Chapter 15

1D Regression

... estimates of the linear regression coefficients are relevant to the linear parameters of a broader class of models than might have been suspected.
Brillinger (1977, p. 509)

After computing $\hat{\beta}$, one may go on to prepare a scatter plot of the points $(\hat{\beta}x_j, y_j)$, $j = 1, ..., n$ and look for a functional form for $g(\cdot)$.
Brillinger (1983, p. 98)

Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response Y given the $(p-1) \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. The scalar Y is a random variable and $\boldsymbol{x}$ is a random vector. A special case of regression is multiple linear regression. In Chapter 2 the multiple linear regression model was $Y_i = w_{i,1}\eta_1 + w_{i,2}\eta_2 + \cdots + w_{i,p}\eta_p + e_i = \boldsymbol{w}_i^T\boldsymbol{\eta} + e_i$ for $i = 1, \ldots, n$. In this chapter, the subscript i is often suppressed and the multiple linear regression model is written as $Y = \alpha + x_1\beta_1 + \cdots + x_{p-1}\beta_{p-1} + e = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$. The primary difference is the separation of the constant term α and the nontrivial predictors $\boldsymbol{x}$. In Chapter 2, $w_{i,1} \equiv 1$ for $i = 1, ..., n$. Taking $Y = Y_i$, $\alpha = \eta_1$, $\beta_j = \eta_{j+1}$, and $x_j = w_{i,j+1}$ and $e = e_i$ for $j = 1, ..., p-1$ shows that the two models are equivalent. The change in notation was made because the distribution of the nontrivial predictors is very important for the theory of the more general regression models.

Definition 15.1: Cook and Weisberg (1999a, p. 414). In a *1D regression model*, the response Y is conditionally independent of $\boldsymbol{x}$ given a single linear combination $\boldsymbol{\beta}^T\boldsymbol{x}$ of the predictors, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T\boldsymbol{x} \text{ or } \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}). \quad (15.1)$$

The 1D regression model is also said to have *1–dimensional structure* or *1D structure*. An important 1D regression model, introduced by Li and Duan (1989), has the form

$$Y = g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e) \tag{15.2}$$

where g is a bivariate (inverse link) function and e is a zero mean error that is independent of $\boldsymbol{x}$. The constant term α may be absorbed by g if desired.

Special cases of the 1D regression model (15.1) include many important *generalized linear models* (GLMs) and the additive error *single index model*

$$Y = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) + e. \tag{15.3}$$

Typically m is the conditional mean or median function. For example if all of the expectations exist, then

$$E[Y|\boldsymbol{x}] = E[m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})|\boldsymbol{x}] + E[e|\boldsymbol{x}] = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}).$$

The *multiple linear regression model* is an important special case where m is the identity function: $m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$. Another important special case of 1D regression is the *response transformation model* where

$$g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e) = t^{-1}(\alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e) \tag{15.4}$$

and t^{-1} is a one to one (typically monotone) function. Hence

$$t(Y) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e.$$

Chapter 16 shows that many *survival models* are 1D regression models, including the Cox (1972) *proportional hazards model*. Li and Duan (1989, p. 1014) note that the class of 1D regression models also includes binary regression models, censored regression models, and certain projection pursuit models.

Definition 15.2. *Regression* is the study of the conditional distribution of $Y|\boldsymbol{x}$. Focus is often on the *mean function* $E(Y|\boldsymbol{x})$ and/or the *variance function* $\mathrm{VAR}(Y|\boldsymbol{x})$. There is a distribution for each value of $\boldsymbol{x} = \boldsymbol{x}_o$ such that $Y|\boldsymbol{x} = \boldsymbol{x}_o$ is defined. For a 1D regression,

$$E(Y|\boldsymbol{x} = \boldsymbol{x}_o) = E(Y|\boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}^T \boldsymbol{x}_o) \equiv M(\boldsymbol{\beta}^T \boldsymbol{x}_o)$$

and

$$\mathrm{VAR}(Y|\boldsymbol{x} = \boldsymbol{x}_o) = \mathrm{VAR}(Y|\boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}^T \boldsymbol{x}_o) \equiv V(\boldsymbol{\beta}^T \boldsymbol{x}_o)$$

where M is the *kernel mean function* and V is the *kernel variance function*.

Notice that the mean and variance functions depend on the *same* linear combination if the 1D regression model is valid. This dependence is typical of GLMs where M and V are known kernel mean and variance functions that depend on the family of GLMs. See Cook and Weisberg (1999a, section 23.1). A *heteroscedastic regression model*

$$Y = M(\boldsymbol{\beta}_1^T \boldsymbol{x}) + \sqrt{V(\boldsymbol{\beta}_2^T \boldsymbol{x})} \;\; e \tag{15.5}$$

is a 1D regression model if $\boldsymbol{\beta}_2 = c\boldsymbol{\beta}_1$ for some scalar c.

In multiple linear regression, the difference between the response Y_i and the estimated conditional mean function $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ is the residual. For more general regression models this difference may not be the residual, and the "discrepancy" $Y_i - M(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$ may not be estimating the error e_i. To guarantee that the residuals are estimating the errors, the following definition is used when possible.

Definition 15.3: Cox and Snell (1968). Let the errors e_i be iid with pdf f and assume that the regression model $Y_i = g(\boldsymbol{x}_i, \boldsymbol{\eta}, e_i)$ has a unique solution for e_i :

$$e_i = h(\boldsymbol{x}_i, \boldsymbol{\eta}, Y_i).$$

Then the ith residual

$$\hat{e}_i = h(\boldsymbol{x}_i, \hat{\boldsymbol{\eta}}, Y_i)$$

where $\hat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$.

Example 15.1. Let $\boldsymbol{\eta} = (\alpha, \boldsymbol{\beta}^T)^T$. If $Y = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) + e$ where m is known, then $e = Y - m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$. Hence $\hat{e}_i = Y_i - m(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$ which is the usual definition of the ith residual for such models.

Dimension reduction can greatly simplify our understanding of the conditional distribution $Y|\boldsymbol{x}$. If a 1D regression model is appropriate, then the $(p-1)$–dimensional vector $\boldsymbol{x}$ can be replaced by the 1–dimensional scalar $\boldsymbol{\beta}^T \boldsymbol{x}$ *with "no loss of information about the conditional distribution."* Cook and Weisberg (1999a, p. 411) define a *sufficient summary plot* (SSP) to be a plot that contains all the sample regression information about the conditional distribution $Y|\boldsymbol{x}$ of the response given the predictors.

Definition 15.4: If the 1D regression model holds, then $Y \perp\!\!\!\perp \boldsymbol{x} | (a + c\boldsymbol{\beta}^T \boldsymbol{x})$ for any constants a and $c \neq 0$. The quantity $a + c\boldsymbol{\beta}^T \boldsymbol{x}$ is called a *sufficient predictor* (SP), and a sufficient summary plot is a plot of any SP versus Y. An *estimated sufficient predictor* (ESP) is $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ where $\hat{\boldsymbol{\beta}}$ is an estimator of $c\boldsymbol{\beta}$ for some nonzero constant c. A *response plot* or *estimated sufficient summary plot* (ESSP) is a plot of any ESP versus Y.

If there is only one predictor x, then the plot of x versus Y is both a sufficient summary plot and a response plot, but generally only a response plot can be made. Since a can be any constant, $\hat{a} = 0$ is often used. The following section shows how to use the OLS regression of Y on $\boldsymbol{x}$ to obtain an ESP.

15.1 Estimating the Sufficient Predictor

Some notation is needed before giving theoretical results. Let $\boldsymbol{x}$, $\boldsymbol{a}$, $\boldsymbol{t}$, and $\boldsymbol{\beta}$ be $(p-1) \times 1$ vectors where only $\boldsymbol{x}$ is random.

