the transformation plot for the selected transformation is the response plot, and a residual plot should also be made. The following example illustrates the procedure, and the plots show $W = t_\lambda(Z)$ on the vertical axis. The label "TZHAT" of the horizontal axis are the "fitted values" $\hat{W}$ that result from using $W = t_\lambda(Z)$ as the "response" in the OLS software.

Example 1.3: 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 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 1.3 are transformation plots of $\hat{W}$ versus $W = Z^\lambda$ 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 1.3a to form along a linear scatter in Figure 1.3c. 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 1.3a shows that a response transformation is needed since the plotted points follow a nonlinear curve while Figure 1.3c 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 1.3a suggests that the log rule is reasonable since $\max(Z)/\min(Z) > 10$.

1.3 The Multivariate Normal Distribution

For much of this book, $\boldsymbol{X}$ is an $n \times p$ design matrix, but this section will usually use the notation $\boldsymbol{X} = (X_1, ..., X_p)^T$ and $\boldsymbol{Y}$ for the random vectors, and $\boldsymbol{x} = (x_1, ..., x_p)^T$ for the observed value of the random vector. This notation will be useful to avoid confusion when studying conditional distributions such as $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$. It can be shown that $\boldsymbol{\Sigma}$ is positive semidefinite and symmetric.

Definition 1.7: Rao (1965, p. 437). A $p \times 1$ random vector $\boldsymbol{X}$ has a $p-$dimensional *multivariate normal distribution* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ iff $\boldsymbol{t}^T\boldsymbol{X}$ has a univariate normal distribution for any $p \times 1$ vector $\boldsymbol{t}$.

If $\boldsymbol{\Sigma}$ is positive definite, then $\boldsymbol{X}$ has a pdf

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(1/2)(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})} \tag{1.5}$$

where $|\boldsymbol{\Sigma}|^{1/2}$ is the square root of the determinant of $\boldsymbol{\Sigma}$. Note that if $p = 1$, then the quadratic form in the exponent is $(z - \mu)(\sigma^2)^{-1}(z - \mu)$ and X has the univariate $N(\mu, \sigma^2)$ pdf. If $\boldsymbol{\Sigma}$ is positive semidefinite but not positive definite, then $\boldsymbol{X}$ has a degenerate distribution. For example, the univariate $N(0, 0^2)$ distribution is degenerate (the point mass at 0).

Definition 1.8. The *population mean* of a random $p \times 1$ vector $\boldsymbol{X} = (X_1, ..., X_p)^T$ is

$$E(\boldsymbol{X}) = (E(X_1), ..., E(X_p))^T$$

and the $p \times p$ *population covariance matrix*

$$\text{Cov}(\boldsymbol{X}) = E(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T = (\sigma_{ij}).$$

That is, the ij entry of $\text{Cov}(\boldsymbol{X})$ is $\text{Cov}(X_i, X_j) = \sigma_{ij}$.

The covariance matrix is also called the variance–covariance matrix and variance matrix. Sometimes the notation $\text{Var}(\boldsymbol{X})$ is used. Note that $\text{Cov}(\boldsymbol{X})$ is a symmetric positive semidefinite matrix. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $p \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{a} + \boldsymbol{X}) = \boldsymbol{a} + E(\boldsymbol{X}) \text{ and } E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}) \tag{1.6}$$

and

$$E(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}E(\boldsymbol{X}) \text{ and } E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}. \tag{1.7}$$

Thus

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}) = \text{Cov}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X})\boldsymbol{A}^T. \tag{1.8}$$

Some important properties of multivariate normal (MVN) distributions are given in the following three theorems. These theorems can be proved using results from Johnson and Wichern (1988, pp. 127-132) or Severini (2005, ch. 8).

Theorem 1.1. a) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and

$$\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}.$$

b) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then any linear combination $\boldsymbol{t}^T\boldsymbol{X} = t_1X_1 + \cdots + t_pX_p \sim N_1(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$. Conversely, if $\boldsymbol{t}^T\boldsymbol{X} \sim N_1(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$ for every $p \times 1$ vector $\boldsymbol{t}$, then $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

c) **The joint distribution of independent normal random variables is MVN.** If $X_1, ..., X_p$ are independent univariate normal $N(\mu_i, \sigma_i^2)$ random vectors, then $\boldsymbol{X} = (X_1, ..., X_p)^T$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)^T$ and $\boldsymbol{\Sigma} = diag(\sigma_1^2, ..., \sigma_p^2)$ (so the off diagonal entries $\sigma_{ij} = 0$ while the diagonal entries of $\boldsymbol{\Sigma}$ are $\sigma_{ii} = \sigma_i^2$).

d) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{A}\boldsymbol{X} \sim N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants and b is a constant, then $\boldsymbol{a} + b\boldsymbol{X} \sim N_p(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$. (Note that $b\boldsymbol{X} = b\boldsymbol{I}_p\boldsymbol{X}$ with $\boldsymbol{A} = b\boldsymbol{I}_p$.)

It will be useful to partition $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. Let $\boldsymbol{X}_1$ and $\boldsymbol{\mu}_1$ be $q \times 1$ vectors, let $\boldsymbol{X}_2$ and $\boldsymbol{\mu}_2$ be $(p-q) \times 1$ vectors, let $\boldsymbol{\Sigma}_{11}$ be a $q \times q$ matrix, let $\boldsymbol{\Sigma}_{12}$ be a $q \times (p-q)$ matrix, let $\boldsymbol{\Sigma}_{21}$ be a $(p-q) \times q$ matrix, and let $\boldsymbol{\Sigma}_{22}$ be a $(p-q) \times (p-q)$ matrix. Then

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Theorem 1.2. a) **All subsets of a MVN are MVN:** $(X_{k_1}, ..., X_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\boldsymbol{\mu}}_i = E(X_{k_i})$ and $\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}(X_{k_i}, X_{k_j})$. In particular, $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

b) If $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent, then $\text{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \boldsymbol{\Sigma}_{12} = E[(\boldsymbol{X}_1 - E(\boldsymbol{X}_1))(\boldsymbol{X}_2 - E(\boldsymbol{X}_2))^T] = \mathbf{0}$, a $q \times (p-q)$ matrix of zeroes.

c) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent iff $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

d) If $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ are independent, then

$$\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N_p \left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right).$$

Theorem 1.3. The conditional distribution of a MVN is MVN. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. That is,

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

Example 1.4. Let $p = 2$ and let $(Y, X)^T$ have a bivariate normal distribution. That is,

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \text{Cov}(Y,X) \\ \text{Cov}(X,Y) & \sigma_X^2 \end{pmatrix} \right).$$

Also, recall that the population correlation between X and Y is given by

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

if $\sigma_X > 0$ and $\sigma_Y > 0$. Then $Y|X = x \sim N(E(Y|X = x), \text{VAR}(Y|X = x))$ where the conditional mean

$$E(Y|X = x) = \mu_Y + \text{Cov}(Y,X)\frac{1}{\sigma_X^2}(x - \mu_X) = \mu_Y + \rho(X,Y)\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}(x - \mu_X)$$

and the conditional variance

$$\begin{aligned} \text{VAR}(Y|X = x) &= \sigma_Y^2 - \text{Cov}(X,Y)\frac{1}{\sigma_X^2}\text{Cov}(X,Y) \\ &= \sigma_Y^2 - \rho(X,Y)\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}\rho(X,Y)\sqrt{\sigma_X^2}\sqrt{\sigma_Y^2} \\ &= \sigma_Y^2 - \rho^2(X,Y)\sigma_Y^2 = \sigma_Y^2[1 - \rho^2(X,Y)]. \end{aligned}$$

Also $aX + bY$ is univariate normal with mean $a\mu_X + b\mu_Y$ and variance

$$a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\ \text{Cov}(X,Y).$$

Remark 1.2. There are several common misconceptions. First, **it is not true that every linear combination $\boldsymbol{t}^T\boldsymbol{X}$ of normal random variables is a normal random variable,** and **it is not true that all uncorrelated normal random variables are independent**. The key condition in Theorem 1.2b and Theorem 1.3c is that the joint distribution of $\boldsymbol{X}$ is MVN. It is possible that $X_1, X_2, ..., X_p$ each has a marginal distribution that is univariate normal, but the joint distribution of $\boldsymbol{X}$ is not MVN. See Seber and Lee (2003, p. 23), and examine the following example from Rohatgi (1976, p. 229). Suppose that the joint pdf of X and Y is a mixture of two bivariate normal distributions both with $EX = EY = 0$ and $\text{VAR}(X) = \text{VAR}(Y) = 1$, but $\text{Cov}(X, Y) = \pm\rho$. Hence $f(x, y) =$

$$\frac{1}{2}\ \frac{1}{2\pi\sqrt{1-\rho^2}}\exp(\frac{-1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)) +$$

$$\frac{1}{2}\ \frac{1}{2\pi\sqrt{1-\rho^2}}\exp(\frac{-1}{2(1-\rho^2)}(x^2 + 2\rho xy + y^2)) \equiv \frac{1}{2}f_1(x, y) + \frac{1}{2}f_2(x, y)$$

where x and y are real and $0 < \rho < 1$. Since both marginal distributions of $f_i(x, y)$ are N(0,1) for $i = 1$ and 2 by Theorem 1.3 a), the marginal distributions of X and Y are N(0,1). Since $\int\int xyf_i(x, y)dxdy = \rho$ for $i = 1$ and $-\rho$ for $i = 2$, X and Y are uncorrelated, but X and Y are not independent since $f(x, y) \neq f_X(x)f_Y(y)$.

Remark 1.3. In Theorem 1.3, suppose that $\boldsymbol{X} = (Y, X_2, ..., X_p)^T$. Let $X_1 = Y$ and $\boldsymbol{X}_2 = (X_2, ..., X_p)^T$. Then $E[Y|\boldsymbol{X}_2] = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ and $\text{VAR}[Y|\boldsymbol{X}_2]$ is a constant that does not depend on $\boldsymbol{X}_2$. Hence $Y|\boldsymbol{X}_2 = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e$ follows the multiple linear regression model.

1.4 Outlier Detection

Outliers are cases that lie far away from the bulk of the data, and outliers can ruin a statistical analysis. For multiple linear regression, the response plot is often useful for outlier detection. Look for gaps in the response plot and for cases far from the identity line. There are no gaps in Figure 1.1, but case 44 is rather far from the identity line. Figure 1.4 has a gap in the response plot.

Next, this section discusses a technique for outlier detection that works well for certain outlier configurations provided bulk of the data consists of more than $n/2$ cases. The technique could fail if there are $g > 2$ groups of about n/g cases per group. First we need to define Mahalanobis distances and the coordinatewise median. Some univariate estimators will be defined first.

1.4.1 The Location Model

The location model is

$$Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{1.9}$$

where $e_1, ..., e_n$ are error random variables, often independent and identically distributed (iid) with zero mean. The location model is used when there is one variable Y, such as height, of interest. The location model is a special case of the multiple linear regression model and of the multivariate location and dispersion model, where there are p variables $x_1, ..., x_p$ of interest, such as height and weight if $p = 2$. Statistical Learning is the analysis of multivariate data, and the location model is an example of univariate data, not an example of multivariate data.

The location model is often summarized by obtaining point estimates and confidence intervals for a location parameter and a scale parameter. Assume that there is a sample $Y_1, \ldots, Y_n$ of size n where the Y_i are iid from a distribution with median MED(Y), mean $E(Y)$, and variance $V(Y)$ if they exist. Also assume that the Y_i have a cumulative distribution function (cdf) F that is known up to a few parameters. For example, Y_i could be normal, exponential, or double exponential. The location parameter μ is often the population mean or median while the scale parameter is often the population standard deviation $\sqrt{V(Y)}$. The ith *case* is Y_i.

Point estimation is one of the oldest problems in statistics and four important statistics for the location model are the sample mean, median, variance, and the median absolute deviation (MAD). Let $Y_1, \ldots, Y_n$ be the random sample; i.e., assume that $Y_1, ..., Y_n$ are iid. The sample mean is a measure of location and estimates the population mean (expected value) $\mu = E(Y)$.

Definition 1.9. The *sample mean*

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}. \tag{1.10}$$

If the data set $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the ith order statistic and the $Y_{(i)}$'s are called the *order statistics*. If the data $Y_1 = 1, Y_2 = 4, Y_3 = 2, Y_4 = 5$, and $Y_5 = 3$, then $\overline{Y} = 3$, $Y_{(i)} = i$ for $i = 1, ..., 5$ and MED(n) = 3 where the sample size $n = 5$. The sample median is a measure of location while the sample standard deviation is a measure of spread. The sample mean and standard deviation are vulnerable to outliers, while the sample median and MAD, defined below, are outlier resistant.

Definition 1.10. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \quad \text{if n is odd,} \tag{1.11}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \text{ if n is even.}$$

The notation $\text{MED}(n) = \text{MED}(n, Y_i) = \text{MED}(Y_1, ..., Y_n)$ will also be used.

Definition 1.11. The *sample variance*

$$S_n^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^n Y_i^2 - n(\overline{Y})^2}{n-1}, \tag{1.12}$$

and the *sample standard deviation* $S_n = \sqrt{S_n^2}$.

Definition 1.12. The *sample median absolute deviation* is

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n). \tag{1.13}$$

Since $\text{MAD}(n) = \text{MAD}(n, Y_i)$ is the median of n distances, at least half of the observations are within a distance $\text{MAD}(n)$ of $\text{MED}(n)$ and at least half of the observations are a distance of $\text{MAD}(n)$ or more away from $\text{MED}(n)$. Like the standard deviation, $\text{MAD}(n)$ is a measure of spread.

Example 1.5. Let the data be $1, 2, 3, 4, 5, 6, 7, 8, 9$. Then $\text{MED}(n) = 5$ and $\text{MAD}(n) = 2 = \text{MED}\{0, 1, 1, 2, 2, 3, 3, 4, 4\}$.

1.4.2 Outlier Detection with Mahalanobis Distances

Now suppose the multivariate data has been collected into an $n \times p$ matrix

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

where the ith row of $\boldsymbol{W}$ is the ith case $\boldsymbol{x}_i^T$ and the jth column $\boldsymbol{v}_j$ of $\boldsymbol{W}$ corresponds to n measurements of the jth random variable X_j for $j = 1, ..., p$. Hence the n rows of the data matrix $\boldsymbol{W}$ correspond to the n cases, while the p columns correspond to measurements on the p random variables $X_1, ..., X_p$. For example, the data may consist of n visitors to a hospital where the $p = 2$ variables *height* and *weight* of each individual were measured.

Definition 1.13. The *coordinatewise median* $\text{MED}(\boldsymbol{W}) = (\text{MED}(X_1), ..., \text{MED}(X_p))^T$ where $\text{MED}(X_i)$ is the sample median of the data in column i corresponding to variable X_i and $\boldsymbol{v}_i$.

Example 1.6. Let the data for X_1 be $1, 2, 3, 4, 5, 6, 7, 8, 9$ while the data for X_2 is $7, 17, 3, 8, 6, 13, 4, 2, 1$. Then $\text{MED}(\boldsymbol{W}) = (\text{MED}(X_1), \text{MED}(X_2))^T = (5, 6)^T$.

For multivariate data, sample Mahalanobis distances play a role similar to that of residuals in multiple linear regression. Let the observed training data be collected in an $n \times p$ matrix $\boldsymbol{W}$. Let the $p \times 1$ column vector $T = T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{W})$ be a dispersion estimator.

Definition 1.14. Let $x_{1j}, ..., x_{nj}$ be measurements on the jth random variable X_j corresponding to the jth column of the data matrix $\boldsymbol{W}$. The jth *sample mean* is $\overline{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj}$. The *sample covariance* S_{ij} estimates $\mathrm{Cov}(X_i, X_j) = \sigma_{ij} = E[(X_i - E(X_i))(X_j - E(X_j))]$, and

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j).$$

$S_{ii} = S_i^2$ is the *sample variance* that estimates the population variance $\sigma_{ii} = \sigma_i^2$. The *sample correlation* r_{ij} estimates the population correlation $\mathrm{Cor}(X_i, X_j) = \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$, and

$$r_{ij} = \frac{S_{ij}}{S_i S_j} = \frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}}.$$

Definition 1.15. Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ be the data where $\boldsymbol{x}_i$ is a $p \times 1$ vector. The **sample mean** or *sample mean vector*

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = (\overline{x}_1, ..., \overline{x}_p)^T = \frac{1}{n}\boldsymbol{W}^T \boldsymbol{1}$$

where $\boldsymbol{1}$ is the $n \times 1$ vector of ones. The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

That is, the ij entry of $\boldsymbol{S}$ is the sample covariance S_{ij}. The *classical estimator of multivariate location and dispersion* is $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$.

It can be shown that $(n-1)\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\, \overline{\boldsymbol{x}}^T =$

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

Hence if the *centering matrix* $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$, then $(n-1)\boldsymbol{S} = \boldsymbol{W}^T \boldsymbol{H} \boldsymbol{W}$.

Definition 1.16. The **sample correlation matrix**

$$\boldsymbol{R} = (r_{ij}).$$

That is, the ij entry of $\boldsymbol{R}$ is the sample correlation r_{ij}.

Let the standardized random variables

$$Z_j = \frac{x_j - \overline{x}_j}{\sqrt{S_{jj}}}$$

for $j = 1, ..., p$. Then the sample correlation matrix $\boldsymbol{R}$ is the sample covariance matrix of the $\boldsymbol{z}_i = (Z_{i1}, ..., Z_{ip})^T$ where $i = 1, ..., n$.

Often it is useful to standardize variables with a robust location estimator and a robust scale estimator. The R function `scale` is useful. The R code below shows how to standardize using

$$Z_j = \frac{x_j - \text{MED}(x_j)}{\text{MAD}(x_j)}$$

for $j = 1, ..., p$. Here $\text{MED}(x_j) = \text{MED}(x_{1j}, ..., x_{nj})$ and $\text{MAD}(x_j) = \text{MAD}(x_{1j}, ..., x_{nj})$ are the sample median and sample median absolute deviation of the data for the jth variable: $x_{1j}, ..., x_{nj}$. See Definitions 1.10 and 1.12. Some of these results are illustrated with the following R code.

```
x <- buxx[,1:3]; cov(x)
                len       nasal     bigonal
len     118299.9257 -191.084603 -104.718925
nasal     -191.0846   18.793905   -1.967121
bigonal   -104.7189   -1.967121   36.796311

cor(x)
                len       nasal     bigonal
len      1.00000000 -0.12815187 -0.05019157
nasal   -0.12815187  1.00000000 -0.07480324
bigonal -0.05019157 -0.07480324  1.00000000
z <- scale(x)
cov(z)
                len       nasal     bigonal
len      1.00000000 -0.12815187 -0.05019157
nasal   -0.12815187  1.00000000 -0.07480324
bigonal -0.05019157 -0.07480324  1.00000000

medd <- apply(x,2,median)
madd <- apply(x,2,mad)/1.4826
z <- scale(x,center=medd,scale=madd)
ddplot4(z)#scaled data still has 5 outliers
```

```
cov(z)    #in the length variable
               len      nasal   bigonal
len     4731.997028 -12.738974 -6.981262
nasal    -12.738974   2.088212 -0.218569
bigonal   -6.981262  -0.218569  4.088479

cor(z)
               len       nasal     bigonal
len      1.00000000 -0.12815187 -0.05019157
nasal   -0.12815187  1.00000000 -0.07480324
bigonal -0.05019157 -0.07480324  1.00000000

apply(z,2,median)
len   nasal bigonal
0       0       0
#scaled data has coord. median = (0,0,0)^T
apply(z,2,mad)/1.4826
len   nasal bigonal
1       1       1 #scaled data has unit MAD
```

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

Rule of Thumb 1.1. Multivariate procedures start to give good results for $n \geq 10p$, especially if the distribution is close to multivariate normal. In particular, we want $n \geq 10p$ for the sample covariance and correlation matrices. For procedures with large sample theory on a large class of distributions, for any value of n, there are always distributions where the results will be poor, but will eventually be good for larger sample sizes. Norman and Streiner (1986, pp. 122, 130, 157) give this rule of thumb and note that some authors recommend $n \geq 30p$. This rule of thumb is much like the rule of thumb that says the central limit theorem normal approximation for $\overline{Y}$ starts to be good for many distributions for $n \geq 30$.