Definition 15.5: Cook and Weisberg (1999a, p. 431). The predictors $\boldsymbol{x}$ satisfy the condition of *linearly related predictors* with 1D structure if

$$E[\boldsymbol{x} | \boldsymbol{\beta}^T \boldsymbol{x}] = \boldsymbol{a} + \boldsymbol{t}\boldsymbol{\beta}^T \boldsymbol{x}. \tag{15.6}$$

If the predictors $\boldsymbol{x}$ satisfy this condition, then for any given predictor x_j,

$$E[x_j | \boldsymbol{\beta}^T \boldsymbol{x}] = a_j + t_j \boldsymbol{\beta}^T \boldsymbol{x}.$$

Notice that $\boldsymbol{\beta}$ is a fixed $(p-1) \times 1$ vector. If $\boldsymbol{x}$ is elliptically contoured (EC) with 1st moments, then the assumption of linearly related predictors holds since

$$E[\boldsymbol{x} | \boldsymbol{b}^T \boldsymbol{x}] = \boldsymbol{a}_b + \boldsymbol{t}_b \boldsymbol{b}^T \boldsymbol{x}$$

for *any* nonzero $(p-1) \times 1$ vector $\boldsymbol{b}$ (see Lemma 14.4). The condition of linearly related predictors is impossible to check since $\boldsymbol{\beta}$ is unknown, but the condition is far weaker than the assumption that $\boldsymbol{x}$ is EC. The stronger EC condition is often used since there are checks for whether this condition is reasonable, eg use the DD plot. The following proposition gives an equivalent

definition of linearly related predictors. Both definitions are frequently used in the dimension reduction literature.

Proposition 15.1. The predictors $\boldsymbol{x}$ are linearly related iff

$$E[\boldsymbol{b}^T\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = a_b + t_b\boldsymbol{\beta}^T\boldsymbol{x} \tag{15.7}$$

for any $(p-1) \times 1$ constant vector $\boldsymbol{b}$ where a_b and t_b are constants that depend on $\boldsymbol{b}$.

Proof. Suppose that the assumption of linearly related predictors holds. Then

$$E[\boldsymbol{b}^T\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = \boldsymbol{b}^T E[\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = \boldsymbol{b}^T\boldsymbol{a} + \boldsymbol{b}^T\boldsymbol{t}\boldsymbol{\beta}^T\boldsymbol{x}.$$

Thus the result holds with $a_b = \boldsymbol{b}^T\boldsymbol{a}$ and $t_b = \boldsymbol{b}^T\boldsymbol{t}$.

Now assume that Equation (15.7) holds. Take $\boldsymbol{b}_i = (0, ..., 0, 1, 0, ..., 0)^T$, the vector of zeroes except for a one in the ith position. Then by Definition 15.5, $E[\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = E[\boldsymbol{I}_{p-1}\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] =$

$$E[\begin{pmatrix} \boldsymbol{b}_1^T\boldsymbol{x} \\ \vdots \\ \boldsymbol{b}_{p-1}^T\boldsymbol{x} \end{pmatrix} \mid \boldsymbol{\beta}^T\boldsymbol{x}] = \begin{pmatrix} a_1 + t_1\boldsymbol{\beta}^T\boldsymbol{x} \\ \vdots \\ a_{p-1} + t_{p-1}\boldsymbol{\beta}^T\boldsymbol{x} \end{pmatrix} \equiv \boldsymbol{a} + \boldsymbol{t}\boldsymbol{\beta}^T\boldsymbol{x}.$$

QED

Following Cook (1998a, p. 143-144), assume that there is an objective function

$$L_n(a, \boldsymbol{b}) = \frac{1}{n}\sum_{i=1}^{n} L(a + \boldsymbol{b}^T\boldsymbol{x}_i, Y_i) \tag{15.8}$$

where $L(u, v)$ is a bivariate function that is a convex function of the first argument u. Assume that the estimate $(\hat{a}, \hat{\boldsymbol{b}})$ of $(a, \boldsymbol{b})$ satisfies

$$(\hat{a}, \hat{\boldsymbol{b}}) = \arg\min_{a,\boldsymbol{b}} L_n(a, \boldsymbol{b}). \tag{15.9}$$

For example, the ordinary least squares (OLS) estimator uses

$$L(a + \boldsymbol{b}^T\boldsymbol{x}, Y) = (Y - a - \boldsymbol{b}^T\boldsymbol{x})^2.$$

Maximum likelihood type estimators such as those used to compute GLMs and Huber's M–estimator also work, as does the Wilcoxon rank estimator. Assume that the population analog $(\alpha^*, \boldsymbol{\beta}^*)$ is the unique minimizer of

$E[L(a+\boldsymbol{b}^T\boldsymbol{x}, Y)]$ where the expectation exists and is with respect to the joint distribution of $(Y, \boldsymbol{x}^T)^T$. For example, $(\alpha^*, \boldsymbol{\beta}^*)$ is unique if $L(u, v)$ is strictly convex in its first argument. The following result is a useful extension of Brillinger (1977, 1983).

Theorem 15.2 (Li and Duan 1989, p. 1016): Assume that the $\boldsymbol{x}$ are linearly related predictors, that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid observations from some joint distribution with $\mathrm{Cov}(\boldsymbol{x}_\mathrm{i})$ nonsingular. Assume $L(u, v)$ is convex in its first argument and that $\boldsymbol{\beta}^*$ is unique. Assume that $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T\boldsymbol{x}$. Then $\boldsymbol{\beta}^* = c\boldsymbol{\beta}$ for some scalar c.

Proof. See Li and Duan (1989) or Cook (1998a, p. 144).

Remark 15.1. This theorem basically means that if the 1D regression model is appropriate and if the condition of linearly related predictors holds, then the (eg OLS) estimator $\hat{\boldsymbol{b}} \equiv \hat{\boldsymbol{\beta}}^* \approx c\boldsymbol{\beta}$. Li and Duan (1989, p. 1031) show that under additional conditions, $(\hat{a}, \hat{\boldsymbol{b}})$ is asymptotically normal. In particular, the OLS estimator frequently has a $\sqrt{n}$ convergence rate. If the OLS estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ satisfies $\hat{\boldsymbol{\beta}} \approx c\boldsymbol{\beta}$ when model (15.1) holds, then the response plot of

$$\hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x} \text{ versus } \mathrm{Y}$$

can be used to visualize the conditional distribution $Y|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$ provided that $c \neq 0$.

Remark 15.2. If $\hat{\boldsymbol{b}}$ is a consistent estimator of $\boldsymbol{\beta}^*$, then certainly

$$\boldsymbol{\beta}^* = c_{\boldsymbol{x}}\boldsymbol{\beta} + \boldsymbol{u}_g$$

where $\boldsymbol{u}_g = \boldsymbol{\beta}^* - c_{\boldsymbol{x}}\boldsymbol{\beta}$ is the bias vector. Moreover, the bias vector $\boldsymbol{u}_g = \boldsymbol{0}$ if $\boldsymbol{x}$ is elliptically contoured under the assumptions of Theorem 15.2. This result suggests that the bias vector might be negligible if the distribution of the predictors is close to being EC. **Often if no strong nonlinearities are present among the predictors,** the bias vector is small enough so that $\hat{\boldsymbol{b}}^T\boldsymbol{x}$ is a useful ESP.

Remark 15.3. Suppose that the 1D regression model is appropriate and $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T\boldsymbol{x}$. Then $Y \perp\!\!\!\perp \boldsymbol{x} | c\boldsymbol{\beta}^T\boldsymbol{x}$ for any nonzero scalar c. If $Y = g(\boldsymbol{\beta}^T\boldsymbol{x}, e)$ and both g and $\boldsymbol{\beta}$ are unknown, then $g(\boldsymbol{\beta}^T\boldsymbol{x}, e) = h_{a,c}(a + c\boldsymbol{\beta}^T\boldsymbol{x}, e)$ where

$$h_{a,c}(w, e) = g(\frac{w - a}{c}, e)$$

for $c \neq 0$. In other words, if g is unknown, we can estimate $c\boldsymbol{\beta}$ but we can not determine c or $\boldsymbol{\beta}$; ie, we can only estimate $\boldsymbol{\beta}$ up to a constant.

A very useful result is that if $Y = m(x)$ for some function m, then m can be visualized with both a plot of x versus Y and a plot of cx versus Y if $c \neq 0$. In fact, there are only three possibilities, if $c > 0$ then the two plots are nearly identical: except the labels of the horizontal axis change. (The two plots are usually not exactly identical since plotting controls to "fill space" depend on several factors and will change slightly.) If $c < 0$, then the plot appears to be flipped about the vertical axis. If $c = 0$, then $m(0)$ is a constant, and the plot is basically a dot plot. Similar results hold if $Y_i = g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i, e_i)$ if the errors e_i are small. OLS often provides a useful estimator of $c\boldsymbol{\beta}$ where $c \neq 0$, but OLS can result in $c = 0$ if g is symmetric about the median of $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$.

Definition 15.6. If the 1D regression model (15.1) holds, and a specific estimator such as OLS is used, then the ESP will be called the OLS ESP and the response plot will be called the OLS response plot.

Example 15.2. Suppose that $\boldsymbol{x}_i \sim N_3(\boldsymbol{0}, \boldsymbol{I}_3)$ and that

$$Y = m(\boldsymbol{\beta}^T \boldsymbol{x}) + e = (x_1 + 2x_2 + 3x_3)^3 + e.$$

Then a 1D regression model holds with $\boldsymbol{\beta} = (1, 2, 3)^T$. Figure 1.11 shows the sufficient summary plot of $\boldsymbol{\beta}^T \boldsymbol{x}$ versus Y, and Figure 1.12 shows the sufficient summary plot of $-\boldsymbol{\beta}^T \boldsymbol{x}$ versus Y. Notice that the functional form m appears to be cubic in both plots and that both plots can be smoothed by eye or with a scatterplot smoother such as *lowess.* The two figures were generated with the following *R/Splus* commands.

```
 X <- matrix(rnorm(300),nrow=100,ncol=3)
SP <- X%*%1:3
 Y <- (SP)^3 + rnorm(100)
      plot(SP,Y)
      plot(-SP,Y)
```

We particularly want to use the OLS estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ to produce an estimated sufficient summary plot. This estimator is obtained from the usual multiple linear regression of Y_i on $\boldsymbol{x}_i$, but *we are not assuming that the multiple linear regression model holds*; however, we are hoping that the 1D

regression model $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T \boldsymbol{x}$ is a useful approximation to the data and that $\hat{\boldsymbol{\beta}} \approx c\boldsymbol{\beta}$ for some nonzero constant c. In addition to Theorem 15.2, nice results exist if the single index model is appropriate. Recall that

$$\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{Y}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))((\boldsymbol{Y} - E(\boldsymbol{Y}))^T].$$

Definition 15.7. Suppose that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid observations and that the positive definite $(p-1) \times (p-1)$ matrix $\mathrm{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_X$ and the $(p-1) \times 1$ vector $\mathrm{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{X,Y}$. Let the OLS estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ be computed from the multiple linear regression of Y on $\boldsymbol{x}$ plus a constant. Then $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ estimates the population quantity $(\alpha_{OLS}, \boldsymbol{\beta}_{OLS})$ where

$$\alpha_{OLS} = E(Y) - \boldsymbol{\beta}_{OLS}^T E(\boldsymbol{x}) \text{ and } \boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}_{\mathrm{X}}^{-1} \boldsymbol{\Sigma}_{\mathrm{X,Y}}. \tag{15.10}$$

The following notation will be useful for studying the OLS estimator. Let the sufficient predictor $\boldsymbol{z} = \boldsymbol{\beta}^T \boldsymbol{x}$ and let $\boldsymbol{w} = \boldsymbol{x} - E(\boldsymbol{x})$. Let $\boldsymbol{r} = \boldsymbol{w} - (\boldsymbol{\Sigma}_X \boldsymbol{\beta}) \boldsymbol{\beta}^T \boldsymbol{w}$.

Theorem 15.3. In addition to the conditions of Definition 15.7, also assume that $Y_i = m(\boldsymbol{\beta}^T \boldsymbol{x}_i) + e_i$ where the zero mean constant variance iid errors e_i are independent of the predictors $\boldsymbol{x}_i$. Then

$$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_{X,Y} = c_{m,X} \boldsymbol{\beta} + \boldsymbol{u}_{m,X} \tag{15.11}$$

where the scalar

$$c_{m,X} = E[\boldsymbol{\beta}^T (\boldsymbol{x} - E(\boldsymbol{x}))\ m(\boldsymbol{\beta}^T \boldsymbol{x})] \tag{15.12}$$

and the bias vector

$$\boldsymbol{u}_{m,X} = \boldsymbol{\Sigma}_X^{-1} E[m(\boldsymbol{\beta}^T \boldsymbol{x}) \boldsymbol{r}]. \tag{15.13}$$

Moreover, $\boldsymbol{u}_{m,X} = \boldsymbol{0}$ if $\boldsymbol{x}$ is from an EC distribution with nonsingular $\boldsymbol{\Sigma}_X$, and $c_{m,X} \neq 0$ unless $\mathrm{Cov}(\boldsymbol{x}, Y) = \boldsymbol{0}$. If the multiple linear regression model holds, then $c_{m,X} = 1$, and $\boldsymbol{u}_{m,X} = \boldsymbol{0}$.

The proof of the above result is outlined in Problem 15.2 using an argument due to Aldrin, Bϕlviken, and Schweder (1993). If the 1D regression model is appropriate, then typically $\mathrm{Cov}(\boldsymbol{x}, Y) \neq \boldsymbol{0}$ unless $\boldsymbol{\beta}^T \boldsymbol{x}$ follows a symmetric distribution and m is symmetric about the median of $\boldsymbol{\beta}^T \boldsymbol{x}$.

Definition 15.8. Let $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ denote the OLS estimate obtained from the OLS multiple linear regression of Y on $\boldsymbol{x}$. The *OLS view* is a response plot of $a + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ versus Y. Typically $a = 0$ or $a = \hat{\alpha}$.

Remark 15.4. All of this awkward notation and theory leads to a rather remarkable result, perhaps first noted by Brillinger (1977, 1983) and called the *1D Estimation Result* by Cook and Weisberg (1999a, p. 432). The result is that if the 1D regression model is appropriate, then *the OLS view will frequently be a useful estimated sufficient summary plot* (ESSP). Hence the OLS predictor $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ is a useful *estimated sufficient predictor* (ESP).

Although the OLS view is frequently a good ESSP if no strong nonlinearities are present in the predictors and if $c_{m,X} \neq 0$ (eg the sufficient summary plot of $\boldsymbol{\beta}^T \boldsymbol{x}$ versus Y is not approximately symmetric), even better estimated sufficient summary plots can be obtained by using ellipsoidal trimming. This topic is discussed in the following section and follows Olive (2002) closely.

15.2 Visualizing 1D Regression

If there are two predictors, even with a distribution that is not EC, Cook and Weisberg (1999a, ch. 8) demonstrate that a 1D regression can be visualized using a three–dimensional plot with Y on the vertical axes and the two predictors on the horizontal and out of page axes. Rotate the plot about the vertical axes. Each combination of the predictors gives a two dimensional "view." Search for the view with a smooth mean function that has the smallest possible variance function and use this view as the estimated sufficient summary plot.