Definition 1.17. The ith *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \quad (1.14)$$

for each point $\boldsymbol{x}_i$. Notice that D_i^2 is a random variable (scalar valued). Let $(T, \boldsymbol{C}) = (T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$. Then

$$D_{\boldsymbol{x}}^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T).$$

Hence D_i^2 uses $\boldsymbol{x} = \boldsymbol{x}_i$.

Let the $p \times 1$ location vector be $\boldsymbol{\mu}$, often the population mean, and let the $p \times p$ dispersion matrix be $\boldsymbol{\Sigma}$, often the population covariance matrix. See Definition 1.8. Notice that if $\boldsymbol{x}$ is a random vector, then the population squared Mahalanobis distance is

$$D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \tag{1.15}$$

and that the term $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$ is the $p-$dimensional analog to the z-score used to transform a univariate $N(\mu, \sigma^2)$ random variable into a $N(0, 1)$ random variable. Hence the sample Mahalanobis distance $D_i = \sqrt{D_i^2}$ is an analog of the absolute value $|Z_i|$ of the sample Z-score $Z_i = (X_i - \overline{X})/\hat{\sigma}$. Also notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

1.4.3 Outlier Detection if $p > n$

Most outlier detection methods work best if $n \geq 20p$, but often data sets have $p > n$, and outliers are a major problem. One of the simplest outlier detection methods uses the Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median $D_i = D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median MED_j: the coordinatewise median computed from the "half set" of cases $\boldsymbol{x}_i$ with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \boldsymbol{I}_p))$ where $\text{MED}_0 = \text{MED}(\boldsymbol{W})$. We often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, ..., D_n) + k\text{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise. Using $k \geq 0$ insures that at least half of the cases get weight 1. This weighting corresponds to the weighting that would be used in a one sided metrically trimmed mean (Huber type skipped mean) of the distances.

Application 1.2. This outlier resistant regression method uses terms from the following definition. Let the ith case $\boldsymbol{w}_i = (Y_i, \boldsymbol{x}_i^T)^T$ where the continuous predictors from $\boldsymbol{x}_i$ are denoted by $\boldsymbol{u}_i$ for $i = 1, ..., n$. Apply the `covmb2` estimator to the $\boldsymbol{u}_i$, and then run the regression method on the m cases $\boldsymbol{w}_i$ corresponding to the `covmb2` set B indices $i_1, ..., i_m$, where $m \geq n/2$.

Definition 1.18. Let the *covmb2 set* B of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set B. Hence

$$T = \frac{\sum_{i=1}^n W_i \boldsymbol{x}_i}{\sum_{i=1}^n W_i} \text{ and } \boldsymbol{C} = \frac{\sum_{i=1}^n W_i (\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^n W_i \ - \ 1}.$$

Example 1.7. Let the clean data (nonoutliers) be i **1** for $i = 1, 2, 3, 4$, and 5 while the outliers are j **1** for $j = 16, 17, 18$, and 19. Here $n = 9$ and **1** is $p \times 1$.

Making a plot of the data for $p = 2$ may be useful. Then the coordinatewise median $\text{MED}_0 = \text{MED}(\boldsymbol{W}) = 5\,\mathbf{1}$. The median Euclidean distance of the data is the Euclidean distance of $5\,\mathbf{1}$ from $1\,\mathbf{1}$ = the Euclidean distance of $5\,\mathbf{1}$ from $9\,\mathbf{1}$. The *median ball* is the hypersphere centered at the coordinatewise median with radius $r = \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p),\ i = 1, ..., n)$ that tends to contain $(n+1)/2$ of the cases if n is odd. Hence the clean data are in the median ball and the outliers are outside of the median ball. The coordinatewise median of the cases with the 5 smallest distances is the coordinatewise median of the clean data: $\text{MED}_1 = 3\ \mathbf{1}$. Then the median Euclidean distance of the data from MED_1 is the Euclidean distance of $3\ \mathbf{1}$ from $1\ \mathbf{1}$ = the Euclidean distance of $3\,\mathbf{1}$ from $5\,\mathbf{1}$. Again the clean cases are the cases with the 5 smallest Euclidean distances. Hence $\text{MED}_j = 3\,\mathbf{1}$ for $j \geq 1$. For $j \geq 1$, if $\boldsymbol{x}_i = j\ \mathbf{1}$, then $D_i = |j-3|\sqrt{p}$. Thus $D_{(1)} = 0$, $D_{(2)} = D_{(3)} = \sqrt{p}$, and $D_{(4)} = D_{(5)} = 2\sqrt{p}$. Hence $\text{MED}(D_1, ..., D_n) = D_{(5)} = 2\sqrt{p} = \text{MAD}(D_1, ..., D_n)$ since the median distance of the D_i from $D_{(5)}$ is $2\sqrt{p} - 0 = 2\sqrt{p}$. Note that the 5 smallest absolute distances $|D_i - D_{(5)}|$ are $0, 0, \sqrt{p}, \sqrt{p}$, and $2\sqrt{p}$. Hence $W_i = 1$ if $D_i \leq 2\sqrt{p} + 10\sqrt{p} = 12\sqrt{p}$. The clean data get weight 1 while the outliers get weight 0 since the smallest distance D_i for the outliers is the Euclidean distance of $3\ \mathbf{1}$ from $16\ \mathbf{1}$ with a $D_i = \|16\ \mathbf{1} - 3\ \mathbf{1}\| = 13\sqrt{p}$. Hence the `covmb2` estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix of the clean data. **Note that the distance for the outliers to get zero weight is proportional to the square root of the dimension** $\sqrt{p}$.

The `covmb2` estimator attempts to give a robust dispersion estimator that reduces the bias by using a big ball about MED_j instead of a ball that contains half of the cases. The weighting is the default method, but you can also plot the squared Euclidean distances and estimate the number $m \geq n/2$ of cases with the smallest distances to be used. The *slpack* function `medout` makes the plot, and the *slpack* function `getB` gives the set B of cases that got weight 1 along with the index `indx` of the case numbers that got weight 1. The function `vecw` stacks the columns of the dispersion matrix $\boldsymbol{C}$ into a vector. Then the elements of the matrix can be plotted.

The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers. An alternative for outlier detection is to replace $\boldsymbol{C}$ by $\boldsymbol{C}_d = diag(\hat{\sigma}_{11}, ..., \hat{\sigma}_{pp})$. For example, use $\hat{\sigma}_{ii} = \boldsymbol{C}_{ii}$. See Ro et al. (2015) and Tarr et al. (2016) for references.

Example 1.8. For the Buxton (1920) data with multiple linear regression, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with

head lengths well over five feet! See Problem 1.13 to reproduce the following plots.

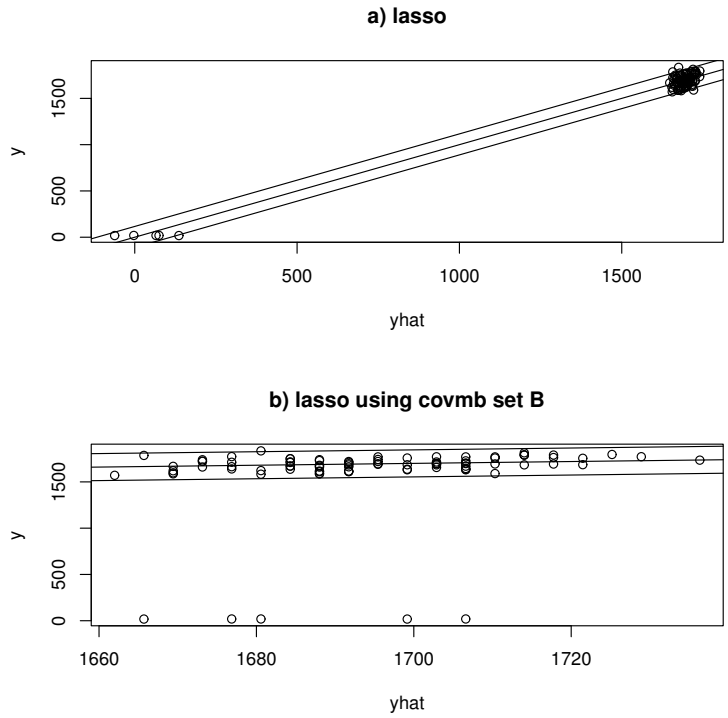

Fig. 1.4 Response plot for lasso and lasso applied to the `covmb2` set B.

Figure 1.4a) shows the response plot for lasso. The identity line passes right through the outliers which are obvious because of the large gap. Figure 1.4b) shows the response plot from lasso for the cases in the `covmb2` set B applied to the predictors, and the set B included all of the clean cases and omitted the 5 outliers. The response plot was made for all of the data, including the outliers. Prediction interval (PI) bands are also included for both plots. Both plots are useful for outlier detection, but the method for plot 1.4b) is better for data analysis: impossible outliers should be deleted or given 0 weight, we do not want to predict that some people are about 0.75 inches tall, and we do want to predict that the people were about 1.6 to 1.8 meters tall. Figure 1.5 shows the DD plot made using `ddplot5`. The five outliers are in the upper right corner.

Also see Problem 1.14 where the `covmb2` set B deleted the 8 cases with the largest D_i, including 5 outliers and 3 clean cases.

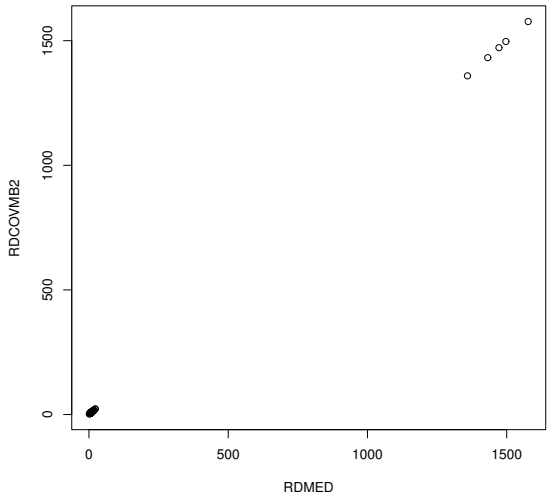

Fig. 1.5 DD plot.

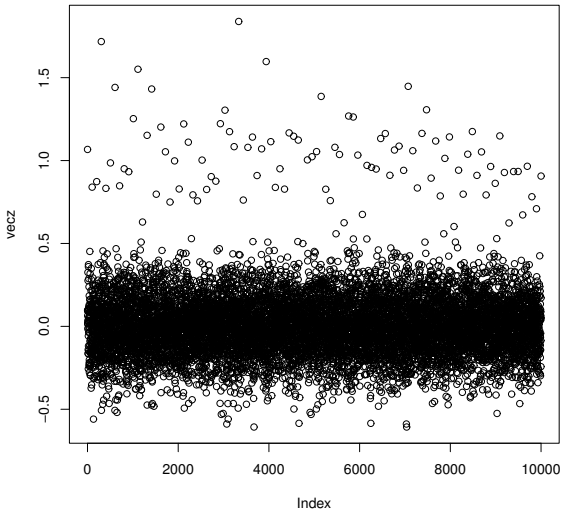

Fig. 1.6 Elements of $\boldsymbol{C}$ for outlier data.

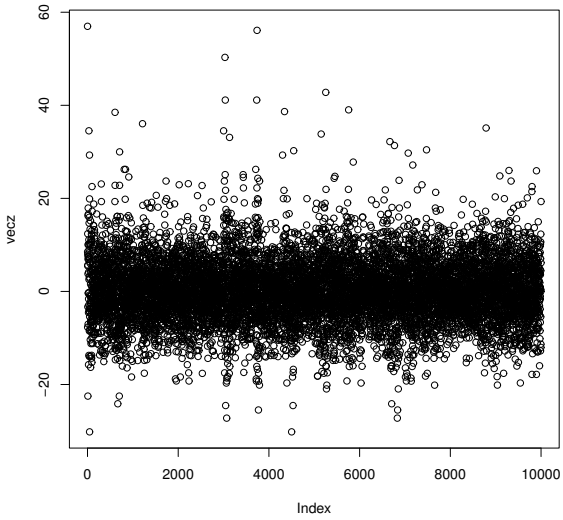

Fig. 1.7 Elements of the classical covariance matrix $\boldsymbol{S}$ for outlier data.

Example 1.9. This example helps illustrate the effect of outliers on classical methods. The artificial data set had $n = 50, p = 100$, and the clean data was iid $N_p(\mathbf{0}, \boldsymbol{I}_p)$. Hence the diagonal elements of the population covariance matrix are 0 and the diagonal elements are 1. Plots of the elements of the sample covariance matrix S and the `covmb2` estimator $\boldsymbol{C}$ are not shown, but were similar to Figure 1.6. Then the first ten cases were contaminated: $\boldsymbol{x}_i \sim N_p(\boldsymbol{\mu}, 100\boldsymbol{I}_p)$ where $\boldsymbol{\mu} = (10, 0, ..., 0)^T$. Figure 1.6 shows that the `covmb2` dispersion matrix $\boldsymbol{C}$ was not much effected by the outliers. The diagonal elements are near 1 and the off diagonal elements are near 0. Figure 1.7 shows that the sample covariance matrix $\boldsymbol{S}$ was greatly effected by the outliers. Several sample covariances are less than -20 and several sample variances are over 40.

R code to used to produce Figures 1.6 and 1.7 is shown below.

```
#n = 50, p = 100
x<-matrix(rnorm(5000),nrow=50,ncol=100)
out<-medout(x) #no outliers, try ddplot5(x)
out <- covmb2(x,msteps=0)
z<-out$cov
plot(diag(z))  #plot the diagonal elements of C
plot(out$center) #plot the elements of  T
vecz <- vecw(z)$vecz
plot(vecz)
```

```
out<-covmb2(x,m=45)
plot(out$center)
plot(diag(out$cov))

#outliers
x[1:10,] <- 10*x[1:10,]
x[1:10,1] <- x[1:10]+10
medout(x) #The 10 outliers are easily detected in
#the plot of the distances from the MED(X).
ddplot5(x) #two widely separated clusters of data
tem <- getB(x,msteps=0)
tem$indx #all 40 clean cases were used
dim(tem$B) #40 by 100
out<-covmb2(x,msteps=0)
z<-out$cov
plot(diag(z))
plot(out$center)
vecz <- vecw(z)$vecz
plot(vecz) #plot the elements of C
#Figure 1.6

#examine the sample covariance matrix and mean
plot(diag(var(x)))
plot(apply(x,2,mean)) #plot elements of xbar
zc <- var(x)
vecz <- vecw(zc)$vecz
plot(vecz) #plot the elements of S
#Figure 1.7

out<-medout(x) #10 outliers
out<-covmb2(x,m=40)
plot(out$center)
plot(diag(out$cov))
```

The `covmb2` estimator can also be used for $n > p$. The *slpack* function `mldsim6` suggests that for 40% outliers, the outliers need to be further away from the bulk of the data (`covmb2(k=5)` needs a larger value of pm) than for the other six estimators if $n \geq 20p$. With some outlier types, `covmb2(k=5)` was often near best. Try the following commands. The other estimators need $n > 2p$, and as n gets close to $2p$, `covmb2` may outperform the other estimators. Also see Problem 1.15.

```
#near point mass on major axis
mldsim6(n=100,p=10,outliers=1,gam=0.25,pm=25)
mldsim6(n=100,p=10,outliers=1,gam=0.4,pm=25) #bad
```

```
mldsim6(n=100,p=40,outliers=1,gam=0.1,pm=100)
mldsim6(n=200,p=60,outliers=1,gam=0.1,pm=100)
#mean shift outliers
mldsim6(n=100,p=40,outliers=3,gam=0.1,pm=10)
mldsim6(n=100,p=40,outliers=3,gam=0.25,pm=20)
mldsim6(n=200,p=60,outliers=3,gam=0.1,pm=10)
#concentration steps can help
mldsim6(n=100,p=10,outliers=3,gam=0.4,pm=10,osteps=0)
mldsim6(n=100,p=10,outliers=3,gam=0.4,pm=10,osteps=9)
```

Elliptically contoured distributions, defined below, are an important class of distributions for multivariate data. The multivariate normal distribution is also an elliptically contoured distribution. This distributions is useful for discriminant analysis in Chapter 5 and for multivariate analysis in Chapter 6.

Definition 1.19: Johnson (1987, pp. 107-108). A $p \times 1$ random vector $\boldsymbol{X}$ has an *elliptically contoured distribution,* also called an *elliptically symmetric distribution,* if $\boldsymbol{X}$ has joint pdf

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z}-\boldsymbol{\mu})], \tag{1.16}$$

and we say $\boldsymbol{X}$ has an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution.

If $\boldsymbol{X}$ has an elliptically contoured (EC) distribution, then the characteristic function of $\boldsymbol{X}$ is

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(i\boldsymbol{t}^T\boldsymbol{\mu})\psi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}) \tag{1.17}$$

for some function ψ. If the second moments exist, then

$$E(\boldsymbol{X}) = \boldsymbol{\mu} \tag{1.18}$$

and

$$\text{Cov}(\boldsymbol{X}) = c_X \boldsymbol{\Sigma} \tag{1.19}$$

where

$$c_X = -2\psi'(0).$$

1.5 Large Sample Theory

The first three subsections will review large sample theory for the univariate case, then multivariate theory will be given.

1.5.1 The CLT and the Delta Method

Large sample theory, also called asymptotic theory, is used to approximate the distribution of an estimator when the sample size n is large. This theory is extremely useful if the exact sampling distribution of the estimator is complicated or unknown. To use this theory, one must determine what the estimator is estimating, the rate of convergence, the asymptotic distribution, and how large n must be for the approximation to be useful. Moreover, the (asymptotic) standard error (SE), an estimator of the asymptotic standard deviation, must be computable if the estimator is to be useful for inference. Often the bootstrap can be used to compute the SE.

Theorem 1.4: the Central Limit Theorem (CLT). Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $\text{VAR}(Y) = \sigma^2$. Let the sample mean $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Hence

$$\sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) = \sqrt{n}\left(\frac{\sum_{i=1}^{n} Y_i - n\mu}{n\sigma}\right) \xrightarrow{D} N(0, 1).$$

Note that the sample mean is estimating the *population mean* μ with a $\sqrt{n}$ convergence rate, the asymptotic distribution is normal, and the $\text{SE} = S/\sqrt{n}$ where S is the *sample standard deviation.* For distributions "close" to the normal distribution, the central limit theorem provides a good approximation if the sample size $n \geq 30$. Hesterberg (2014, pp. 41, 66) suggests $n \geq 5000$ is needed for moderately skewed distributions. A special case of the CLT is proven after Theorem 1.17.

Notation. The notation $X \sim Y$ and $X \stackrel{D}{=} Y$ both mean that the random variables X and Y have the same distribution. Hence $F_X(x) = F_Y(y)$ for all real y. The notation $Y_n \xrightarrow{D} X$ means that for large n we can approximate the cdf of Y_n by the cdf of X. The distribution of X is the limiting distribution or asymptotic distribution of Y_n. For the CLT, notice that

$$Z_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) = \left(\frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}}\right)$$

is the z–score of $\overline{Y}$. If $Z_n \xrightarrow{D} N(0, 1)$, then the notation $Z_n \approx N(0, 1)$, also written as $Z_n \sim AN(0, 1)$, means approximate the cdf of Z_n by the standard normal cdf. See Definition 1.20. Similarly, the notation

$$\overline{Y}_n \approx N(\mu, \sigma^2/n),$$

also written as $\overline{Y}_n \sim AN(\mu, \sigma^2/n)$, means approximate the cdf of $\overline{Y}_n$ as if $\overline{Y}_n \sim N(\mu, \sigma^2/n)$. The distribution of X does not depend on n, but the approximate distribution $\overline{Y}_n \approx N(\mu, \sigma^2/n)$ does depend on n.

The two main applications of the CLT are to give the limiting distribution of $\sqrt{n}(\overline{Y}_n - \mu)$ and the limiting distribution of $\sqrt{n}(Y_n/n - \mu_X)$ for a random variable Y_n such that $Y_n = \sum_{i=1}^n X_i$ where the X_i are iid with $E(X) = \mu_X$ and $\text{VAR}(X) = \sigma_X^2$.