For higher dimensions, Cook and Nachtsheim (1994) and Cook (1998a, p. 152) demonstrate that the bias $\boldsymbol{u}_{m,X}$ can often be made small by ellipsoidal trimming. To perform ellipsoidal trimming, an estimator $(T, \boldsymbol{C})$ is computed where T is a $(p-1) \times 1$ multivariate location estimator and $\boldsymbol{C}$ is a $(p-1) \times (p-1)$ symmetric positive definite dispersion estimator. Then the ith squared Mahalanobis distance is the random variable

$$D_i^2 = (\boldsymbol{x}_i - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x}_i - T) \tag{15.14}$$

for each vector of observed predictors $\boldsymbol{x}_i$. If the ordered distances $D_{(j)}$ are unique, then j of the $\boldsymbol{x}_i$ are in the hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T C^{-1} (\boldsymbol{x} - T) \leq D_{(j)}^2\}. \tag{15.15}$$

The ith case $(Y_i, \boldsymbol{x}_i^T)^T$ is trimmed if $D_i > D_{(j)}$. Thus if $j \approx 0.9n$, then about 10% of the cases are trimmed.

We suggest that the estimator $(T, \boldsymbol{C})$ should be the classical sample mean and covariance matrix $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ or a robust estimator such as `covfch`. When $j \approx n/2$, the `covfch` estimator attempts to make the volume of the hyperellipsoid given by Equation (15.15) small.

Ellipsoidal trimming seems to work for at least three reasons. The trimming divides the data into two groups: the *trimmed cases* and the *remaining cases* $(\boldsymbol{x}_M, Y_M)$ where $M\%$ is the amount of trimming, eg $M = 10$ for 10% trimming. If the distribution of the predictors $\boldsymbol{x}$ is EC then the distribution of $\boldsymbol{x}_M$ still retains enough symmetry so that the bias vector is approximately zero. If the distribution of $\boldsymbol{x}$ is not EC, then the distribution of $\boldsymbol{x}_M$ will often have enough symmetry so that the bias vector is small. In particular, trimming often removes strong nonlinearities from the predictors and the weighted predictor distribution is more nearly elliptically symmetric than the predictor distribution of the entire data set (recall Winsor's principle: "all data are roughly Gaussian in the middle"). Secondly, under heavy trimming, the mean function of the remaining cases may be more linear than the mean function of the entire data set. Thirdly, if $|c|$ is very large, then the bias vector may be small relative to $c\boldsymbol{\beta}$. Trimming sometimes inflates $|c|$. From Theorem 15.3, any of these three reasons should produce a better estimated sufficient predictor.

Example 15.3. Cook and Weisberg (1999a, p. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. The variables are the *muscle mass* M in grams, the *length* L and *height* H of the shell in mm, the *shell width* W and the *shell mass* S. The robust and classical Mahalanobis distances were calculated, and Figure 15.1 shows a scatterplot matrix of the mussel data, the RD_i's, and the MD_i's. Notice that many of the subplots are nonlinear. The cases marked by open circles were given weight zero by the `cov.mcd` algorithm, and the linearity of the retained cases has increased. Note that only one trimming proportion is shown and that a heavier trimming proportion would increase the linearity of the cases that were not trimmed.

The two ideas of using ellipsoidal trimming to reduce the bias and choosing a view with a smooth mean function and smallest variance function can be combined into a graphical method for finding the estimated sufficient summary plot and the estimated sufficient predictor. Trim the $M\%$ of the cases with the largest Mahalanobis distances, and then compute the OLS estima-

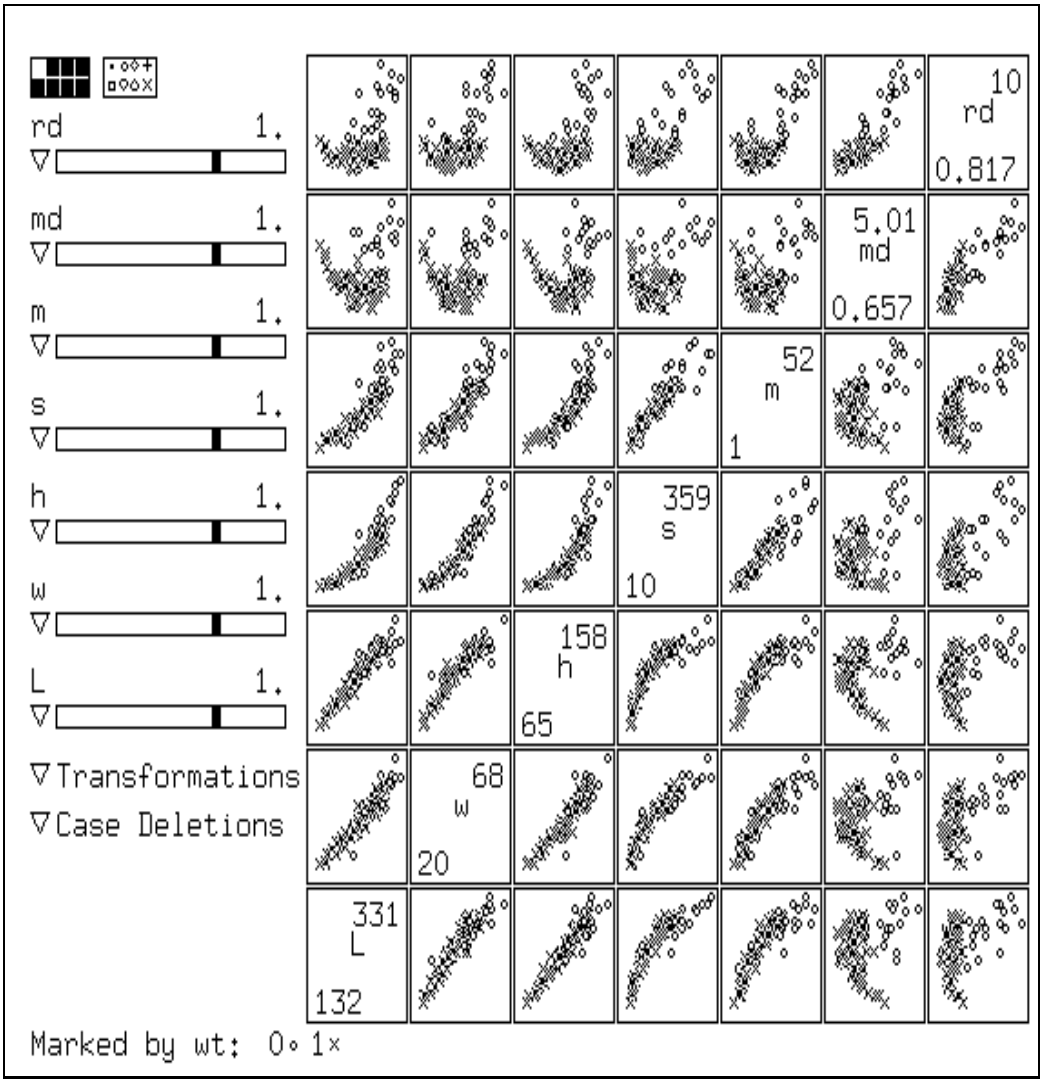

Figure 15.1: Scatterplot for Mussel Data, o Corresponds to Trimmed Cases

tor $(\hat{\alpha}_M, \hat{\boldsymbol{\beta}}_M)$ from the cases that remain. Use $M = 0$, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten plots of $\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}$ versus Y using all n cases. In analogy with the Cook and Weisberg procedure for visualizing 1D structure with two predictors, the plots will be called "trimmed views." Notice that $M = 0$ corresponds to the OLS view.

Definition 15.9. The *best trimmed view* is the trimmed view with a smooth mean function and the smallest variance function and is the estimated sufficient summary plot. If $M^* = E$ is the percentage of cases trimmed that corresponds to the best trimmed view, then $\hat{\boldsymbol{\beta}}_E^T \boldsymbol{x}$ or $\hat{\alpha}_E + \hat{\boldsymbol{\beta}}_E^T \boldsymbol{x}$ is the estimated sufficient predictor.