Example 1.10. a) Let $Y_1, ..., Y_n$ be iid Ber(ρ). Then $E(Y) = \rho$ and $\text{VAR}(Y) = \rho(1-\rho)$. (The Bernoulli ($\rho$) distribution is the binomial (1,ρ) distribution.) Hence

$$\sqrt{n}(\overline{Y}_n - \rho) \xrightarrow{D} N(0, \rho(1-\rho))$$

by the CLT.

b) Now suppose that $Y_n \sim BIN(n, \rho)$. Then $Y_n \stackrel{D}{=} \sum_{i=1}^n X_i$ where $X_1, ..., X_n$ are iid Ber(ρ). Hence

$$\sqrt{n}\left(\frac{Y_n}{n} - \rho\right) \xrightarrow{D} N(0, \rho(1-\rho))$$

since

$$\sqrt{n}\left(\frac{Y_n}{n} - \rho\right) \stackrel{D}{=} \sqrt{n}(\overline{X}_n - \rho) \xrightarrow{D} N(0, \rho(1-\rho))$$

by a).

c) Now suppose that $Y_n \sim BIN(k_n, \rho)$ where $k_n \to \infty$ as $n \to \infty$. Then

$$\sqrt{k_n}\left(\frac{Y_n}{k_n} - \rho\right) \approx N(0, \rho(1-\rho))$$

or

$$\frac{Y_n}{k_n} \approx N\left(\rho, \frac{\rho(1-\rho)}{k_n}\right) \quad \text{or} \quad \mathrm{Y_n} \approx \mathrm{N}\left(\mathrm{k_n}\rho, \mathrm{k_n}\rho(1-\rho)\right).$$

Theorem 1.5: the Delta Method. If g does not depend on n, $g'(\theta) \neq 0$, and

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2),$$

then

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{D} N(0, \sigma^2[g'(\theta)]^2).$$

Example 1.11. Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $\text{VAR}(Y) = \sigma^2$. Then by the CLT,

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Let $g(\mu) = \mu^2$. Then $g'(\mu) = 2\mu \neq 0$ for $\mu \neq 0$. Hence

$$\sqrt{n}((\overline{Y}_n)^2 - \mu^2) \xrightarrow{D} N(0, 4\sigma^2\mu^2)$$

for $\mu \neq 0$ by the delta method.

Example 1.12. Let $X \sim \text{Binomial}(n, p)$ where the positive integer n is large and $0 < p < 1$. Find the limiting distribution of $\sqrt{n}\ \left[\left(\frac{X}{n}\right)^2 - \ p^2 \right]$.

Solution. Example 1.10b gives the limiting distribution of $\sqrt{n}(\frac{X}{n} - p)$. Let $g(p) = p^2$. Then $g'(p) = 2p$ and by the delta method,

$$\sqrt{n}\ \left[\left(\frac{X}{n}\right)^2 - p^2 \right] = \sqrt{n}\ \left(\ g\left(\frac{X}{n}\right) - g(p) \ \right) \xrightarrow{D}$$

$$N(0, p(1-p)(g'(p))^2) = N(0, p(1-p)4p^2) = N(0, 4p^3(1-p)).$$

Example 1.13. Let $X_n \sim \text{Poisson}(n\lambda)$ where the positive integer n is large and $\lambda > 0$.

a) Find the limiting distribution of $\sqrt{n}\ \left(\ \frac{X_n}{n} - \ \lambda \ \right)$.

b) Find the limiting distribution of $\sqrt{n}\ \left[\ \sqrt{\frac{X_n}{n}} - \ \sqrt{\lambda} \ \right]$.

Solution. a) $X_n \stackrel{D}{=} \sum_{i=1}^n Y_i$ where the Y_i are iid Poisson(λ). Hence $E(Y) = \lambda = Var(Y)$. Thus by the CLT,

$$\sqrt{n}\ \left(\ \frac{X_n}{n} - \lambda \ \right) \stackrel{D}{=} \sqrt{n}\ \left(\ \frac{\sum_{i=1}^n Y_i}{n} - \lambda \ \right) \xrightarrow{D} N(0, \lambda).$$

b) Let $g(\lambda) = \sqrt{\lambda}$. Then $g'(\lambda) = \frac{1}{2\sqrt{\lambda}}$ and by the delta method,

$$\sqrt{n}\ \left[\ \sqrt{\frac{X_n}{n}} - \ \sqrt{\lambda} \ \right] = \sqrt{n}\ \left(\ g\left(\frac{X_n}{n}\right) - g(\lambda) \ \right) \xrightarrow{D}$$

$$N(0, \lambda\ (g'(\lambda))^2) = N\left(0, \lambda \frac{1}{4\lambda}\right) = N\left(0, \frac{1}{4}\right).$$

Example 1.14. Let $Y_1, ..., Y_n$ be independent and identically distributed (iid) from a Gamma(α, β) distribution.

a) Find the limiting distribution of $\sqrt{n}\ \left(\ \overline{Y} - \ \alpha\beta \ \right)$.

b) Find the limiting distribution of $\sqrt{n}\ \left(\ (\overline{Y})^2 - \ c \ \right)$ for appropriate constant c.

Solution: a) Since $E(Y) = \alpha\beta$ and $V(Y) = \alpha\beta^2$, by the CLT $\sqrt{n}\ (\ \overline{Y} - \ \alpha\beta\ ) \xrightarrow{D} N(0, \alpha\beta^2)$.

b) Let $\mu = \alpha\beta$ and $\sigma^2 = \alpha\beta^2$. Let $g(\mu) = \mu^2$ so $g'(\mu) = 2\mu$ and $[g'(\mu)]^2 = 4\mu^2 = 4\alpha^2\beta^2$. Then by the delta method, $\sqrt{n}\ (\ (\overline{Y})^2 - \ c\ ) \xrightarrow{D} N(0, \sigma^2[g'(\mu)]^2) = N(0, 4\alpha^3\beta^4)$ where $c = \mu^2 = \alpha^2\beta^2$.

1.5.2 Modes of Convergence and Consistency

Definition 1.20. Let $\{Z_n, n = 1, 2, ...\}$ be a sequence of random variables with cdfs F_n, and let X be a random variable with cdf F. Then Z_n **converges in distribution to** X, written

$$Z_n \xrightarrow{D} X,$$

or Z_n *converges in law to* X, written $Z_n \xrightarrow{L} X$, if

$$\lim_{n\to\infty} F_n(t) = F(t)$$

at each continuity point t of F. The distribution of X is called the **limiting distribution** or the **asymptotic distribution** of Z_n.

An important fact is that **the limiting distribution does not depend on the sample size** n. Notice that the CLT and delta method give the limiting distributions of $Z_n = \sqrt{n}(\overline{Y}_n - \mu)$ and $Z_n = \sqrt{n}(g(T_n) - g(\theta))$, respectively.

Convergence in distribution is useful if the distribution of X_n is unknown or complicated and the distribution of X is easy to use. Then for large n we can approximate the probability that X_n is in an interval by the probability that X is in the interval. To see this, notice that if $X_n \xrightarrow{D} X$, then $P(a < X_n \leq b) = F_n(b) - F_n(a) \to F(b) - F(a) = P(a < X \leq b)$ if F is continuous at a and b.

Warning: convergence in distribution says that the cdf $F_n(t)$ of X_n gets close to the cdf of $F(t)$ of X as $n \to \infty$ provided that t is a continuity point of F. Hence for any $\epsilon > 0$ there exists N_t such that if $n > N_t$, then $|F_n(t) - F(t)| < \epsilon$. Notice that N_t depends on the value of t. Convergence in distribution does not imply that the random variables $X_n \equiv X_n(\omega)$ converge to the random variable $X \equiv X(\omega)$ for all ω.

Example 1.15. Suppose that $X_n \sim U(-1/n, 1/n)$. Then the cdf $F_n(x)$ of X_n is

$$F_n(x) = \begin{cases} 0, & x \leq \frac{-1}{n} \\ \frac{nx}{2} + \frac{1}{2}, & \frac{-1}{n} \leq x \leq \frac{1}{n} \\ 1, & x \geq \frac{1}{n}. \end{cases}$$

Sketching $F_n(x)$ shows that it has a line segment rising from 0 at $x = -1/n$ to 1 at $x = 1/n$ and that $F_n(0) = 0.5$ for all $n \geq 1$. Examining the cases $x < 0$, $x = 0$, and $x > 0$ shows that as $n \to \infty$,

$$F_n(x) \to \begin{cases} 0, & x < 0 \\ \frac{1}{2} & x = 0 \\ 1, & x > 0. \end{cases}$$

Notice that the right hand side is not a cdf since right continuity does not hold at $x = 0$. Notice that if X is a random variable such that $P(X = 0) = 1$, then X has cdf

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0. \end{cases}$$

Since $x = 0$ is the only discontinuity point of $F_X(x)$ and since $F_n(x) \to F_X(x)$ for all continuity points of $F_X(x)$ (i.e. for $x \neq 0$),

$$X_n \xrightarrow{D} X.$$

Example 1.16. Suppose $Y_n \sim U(0, n)$. Then $F_n(t) = t/n$ for $0 < t \leq n$ and $F_n(t) = 0$ for $t \leq 0$. Hence $\lim_{n\to\infty} F_n(t) = 0$ for $t \leq 0$. If $t > 0$ and $n > t$, then $F_n(t) = t/n \to 0$ as $n \to \infty$. Thus $\lim_{n\to\infty} F_n(t) = 0$ for all t, and Y_n does not converge in distribution to any random variable Y since $H(t) \equiv 0$ is not a cdf.

Definition 1.21. A sequence of random variables X_n *converges in distribution to a constant* $\tau(\theta)$, written

$$X_n \xrightarrow{D} \tau(\theta), \quad \text{if} \quad \mathrm{X_n} \xrightarrow{\mathrm{D}} \mathrm{X}$$

where $P(X = \tau(\theta)) = 1$. The distribution of the random variable X is said to be *degenerate at* $\tau(\theta)$ or to be a *point mass at* $\tau(\theta)$.

Definition 1.22. A sequence of random variables X_n *converges in probability to a constant* $\tau(\theta)$, written

$$X_n \xrightarrow{P} \tau(\theta),$$

if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - \tau(\theta)| < \epsilon) = 1 \quad \text{or, equivalently,} \quad \lim_{\mathrm{n}\to\infty} \mathrm{P}(|\mathrm{X_n} - \tau(\theta)| \geq \epsilon) = 0.$$

The sequence X_n **converges in probability to** X, written

$$X_n \xrightarrow{P} X,$$

if $X_n - X \xrightarrow{P} 0$.

Notice that $X_n \xrightarrow{P} X$ if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1, \quad \text{or, equivalently,} \quad \lim_{\mathrm{n}\to\infty} \mathrm{P}(|\mathrm{X_n} - \mathrm{X}| \geq \epsilon) = 0.$$

Definition 1.23. Let the *parameter space* Θ be the set of possible values of θ. A sequence of estimators T_n of $\tau(\theta)$ is **consistent** for $\tau(\theta)$ if

$$T_n \xrightarrow{P} \tau(\theta)$$

for every $\theta \in \Theta$. If T_n is consistent for $\tau(\theta)$, then T_n is a **consistent estimator** of $\tau(\theta)$.

Consistency is a weak property that is usually satisfied by good estimators. T_n is a consistent estimator for $\tau(\theta)$ if the probability that T_n falls in any neighborhood of $\tau(\theta)$ goes to one, regardless of the value of $\theta \in \Theta$.

Definition 1.24. For a real number $r > 0$, Y_n *converges in rth mean* to a random variable Y, written

$$Y_n \xrightarrow{r} Y,$$

if

$$E(|Y_n - Y|^r) \to 0$$

as $n \to \infty$. In particular, if $r = 2$, Y_n **converges in quadratic mean** to Y, written

$$Y_n \xrightarrow{2} Y \quad \text{or} \quad \mathrm{Y_n} \xrightarrow{\mathrm{qm}} \mathrm{Y},$$

if

$$E[(Y_n - Y)^2] \to 0$$

as $n \to \infty$.

Theorem 1.6: Generalized Chebyshev's Inequality. Let $u : \mathbb{R} \to [0, \infty)$ be a nonnegative function. If $E[u(Y)]$ exists then for any $c > 0$,

$$P[u(Y) \geq c] \leq \frac{E[u(Y)]}{c}.$$

If $\mu = E(Y)$ exists, then taking $u(y) = |y - \mu|^r$ and $\tilde{c} = c^r$ gives
Markov's Inequality: for $r > 0$ and any $c > 0$,

$$P[|Y - \mu| \geq c] = P[|Y - \mu|^r \geq c^r] \leq \frac{E[|Y - \mu|^r]}{c^r}.$$

If $r = 2$ and $\sigma^2 = \mathrm{VAR}(Y)$ exists, then we obtain
Chebyshev's Inequality:

$$P[|Y - \mu| \geq c] \leq \frac{\mathrm{VAR}(Y)}{c^2}.$$

Proof. The proof is given for pdfs. For pmfs, replace the integrals by sums. Now

$$E[u(Y)] = \int_{\mathbb{R}} u(y)f(y)dy = \int_{\{y:u(y)\geq c\}} u(y)f(y)dy + \int_{\{y:u(y)<c\}} u(y)f(y)dy$$

$$\geq \int_{\{y:u(y)\geq c\}} u(y)f(y)dy$$

since the integrand $u(y)f(y) \geq 0$. Hence

$$E[u(Y)] \geq c \int_{\{y:u(y)\geq c\}} f(y)dy = cP[u(Y) \geq c]. \ \square$$

The following theorem gives sufficient conditions for T_n to be a consistent estimator of $\tau(\theta)$. Notice that $E_\theta[(T_n - \tau(\theta))^2] = MSE_{\tau(\theta)}(T_n) \to 0$ for all $\theta \in \Theta$ is equivalent to $T_n \stackrel{qm}{\to} \tau(\theta)$ for all $\theta \in \Theta$.

Theorem 1.7. a) If

$$\lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) = 0$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

b) If

$$\lim_{n\to\infty} \mathrm{VAR}_\theta(T_n) = 0 \ \text{ and } \ \lim_{n\to\infty} E_\theta(T_n) = \tau(\theta)$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

Proof. a) Using Theorem 1.6 with $Y = T_n$, $u(T_n) = (T_n - \tau(\theta))^2$ and $c = \epsilon^2$ shows that for any $\epsilon > 0$,

$$P_\theta(|T_n - \tau(\theta)| \geq \epsilon) = P_\theta[(T_n - \tau(\theta))^2 \geq \epsilon^2] \leq \frac{E_\theta[(T_n - \tau(\theta))^2]}{\epsilon^2}.$$

Hence

$$\lim_{n\to\infty} E_\theta[(T_n - \tau(\theta))^2] = \lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) \to 0$$

is a sufficient condition for T_n to be a consistent estimator of $\tau(\theta)$.

b) Recall that

$$MSE_{\tau(\theta)}(T_n) = \mathrm{VAR}_\theta(T_n) + [\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n})]^2$$

where $\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n}) = \mathrm{E}_\theta(\mathrm{T_n}) - \tau(\theta)$. Since $MSE_{\tau(\theta)}(T_n) \to 0$ if both $\mathrm{VAR}_\theta(T_n) \to 0$ and $\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n}) = \mathrm{E}_\theta(\mathrm{T_n}) - \tau(\theta) \to 0$, the result follows from a). $\square$

The following result shows estimators that converge at a $\sqrt{n}$ rate are consistent. Use this result and the delta method to show that $g(T_n)$ is a consistent

estimator of $g(\theta)$. Note that b) follows from a) with $X_\theta \sim N(0, v(\theta))$. The WLLN shows that $\overline{Y}$ is a consistent estimator of $E(Y) = \mu$ if $E(Y)$ exists.

Theorem 1.8. a) Let X_θ be a random variable with distribution depending on θ, and $0 < \delta \leq 1$. If

$$n^\delta(T_n - \tau(\theta)) \xrightarrow{D} X_\theta$$

then $T_n \xrightarrow{P} \tau(\theta)$.

b) If

$$\sqrt{n}(T_n - \tau(\theta)) \xrightarrow{D} N(0, v(\theta))$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

Definition 1.25. A sequence of random variables X_n *converges almost everywhere* (or *almost surely*, or *with probability 1*) to X if

$$P(\lim_{n\to\infty} X_n = X) = 1.$$

This type of convergence will be denoted by

$$X_n \xrightarrow{ae} X.$$

Notation such as "X_n converges to X ae" will also be used. Sometimes "ae" will be replaced with "as" or "wp1." We say that X_n *converges almost everywhere* to $\tau(\theta)$, written

$$X_n \xrightarrow{ae} \tau(\theta),$$

if $P(\lim_{n\to\infty} X_n = \tau(\theta)) = 1$.

Theorem 1.9. Let Y_n be a sequence of iid random variables with $E(Y_i) = \mu$. Then

a) **Strong Law of Large Numbers (SLLN):** $\overline{Y}_n \xrightarrow{ae} \mu$, and

b) **Weak Law of Large Numbers (WLLN):** $\overline{Y}_n \xrightarrow{P} \mu$.

Proof of WLLN when $V(Y_i) = \sigma^2$: By Chebyshev's inequality, for every $\epsilon > 0$,

$$P(|\overline{Y}_n - \mu| \geq \epsilon) \leq \frac{V(\overline{Y}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

as $n \to \infty$. □

In proving consistency results, there is an infinite sequence of estimators that depend on the sample size n. Hence the subscript n will be added to the estimators.

Definition 1.26. Lehmann (1999, pp. 53-54): a) A sequence of random variables W_n is *tight* or *bounded in probability,* written $W_n = O_P(1)$, if for every $\epsilon > 0$ there exist positive constants D_ϵ and N_ϵ such that

$$P(|W_n| \leq D_\epsilon) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$. Also $W_n = O_P(X_n)$ if $|W_n/X_n| = O_P(1)$. Similarly, $W_n = O_P(n^{-1/2})$ if $|\sqrt{n}\ W_n| = O_P(1)$.

b) The sequence $W_n = o_P(n^{-\delta})$ if $n^\delta W_n = o_P(1)$ which means that

$$n^\delta W_n \xrightarrow{P} 0.$$

c) W_n has the *same order as* X_n *in probability*, written $W_n \asymp_P X_n$, if for every $\epsilon > 0$ there exist positive constants N_ϵ and $0 < d_\epsilon < D_\epsilon$ such that

$$P\left(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon\right) = P\left(\frac{1}{D_\epsilon} \leq \left|\frac{X_n}{W_n}\right| \leq \frac{1}{d_\epsilon}\right) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$.

d) Similar notation is used for a $k \times r$ matrix $\boldsymbol{A}_n = \boldsymbol{A} = [a_{i,j}(n)]$ if each element $a_{i,j}(n)$ has the desired property. For example, $\boldsymbol{A} = O_P(n^{-1/2})$ if each $a_{i,j}(n) = O_P(n^{-1/2})$.

Definition 1.27. Let $W_n = \|\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}\|$.

a) If $W_n \asymp_P n^{-\delta}$ for some $\delta > 0$, then both W_n and $\hat{\boldsymbol{\mu}}_n$ have (tightness) **rate** n^δ.

b) If there exists a constant κ such that

$$n^\delta(W_n - \kappa) \xrightarrow{D} X$$

for some nondegenerate random variable X, then both W_n and $\hat{\boldsymbol{\mu}}_n$ have *convergence rate* n^δ.

Theorem 1.10. Suppose there exists a constant κ such that

$$n^\delta(W_n - \kappa) \xrightarrow{D} X.$$

a) Then $W_n = O_P(n^{-\delta})$.

b) If X is not degenerate, then $W_n \asymp_P n^{-\delta}$.

The above result implies that if W_n has convergence rate n^δ, then W_n has tightness rate n^δ, and the term "tightness" will often be omitted. Part a) is proved, for example, in Lehmann (1999, p. 67).

The following result shows that if $W_n \asymp_P X_n$, then $X_n \asymp_P W_n$, $W_n = O_P(X_n)$, and $X_n = O_P(W_n)$. Notice that if $W_n = O_P(n^{-\delta})$, then n^δ is a lower bound on the rate of W_n. As an example, if the CLT holds then $\overline{Y}_n = O_P(n^{-1/3})$, but $\overline{Y}_n \asymp_P n^{-1/2}$.

Theorem 1.11. a) If $W_n \asymp_P X_n$, then $X_n \asymp_P W_n$.

b) If $W_n \asymp_P X_n$, then $W_n = O_P(X_n)$.

c) If $W_n \asymp_P X_n$, then $X_n = O_P(W_n)$.

d) $W_n \asymp_P X_n$ iff $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$.

Proof. a) Since $W_n \asymp_P X_n$,

$$P\left(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon\right) = P\left(\frac{1}{D_\epsilon} \leq \left|\frac{X_n}{W_n}\right| \leq \frac{1}{d_\epsilon}\right) \geq 1-\epsilon$$

for all $n \geq N_\epsilon$. Hence $X_n \asymp_P W_n$.

b) Since $W_n \asymp_P X_n$,

$$P(|W_n| \leq |X_n D_\epsilon|) \geq P\left(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon\right) \geq 1-\epsilon$$

for all $n \geq N_\epsilon$. Hence $W_n = O_P(X_n)$.

c) Follows by a) and b).

d) If $W_n \asymp_P X_n$, then $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$ by b) and c). Now suppose $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$. Then

$$P(|W_n| \leq |X_n| D_{\epsilon/2}) \geq 1 - \epsilon/2$$

for all $n \geq N_1$, and

$$P(|X_n| \leq |W_n| 1/d_{\epsilon/2}) \geq 1 - \epsilon/2$$

for all $n \geq N_2$. Hence

$$P(A) \equiv P\left(\left|\frac{W_n}{X_n}\right| \leq D_{\epsilon/2}\right) \geq 1 - \epsilon/2$$

and

$$P(B) \equiv P\left(d_{\epsilon/2} \leq \left|\frac{W_n}{X_n}\right|\right) \geq 1 - \epsilon/2$$

for all $n \geq N = \max(N_1, N_2)$. Since $P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1$,

$$P(A \cap B) = P(d_{\epsilon/2} \leq \left|\frac{W_n}{X_n}\right| \leq D_{\epsilon/2}) \geq 1 - \epsilon/2 + 1 - \epsilon/2 - 1 = 1 - \epsilon$$

for all $n \geq N$. Hence $W_n \asymp_P X_n$. □

The following result is used to prove the following Theorem 1.13 which says that if there are K estimators $T_{j,n}$ of a parameter $\boldsymbol{\beta}$, such that $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ where $0 < \delta \leq 1$, and if T_n^* picks one of these estimators, then $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$.

Theorem 1.12: Pratt (1959). Let $X_{1,n}, ..., X_{K,n}$ each be $O_P(1)$ where K is fixed. Suppose $W_n = X_{i_n,n}$ for some $i_n \in \{1, ..., K\}$. Then

$$W_n = O_P(1). \tag{1.20}$$

Proof.

$$P(\max\{X_{1,n}, ..., X_{K,n}\} \leq x) = P(X_{1,n} \leq x, ..., X_{K,n} \leq x) \leq$$

$$F_{W_n}(x) \leq P(\min\{X_{1,n}, ..., X_{K,n}\} \leq x) = 1 - P(X_{1,n} > x, ..., X_{K,n} > x).$$

Since K is finite, there exists $B > 0$ and N such that $P(X_{i,n} \leq B) > 1 - \epsilon/2K$ and $P(X_{i,n} > -B) > 1 - \epsilon/2K$ for all $n > N$ and $i = 1, ..., K$. Bonferroni's inequality states that $P(\cap_{i=1}^{K} A_i) \geq \sum_{i=1}^{K} P(A_i) \ - \ (K-1)$. Thus

$$F_{W_n}(B) \geq P(X_{1,n} \leq B, ..., X_{K,n} \leq B) \geq$$

$$K(1 - \epsilon/2K) - (K-1) = K - \epsilon/2 - K + 1 = 1 - \epsilon/2$$

and

$$-F_{W_n}(-B) \geq -1 + P(X_{1,n} > -B, ..., X_{K,n} > -B) \geq$$

$$-1 + K(1 - \epsilon/2K) - (K-1) = -1 + K - \epsilon/2 - K + 1 = -\epsilon/2.$$

Hence

$$F_{W_n}(B) - F_{W_n}(-B) \geq 1 - \epsilon \ \text{ for } \ \text{n} > \text{N}. \ \square$$

Theorem 1.13. Suppose $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ for $j = 1, ..., K$ where $0 < \delta \leq 1$. Let $T_n^* = T_{i_n,n}$ for some $i_n \in \{1, ..., K\}$ where, for example, $T_{i_n,n}$ is the $T_{j,n}$ that minimized some criterion function. Then

$$\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{1.21}$$

Proof. Let $X_{j,n} = n^{\delta}\|T_{j,n} - \boldsymbol{\beta}\|$. Then $X_{j,n} = O_P(1)$ so by Proposition 1.10, $n^{\delta}\|T_n^* - \boldsymbol{\beta}\| = O_P(1)$. Hence $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$. $\square$

1.5.3 Slutsky's Theorem and Related Results

Theorem 1.14: Slutsky's Theorem. Suppose $Y_n \xrightarrow{D} Y$ and $W_n \xrightarrow{P} w$ for some constant w. Then

a) $Y_n + W_n \xrightarrow{D} Y + w$,

b) $Y_n W_n \xrightarrow{D} wY$, and

c) $Y_n/W_n \xrightarrow{D} Y/w$ if $w \neq 0$.

Theorem 1.15. a) If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.

b) If $X_n \xrightarrow{ae} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

c) If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

d) $X_n \xrightarrow{P} \tau(\theta)$ iff $\text{X}_\text{n} \xrightarrow{\text{D}} \tau(\theta)$.

e) If $X_n \xrightarrow{P} \theta$ and τ is continuous at θ, then $\tau(X_n) \xrightarrow{P} \tau(\theta)$.

f) If $X_n \xrightarrow{D} \theta$ and τ is continuous at θ, then $\tau(X_n) \xrightarrow{D} \tau(\theta)$.

Suppose that for all $\theta \in \Theta$, $T_n \xrightarrow{D} \tau(\theta)$, $T_n \xrightarrow{r} \tau(\theta)$, or $T_n \xrightarrow{ae} \tau(\theta)$. Then T_n is a consistent estimator of $\tau(\theta)$ by Theorem 1.15. We are assuming that the function τ does not depend on n.

Example 1.17. Let $Y_1, ..., Y_n$ be iid with mean $E(Y_i) = \mu$ and variance $V(Y_i) = \sigma^2$. Then the sample mean $\overline{Y}_n$ is a consistent estimator of μ since i) the SLLN holds (use Theorems 1.9 and 1.15), ii) the WLLN holds, and iii) the CLT holds (use Theorem 1.8). Since

$$\lim_{n\to\infty} \mathrm{VAR}_\mu(\overline{Y}_n) = \lim_{n\to\infty} \sigma^2/n = 0 \quad \text{and} \quad \lim_{n\to\infty} E_\mu(\overline{Y}_n) = \mu,$$

$\overline{Y}_n$ is also a consistent estimator of μ by Theorem 1.7b. By the delta method and Theorem 1.8b, $T_n = g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if $g'(\mu) \neq 0$ for all $\mu \in \Theta$. By Theorem 1.15e, $g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if g is continuous at μ for all $\mu \in \Theta$.

Theorem 1.16. Assume that the function g does not depend on n.

a) **Generalized Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is such that $P[X \in C(g)] = 1$ where $C(g)$ is the set of points where g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

b) **Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

Remark 1.4. For Theorem 1.15, a) follows from Slutsky's Theorem by taking $Y_n \equiv X = Y$ and $W_n = X_n - X$. Then $Y_n \xrightarrow{D} Y = X$ and $W_n \xrightarrow{P} 0$. Hence $X_n = Y_n + W_n \xrightarrow{D} Y + 0 = X$. The convergence in distribution parts of b) and c) follow from a). Part f) follows from d) and e). Part e) implies that if T_n is a consistent estimator of θ and τ is a continuous function, then $\tau(T_n)$ is a consistent estimator of $\tau(\theta)$. Theorem 1.16 says that convergence in distribution is preserved by continuous functions, and even some discontinuities are allowed as long as the set of continuity points is assigned probability 1 by the asymptotic distribution. Equivalently, the set of discontinuity points is assigned probability 0.

Example 1.18. (Ferguson 1996, p. 40): If $X_n \xrightarrow{D} X$, then $1/X_n \xrightarrow{D} 1/X$ if X is a continuous random variable since $P(X = 0) = 0$ and $x = 0$ is the only discontinuity point of $g(x) = 1/x$.

Example 1.19. Show that if $Y_n \sim t_n$, a t distribution with n degrees of freedom, then $Y_n \xrightarrow{D} Z$ where $Z \sim N(0, 1)$.

Solution: $Y_n \stackrel{D}{=} Z/\sqrt{V_n/n}$ where $Z \perp\!\!\!\perp V_n \sim \chi^2_n$. If $W_n = \sqrt{V_n/n} \xrightarrow{P} 1$, then the result follows by Slutsky's Theorem. But $V_n \stackrel{D}{=} \sum_{i=1}^n X_i$ where the

iid $X_i \sim \chi^2_1$. Hence $V_n/n \xrightarrow{P} 1$ by the WLLN and $\sqrt{V_n/n} \xrightarrow{P} 1$ by Theorem 1.15e.

Theorem 1.17: Continuity Theorem. Let Y_n be sequence of random variables with characteristic functions $\phi_n(t)$. Let Y be a random variable with characteristic function (cf) $\phi(t)$.

a)

$$Y_n \xrightarrow{D} Y \text{ iff } \phi_n(t) \to \phi(t) \ \forall t \in \mathbb{R}.$$

b) Also assume that Y_n has moment generating function (mgf) m_n and Y has mgf m. Assume that all of the mgfs m_n and m are defined on $|t| \leq d$ for some $d > 0$. Then if $m_n(t) \to m(t)$ as $n \to \infty$ for all $|t| < c$ where $0 < c < d$, then $Y_n \xrightarrow{D} Y$.

Application: Proof of a Special Case of the CLT. Following Rohatgi (1984, pp. 569-9), let $Y_1, ..., Y_n$ be iid with mean μ, variance σ^2, and mgf $m_Y(t)$ for $|t| < t_o$. Then

$$Z_i = \frac{Y_i - \mu}{\sigma}$$

has mean 0, variance 1, and mgf $m_Z(t) = \exp(-t\mu/\sigma)m_Y(t/\sigma)$ for $|t| < \sigma t_o$. We want to show that

$$W_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1).$$

Notice that $W_n =$

$$n^{-1/2}\sum_{i=1}^{n} Z_i = n^{-1/2}\sum_{i=1}^{n}\left(\frac{Y_i - \mu}{\sigma}\right) = n^{-1/2}\frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma} = \frac{n^{-1/2}}{\frac{1}{n}}\ \frac{\overline{Y}_n - \mu}{\sigma}.$$

Thus

$$m_{W_n}(t) = E(e^{tW_n}) = E[\exp(tn^{-1/2}\sum_{i=1}^{n} Z_i)] = E[\exp(\sum_{i=1}^{n} tZ_i/\sqrt{n})]$$

$$= \prod_{i=1}^{n} E[e^{tZ_i/\sqrt{n}}] = \prod_{i=1}^{n} m_Z(t/\sqrt{n}) = [m_Z(t/\sqrt{n})]^n.$$

Set $\psi(x) = \log(m_Z(x))$. Then

$$\log[m_{W_n}(t)] = n\log[m_Z(t/\sqrt{n})] = n\psi(t/\sqrt{n}) = \frac{\psi(t/\sqrt{n})}{\frac{1}{n}}.$$

Now $\psi(0) = \log[m_Z(0)] = \log(1) = 0$. Thus by L'Hôpital's rule (where the derivative is with respect to n), $\lim_{n\to\infty} \log[m_{W_n}(t)] =$

$$\lim_{n\to\infty} \frac{\psi(t/\sqrt{n}\,)}{\frac{1}{n}} = \lim_{n\to\infty} \frac{\psi'(t/\sqrt{n}\,)[\frac{-t/2}{n^{3/2}}]}{(\frac{-1}{n^2})} = \frac{t}{2} \lim_{n\to\infty} \frac{\psi'(t/\sqrt{n}\,)}{\frac{1}{\sqrt{n}}}.$$

Now

$$\psi'(0) = \frac{m'_Z(0)}{m_Z(0)} = E(Z_i)/1 = 0,$$

so L'Hôpital's rule can be applied again, giving $\lim_{n\to\infty} \log[m_{W_n}(t)] =$

$$\frac{t}{2} \lim_{n\to\infty} \frac{\psi''(t/\sqrt{n}\,)[\frac{-t}{2n^{3/2}}]}{(\frac{-1}{2n^{3/2}})} = \frac{t^2}{2} \lim_{n\to\infty} \psi''(t/\sqrt{n}\,) = \frac{t^2}{2}\psi''(0).$$

Now

$$\psi''(t) = \frac{d}{dt}\frac{m'_Z(t)}{m_Z(t)} = \frac{m''_Z(t)m_Z(t) - (m'_Z(t))^2}{[m_Z(t)]^2}.$$

So

$$\psi''(0) = m''_Z(0) - [m'_Z(0)]^2 = E(Z_i^2) - [E(Z_i)]^2 = 1.$$

Hence $\lim_{n\to\infty} \log[m_{W_n}(t)] = t^2/2$ and

$$\lim_{n\to\infty} m_{W_n}(t) = \exp(t^2/2)$$

which is the N(0,1) mgf. Thus by the continuity theorem,

$$W_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1). \ \square$$

1.5.4 Multivariate Limit Theorems

Many of the univariate results of the previous 3 subsections can be extended to random vectors. For the limit theorems, the vector $\boldsymbol{X}$ is typically a $k \times 1$ column vector and $\boldsymbol{X}^T$ is a row vector. Let $\|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_k^2}$ be the Euclidean norm of $\boldsymbol{x}$.

Definition 1.28. Let $\boldsymbol{X}_n$ be a sequence of random vectors with joint cdfs $F_n(\boldsymbol{x})$ and let $\boldsymbol{X}$ be a random vector with joint cdf $F(\boldsymbol{x})$.

a) $\boldsymbol{X}_n$ **converges in distribution** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, if $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ as $n \to \infty$ for all points $\boldsymbol{x}$ at which $F(\boldsymbol{x})$ is continuous. The distribution of $\boldsymbol{X}$ is the **limiting distribution** or **asymptotic distribution** of $\boldsymbol{X}_n$.

b) $\boldsymbol{X}_n$ **converges in probability** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \to 0$ as $n \to \infty$.

c) Let $r > 0$ be a real number. Then $\boldsymbol{X}_n$ **converges in rth mean** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{r} \boldsymbol{X}$, if $E(\|\boldsymbol{X}_n - \boldsymbol{X}\|^r) \to 0$ as $n \to \infty$.

d) $\boldsymbol{X}_n$ **converges almost everywhere** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{ae} \boldsymbol{X}$, if $P(\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}) = 1$.

Theorems 1.18 and 1.19 below are the multivariate extensions of the limit theorems in subsection 1.5.1. When the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta}))$ is multivariate normal $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$, approximate the joint cdf of $\boldsymbol{Z}_n$ with the joint cdf of the $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$ distribution. Thus to find probabilities, manipulate $\boldsymbol{Z}_n$ as if $\boldsymbol{Z}_n \approx N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. To see that the CLT is a special case of the MCLT below, let $k = 1$, $E(X) = \mu$, and $V(X) = \boldsymbol{\Sigma}_{\boldsymbol{X}} = \sigma^2$.

Theorem 1.18: the Multivariate Central Limit Theorem (MCLT). If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}}$, then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$$

where the sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

To see that the delta method is a special case of the multivariate delta method, note that if T_n and parameter θ are real valued, then $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = g'(\theta)$.

Theorem 1.19: the Multivariate Delta Method. If $\boldsymbol{g}$ does not depend on n and

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

then

$$\sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{D} N_d(\boldsymbol{0}, \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} \boldsymbol{\Sigma} \boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})})$$

where the $d \times k$ Jacobian matrix of partial derivatives

$$\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} g_1(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial \theta_k} g_1(\boldsymbol{\theta}) \\ \vdots & & \vdots \\ \frac{\partial}{\partial \theta_1} g_d(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial \theta_k} g_d(\boldsymbol{\theta}) \end{bmatrix}.$$

Here the mapping $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^d$ needs to be differentiable in a neighborhood of $\boldsymbol{\theta} \in \mathbb{R}^k$.

Definition 1.29. If the estimator $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$, then $\boldsymbol{g}(\boldsymbol{T}_n)$ is a **consistent estimator** of $\boldsymbol{g}(\boldsymbol{\theta})$.

Theorem 1.20. If $0 < \delta \leq 1$, $\boldsymbol{X}$ is a random vector, and

$$n^{\delta}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{D} \boldsymbol{X},$$

then $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$.

Theorem 1.21. If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid, $E(\|\boldsymbol{X}\|) < \infty$, and $E(\boldsymbol{X}) = \boldsymbol{\mu}$, then
a) WLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{P} \boldsymbol{\mu}$, and
b) SLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{ae} \boldsymbol{\mu}$.

Theorem 1.22: Continuity Theorem. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors with characteristic functions $\phi_n(\boldsymbol{t})$, and let $\boldsymbol{X}$ be a $k \times 1$ random vector with cf $\phi(\boldsymbol{t})$. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \phi_{\mathrm{n}}(\boldsymbol{t}) \to \phi(\boldsymbol{t})$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

Theorem 1.23: Cramér Wold Device. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{X}$ be a $k \times 1$ random vector. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \boldsymbol{t}^{\mathrm{T}} \boldsymbol{X}_{\mathrm{n}} \xrightarrow{\mathrm{D}} \boldsymbol{t}^{\mathrm{T}} \boldsymbol{X}$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

Application: Proof of the MCLT Theorem 1.18. Note that for fixed $\boldsymbol{t}$, the $\boldsymbol{t}^T \boldsymbol{X}_i$ are iid random variables with mean $\boldsymbol{t}^T \boldsymbol{\mu}$ and variance $\boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t}$. Hence by the CLT, $\boldsymbol{t}^T \sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N(0, \boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t})$. The right hand side has distribution $\boldsymbol{t}^T \boldsymbol{X}$ where $\boldsymbol{X} \sim N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. Hence by the Cramér Wold Device, $\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. □

Theorem 1.24. a) If $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, then $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$.
b)

$$\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta}) \text{ iff } \boldsymbol{X}_{\mathrm{n}} \xrightarrow{\mathrm{D}} \boldsymbol{g}(\boldsymbol{\theta}).$$

Let $g(n) \geq 1$ be an increasing function of the sample size n: $g(n) \uparrow \infty$, e.g. $g(n) = \sqrt{n}$. See White (1984, p. 15). If a $k \times 1$ random vector $\boldsymbol{T}_n - \boldsymbol{\mu}$ converges to a nondegenerate multivariate normal distribution with convergence rate $\sqrt{n}$, then $\boldsymbol{T}_n$ has (tightness) rate $\sqrt{n}$.

Definition 1.30. Let $\boldsymbol{A}_n = [a_{i,j}(n)]$ be an $r \times c$ random matrix.
a) $\boldsymbol{A}_n = O_P(X_n)$ if $a_{i,j}(n) = O_P(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
b) $\boldsymbol{A}_n = o_p(X_n)$ if $a_{i,j}(n) = o_p(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
c) $\boldsymbol{A}_n \asymp_P (1/(g(n))$ if $a_{i,j}(n) \asymp_P (1/(g(n))$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
d) Let $\boldsymbol{A}_{1,n} = \boldsymbol{T}_n - \boldsymbol{\mu}$ and $\boldsymbol{A}_{2,n} = \boldsymbol{C}_n - c\boldsymbol{\Sigma}$ for some constant $c > 0$. If $\boldsymbol{A}_{1,n} \asymp_P (1/(g(n))$ and $\boldsymbol{A}_{2,n} \asymp_P (1/(g(n))$, then $(\boldsymbol{T}_n, \boldsymbol{C}_n)$ has (tightness) rate $g(n)$.