The following examples illustrate the *R/Splus regpack* function `trviews` that is used to produce the ESSP. If *R* is used instead of *Splus*, the command

```
library(MASS)
```

needs to be entered to access the function `cov.mcd` called by `trviews`. The robust estimators `cov.fch` and `cov.mba`can also be used. The function `trviews` is used in Problem 15.6. The estimator can be used to simultaneously detect whether the data is following a multiple linear regression model or some other single index model. Plot $\hat{\alpha}_E + \hat{\boldsymbol{\beta}}_E^T \boldsymbol{x}$ versus Y and add the identity line. If the plotted points follow the identity line then the MLR model is reasonable, but if the plotted points follow a nonlinear mean function, then a nonlinear single index model may be reasonable.

Example 15.2 continued. The command

```
trviews(X, Y)
```

produced the following output.

```
Intercept       X1       X2       X3
0.6701255 3.133926 4.031048 7.593501
Intercept       X1       X2       X3
 1.101398 8.873677 12.99655 18.29054
Intercept       X1       X2       X3
0.9702788 10.71646 15.40126 23.35055
Intercept       X1       X2       X3
0.5937255 13.44889 23.47785 32.74164
```

```
Intercept       X1       X2       X3
 1.086138 12.60514 25.06613 37.25504
Intercept       X1       X2       X3
 4.621724 19.54774 34.87627 48.79709
Intercept       X1       X2       X3
 3.165427 22.85721 36.09381 53.15153
Intercept       X1       X2       X3
 5.829141 31.63738 56.56191 82.94031
Intercept       X1       X2       X3
 4.241797 36.24316 70.94507 105.3816
Intercept       X1       X2       X3
 6.485165 41.67623 87.39663 120.8251
```

The function generates 10 trimmed views. The first plot trims 90% of the cases while the last plot does not trim any of the cases and is the OLS view. To advance a plot, press the right button on the mouse (in *R*, highlight `stop` rather than `continue`). After all of the trimmed views have been generated, the output is presented. For example, the 5th line of numbers in the output corresponds to $\hat{\alpha}_{50} = 1.086138$ and $\hat{\boldsymbol{\beta}}_{50}^T$ where 50% trimming was used. The second line of numbers corresponds to 80% trimming while the last line corresponds to 0% trimming and gives the OLS estimate $(\hat{\alpha}_0, \hat{\boldsymbol{\beta}}_0^T) = (\hat{a}, \hat{\boldsymbol{b}})$. The trimmed views with 50% and 90% trimming were very good. We decided that the view with 50% trimming was the best. Hence $\hat{\boldsymbol{\beta}}_E = (12.60514, 25.06613, 37.25504)^T \approx 12.5\boldsymbol{\beta}$. The best view is shown in Figure 15.2 and is nearly identical to the sufficient summary plot shown in Figure 1.11. Notice that the OLS estimate $= (41.68, 87.40, 120.83)^T \approx 42\boldsymbol{\beta}$. The OLS view is shown in Figure 1.13, and is again very similar to the sufficient summary plot, but it is not quite as smooth as the best trimmed view.

The plot of the estimated sufficient predictor versus the sufficient predictor is also informative. Of course this plot can usually only be generated for simulated data since $\boldsymbol{\beta}$ is generally unknown. If the plotted points are highly correlated (with $|\text{corr}(\text{ESP},\text{SP})| > 0.95$) and follow a line through the origin, then the estimated sufficient summary plot is nearly as good as the sufficient summary plot. The simulated data used $\boldsymbol{\beta} = (1, 2, 3)^T$, and the commands

```
SP <- X %*% 1:3
ESP <- X %*% c(12.60514, 25.06613, 37.25504)
plot(ESP,SP)
```

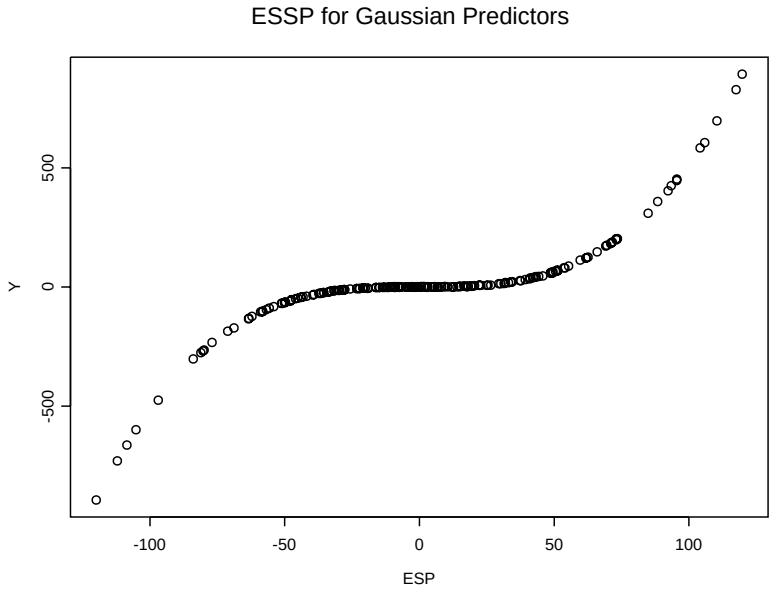

Figure 15.2: Best View for Estimating $m(u) = u^3$

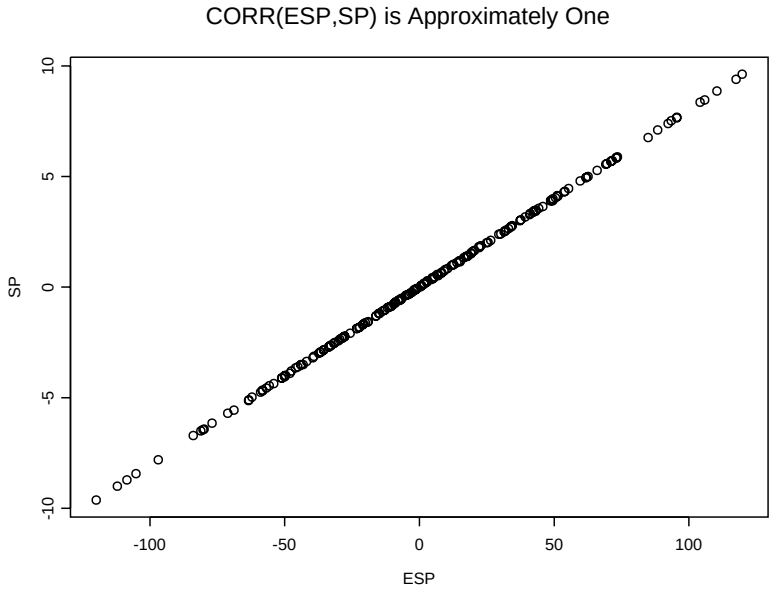

Figure 15.3: The angle between the SP and the ESP is nearly zero.

generated the plot shown in Figure 15.3.

Example 15.4. An artificial data set with 200 trivariate vectors $\boldsymbol{x}_i$ was generated. The marginal distributions of $x_{i,j}$ are iid lognormal for $j = 1, 2$, and 3. Since the response $Y_i = \sin(\boldsymbol{\beta}^T\boldsymbol{x}_i)/\boldsymbol{\beta}^T\boldsymbol{x}_i$ where $\boldsymbol{\beta} = (1, 2, 3)^T$, the random vector $\boldsymbol{x}_i$ is not elliptically contoured and the function m is strongly nonlinear. Figure 15.5 shows the OLS view where $\hat{\boldsymbol{\beta}}_0^T = (0.0032, 0.0011, 0.0047)^T$ and Figure 15.4 shows the best trimmed view where $\hat{\boldsymbol{\beta}}_{90}^T = (0.086, 0.182, 0.338)^T \approx 0.1\boldsymbol{\beta}$, roughly. Notice that it is difficult to visualize the mean function with the OLS view, and notice that the correlation between Y and the ESP is very low. By focusing on a part of the data where the correlation is high, it may be possible to improve the estimated sufficient summary plot. For example, in Figure 15.4, temporarily omit cases that have ESP less than 0.3 and greater than 0.75. From the untrimmed cases, obtained the ten trimmed estimates $\hat{\boldsymbol{\beta}}_{90}, ..., \hat{\boldsymbol{\beta}}_0$. Then using *all of the data*, obtain the ten views. The best view could be used as the ESSP.

Application 15.1. Suppose that a 1D regression analysis is desired on a data set, use the trimmed views as an exploratory data analysis technique to visualize the conditional distribution $Y|\boldsymbol{\beta}^T\boldsymbol{x}$. The best trimmed view is an estimated sufficient summary plot. If the single index model (15.3) holds, the function m can be estimated from this plot using parametric models or scatterplot smoothers such as `lowess`. Notice that Y can be predicted visually using *up and over lines*.

Application 15.2. The best trimmed view can also be used as a diagnostic for linearity and monotonicity.

For example in Figure 15.2, if ESP = 0, then $\hat{Y} = 0$ and if ESP = 100, then $\hat{Y} = 500$. Figure 15.2 suggests that the mean function is monotone but not linear, and Figure 15.4 suggests that the mean function is neither linear nor monotone.

Application 15.3. Assume that a known 1D regression model is assumed for the data. Then the best trimmed view can be used as a diagnostic for whether the assumed model is appropriate.

The trimmed views are sometimes useful even when the assumption of linearly related predictors fails. OLS frequently performs well if there are no strong nonlinearities present in the predictors.

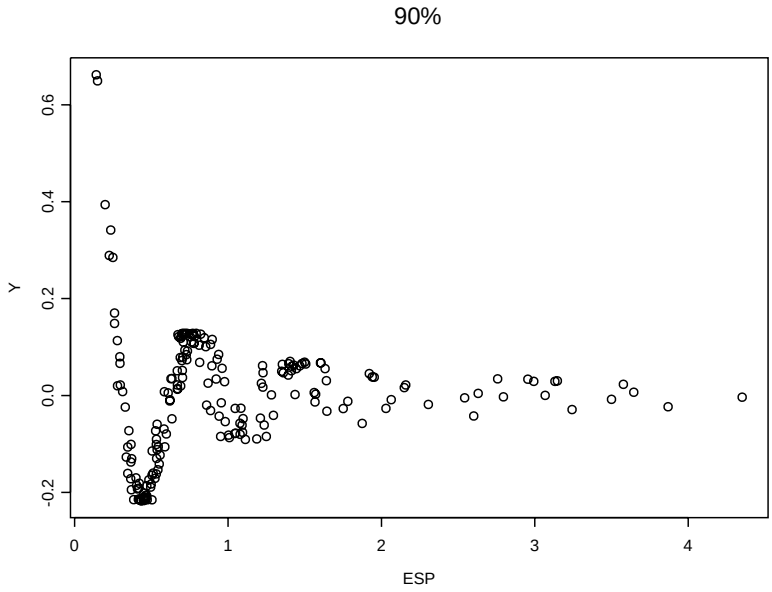

Figure 15.4: OLS View with 90% Trimming

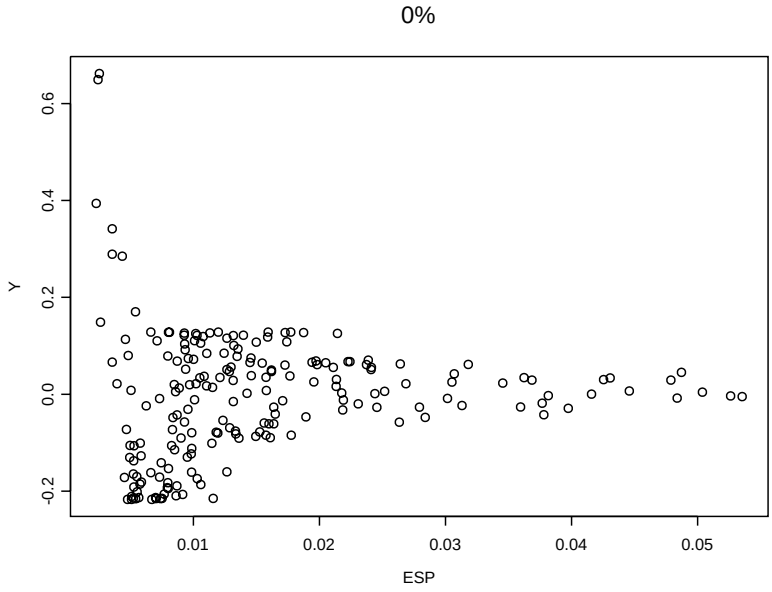

Figure 15.5: OLS View with 0% Trimming

15.3 Predictor Transformations

As a general rule, inferring about the distribution of $Y|\boldsymbol{X}$ from a lower dimensional plot should be avoided when there are strong nonlinearities among the predictors.
Cook and Weisberg (1999b, p. 34)

Even if the multiple linear regression model is valid, a model based on a subset of the predictor variables depends on the predictor distribution. If the predictors are linearly related (eg EC), then the submodel mean and variance functions are generally well behaved, but otherwise the submodel mean function could be nonlinear and the submodel variance function could be nonconstant. For 1D regression models, the presence of strong nonlinearities among the predictors can invalidate inferences. A necessary condition for $\boldsymbol{x}$ to have an EC distribution (or for no strong nonlinearities to be present among the predictors) is for each marginal plot of the scatterplot matrix of the predictors to have a linear or ellipsoidal shape if n is large.

One of the most useful techniques in regression is to remove gross nonlinearities in the predictors by using predictor transformations. Power transformations are particularly effective. A multivariate version of the Box–Cox transformation due to Velilla (1993) can cause the distribution of the transformed predictors to be closer to multivariate normal, and the Cook and Nachtsheim (1994) procedure can cause the distribution to be closer to elliptical symmetry. Marginal Box-Cox transformations also seem to be effective. Power transformations can also be selected with slider bars in *Arc*.

There are several rules for selecting marginal transformations visually. (Also see discussion in Section 3.1.) First, use theory if available. Suppose that variable X_2 is on the vertical axis and X_1 is on the horizontal axis and that the plot of X_1 versus X_2 is nonlinear. The *unit rule* says that if X_1 and X_2 have the same units, then try the same transformation for both X_1 and X_2.

Power transformations are also useful. Assume that all values of X_1 and X_2 are positive. Let λ be the power of the transformation. Then the following four rules are often used.

The *log rule* states that positive predictors that have the ratio between their largest and smallest values greater than ten should be transformed to logs. See Cook and Weisberg (1999a, p. 87).

Secondly, if it is known that $X_2 \approx X_1^{\lambda}$ and the ranges of X_1 and X_2 are

such that this relationship is one to one, then

$$X_1^{\lambda} \approx X_2 \text{ and } X_2^{1/\lambda} \approx X_1.$$

Hence either the transformation X_1^{λ} or $X_2^{1/\lambda}$ will linearize the plot. This relationship frequently occurs if there is a volume present. For example let X_2 be the volume of a sphere and let X_1 be the circumference of a sphere. The plot of $\log(X_1)$ versus $\log(X_2)$ will also be linear.

Thirdly, the *bulging rule* states that changes to the power of X_2 and the power of X_1 can be determined by the direction that the bulging side of the curve points. If the curve is hollow up (the bulge points down), decrease the power of X_2. If the curve is hollow down (the bulge points up), increase the power of X_2 If the curve bulges towards large values of X_1 increase the power of X_1. If the curve bulges towards small values of X_1 decrease the power of X_1. See Tukey (1977, p. 173–176).