Theorem 1.25: Continuous Mapping Theorem. Let $\boldsymbol{X}_n \in \mathbb{R}^k$. If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and if the function $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^j$ is continuous, then $\boldsymbol{g}(\boldsymbol{X}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{X})$.

The following two theorems are taken from Severini (2005, pp. 345-349, 354).

Theorem 1.26. Let $\boldsymbol{X}_n = (X_{1n}, ..., X_{kn})^T$ be a sequence of $k \times 1$ random vectors, let $\boldsymbol{Y}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{X} = (X_1, ..., X_k)^T$ be a $k \times 1$ random vector. Let $\boldsymbol{W}_n$ be a sequence of $k \times k$ nonsingular random matrices, and let $\boldsymbol{C}$ be a $k \times k$ constant nonsingular matrix.

a) $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$ iff $X_{in} \xrightarrow{P} X_i$ for $i = 1, ..., k$.

b) **Slutsky's Theorem**: If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{Y}_n \xrightarrow{P} \boldsymbol{c}$ for some constant $k \times 1$ vector $\boldsymbol{c}$, then i) $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow{D} \boldsymbol{X} + \boldsymbol{c}$ and

ii) $\boldsymbol{Y}_n^T \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{c}^T \boldsymbol{X}$.

c) If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{C}$, then $\boldsymbol{W}_n \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}\boldsymbol{X}$, $\boldsymbol{X}_n^T \boldsymbol{W}_n \xrightarrow{D} \boldsymbol{X}^T \boldsymbol{C}$, $\boldsymbol{W}_n^{-1} \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}^{-1} \boldsymbol{X}$, and $\boldsymbol{X}_n^T \boldsymbol{W}_n^{-1} \xrightarrow{D} \boldsymbol{X}^T \boldsymbol{C}^{-1}$.

Theorem 1.27. Let W_n, X_n, Y_n, and Z_n be sequences of random variables such that $Y_n > 0$ and $Z_n > 0$. (Often Y_n and Z_n are deterministic, e.g. $Y_n = n^{-1/2}$.)

a) If $W_n = O_P(1)$ and $X_n = O_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = O_P(1)$, thus $O_P(1) + O_P(1) = O_P(1)$ and $O_P(1)O_P(1) = O_P(1)$.

b) If $W_n = O_P(1)$ and $X_n = o_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = o_P(1)$, thus $O_P(1) + o_P(1) = O_P(1)$ and $O_P(1)o_P(1) = o_P(1)$.

c) If $W_n = O_P(Y_n)$ and $X_n = O_P(Z_n)$, then $W_n + X_n = O_P(\max(Y_n, Z_n))$ and $W_n X_n = O_P(Y_n Z_n)$, thus $O_P(Y_n) + O_P(Z_n) = O_P(\max(Y_n, Z_n))$ and $O_P(Y_n)O_P(Z_n) = O_P(Y_n Z_n)$.

Theorem 1.28. i) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A}\boldsymbol{\mu}) \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

ii) Let $\boldsymbol{\Sigma} > 0$. Assume n is large enough so that $\boldsymbol{C} > 0$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ where $s > 0$ is some constant, then $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_P(1)$, so $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a consistent estimator of $s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

iii) Let $\boldsymbol{\Sigma} > 0$. Assume n is large enough so that $\boldsymbol{C} > 0$. If $\sqrt{n}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and if $\boldsymbol{C}$ is a consistent estimator of $\boldsymbol{\Sigma}$, then $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. In particular,
$n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$.

Proof: ii) $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$
$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$
$= (\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \boldsymbol{\mu}) + (\boldsymbol{x} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - T)$
$+ (\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}] (\boldsymbol{\mu} - T) + (\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \boldsymbol{\mu})$
$+ (\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}] (\boldsymbol{\mu} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(1)$.

(Note that $D^2_{\boldsymbol{x}}(T,\boldsymbol{C}) = s^{-1}D^2_{\boldsymbol{x}}(\boldsymbol{\mu},\boldsymbol{\Sigma}) + O_P(n^{-\delta})$ if $(T,\boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^{δ} where $0 < \delta \leq 0.5$ if $[\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}] = O_P(n^{-\delta})$.)

Alternatively, $D^2_{\boldsymbol{x}}(T,\boldsymbol{C})$ is a continuous function of $(T,\boldsymbol{C})$ if $\boldsymbol{C} > 0$ for $n > 10p$. Hence $D^2_{\boldsymbol{x}}(T,\boldsymbol{C}) \xrightarrow{P} D^2_{\boldsymbol{x}}(\mu, s\boldsymbol{\Sigma})$.

iii) Note that $\boldsymbol{Z}_n = \sqrt{n}\ \boldsymbol{\Sigma}^{-1/2}(T-\boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Thus $\boldsymbol{Z}_n^T\boldsymbol{Z}_n = n(T-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(T-\boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. Now $n(T-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(T-\boldsymbol{\mu}) =$
$n(T-\boldsymbol{\mu})^T[\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}](T-\boldsymbol{\mu}) = n(T-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(T-\boldsymbol{\mu}) +$
$n(T-\boldsymbol{\mu})^T[\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}](T-\boldsymbol{\mu}) = n(T-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(T-\boldsymbol{\mu}) + o_P(1) \xrightarrow{D} \chi^2_p$ since $\sqrt{n}(T-\boldsymbol{\mu})^T[\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}]\sqrt{n}(T-\boldsymbol{\mu}) = O_P(1)o_P(1)O_P(1) = o_P(1)$. $\square$

Example 1.20. Suppose that $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ for $n = 1, 2, ...$ Suppose $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$, and $\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y}$ where $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. Then

$$\begin{bmatrix} \boldsymbol{x}_n \\ \boldsymbol{y}_n \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$$

by Theorem 1.22. To see this, let $\boldsymbol{t} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T)^T$, $\boldsymbol{z}_n = (\boldsymbol{x}_n^T, \boldsymbol{y}_n^T)^T$, and $\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T$. Since $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ and $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$, the characteristic function

$$\phi_{\boldsymbol{z}_n}(\boldsymbol{t}) = \phi_{\boldsymbol{x}_n}(\boldsymbol{t}_1)\phi_{\boldsymbol{y}_n}(\boldsymbol{t}_2) \to \phi_{\boldsymbol{x}}(\boldsymbol{t}_1)\phi_{\boldsymbol{y}}(\boldsymbol{t}_2) = \phi_{\boldsymbol{z}}(\boldsymbol{t}).$$

Hence $\boldsymbol{g}(\boldsymbol{z}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{z})$ by Theorem 1.25.

Remark 1.5. In the above example, we can show $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$ instead of assuming $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. See Ferguson (1996, p. 42).

1.6 Mixture Distributions

Mixture distributions are useful for model and variable selection since $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{I_j,0}$, and the lasso estimator $\hat{\boldsymbol{\beta}}_L$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{L,\lambda_i}$ for $i = 1, ..., M$. See Sections 2.3, 3.2, and 3.6. A random vector $\boldsymbol{u}$ has a mixture distribution if $\boldsymbol{u}$ equals a random vector $\boldsymbol{u}_j$ with probability π_j for $j = 1, ..., J$. See Definition 1.8 for the population mean and population covariance matrix of a random vector.

Definition 1.31. The distribution of a $g \times 1$ random vector $\boldsymbol{u}$ is a mixture distribution if the cumulative distribution function (cdf) of $\boldsymbol{u}$ is

$$F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t}) \tag{1.22}$$

where the probabilities π_j satisfy $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \geq 2$, and $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ is the cdf of a $g \times 1$ random vector $\boldsymbol{u}_j$. Then $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j.

Theorem 1.29. Suppose $E(h(\boldsymbol{u}))$ and the $E(h(\boldsymbol{u}_j))$ exist. Then

$$E(h(\boldsymbol{u})) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]. \tag{1.23}$$

Hence

$$E(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j], \tag{1.24}$$

and $Cov(\boldsymbol{u}) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})E(\boldsymbol{u}^T) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j \boldsymbol{u}_j^T] - E(\boldsymbol{u})[E(\boldsymbol{u})]^T =$

$$\sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T. \tag{1.25}$$

If $E(\boldsymbol{u}_j) = \boldsymbol{\theta}$ for $j = 1, ..., J$, then $E(\boldsymbol{u}) = \boldsymbol{\theta}$ and

$$Cov(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j).$$

This theorem is easy to prove if the $\boldsymbol{u}_j$ are continuous random vectors with (joint) probability density functions (pdfs) $f_{\boldsymbol{u}_j}(\boldsymbol{t})$. Then $\boldsymbol{u}$ is a continuous random vector with pdf

$$f_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j f_{\boldsymbol{u}_j}(\boldsymbol{t}), \quad \text{and} \quad \mathrm{E(h(\boldsymbol{u}))} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{h}(\boldsymbol{t}) \mathrm{f}_{\boldsymbol{u}}(\boldsymbol{t}) \mathrm{d}\boldsymbol{t}$$

$$= \sum_{j=1}^{J} \pi_j \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{t}) f_{\boldsymbol{u}_j}(\boldsymbol{t}) d\boldsymbol{t} = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]$$

where $E[h(\boldsymbol{u}_j)]$ is the expectation with respect to the random vector $\boldsymbol{u}_j$. Note that

$$E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \sum_{k=1}^{J} \pi_j \pi_k E(\boldsymbol{u}_j)[E(\boldsymbol{u}_k)]^T. \tag{1.26}$$

Alternatively, with respect to a Riemann Stieltjes integral, $E[h(\boldsymbol{u})] = \int h(\boldsymbol{t}) dF(\boldsymbol{t})$ provided the expected value exists, and the integral is a linear operator with respect to both h and F. Hence for a mixture distribution, $E[h(\boldsymbol{u})] = \int h(\boldsymbol{t}) dF(\boldsymbol{t}) =$

$$\int h(\boldsymbol{t})\ d\left[\sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})\right] = \sum_{j=1}^{J} \pi_j \int h(\boldsymbol{t}) dF_{\boldsymbol{u}_j}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)].$$

1.7 A Review of Multiple Linear Regression

The following review follows Olive (2017a: ch. 2) closely. Several of the results in this section will be covered in more detail or proven in Chapter 2.

Definition 1.32. Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response variable Y given the vector of predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$.

Definition 1.33. A **quantitative variable** takes on numerical values while a **qualitative variable** takes on categorical values.

Definition 1.34. Suppose that the response variable Y and at least one predictor variable x_i are quantitative. Then the **multiple linear regression (MLR) model** is

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{1.27}$$

for $i = 1, \ldots, n$. Here n is the *sample size* and the random variable e_i is the ith *error*. Suppressing the subscript i, the model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$.

In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1.28}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent 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,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{1.29}$$

Often the first column of $\boldsymbol{X}$ is $X_1 = \boldsymbol{1}$, the $n \times 1$ vector of ones. The ith **case** $(\boldsymbol{x}_i^T, Y_i) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_i)$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element of $\boldsymbol{Y}$ (if $x_{i1} \equiv 1$, then x_{i1} could be omitted). In the MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, the Y and e are random variables, but we only have observed values Y_i and $\boldsymbol{x}_i$. If the e_i are **iid** (independent and identically distributed) with zero mean $E(e_i) = 0$ and variance $\text{VAR}(e_i) = V(e_i) = \sigma^2$, then regression is used to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2.

Definition 1.35. The **constant variance MLR model** uses the assumption that the errors $e_1, ..., e_n$ are iid with mean $E(e_i) = 0$ and variance $\text{VAR}(e_i) = \sigma^2 < \infty$. Also assume that the errors are independent of the predictor variables $\boldsymbol{x}_i$. The predictor variables $\boldsymbol{x}_i$ are assumed to be fixed and measured without error. The cases $(\boldsymbol{x}_i^T, Y_i)$ are independent for $i = 1, ..., n$.

If the predictor variables are random variables, then the above MLR model is conditional on the observed values of the $\boldsymbol{x}_i$. That is, observe the $\boldsymbol{x}_i$ and then act as if the observed $\boldsymbol{x}_i$ are fixed.

Definition 1.36. The **unimodal MLR model** has the same assumptions as the constant variance MLR model, as well as the assumption that the zero mean constant variance errors $e_1, ..., e_n$ are iid from a unimodal distribution that is not highly skewed. Note that $E(e_i) = 0$ and $V(e_i) = \sigma^2 < \infty$.

Definition 1.37. The *normal MLR model* or **Gaussian MLR model** has the same assumptions as the unimodal MLR model but adds the assumption that the errors $e_1, ..., e_n$ are iid $N(0, \sigma^2)$ random variables. That is, the e_i are iid normal random variables with zero mean and variance σ^2.

The unknown coefficients for the above 3 models are usually estimated using (ordinary) least squares (OLS).

Notation. The symbol $A \equiv B = f(c)$ means that A and B are equivalent and equal, and that $f(c)$ is the formula used to compute A and B.

Definition 1.38. Given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the corresponding vector of *predicted values* or *fitted values* is $\widehat{\boldsymbol{Y}} \equiv \widehat{\boldsymbol{Y}}(\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{b}$. Thus the ith fitted value

$$\hat{Y}_i \equiv \hat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T \boldsymbol{b} = x_{i,1} b_1 + \cdots + x_{i,p} b_p.$$

The vector of *residuals* is $\boldsymbol{r} \equiv \boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\boldsymbol{b})$. Thus ith residual $r_i \equiv r_i(\boldsymbol{b}) = Y_i - \hat{Y}_i(\boldsymbol{b}) = Y_i - x_{i,1} b_1 - \cdots - x_{i,p} b_p$.

Most regression methods attempt to find an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ which minimizes some criterion function $Q(\boldsymbol{b})$ of the residuals.

Definition 1.39. The *ordinary least squares (OLS) estimator* $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes

$$Q_{OLS}(\boldsymbol{b}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{b}), \tag{1.30}$$

$$\text{and} \quad \hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}.$$

The vector of *predicted* or *fitted values* $\widehat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$ provided the inverse exists. Typically the subscript OLS is omitted, and the least squares *regression equation* is

$\hat{Y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$ where $x_1 \equiv 1$ if the model contains a constant.

Definition 1.40. For MLR, the *response plot* is a plot of the ESP = fitted values $= \hat{Y}_i$ versus the response Y_i, while the *residual plot* is a plot of the ESP $= \hat{Y}_i$ versus the residuals r_i.

Theorem 1.30. Suppose that the regression estimator $\boldsymbol{b}$ of $\boldsymbol{\beta}$ is used to find the residuals $r_i \equiv r_i(\boldsymbol{b})$ and the fitted values $\widehat{Y}_i \equiv \widehat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T \boldsymbol{b}$. Then in the response plot of $\widehat{Y}_i$ versus Y_i, the vertical deviations from the identity line (that has unit slope and zero intercept) are the residuals $r_i(\boldsymbol{b})$.

Proof. The identity line in the response plot is $Y = \boldsymbol{x}^T \boldsymbol{b}$. Hence the vertical deviation is $Y_i - \boldsymbol{x}_i^T \boldsymbol{b} = r_i(\boldsymbol{b})$. □

The results in the following theorem are properties of least squares (OLS), not of the underlying MLR model. Definitions 1.38 and 1.39 define the hat matrix $\boldsymbol{H}$, vector of fitted values $\hat{\boldsymbol{Y}}$, and vector of residuals $\boldsymbol{r}$. Parts f) and g) make residual plots useful. If the plotted points are linear with roughly constant variance and the correlation is zero, then the plotted points scatter about the $r = 0$ line with no other pattern. If the plotted points in a residual plot of w versus r do show a pattern such as a curve or a right opening megaphone, zero correlation will usually force symmetry about either the $r = 0$ line or the $w = \text{median}(w)$ line. Hence departures from the ideal plot of random scatter about the $r = 0$ line are often easy to detect.

Let the $n \times p$ design matrix of predictor variables be

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

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

Warning: If $n > p$, as is usually the case for the full rank linear model, $\boldsymbol{X}$ is not square, so $(\boldsymbol{X}^T \boldsymbol{X})^{-1} \neq \boldsymbol{X}^{-1} (\boldsymbol{X}^T)^{-1}$ since $\boldsymbol{X}^{-1}$ does not exist.

Theorem 1.31. Suppose that $\boldsymbol{X}$ is an $n \times p$ matrix of full rank p. Then

a) $\boldsymbol{H}$ is symmetric: $\boldsymbol{H} = \boldsymbol{H}^T$.

b) $\boldsymbol{H}$ is idempotent: $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$.

c) $\boldsymbol{X}^T \boldsymbol{r} = \boldsymbol{0}$ so that $\boldsymbol{v}_j^T \boldsymbol{r} = 0$.

d) If there is a constant $\boldsymbol{v}_1 = \boldsymbol{1}$ in the model, then the sum of the residuals is zero: $\sum_{i=1}^n r_i = 0$.

e) $\boldsymbol{r}^T \hat{\boldsymbol{Y}} = 0$.

f) If there is a constant in the model, then the sample correlation of the fitted values and the residuals is 0: $\text{corr}(\boldsymbol{r}, \hat{\boldsymbol{Y}}) = 0$.

g) If there is a constant in the model, then the sample correlation of the jth predictor with the residuals is 0: $\text{corr}(\boldsymbol{r}, \boldsymbol{v}_j) = 0$ for $j = 1, ..., p$.

Proof. a) $\boldsymbol{X}^T\boldsymbol{X}$ is symmetric since $(\boldsymbol{X}^T\boldsymbol{X})^T = \boldsymbol{X}^T(\boldsymbol{X}^T)^T = \boldsymbol{X}^T\boldsymbol{X}$. Hence $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is symmetric since the inverse of a symmetric matrix is symmetric. (Recall that if $\boldsymbol{A}$ has an inverse then $(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T$.) Thus using $(\boldsymbol{A}^T)^T = \boldsymbol{A}$ and $(\boldsymbol{ABC})^T = \boldsymbol{C}^T\boldsymbol{B}^T\boldsymbol{A}^T$ shows that

$$\boldsymbol{H}^T = \boldsymbol{X}^T[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]^T(\boldsymbol{X}^T)^T = \boldsymbol{H}.$$

b) $\boldsymbol{HH} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H}$ since $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_p$, the $p \times p$ identity matrix.

c) $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{X}^T(\boldsymbol{I}_p - \boldsymbol{H})\boldsymbol{Y} = [\boldsymbol{X}^T - \boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Y} = [\boldsymbol{X}^T - \boldsymbol{X}^T]\boldsymbol{Y} = \boldsymbol{0}$. Since $\boldsymbol{v}_j$ is the jth column of $\boldsymbol{X}$, $\boldsymbol{v}_j^T$ is the jth row of $\boldsymbol{X}^T$ and $\boldsymbol{v}_j^T\boldsymbol{r} = 0$ for $j = 1, ..., p$.

d) Since $\boldsymbol{v}_1 = \boldsymbol{1}$, $\boldsymbol{v}_1^T\boldsymbol{r} = \sum_{i=1}^n r_i = 0$ by c).

e) $\boldsymbol{r}^T\hat{\boldsymbol{Y}} = [(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}]^T\boldsymbol{HY} = \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{HY} = \boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H})\boldsymbol{Y} = 0$.

f) The sample correlation between W and Z is $\text{corr}(W, Z) =$

$$\frac{\sum_{i=1}^n (w_i - \overline{w})(z_i - \overline{z})}{(n-1)s_w s_z} = \frac{\sum_{i=1}^n (w_i - \overline{w})(z_i - \overline{z})}{\sqrt{\sum_{i=1}^n (w_i - \overline{w})^2 \sum_{i=1}^n (z_i - \overline{z})^2}}$$

where s_m is the sample standard deviation of m for $m = w, z$. So the result follows if $A = \sum_{i=1}^n (\hat{Y}_i - \overline{\hat{Y}})(r_i - \overline{r}) = 0$. Now $\overline{r} = 0$ by d), and thus

$$A = \sum_{i=1}^n \hat{Y}_i r_i - \overline{\hat{Y}} \sum_{i=1}^n r_i = \sum_{i=1}^n \hat{Y}_i r_i$$

by d) again. But $\sum_{i=1}^n \hat{Y}_i r_i = \boldsymbol{r}^T\hat{\boldsymbol{Y}} = 0$ by e).

g) Following the argument in f), the result follows if $A = \sum_{i=1}^n (x_{i,j} - \overline{x}_j)(r_i - \overline{r}) = 0$ where $\overline{x}_j = \sum_{i=1}^n x_{i,j}/n$ is the sample mean of the jth predictor. Now $\overline{r} = \sum_{i=1}^n r_i/n = 0$ by d), and thus

$$A = \sum_{i=1}^n x_{i,j} r_i - \overline{x}_j \sum_{i=1}^n r_i = \sum_{i=1}^n x_{i,j} r_i$$

by d) again. But $\sum_{i=1}^n x_{i,j} r_i = \boldsymbol{v}_j^T\boldsymbol{r} = 0$ by c). □

1.7.1 The ANOVA F Test

After fitting least squares and checking the response and residual plots to see that an MLR model is reasonable, the next step is to check whether there is

an MLR relationship between Y and the nontrivial predictors x_2, ..., x_p. If at least one of these predictors is useful, then the OLS fitted values $\hat{Y}_i$ should be used. If none of the nontrivial predictors is useful, then $\overline{Y}$ will give as good predictions as $\hat{Y}_i$. Here the *sample mean* $\overline{Y}$ is given by Definition 1.9. In the definition below, SSE is the sum of squared residuals and a residual $r_i = \hat{e}_i =$ "errorhat." In the literature "errorhat" is often rather misleadingly abbreviated as "error."