Finally, Cook and Weisberg (1999a, p. 86) give the following rule.
To spread *small* values of a variable, make λ *smaller*.
To spread *large* values of a variable, make λ *larger*.

For example, in Figure 15.10c, small values of Y and large values of FESP need spreading, and using $\log(Y)$ would make the plot more linear.

15.4 Variable Selection

A standard problem in 1D regression is variable selection, also called subset or model selection. Assume that model (15.1) holds, that a constant is always included, and that $\boldsymbol{x} = (x_1, ..., x_{p-1})^T$ are the $p-1$ nontrivial predictors, which we assume to be of full rank. Then *variable selection* is a search for a subset of predictor variables that can be deleted without important loss of information. This section follows Olive and Hawkins (2005) closely.

Variable selection for the 1D regression model is very similar to variable selection for the multiple linear regression model (see Section 3.4). To clarify ideas, assume that there exists a subset S of predictor variables such that if $\boldsymbol{x}_S$ is in the 1D model, then none of the other predictors are needed in the model. Write E for these ('extraneous') variables not in S, partitioning $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$. Then

$$SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S. \quad (15.16)$$

The extraneous terms that can be eliminated given that the subset S is in the model have zero coefficients.

Now suppose that I is a candidate subset of predictors, that $S \subseteq I$ and that O is the set of predictors not in I. Then

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

(if I includes predictors from E, these will have zero coefficient). For any subset I that contains the subset S of relevant predictors, the correlation

$$\text{corr}(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}, \alpha + \boldsymbol{\beta}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{I,i}}) = 1. \tag{15.17}$$

This observation, which is true regardless of the explanatory power of the model, suggests that variable selection for 1D regression models is simple in principle. For each value of $j = 1, 2, ..., p-1$ nontrivial predictors, keep track of subsets I that provide the largest values of corr(ESP,ESP(I)). Any such subset for which the correlation is high is worth closer investigation and consideration. To make this advice more specific, use the *rule of thumb* that a candidate subset of predictors I is worth considering if the sample correlation of ESP and ESP(I) satisfies

$$\text{corr}(\hat{\alpha} + \hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}, \hat{\alpha}_{\mathrm{I}} + \hat{\boldsymbol{\beta}}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{I,i}}) = \text{corr}(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}, \hat{\boldsymbol{\beta}}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{I,i}}) \geq 0.95. \tag{15.18}$$

The difficulty with this approach is that fitting all of the possible submodels involves substantial computation. An exception to this difficulty is multiple linear regression where there are efficient "leaps and bounds" algorithms for searching all subsets when OLS is used (see Furnival and Wilson 1974). Since OLS often gives a useful ESP, the following all subsets procedure can be used for 1D models when $p < 20$.

- Fit a full model using the methods appropriate to that 1D problem to find the ESP $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$.
- Find the OLS ESP $\hat{\alpha}_{OLS} + \hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x}$.
- If the 1D ESP and the OLS ESP have "a strong linear relationship" (for example $|\text{corr}(\text{ESP}, \text{OLS ESP})| \geq 0.95$), then infer that the 1D problem is one in which OLS may serve as an adequate surrogate for the correct 1D model fitting procedure.

- Use computationally fast OLS variable selection procedures such as forward selection, backward elimination and the leaps and bounds algorithm along with the Mallows (1973) C_p criterion to identify predictor subsets I containing k variables (including the constant) with $C_p(I) \leq \min(2k, p)$.

- Perform a final check on the subsets that satisfy the C_p screen by using them to fit the 1D model.

For a 1D model, the response, ESP and vertical discrepancies $V = Y - ESP$ are important. When the multiple linear regression (MLR) model holds, the fitted values are the ESP: $\hat{Y} = ESP$, and the vertical discrepancies are the residuals.

Definition 15.10. a) The plot of $\tilde{\alpha}_I + \tilde{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ versus $\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ is called an *EE plot* (often called an FF plot for MLR).
b) The plot of discrepancies $Y_i - \tilde{\alpha}_I - \tilde{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ versus $Y_i - \tilde{\alpha} - \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ is called a *VV plot* (often called an RR plot for MLR).
c) The plots of $\tilde{\alpha}_I + \tilde{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ versus Y_i and of $\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Y_i are called *estimated sufficient summary plots* or *response plots*.

Many numerical methods such as forward selection, backward elimination, stepwise and all subset methods using the C_p criterion (Jones 1946, Mallows 1973), have been suggested for variable selection. The four plots in Definition 15.10 contain valuable information to supplement the raw numerical results of these selection methods. Particular uses include:

- The key to understanding which plots are the most useful is the observation that a wz plot *is used to visualize the conditional distribution of z given w. Since a 1D regression is the study of the conditional distribution* of Y given $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$, *the response plot is used to visualize this conditional distribution and should always be made.* A major problem with variable selection is that deleting important predictors can change the functional form m of the model. In particular, if a multiple linear regression model is appropriate for the full model, linearity may be destroyed if important predictors are deleted. When the single index model (15.3) holds, m can be visualized with a response plot. Adding visual aids such as the estimated parametric mean function

$m(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})$ can be useful. If an estimated nonparametric mean function $\hat{m}(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})$ such as lowess follows the parametric curve closely, then often numerical goodness of fit tests will suggest that the model is good. See Chambers, Cleveland, Kleiner, and Tukey (1983, p. 280) and Cook and Weisberg (1999a, p. 425, 432). For variable selection, *the response plots from the full model and submodel should be very similar if the submodel is good.*

- Sometimes outliers will influence numerical methods for variable selection. Outliers tend to stand out in at least one of the plots. An EE plot is useful for variable selection because the correlation of ESP(I) and ESP is important. The EE plot can be used to quickly check that the correlation is high, that the plotted points fall about some line, that the line is the identity line, and that the correlation is high because the relationship is linear, rather than because of outliers.

- Numerical methods may include too many predictors. Investigators can examine the p–values for individual predictors, but the assumptions needed to obtain valid p–values are often violated; however, the OLS t tests for individual predictors are meaningful since deleting a predictor changes the C_p value by $t^2 - 2$ where t is the test statistic for the predictor. See Section 15.5, Daniel and Wood (1980, p. 100-101) and the following two remarks.

Remark 15.5. Variable selection with the C_p criterion is closely related to the partial F test that uses test statistic F_I. Suppose that the full model contains p predictors including a constant and the submodel I includes k predictors including a constant. If $n \geq 10p$, then the submodel I is "interesting" if $C_p(I) \leq \min(2k, p)$.

To see this claim notice that *the following results are properties of OLS and hold even if the data does not follow a 1D model.* If the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then

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

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

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k \tag{15.19}$$

where MSE is the "residual" mean square for the full model. Let $\text{ESP}(I) = \hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_I$ be the ESP for the submodel and let $V_I = Y - ESP(I)$ so that $V_{I,i} = Y_i - \hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$. Let ESP and V denote the corresponding quantities for the full model. Using Proposition 3.2 and Remark 3.2 with $\text{corr}(r, r_I)$ replaced by $\text{corr}(V, V_I)$, it can be shown that

$$corr(V, V_I) = \sqrt{\frac{SSE}{SSE(I)}} = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}.$$

It can also be shown that $C_p(I) \leq 2k$ corresponds to $\text{corr}(V, V_I) \geq d_n$ where

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

Notice that for a fixed value of k, the submodel I_k that minimizes $C_p(I)$ also maximizes $corr(V, V_I)$. If $C_p(I) \leq 2k$ and $n \geq 10p$, then $0.948 \leq \text{corr}(V, V_I)$, and both $\text{corr}(V, V_I) \to 1.0$ and $\text{corr}(\text{OLS ESP, OLS ESP}(I)) \to 1.0$ as $n \to \infty$. Hence the plotted points in both the VV plot and the EE plot will cluster about the identity line (see Proposition 3.2).