Definition 1.41. Assume that a constant is in the MLR model.
a) The *total sum of squares*

$$SSTO = \sum_{i=1}^{n} (Y_i - \overline{Y})^2. \tag{1.31}$$

b) The *regression sum of squares*

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2. \tag{1.32}$$

c) The residual sum of squares or *error sum of squares* is

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \tag{1.33}$$

The result in the following theorem is a property of least squares (OLS), not of the underlying MLR model. An obvious application is that given any two of SSTO, SSE, and SSR, the 3rd sum of squares can be found using the formula $SSTO = SSE + SSR$.

Theorem 1.32. Assume that a constant is in the MLR model. Then $SSTO = SSE + SSR$.

Proof.

$$SSTO = \sum_{i=1}^{n} (Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})^2 = SSE + SSR + 2\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}).$$

Hence the result follows if

$$A \equiv \sum_{i=1}^{n} r_i (\hat{Y}_i - \overline{Y}) = 0.$$

But

$$A = \sum_{i=1}^{n} r_i \hat{Y}_i - \overline{Y} \sum_{i=1}^{n} r_i = 0$$

by Theorem 1.31 d) and e). □

Definition 1.42. Assume that a constant is in the MLR model and that SSTO $\neq 0$. The **coefficient of multiple determination**

$$R^2 = [\mathrm{corr}(\mathrm{Y_i}, \hat{\mathrm{Y}}_\mathrm{i})]^2 = \frac{\mathrm{SSR}}{\mathrm{SSTO}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SSTO}}$$

where $\mathrm{corr}(\mathrm{Y_i}, \hat{\mathrm{Y}}_\mathrm{i})$ is the sample correlation of Y_i and $\hat{Y}_i$.

Warnings: i) $0 \leq R^2 \leq 1$, but small R^2 does not imply that the MLR model is bad.

ii) If the MLR model contains a constant, then there are several equivalent formulas for R^2. If the model does not contain a constant, then R^2 depends on the software package.

iii) R^2 does not have much meaning unless the response plot and residual plot both look good.

iv) R^2 tends to be too high if n is small.

v) R^2 tends to be too high if there are two or more separated clusters of data in the response plot.

vi) R^2 is too high if the number of predictors p is close to n.

vii) In large samples R^2 will be large (close to one) if σ^2 is small compared to the sample variance S_Y^2 of the response variable Y. R^2 is also large if the sample variance of $\hat{Y}$ is close to S_Y^2. Thus R^2 is sometimes interpreted as the proportion of the variability of Y explained by conditioning on $\boldsymbol{x}$, but warnings i) - v) suggest that R^2 may not have much meaning.

The following 2 theorems suggest that R^2 does not behave well when many predictors that are not needed in the model are included in the model. Such a variable is sometimes called a noise variable and the MLR model is "fitting noise." Theorem 1.34 appears, for example, in Cramér (1946, pp. 414-415), and suggests that R^2 should be considerably larger than p/n if the predictors are useful. Note that if $n = 10p$ and $p \geq 2$, then under the conditions of Theorem 1.34, $E(R^2) \leq 0.1$.

Theorem 1.33. Assume that a constant is in the MLR model. Adding a variable to the MLR model does not decrease (and usually increases) R^2.

Theorem 1.34. Assume that a constant β_1 is in the MLR model, that $\beta_2 = \cdots = \beta_p = 0$ and that the e_i are iid $N(0, \sigma^2)$. Hence the Y_i are iid $N(\beta_1, \sigma^2)$. Then

a) R^2 follows a beta distribution: $R^2 \sim \mathrm{beta}(\frac{\mathrm{p-1}}{2}, \frac{\mathrm{n-p}}{2})$.

b)

$$E(R^2) = \frac{p-1}{n-1}.$$

c)

$$\mathrm{VAR}(R^2) = \frac{2(p-1)(n-p)}{(n-1)^2(n+1)}.$$

Notice that each SS/n estimates the variability of some quantity. $SSTO/n \approx S_Y^2$, $SSE/n \approx S_e^2 = \sigma^2$, and $SSR/n \approx S_{\hat{Y}}^2$.

Definition 1.43. Assume that a constant is in the MLR model. Associated with each SS in Definition 1.41 is a degrees of freedom (df) and a mean square $= SS/df$. For SSTO, df $= n-1$ and $MSTO = SSTO/(n-1)$. For SSR, df $= p-1$ and $MSR = SSR/(p-1)$. For SSE, $df = n-p$ and $MSE = SSE/(n-p)$.

Under mild conditions, if the MLR model is appropriate, then MSE is a $\sqrt{n}$ consistent estimator of σ^2 by Su and Cook (2012).

The ANOVA F test tests whether any of the nontrivial predictors $x_2, ..., x_p$ are needed in the OLS MLR model, that is, whether Y_i should be predicted by the OLS fit $\hat{Y}_i = \hat{\beta}_1 + x_{i,2}\hat{\beta}_2 + \cdots + x_{i,p}\hat{\beta}_p$ or with the sample mean $\overline{Y}$. ANOVA stands for analysis of variance, and the computer output needed to perform the test is contained in the ANOVA table. Below is an ANOVA table given in symbols. Sometimes "Regression" is replaced by "Model" and "Residual" by "Error."

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Regression	$p-1$	SSR	MSR	F_0=MSR/MSE	for H_0:
Residual	$n-p$	SSE	MSE		$\beta_2 = \cdots = \beta_p = 0$

Remark 1.6. Recall that for a 4 step test of hypotheses, the p–value is the probability of getting a test statistic as extreme as the test statistic actually observed and that H_0 is rejected if the p–value $< \delta$. As a benchmark for this textbook, use $\delta = 0.05$ if δ is not given. The 4th step is the nontechnical conclusion which is crucial for presenting your results to people who are not familiar with MLR. Replace Y and $x_2, ..., x_p$ by the actual variables used in the MLR model.

Notation. The p–value $\equiv$ pvalue given by output tends to only be correct for the normal MLR model. Hence the output is usually only giving an estimate of the pvalue, which will often be denoted by *pval*. So reject H_0 if pval $\leq \delta$. Often

$$\text{pval} - \text{pvalue} \xrightarrow{\text{P}} 0$$

(converges to 0 in probability, so pval is a consistent estimator of pvalue) as the sample size $n \to \infty$. See Section 1.4. Then the computer output pval is a good estimator of the unknown pvalue. We will use $Fo \equiv F_0$, $Ho \equiv H_0$, and $Ha \equiv H_A \equiv H_1$.

The 4 step ANOVA F test of hypotheses is below.
i) State the hypotheses $H_0 : \beta_2 = \cdots = \beta_p = 0$ H_A: not H_0.

ii) Find the test statistic $F_0 = MSR/MSE$ or obtain it from output.
iii) Find the pval from output or use the F–table: pval =

$$P(F_{p-1,n-p} > F_0).$$

iv) State whether you reject H_0 or fail to reject H_0. If H_0 is rejected, conclude that there is an MLR relationship between Y and the predictors x_2, ..., x_p. If you fail to reject H_0, conclude that there is not an MLR relationship between Y and the predictors x_2, ..., x_p. (Or there is not enough evidence to conclude that there is an MLR relationship between Y and the predictors.)

Some assumptions are needed on the ANOVA F test. Assume that both the response and residual plots look good. It is crucial that there are no outliers. Then a rule of thumb is that if $n - p$ is large, then the ANOVA F test p–value is approximately correct. An analogy can be made with the central limit theorem, $\overline{Y}$ is a good estimator for μ if the Y_i are iid $N(\mu, \sigma^2)$ and also a good estimator for μ if the data are iid with mean μ and variance σ^2 if n is large enough.

If all of the $\boldsymbol{x}_i$ are different (no replication) and if the number of predictors $p = n$, then the OLS fit $\hat{Y}_i = Y_i$ and $R^2 = 1$. Notice that H_0 is rejected if the statistic F_0 is large. More precisely, reject H_0 if

$$F_0 > F_{p-1,n-p,1-\delta}$$

where

$$P(F \leq F_{p-1,n-p,1-\delta}) = 1 - \delta$$

when $F \sim F_{p-1,n-p}$. Since R^2 increases to 1 while $(n-p)/(p-1)$ decreases to 0 as p increases to n, Theorem 1.35a below implies that if p is large then the F_0 statistic may be small even if some of the predictors are very good. It is a good idea to use $n \geq 10p$ or at least $n \geq 5p$ if possible.

Theorem 1.35. Assume that the MLR model has a constant β_1.
a)

$$F_0 = \frac{MSR}{MSE} = \frac{R^2}{1 - R^2} \quad \frac{n-p}{p-1}.$$

b) If the errors e_i are iid $N(0, \sigma^2)$, and if $H_0 : \beta_2 = \cdots = \beta_p = 0$ is true, then F_0 has an F distribution with $p - 1$ numerator and $n - p$ denominator degrees of freedom: $F_0 \sim F_{p-1,n-p}$.

c) If the errors are iid with mean 0 and variance σ^2, if the error distribution is close to normal, and if $n - p$ is large enough, and if H_0 is true, then $F_0 \approx F_{p-1,n-p}$ in that the p-value from the software (pval) is approximately correct.

Remark 1.7. When a constant is not contained in the model (i.e. $x_{i,1}$ is not equal to 1 for all i), then the computer output still produces an ANOVA

table with the test statistic and p–value, and nearly the same 4 step test of hypotheses can be used. The hypotheses are now $H_0 : \beta_1 = \cdots = \beta_p = 0$ H_A: not H_0, and you are testing whether or not there is an MLR relationship between Y and $x_1, ..., x_p$. An MLR model without a constant (no intercept) is sometimes called a "regression through the origin." See Section 1.7.5.

1.7.2 The Partial F Test

Suppose that there is data on variables Z, $w_1, ..., w_r$ and that a useful MLR model has been made using $Y = t(Z)$, $x_1 \equiv 1$, $x_2, ..., x_p$ where each x_i is some function of $w_1, ..., w_r$. This useful model will be called the full model. It is important to realize that the full model does not need to use every variable w_j that was collected. For example, variables with outliers or missing values may not be used. Forming a useful full model is often very difficult, and it is often not reasonable to assume that the candidate full model is good based on a single data set, especially if the model is to be used for prediction.

Even if the full model is useful, the investigator will often be interested in checking whether a model that uses fewer predictors will work just as well. For example, perhaps x_p is a very expensive predictor but is not needed given that $x_1, ..., x_{p-1}$ are in the model. Also a model with fewer predictors tends to be easier to understand.

Definition 1.44. Let the **full model** use Y, $x_1 \equiv 1$, $x_2, ..., x_p$ and let the **reduced model** use Y, x_1, $x_{i_2}, ..., x_{i_q}$ where $\{i_2, ..., i_q\} \subset \{2, ..., p\}$.

The partial F test is used to test whether the reduced model is good in that it can be used instead of the full model. It is crucial that the reduced and full models be selected before looking at the data. If the reduced model is selected after looking at the full model output and discarding the worst variables, then the p–value for the partial F test will be too high. If the data needs to be looked at to build the full model, as is often the case, data splitting is useful.

For (ordinary) least squares, usually a constant is used, and we are assuming that both the full model and the reduced model contain a constant. The partial F test has null hypothesis $H_0 : \beta_{i_{q+1}} = \cdots = \beta_{i_p} = 0$, and alternative hypothesis H_A : at least one of the $\beta_{i_j} \neq 0$ for $j > q$. The null hypothesis is equivalent to H_0: "the reduced model is good." Since only the full model and reduced model are being compared, the alternative hypothesis is equivalent to H_A: "the reduced model is not as good as the full model, so use the full model," or more simply, H_A : "use the full model."

To perform the partial F test, fit the full model and the reduced model and obtain the ANOVA table for each model. The quantities df_F, SSE(F) and MSE(F) are for the full model and the corresponding quantities from

the reduced model use an R instead of an F. Hence SSE(F) and SSE(R) are the residual sums of squares for the full and reduced models, respectively. Shown below is output only using symbols.
Full model

Source df	SS	MS	F_0 and p-value
Regression $p-1$	SSR	MSR	F_0=MSR/MSE
Residual $df_F = n-p$	SSE(F)	MSE(F)	for $H_0 : \beta_2 = \cdots = \beta_p = 0$

Reduced model

Source df	SS	MS	F_0 and p-value
Regression $q-1$	SSR	MSR	F_0=MSR/MSE
Residual $df_R = n-q$	SSE(R)	MSE(R)	for $H_0 : \beta_2 = \cdots = \beta_q = 0$

The 4 step partial F test of hypotheses is below. i) State the hypotheses. H_0: the reduced model is good H_A: use the full model
ii) Find the test statistic. $F_R =$

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

iii) Find the pval = P($F_{df_R - df_F, df_F} > F_R$). (Here $df_R - df_F = p - q$ = number of parameters set to 0, and $df_F = n - p$, while pval is the estimated p–value.)
iv) State whether you reject H_0 or fail to reject H_0. Reject H_0 if the pval $\leq \delta$ and conclude that the full model should be used. Otherwise, fail to reject H_0 and conclude that the reduced model is good.

Sometimes software has a shortcut. In particular, the R software uses the `anova` command. As an example, assume that the full model uses x_2 and x_3 while the reduced model uses x_2. Both models contain a constant. Then the following commands will perform the partial F test. (On the computer screen the second command looks more like
red <− lm(y∼x2).)

```
full <- lm(y~x2+x3)
red <- lm(y~x2)
anova(red,full)
```

For an $n \times 1$ vector $\boldsymbol{a}$, let

$$\|\boldsymbol{a}\| = \sqrt{a_1^2 + \cdots + a_n^2} = \sqrt{\boldsymbol{a}^T \boldsymbol{a}}$$

be the Euclidean norm of $\boldsymbol{a}$. If $\boldsymbol{r}$ and $\boldsymbol{r}_R$ are the vector of residuals from the full and reduced models, respectively, notice that $SSE(F) = \|\boldsymbol{r}\|^2$ and $SSE(R) = \|\boldsymbol{r}_R\|^2$.

The following theorem suggests that H_0 is rejected in the partial F test if the change in residual sum of squares $SSE(R) - SSE(F)$ is large compared to $SSE(F)$. If the change is small, then F_R is small and the test suggests that the reduced model can be used.

Theorem 1.36. Let R^2 and R_R^2 be the multiple coefficients of determination for the full and reduced models, respectively. Let $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Y}}_R$ be the vectors of fitted values for the full and reduced models, respectively. Then the test statistic in the partial F test is

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

$$\left[\frac{\|\hat{\boldsymbol{Y}}\|^2 - \|\hat{\boldsymbol{Y}}_R\|^2}{df_R - df_F}\right] / MSE(F) =$$

$$\frac{SSE(R) - SSE(F)}{SSE(F)} \ \frac{n-p}{p-q} = \frac{R^2 - R_R^2}{1 - R^2} \ \frac{n-p}{p-q}.$$

Definition 1.45. An **FF plot** is a plot of fitted values from 2 different models or fitting methods. An **RR plot** is a plot of residuals from 2 different models or fitting methods.

Six plots are useful diagnostics for the partial F test: the RR plot with the full model residuals on the vertical axis and the reduced model residuals on the horizontal axis, the FF plot with the full model fitted values on the vertical axis, and always make the response and residual plots for the full and reduced models. Suppose that the full model is a useful MLR model. If the reduced model is good, then the response plots from the full and reduced models should be very similar, visually. Similarly, the residual plots from the full and reduced models should be very similar, visually. Finally, the correlation of the plotted points in the RR and FF plots should be high, ≥ 0.95, say, and the plotted points in the RR and FF plots should cluster tightly about the identity line. Add the identity line to both the RR and FF plots as a visual aid. Also add the OLS line from regressing $\boldsymbol{r}$ on $\boldsymbol{r}_R$ to the RR plot (the OLS line is the identity line in the FF plot). If the reduced model is good, then the OLS line should nearly coincide with the identity line in that it should be difficult to see that the two lines intersect at the origin. If the FF plot looks good but the RR plot does not, the reduced model may be good if the main goal of the analysis is to predict Y. These plots are also useful for other methods such as lasso.

1.7.3 The Wald t Test

Often investigators hope to examine β_k in order to determine the importance of the predictor x_k in the model; however, β_k is the coefficient for x_k given that the other predictors are in the model. Hence β_k depends strongly on the other predictors in the model. Suppose that the model has an intercept: $x_1 \equiv 1$. The predictor x_k is highly correlated with the other predictors if the OLS regression of x_k on $x_1, ..., x_{k-1}, x_{k+1}, ..., x_p$ has a high coefficient of determination R_k^2. If this is the case, then often x_k is not needed in the model given that the other predictors are in the model. If at least one R_k^2 is high for $k \geq 2$, then there is multicollinearity among the predictors.

As an example, suppose that $Y =$ *height*, $x_1 \equiv 1$, $x_2 =$ *left leg length*, and $x_3 =$ *right leg length*. Then x_2 should not be needed given x_3 is in the model and $\beta_2 = 0$ is reasonable. Similarly $\beta_3 = 0$ is reasonable. On the other hand, if the model only contains x_1 and x_2, then x_2 is extremely important with β_2 near 2. If the model contains x_1, x_2, x_3, $x_4 =$ *height at shoulder*, $x_5 =$ *right arm length*, $x_6 =$ *head length*, and $x_7 =$ *length of back*, then R_i^2 may be high for each $i \geq 2$. Hence x_i is not needed in the MLR model for Y given that the other predictors are in the model.

Definition 1.46. The 100 $(1 - \delta)$ % CI for β_k is $\hat{\beta}_k \pm t_{n-p,1-\delta/2}\ se(\hat{\beta}_k)$. If the degrees of freedom $d = n - p \geq 30$, the N(0,1) cutoff $z_{1-\delta/2}$ may be used.

Know how to do the 4 step Wald t–test of hypotheses.
i) State the hypotheses $H_0 : \beta_k = 0$ $H_A : \beta_k \neq 0$.
ii) Find the test statistic $t_{o,k} = \hat{\beta}_k / se(\hat{\beta}_k)$ or obtain it from output.
iii) Find pval from output or use the t–table: pval =

$$2P(t_{n-p} < -|t_{o,k}|) = 2P(t_{n-p} > |t_{o,k}|).$$

Use the normal table or the $d = Z$ line in the t–table if the degrees of freedom $d = n - p \geq 30$. Again pval is the estimated p–value.
iv) State whether you reject H_0 or fail to reject H_0 and give a nontechnical sentence restating your conclusion in terms of the story problem.

Recall that H_0 is rejected if the pval $\leq \delta$. As a benchmark for this textbook, use $\delta = 0.05$ if δ is not given. If H_0 is rejected, then conclude that x_k is needed in the MLR model for Y given that the other predictors are in the model. If you fail to reject H_0, then conclude that x_k is not needed in the MLR model for Y given that the other predictors are in the model. (Or there is not enough evidence to conclude that x_k is needed in the MLR model given that the other predictors are in the model.) Note that x_k could be a very useful individual predictor, but may not be needed if other predictors are added to the model.

1.7.4 The OLS Criterion

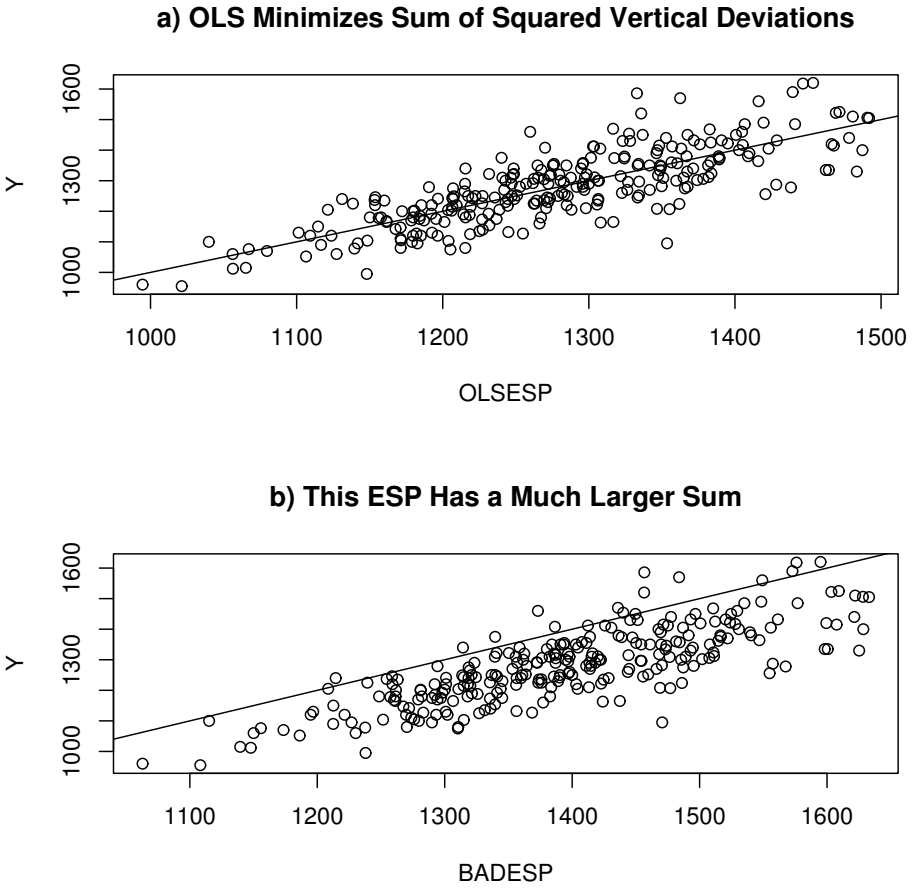

Fig. 1.8 The OLS Fit Minimizes the Sum of Squared Residuals

The OLS estimator $\hat{\boldsymbol{\beta}}$ minimizes the OLS criterion

$$Q_{OLS}(\boldsymbol{\eta}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{\eta})$$

where the residual $r_i(\boldsymbol{\eta}) = Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta}$. In other words, let $r_i = r_i(\hat{\boldsymbol{\beta}})$ be the OLS residuals. Then $\sum_{i=1}^{n} r_i^2 \leq \sum_{i=1}^{n} r_i^2(\boldsymbol{\eta})$ for any $p \times 1$ vector $\boldsymbol{\eta}$, and the equality holds (if and only if) iff $\boldsymbol{\eta} = \hat{\boldsymbol{\beta}}$ if the $n \times p$ design matrix $\boldsymbol{X}$ is of full rank $p \leq n$.

In particular, if $\boldsymbol{X}$ has full rank p, then $\sum_{i=1}^{n} r_i^2 < \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2$ even if the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ is a good approximation to the data.

Warning: Often $\boldsymbol{\eta}$ is replaced by $\boldsymbol{\beta}$: $Q_{OLS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta})$. This notation is often used in Statistics when there are estimating equations. For example, maximum likelihood estimation uses the log likelihood $\log(L(\boldsymbol{\theta}))$ where $\boldsymbol{\theta}$ is the vector of unknown parameters and the dummy variable in the log likelihood.

Example 1.21. When a model depends on the predictors $\boldsymbol{x}$ only through the linear combination $\boldsymbol{x}^T\boldsymbol{\beta}$, then $\boldsymbol{x}^T\boldsymbol{\beta}$ is called a sufficient predictor and $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ is called an estimated sufficient predictor (ESP). For OLS the model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, and the fitted value $\hat{Y} = ESP$. To illustrate the OLS criterion graphically, consider the Gladstone (1905) data where we used *brain weight* as the response. A constant, $x_2 = age$, $x_3 = sex$, and $x_4 = (size)^{1/3}$ were used as predictors after deleting five "infants" from the data set. In Figure 1.8a, the OLS response plot of the OLS ESP $= \hat{Y}$ versus Y is shown. The vertical deviations from the identity line are the residuals, and OLS minimizes the sum of squared residuals. If any other ESP $\boldsymbol{x}^T\boldsymbol{\eta}$ is plotted versus Y, then the vertical deviations from the identity line are the residuals $r_i(\boldsymbol{\eta})$. For this data, the OLS estimator $\hat{\boldsymbol{\beta}} = (498.726, -1.597, 30.462, 0.696)^T$. Figure 1.8b shows the response plot using the ESP $\boldsymbol{x}^T\boldsymbol{\eta}$ where $\boldsymbol{\eta} = (498.726, -1.597, 30.462, 0.796)^T$. Hence only the coefficient for x_4 was changed; however, the residuals $r_i(\boldsymbol{\eta})$ in the resulting plot are much larger in magnitude on average than the residuals in the OLS response plot. With slightly larger changes in the OLS ESP, the resulting $\boldsymbol{\eta}$ will be such that the squared residuals are massive.

Theorem 1.37. The OLS estimator $\hat{\boldsymbol{\beta}}$ is the unique minimizer of the OLS criterion if $\boldsymbol{X}$ has full rank $p \leq n$.

Proof: Seber and Lee (2003, pp. 36-37). Recall that the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and notice that $(\boldsymbol{I}-\boldsymbol{H})^T = \boldsymbol{I}-\boldsymbol{H}$, that $(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{H} = \boldsymbol{0}$ and that $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$. Let $\boldsymbol{\eta}$ be any $p \times 1$ vector. Then

$$(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}) = (\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y})^T(\boldsymbol{H}\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{X}\boldsymbol{\eta}) =$$

$$\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{H}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}) = \boldsymbol{0}.$$

$$\text{Thus } \mathrm{Q_{OLS}}(\boldsymbol{\eta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}\|^2 =$$

$$\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}\|^2 + 2(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}).$$

Hence

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}\|^2. \tag{1.34}$$

So

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2 \geq \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$$

with equality iff

$$\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\eta}) = \boldsymbol{0}$$

iff $\hat{\boldsymbol{\beta}} = \boldsymbol{\eta}$ since $\boldsymbol{X}$ is full rank. $\square$

Alternatively calculus can be used. Notice that $r_i(\boldsymbol{\eta}) = Y_i - x_{i,1}\eta_1 - x_{i,2}\eta_2 - \cdots - x_{i,p}\eta_p$. Recall that $\boldsymbol{x}_i^T$ is the ith row of $\boldsymbol{X}$ while $\boldsymbol{v}_j$ is the jth column. Since $Q_{OLS}(\boldsymbol{\eta}) =$

$$\sum_{i=1}^{n}(Y_i - x_{i,1}\eta_1 - x_{i,2}\eta_2 - \cdots - x_{i,p}\eta_p)^2,$$

the jth partial derivative

$$\frac{\partial Q_{OLS}(\boldsymbol{\eta})}{\partial \eta_j} = -2\sum_{i=1}^{n} x_{i,j}(Y_i - x_{i,1}\eta_1 - x_{i,2}\eta_2 - \cdots - x_{i,p}\eta_p) = -2(\boldsymbol{v}_j)^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta})$$

for $j = 1, ..., p$. Combining these equations into matrix form, setting the derivative to zero and calling the solution $\hat{\boldsymbol{\beta}}$ gives

$$\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{0},$$

or

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}. \tag{1.35}$$

Equation (1.35) is known as the **normal equations**. If $\boldsymbol{X}$ has full rank then $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. To show that $\hat{\boldsymbol{\beta}}$ is the global minimizer of the OLS criterion, use the argument following Equation (1.34).

1.7.5 The No Intercept MLR Model

The *no intercept MLR model*, also known as *regression through the origin*, is still $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, but there is no intercept in the model, so $\boldsymbol{X}$ does not contain a column of ones $\boldsymbol{1}$. Hence the intercept term $\beta_1 = \beta_1(1)$ is replaced by $\beta_1 x_{i1}$. Software gives output for this model if the "no intercept" or "intercept = F" option is selected. For the no intercept model, the assumption $E(\boldsymbol{e}) = \boldsymbol{0}$ is important, and this assumption is rather strong.

Many of the usual MLR results still hold: $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, the vector of *predicted fitted values* $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ provided the inverse exists, and the vector of residuals is $\boldsymbol{r} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$. The response plot and residual plot are made in the same way and should be made before performing inference.

The main difference in the output is the ANOVA table. The ANOVA F test in Section 1.7.1 tests $H_0 : \beta_2 = \cdots = \beta_p = 0$. The test in this subsection tests $H_0 : \beta_1 = \cdots = \beta_p = 0 \equiv H_0 : \boldsymbol{\beta} = \boldsymbol{0}$. The following definition and test follows Guttman (1982, p. 147) closely.

Definition 1.47. Assume that $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where the e_i are iid. Assume that it is desired to test $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ versus $H_A : \boldsymbol{\beta} \neq \boldsymbol{0}$.

a) The *uncorrected total sum of squares*

$$SST = \sum_{i=1}^{n} Y_i^2. \tag{1.36}$$

b) The *model sum of squares*

$$SSM = \sum_{i=1}^{n} \hat{Y}_i^2. \tag{1.37}$$

c) The residual sum of squares or *error sum of squares* is

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \tag{1.38}$$

d) The degrees of freedom (df) for SSM is p, the df for SSE is $n - p$ and the df for SST is n. The mean squares are MSE = SSE/$(n - p)$ and MSM = SSM/p.

The ANOVA table given for the "no intercept" or "intercept = F" option is below.

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Model	p	SSM	MSM	F_0=MSM/MSE	for H_0:
Residual	$n - p$	SSE	MSE		$\boldsymbol{\beta} = \boldsymbol{0}$

The 4 step no intercept ANOVA F **test for** $\boldsymbol{\beta} = \boldsymbol{0}$ is below.
i) State the hypotheses $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$, $H_A : \boldsymbol{\beta} \neq \boldsymbol{0}$.
ii) Find the test statistic $F_0 = MSM/MSE$ or obtain it from output.
iii) Find the pval from output or use the F–table: pval = $P(F_{p,n-p} > F_0)$.
iv) State whether you reject H_0 or fail to reject H_0. If H_0 is rejected, conclude that there is an MLR relationship between Y and the predictors x_1, ..., x_p. If you fail to reject H_0, conclude that there is not an MLR relationship between Y and the predictors x_1, ..., x_p. (Or there is not enough evidence to conclude that there is an MLR relationship between Y and the predictors.)

1.8 Summary

1) Statistical Learning techniques extract information from multivariate data. A **case** or **observation** consists of k random variables measured for one

person or thing. The ith case $\boldsymbol{z}_i = (z_{i1}, ..., z_{ik})^T$. The **training data** consists of $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$. A statistical model or method is fit (trained) on the training data. The **test data** consists of $\boldsymbol{z}_{n+1}, ..., \boldsymbol{z}_{n+m}$, and the test data is often used to evaluate the quality of the fitted model.

2) The focus of *supervised learning* is predicting a future value of the response variable Y_f given $\boldsymbol{x}_f$ and the training data $(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$. The focus of *unsupervised learning* is to group $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ into clusters. *Data mining* is looking for relationships in large data sets.

3) For classical regression and multivariate analysis, we often want $n \geq 10p$, and a model with $n < 5p$ is overfitting: the model does not have enough data to estimate parameters accurately if $\boldsymbol{x}$ is $p \times 1$. Statistical Learning methods often use a model with a crude degrees of freedom d, where $n \geq Jd$ with $J \geq 5$ and preferably $J \geq 10$. A model is underfitting if it omits important predictors. Fix p, if the probability that a model underfits goes to 0 as the sample size $n \to \infty$, then overfitting may not be too serious if $n \geq Jd$. Underfitting can cause the model to fail to hold.

4) There are several important Statistical Learning principles.
i) There is more interest in prediction or classification, e.g. producing $\hat{Y}_f$, than in other types of inference.
ii) Often the focus is on extracting useful information when n/p is not large, e.g. $p > n$. If d is a crude estimator of the fitted model degrees of freedom, we want n/d large. A *sparse model* has few nonzero coefficients. We can have sparse population models and sparse fitted models. Sometimes sparse fitted models are useful even if the population model is *dense* (not sparse). Often the number of nonzero coefficients of a *sparse fitted model* $= d$.
iii) Interest is in how well the method performs on test data. Performance on training data is overly optimistic for estimating performance on test data.
iv) Some methods are *flexible* while others are *unflexible*. For unflexible methods, the sufficient predictor is often a hyperplane $SP = \boldsymbol{x}^T\boldsymbol{\beta}$ and often the mean function $E(Y|\boldsymbol{x}) = M(\boldsymbol{x}^T\boldsymbol{\beta})$ where the function M is known but the $p \times 1$ vector of parameters $\boldsymbol{\beta}$ is unknown and must be estimated (GLMs). Flexible methods tend to be useful for more complicated regression methods where $E(Y|\boldsymbol{x}) = m(\boldsymbol{x})$ for an unknown function m or $SP \neq \boldsymbol{x}^T\boldsymbol{\beta}$ (GAMs). Flexibility tends to increase with d.

5) *Regression* investigates how the response variable Y changes with the value of a $p \times 1$ vector $\boldsymbol{x}$ of predictors. For a *1D regression model*, Y is conditionally independent of $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$, written $Y \perp\!\!\!\perp \boldsymbol{x} | h(\boldsymbol{x})$, where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. The *estimated sufficient predictor* $ESP = \hat{h}(\boldsymbol{x})$. A **response plot** is a plot of the ESP versus the response Y. Often $SP = \boldsymbol{x}^T\boldsymbol{\beta}$ and $ESP = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. A *residual plot* is a plot of the ESP versus the residuals. Tip: if the model for Y (more accurately for $Y|\boldsymbol{x}$) depends on $\boldsymbol{x}$ only through the real valued function $h(\boldsymbol{x})$, then $SP = h(\boldsymbol{x})$.

6) a) The **log rule** states that a positive variable that has the ratio between the largest and smallest values greater than ten should be transformed to logs. So $W > 0$ and $\max(W)/\min(W) > 10$ suggests using $\log(W)$.

b) The **ladder rule**: to spread *small* values of a variable, make λ *smaller*, to spread *large* values of a variable, make λ *larger*.

7) Let the ladder of powers $\Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\}$. Let $t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$. Consider the additive error regression model $Y = m(\boldsymbol{x}) + e$. Then the response transformation model is $Y = t_\lambda(Z) = m_\lambda(\boldsymbol{x}) + e$. 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. Make the transformations for $\lambda \in \Lambda_L$, and choose the transformation with the best transformation plot where the plotted points scatter about the identity line.

8) For the location model, the sample mean $\overline{Y} = \dfrac{\sum_{i=1}^n Y_i}{n}$, the sample variance $S_n^2 = \dfrac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1}$, and the sample standard deviation $S_n = \sqrt{S_n^2}$. If the data $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the ith order statistic and the $Y_{(i)}$'s are called the *order statistics*. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \quad \text{if n is odd,}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \quad \text{if n is even.}$$

The notation $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ will also be used. The *sample median absolute deviation* is $\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|,\ i = 1, \ldots, n)$.

9) Suppose the multivariate data has been collected into an $n \times p$ matrix

$$\boldsymbol{W} = \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}.$$

The *coordinatewise median* $\text{MED}(\boldsymbol{W}) = (\text{MED}(X_1), ..., \text{MED}(X_p))^T$ where $\text{MED}(X_i)$ is the sample median of the data in column i corresponding to variable X_i. The **sample mean** $\overline{\boldsymbol{x}} = \dfrac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i = (\overline{X}_1, ..., \overline{X}_p)^T$ where $\overline{X}_i$ is the sample mean of the data in column i corresponding to variable X_i. The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

That is, the ij entry of $\boldsymbol{S}$ is the sample covariance S_{ij}. The *classical estimator of multivariate location and dispersion* is $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$.

10) Let $(T, \boldsymbol{C}) = (T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ be an estimator of multivariate location and dispersion. The ith *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the ith *squared Mahalanobis distance* is $D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W}))$.

11) The squared Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median is $D_i^2 = D_i^2(\mathrm{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median MED_j: the coordinatewise median computed from the cases $\boldsymbol{x}_i$ with $D_i^2 \leq \mathrm{MED}(D_i^2(\mathrm{MED}_{j-1}, \boldsymbol{I}_p))$ where $\mathrm{MED}_0 = \mathrm{MED}(\boldsymbol{W})$. Often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\mathrm{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \mathrm{MED}(D_1, ..., D_n) + k\mathrm{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise.

12) Let the *covmb2 set* B of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set B. Hence

$$T = \frac{\sum_{i=1}^n W_i \boldsymbol{x}_i}{\sum_{i=1}^n W_i} \text{ and } \boldsymbol{C} = \frac{\sum_{i=1}^n W_i (\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^n W_i \ - \ 1}.$$

The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers.

13) If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $p \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{X}+\boldsymbol{Y}) = E(\boldsymbol{X})+E(\boldsymbol{Y}),\ E(\boldsymbol{a}+\boldsymbol{Y}) = \boldsymbol{a}+E(\boldsymbol{Y}),\ \&\ E(\boldsymbol{AXB}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}.$$

Also

$$\mathrm{Cov}(\boldsymbol{a} + \boldsymbol{AX}) = \mathrm{Cov}(\boldsymbol{AX}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{X})\boldsymbol{A}^T.$$

Note that $E(\boldsymbol{AY}) = \boldsymbol{A}E(\boldsymbol{Y})$ and $\mathrm{Cov}(\boldsymbol{AY}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{Y})\boldsymbol{A}^T$.

14) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

15) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{AX} \sim N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants, then $\boldsymbol{X} + \boldsymbol{a} \sim N_p(\boldsymbol{\mu} + \boldsymbol{a}, \boldsymbol{\Sigma})$.

16) Let $\boldsymbol{X}_n$ be a sequence of random vectors with joint cdfs $F_n(\boldsymbol{x})$ and let $\boldsymbol{X}$ be a random vector with joint cdf $F(\boldsymbol{x})$.

a) $\boldsymbol{X}_n$ **converges in distribution** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, if $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ as $n \to \infty$ for all points $\boldsymbol{x}$ at which $F(\boldsymbol{x})$ is continuous. The distribution of $\boldsymbol{X}$ is the **limiting distribution** or **asymptotic distribution** of $\boldsymbol{X}_n$. Note that $\boldsymbol{X}$ does not depend on n.

b) $\boldsymbol{X}_n$ **converges in probability** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \to 0$ as $n \to \infty$.

17) Multivariate Central Limit Theorem (MCLT): If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$, then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$$

where the sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

18) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A}\boldsymbol{\mu}) \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

18) Suppose $\boldsymbol{A}$ is a conformable constant matrix and $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$. Then $\boldsymbol{A}\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{A}\boldsymbol{X}$.

19) A $g \times 1$ random vector $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j if $\boldsymbol{u}$ is equal to $\boldsymbol{u}_j$ with probability π_j. The cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$ where the probabilities π_j satisfy $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \geq 2$, and $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ is the cdf of a $g \times 1$ random vector $\boldsymbol{u}_j$. Then $E(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j]$ and $Cov(\boldsymbol{u}) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})E(\boldsymbol{u}^T) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j \boldsymbol{u}_j^T] - E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T$. If $E(\boldsymbol{u}_j) = \boldsymbol{\theta}$ for $j = 1, ..., J$, then $E(\boldsymbol{u}) = \boldsymbol{\theta}$ and $Cov(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j)$. Note that $E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \sum_{k=1}^{J} \pi_j \pi_k E(\boldsymbol{u}_j)[E(\boldsymbol{u}_k)]^T$.

1.9 Complements

Graphical response transformation methods similar to those in Section 1.2 include Cook and Olive (2001) and Olive (2004, 2017a: section 3.2). A numerical method is given by Zhang and Yang (2017).

Section 1.5 followed Olive (2014, ch. 8) closely, which is a good Master's level treatment of large sample theory. Olive (2023d) is an online text. There are several PhD level texts on large sample theory including, in roughly increasing order of difficulty, Lehmann (1999), Ferguson (1996), Sen and Singer (1993), and Serfling (1980). White (1984) considers asymptotic theory for econometric applications.

For a nonsingular matrix, the inverse of the matrix, the determinant of the matrix, and the eigenvalues of the matrix are continuous functions of the matrix. Hence if $\hat{\boldsymbol{\Sigma}}$ is a consistent estimator of $\boldsymbol{\Sigma}$, then the inverse, determinant, and eigenvalues of $\hat{\boldsymbol{\Sigma}}$ are consistent estimators of the inverse,

determinant, and eigenvalues of $\boldsymbol{\Sigma} > 0$. See, for example, Bhatia et al. (1990), Stewart (1969), and Severini (2005, pp. 348-349).

Outliers

The outlier detection methods of Section 1.4 are due to Olive (2017b, section 4.7). For competing outlier detection methods, see Boudt et al. (2017). Also, google "novelty detection," "anomaly detection," and "artefact identification."

Big Data Sets

Sometimes n is huge and p is small. Then importance sampling and sequential analysis with sample size less than 1000 can be useful for inference for regression and time series models. Sometimes n is much smaller than p, for example with microarrays. Sometimes both n and p are large.

1.10 Problems

crancap	hdlen	hdht	Data for 1.1
1485	175	132	
1450	191	117	
1460	186	122	
1425	191	125	
1430	178	120	
1290	180	117	
90	75	51	

1.1*. The table ($\boldsymbol{W}$) above represents 3 head measurements on 6 people and one ape. Let X_1 = *cranial capacity*, X_2 = *head length*, and X_3 = *head height*. Let $\boldsymbol{x} = (X_1, X_2, X_3)^T$. Several multivariate location estimators, including the coordinatewise median and sample mean, are found by applying a univariate location estimator to each random variable and then collecting the results into a vector. a) Find the coordinatewise median MED($\boldsymbol{W}$).

b) Find the sample mean $\overline{\boldsymbol{x}}$.

1.2. The table $\boldsymbol{W}$ shown below represents 4 measurements on 5 people.

age	breadth	cephalic	size
39.00	149.5	81.9	3738
35.00	152.5	75.9	4261
35.00	145.5	75.4	3777
19.00	146.0	78.1	3904
0.06	88.5	77.6	933

a) Find the sample mean $\overline{\boldsymbol{x}}$.

b) Find the coordinatewise median MED($\boldsymbol{W}$).

1.3. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors from a multivariate t-distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with d degrees of freedom. Then $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{x}) = \dfrac{d}{d-2}\boldsymbol{\Sigma}$ for $d > 2$. Assuming $d > 2$, find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

1.4. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors where $E(\boldsymbol{x}_i) = e^{0.5}\mathbf{1}$ and $\mathrm{Cov}(\boldsymbol{x}_i) = (e^2 - e)\boldsymbol{I}_p$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

1.5. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid 2×1 random vectors from a multivariate lognormal $\mathrm{LN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Let $\boldsymbol{x}_i = (X_{i1}, X_{i2})^T$. Following Press (2005, pp. 149-150), $E(X_{ij}) = \exp(\mu_j + \sigma_j^2/2)$,
$V(X_{ij}) = \exp(\sigma_j^2)[\exp(\sigma_j^2) - 1]\exp(2\mu_j)$ for $j = 1, 2$, and
$\mathrm{Cov}(X_{i1}, X_{i2}) = \exp[\mu_1 + \mu_2 + 0.5(\sigma_1^2 + \sigma_2^2) + \sigma_{12}][\exp(\sigma_{12}) - 1]$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

1.6. The most used Poisson regression model is $Y|\boldsymbol{x} \sim \mathrm{Poisson}(\exp(\boldsymbol{x}^T\boldsymbol{\beta}))$. What is the sufficient predictor $SP = h(\boldsymbol{x})$?

1.7. Let Z be the variable of interest and let $Y = t(z)$ be the response variable for the multiple linear regression model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$. For the four transformation plots shown in Figure 1.9, $n = 1000$, and $p = 4$. The fitting method was the elastic net. What response transformation should be used?

1.8. The data set follows the multiple linear regression model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ with $n = 100$ and $p = 101$. The response plots for two methods are shown in Figure 1.10. Which method fits the data better, lasso or ridge regression? For ridge regression, is anything wrong with yhat $= \hat{Y}$.

1.9. For the Buxton (1920) data with multiple linear regression, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet! The response plot shown in Figure 1.4a) is for lasso. The response plot in Figure 1.4b) did lasso for the cases in the `covmb2` set B applied to the predictors and set B included all of the clean cases and omitted the 5 outliers. The response plot was made for all of the data, including the outliers. Both plots include the identity line and prediction interval bands.

Which method is better: Fig. 1.4 a) or Fig. 1.4 b) for data analysis?

R Problem

Use the command *source("G:/slpack.txt")* **to download the functions** and the command *source("G:/sldata.txt")* **to download the data. See Preface or Section 8.1.** Typing the name of the `slpack` func-

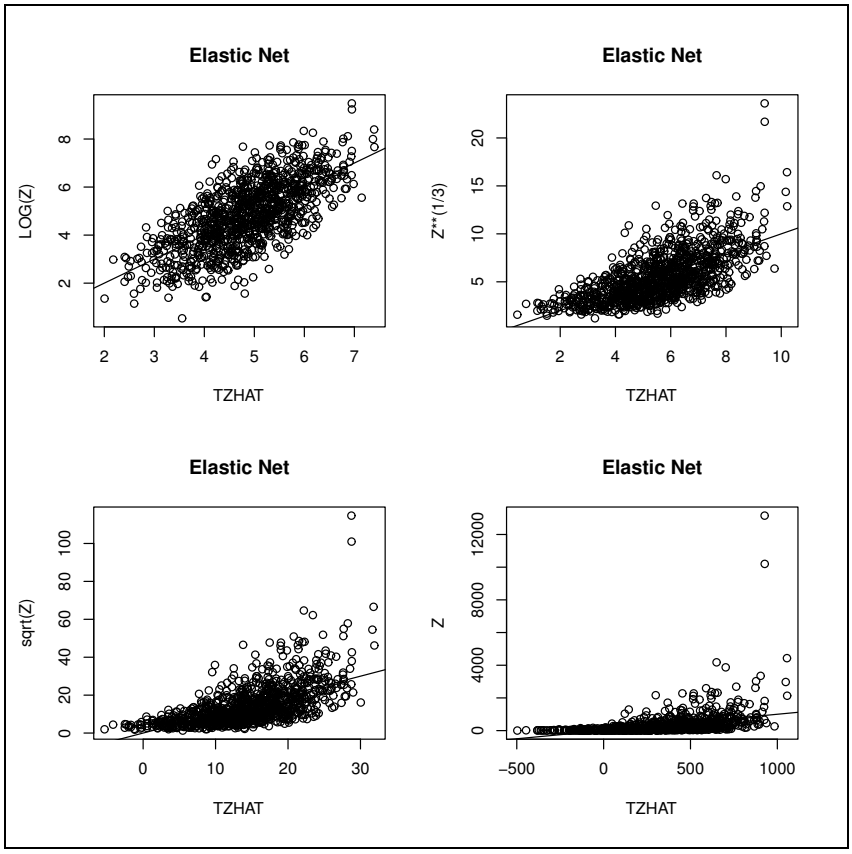

Fig. 1.9 Elastic Net Transformation Plots for Problem 1.7.

tion, e.g. *tplot2*, will display the code for the function. Use the `args` command, e.g. *args(tplot2)*, to display the needed arguments for the function. For the following problem, the R command can be copied and pasted from (http://parker.ad.siu.edu/Olive/slrhw.txt) into R.

1.10. This problem uses some of the R commands at the end of Section 1.2.1. A problem with response and residual plots is that there can be a lot of black in the plot if the sample size n is large (more than a few thousand). A variant of the response plot for the additive error regression model $Y = m(\boldsymbol{x})+e$ would plot the identity line, the two lines parallel to the identity line corresponding to the Section 2.1 large sample $100(1-\delta)\%$ prediction intervals for Y_f that depends on $\hat{Y}_f$. Then plot points corresponding to training data cases that do not lie in their $100(1-\delta)\%$ PI. We will use $\delta = 0.01, n = 100000$, and $p = 8$.

a) Copy and paste the commands for this part into R. They make the usual response plot with a lot of black. Do not include the plot in *Word*.

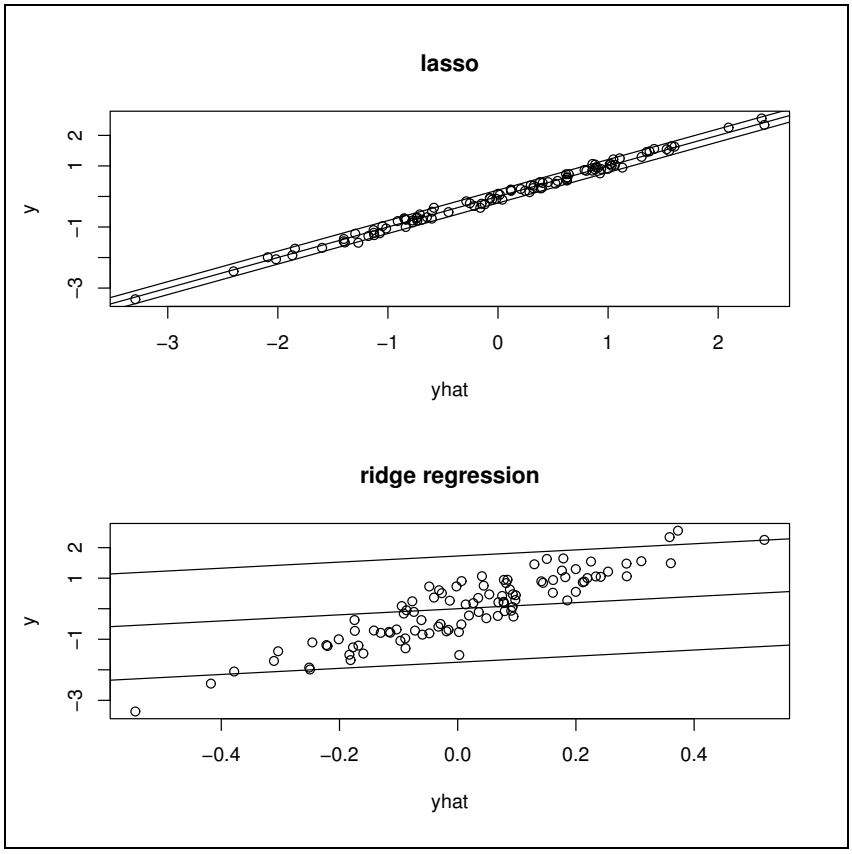

Fig. 1.10 Response Plots for Problem 1.8.

b) Copy and paste the commands for this part into *R*. They make the response plot with the points within the pointwise 99% prediction interval bands omitted. Include this plot in *Word*. For example, left click on the plot and hit the *Ctrl* and *c* keys at the same time to make a copy. Then paste the plot into *Word*, e.g., get into *Word* and hit the *Ctrl* and *v* keys at the same time.

c) The additive error regression model is a 1D regression model. What is the sufficient predictor $= h(\boldsymbol{x})$?

1.11. The *slpack* function `tplot2` makes transformation plots for the multiple linear regression model $Y = t(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e$. Type = 1 for full model OLS and should not be used if $n < 5p$, type = 2 for elastic net, 3 for lasso, 4 for ridge regression, 5 for PLS, 6 for PCR, and 7 for forward selection with C_p if $n \geq 10p$ and EBIC if $n < 10p$. These methods are discussed in Chapter 3.

Copy and paste the three library commands near the top of *slrhw* into *R*.

For parts a) and b), $n = 100, p = 4$ and $Y = \log(Z) = 0x_1 + x_2 + 0x_3 + 0x_4 + e = x_2 + e$. ($Y$ and Z are swapped in the R code.)

a) Copy and paste the commands for this part into *R*. This makes the response plot for the elastic net using $Y = Z$ and $\boldsymbol{x}$ when the linear model needs $Y = \log(Z)$. Do not include the plot in *Word*, but explain why the plot suggests that something is wrong with the model $Z = \boldsymbol{x}^T\boldsymbol{\beta} + e$.

b) Copy and paste the command for this part into *R*. Right click *Stop* 3 times until the horizontal axis has log(z). This is the response plot for the true model $Y = \log(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e = x_2 + e$. Include the plot in *Word*. Right click *Stop* 3 more times so that the cursor returns in the command window.

c) Is the response plot linear?

For the remaining parts, $n = p - 1 = 100$ and $Y = \log(Z) = 0x_1 + x_2 + 0x_3 + \cdots + 0x_{101} + e = x_2 + e$. Hence the model is sparse.

d) Copy and paste the commands for this part into *R*. Right click *Stop* 3 times until the horizontal axis has log(z). This is the response plot for the true model $Y = \log(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e = x_2 + e$. Include the plot in *Word*. Right click *Stop* 3 more times so that the cursor returns in the command window.

e) Is the plot linear?

f) Copy and paste the commands for this part into *R*. Right click *Stop* 3 times until the horizontal axis has log(z). This is the response plot for the true model $Y = \log(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e = x_2 + e$. Include the plot in *Word*. Right click *Stop* 3 more times so that the cursor returns in the command window. PLS is probably overfitting since the identity line nearly interpolates the fitted points.

1.12. Get the *R* commands for this problem. The data is such that $Y = 2 + x_2 + x_3 + x_4 + e$ where the zero mean errors are iid [exponential(2) - 2]. Hence the residual and response plots should show high skew. Note that $\boldsymbol{\beta} = (2, 1, 1, 1)^T$. The *R* code uses 3 nontrivial predictors and a constant, and the sample size $n = 1000$.

a) Copy and paste the commands for part a) of this problem into *R*. Include the response plot in *Word*. Is the lowess curve fairly close to the identity line?

b) Copy and paste the commands for part b) of this problem into *R*. Include the residual plot in *Word*: press the *Ctrl* and *c* keys as the same time. Then use the menu command "Paste" in *Word*. Is the lowess curve fairly close to the $r = 0$ line? The lowess curve is a flexible scatterplot smoother.

c) The output out$coef gives $\hat{\boldsymbol{\beta}}$. Write down $\hat{\boldsymbol{\beta}}$ or copy and paste $\hat{\boldsymbol{\beta}}$ into *Word*. Is $\hat{\boldsymbol{\beta}}$ close to $\boldsymbol{\beta}$?

1.13. For the Buxton (1920) data with multiple linear regression, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet!

a) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. The identity line passes right through the outliers which are obvious because of the large gap. Prediction interval (PI) bands are also included in the plot.

b) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. This did lasso for the cases in the `covmb2` set B applied to the predictors which included all of the clean cases and omitted the 5 outliers. The response plot was made for all of the data, including the outliers.

c) Copy and paste the commands for this problem into *R*. Include the DD plot in *Word*. The outliers are in the upper right corner of the plot.

1.14. Consider the Gladstone (1905) data set that has 12 variables on 267 persons after death. There are 5 infants in the data set. The response variable was *brain weight*. Head measurements were *breadth, circumference, head height, length,* and *size* as well as *cephalic index* and *brain weight*. *Age*, *height*, and three categorical variables *cause*, *ageclass* (0: under 20, 1: 20-45, 2: over 45) and *sex* were also given. The constant x_1 was the first variable. The variables *cause* and *ageclass* were not coded as factors. Coding as factors might improve the fit.

a) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. The identity line passes right through the infants which are obvious because of the large gap. Prediction interval (PI) bands are also included in the plot.

b) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. This did lasso for the cases in the `covmb2` set B applied to the nontrivial predictors which are not categorical (omit the *constant, cause, ageclass* and *sex*) which omitted 8 cases, including the 5 infants. The response plot was made for all of the data.

c) Copy and paste the commands for this problem into *R*. Include the DD plot in *Word*. The infants are in the upper right corner of the plot.

1.15. The *slpack* function `mldsim6` compares 7 estimators: FCH, RFCH, CMVE, RCMVE, RMVN, `covmb2`, and MB described in Olive (2017b, ch. 4). Most of these estimators need $n > 2p$, need a nonsingular dispersion matrix, and work best with $n > 10p$. The function generates data sets and counts how many times the minimum Mahalanobis distance $D_i(T, \boldsymbol{C})$ of the outliers is larger than the maximum distance of the clean data. The value pm controls how far the outliers need to be from the bulk of the data, and pm roughly needs to increase with $\sqrt{p}$.

For data sets with $p > n$ possible, the function `mldsim7` used the Euclidean distances $D_i(T, \boldsymbol{I}_p)$ and the Mahalanobis distances $D_i(T, \boldsymbol{C}_d)$ where $\boldsymbol{C}_d$ is the diagonal matrix with the same diagonal entries as $\boldsymbol{C}$ where $(T, \boldsymbol{C})$ is the `covmb2` estimator using j concentration type steps. Dispersion matrices are effected more by outliers than good robust location estimators, so when the outlier proportion is high, it is expected that the Euclidean distances $D_i(T, \boldsymbol{I}_p)$ will outperform the Mahalanobis distance $D_i(T, \boldsymbol{C}_d)$ for

many outlier configurations. Again the function counts the number of times the minimum outlier distance is larger than the maximum distance of the clean data.

Both functions used several outlier types. The simulations generated 100 data sets. The clean data had $\boldsymbol{x}_i \sim N_p(\mathbf{0}, diag(1, ..., p))$. Type 1 had outliers in a tight cluster (near point mass) at the major axis $(0, ..., 0, pm)^T$. Type 2 had outliers in a tight cluster at the minor axis $(pm, 0, ..., 0)^T$. Type 3 had mean shift outliers $\boldsymbol{x}_i \sim N_p((pm, ..., pm)^T, diag(1, ..., p))$. Type 4 changed the pth coordinate of the outliers to pm. Type 5 changed the 1st coordinate of the outliers to pm. (If the outlier $\boldsymbol{x}_i = (x_{1i}, ..., x_{pi})^T$, then $x_{i1} = pm$.)

Table 1.2 Number of Times All Outlier Distances > Clean Distances, otype=1

n	p	γ	osteps	pm	FCH	RFCH	CMVE	RCMVE	RMVN	covmb2	MB
100	10	0.25	0	20	85	85	85	85	86	67	89

a) Table 1.2 suggests with osteps = 0, `covmb2` had the worst count. When pm is increased to 25, all counts become 100. Copy and paste the commands for this part into R and make a table similar to Table 1.2, but now osteps=9 and $p = 45$ is close to $n/2$ for the second line where $pm = 60$. Your table should have 2 lines from output.

Table 1.3 Number of Times All Outlier Distances > Clean Distances, otype=1

n	p	γ	osteps	pm	covmb2	diag
100	1000	0.4	0	1000	100	41
100	1000	0.4	9	600	100	42

b) Copy and paste the commands for this part into R and make a table similar to Table 1.3, but type 2 outliers are used.

c) When you have two reasonable outlier detectors, there are outlier configurations where one will beat the other. Simulations by Wang (2018) suggest that "covmb2" using $D_i(T, \boldsymbol{I}_p)$ outperforms "diag" using $D_i(T, \boldsymbol{C}_d)$ for many outlier configurations, but there are some exceptions. Copy and paste the commands for this part into R and make a table similar to Table 1.3, but type 3 outliers are used.