Remark 15.6. Suppose that the OLS ESP and the standard ESP are highly correlated: $|\text{corr}(\text{ESP}, \text{OLS ESP})| \geq 0.95$. Then often OLS variable selection can be used for the 1D data, and using the p–values from OLS output seems to be a useful benchmark. To see this, suppose that $n > 5p$ and first consider the model I_i that deletes the predictor X_i. Then model I_i has $k = p - 1$ predictors including the constant, and the test statistic is t_i where

$$t_i^2 = F_{I_i}.$$

Using (15.19) and $C_p(I_{full}) = p$, it can be shown that

$$C_p(I_i) = C_p(I_{full}) + (t_i^2 - 2).$$

Using the screen $C_p(I) \leq \min(2k, p)$ suggests that the predictor X_i should not be deleted if

$$|t_i| > \sqrt{2} \approx 1.414.$$

If $|t_i| < \sqrt{2}$ then the predictor can probably be deleted since C_p decreases.

More generally, it can be shown that $C_p(I) \leq 2k$ iff

$$F_I \leq \frac{p}{p-k}.$$

Now k is the number of terms in the model including a constant while $p-k$ is the number of terms set to 0. As $k \to 0$, the partial F test will reject Ho: $\boldsymbol{\beta}_O = \mathbf{0}$ (ie, say that the full model should be used instead of the submodel I) unless F_I is not much larger than 1. If p is very large and $p-k$ is very small, then the change in SS F test will tend to suggest that there is a model I that is about as good as the full model even though model I deletes $p-k$ predictors.

The $C_p(I) \leq k$ screen tends to overfit. We simulated multiple linear regression and single index model data sets with $p = 8$ and $n = 50, 100, 1000$ and 10000. The true model S satisfied $C_p(S) \leq k$ for about 60% of the simulated data sets, but S satisfied $C_p(S) \leq 2k$ for about 97% of the data sets.

In many settings, not all of which meet the Li–Duan sufficient conditions, the full model OLS ESP is a good estimator of the sufficient predictor. If the fitted full 1D model $Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$ is a useful approximation to the data and if $\hat{\boldsymbol{\beta}}_{OLS}$ is a good estimator of $c\boldsymbol{\beta}$ where $c \neq 0$, then a subset I will produce a response plot similar to the response plot of the full model if corr(OLS ESP, OLS ESP(I)) ≥ 0.95. Hence the response plots based on the full and submodel ESP can both be used to visualize the conditional distribution of Y.

Assuming that a 1D model holds, a common assumption made for variable selection is that the fitted full model ESP is a good estimator of the sufficient predictor, and the usual numerical and graphical checks on this assumption should be made. To see that this assumption is weaker than the assumption that the OLS ESP is good, notice that if a 1D model holds but $\hat{\boldsymbol{\beta}}_{OLS}$ estimates $c\boldsymbol{\beta}$ where $c = 0$, then the $C_p(I)$ criterion could wrongly suggest that all subsets I have $C_p(I) \leq 2k$. Hence we also need to check that $c \neq 0$.

There are several methods are for checking the OLS ESP, including: a) if an ESP from an alternative fitting method is believed to be useful, check that the ESP and the OLS ESP have a strong linear relationship: for example that |corr(ESP, OLS ESP)| ≥ 0.95. b) Often examining the OLS response plot shows that a 1D model is reasonable. For example, if the data are tightly clustered about a smooth curve, then a single index model may be appropriate. c) Verify that a 1D model is appropriate using graphical techniques given by Cook and Weisberg (1999a, p. 434-441). d) Verify that $\boldsymbol{x}$ has an EC distribution with nonsingular covariance matrix and that the mean function $m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$ is not symmetric about the median of the distribution of

$\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$. Then results from Li and Duan (1989) suggest that $c \neq 0$.

Condition a) is both the most useful (being a direct performance check) and the easiest to check. A standard fitting method should be used when available (eg, for parametric 1D models such as GLMs). Conditions c) and d) need $\boldsymbol{x}$ to have a continuous multivariate distribution while the predictors can be factors for a) and b). Using trimmed views results in an ESP that can sometimes cause condition b) to hold when d) is violated.

To summarize, variable selection procedures, originally meant for MLR, can often be used for 1D data. If the fitted full 1D model $Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$ is a useful approximation to the data and if $\hat{\boldsymbol{\beta}}_{OLS}$ is a good estimator of $c\boldsymbol{\beta}$ where $c \neq 0$, then a subset I is good if corr(OLS ESP, OLS ESP(I)) ≥ 0.95. If n is large enough, Remark 15.5 implies that this condition will hold if $C_p(I) \leq 2k$ or if $F_I \leq 1$. This result suggests that within the (large) subclass of 1D models where the OLS ESP is useful, the OLS partial F test is robust (asymptotically) to model misspecifications in that $F_I \leq 1$ correctly suggests that submodel I is good. The OLS t tests for individual predictors are also meaningful since if $|t| < \sqrt{2}$ then the predictor can probably be deleted since C_p decreases while if $|t| \geq 2$ then the predictor is probably useful even when the other predictors are in the model. Section 15.5 provides related theory, and the following examples help illustrate the above discussion.

Example 15.5. This example illustrates that the plots are useful for general 1D regression models such as the response transformation model. Cook and Weisberg (1999a, p. 351, 433, 447, 463) describe a data set on 82 mussels. The response Y is the *muscle mass* in grams, and the four predictors are the *logarithms of the shell length, width, height and mass*. The logarithm transformation was used to remove strong nonlinearities that were evident in a scatterplot matrix of the untransformed predictors. The C_p criterion suggests using log(width) and log(shell mass) as predictors. The EE and VV plots are shown in Figure 15.6ab. The response plots based on the full and submodel are shown in Figure 15.6cd and are nearly identical, but not linear.

When log(muscle mass) is used as the response, the C_p criterion suggests using log(height) and log(shell mass) as predictors (the correlation between log(height) and log(width) is very high). Figure 15.7a shows the RR plot and 2 outliers are evident. These outliers correspond to the two outliers in the response plot shown in Figure 15.7b. After deleting the outliers, the C_p criterion still suggested using log(height) and log(shell mass) as predictors.

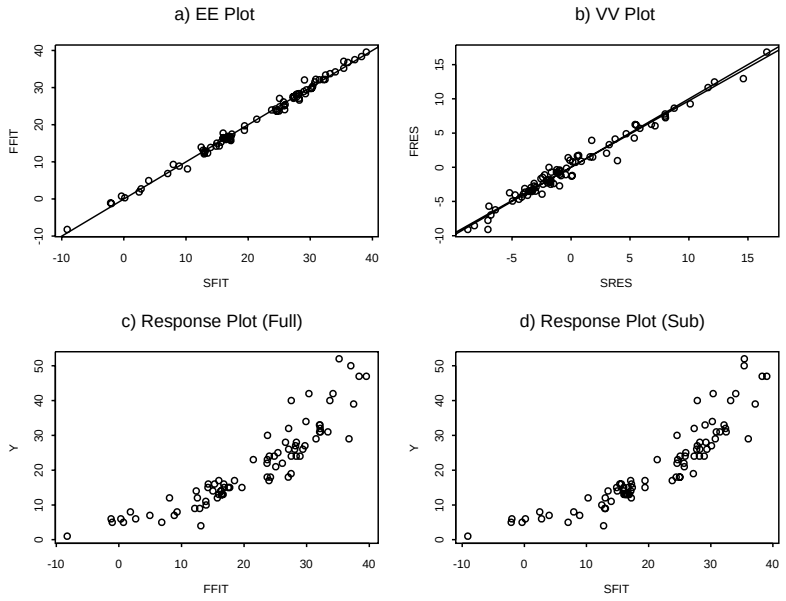

Figure 15.6: Mussel Data with Muscle Mass as the Response

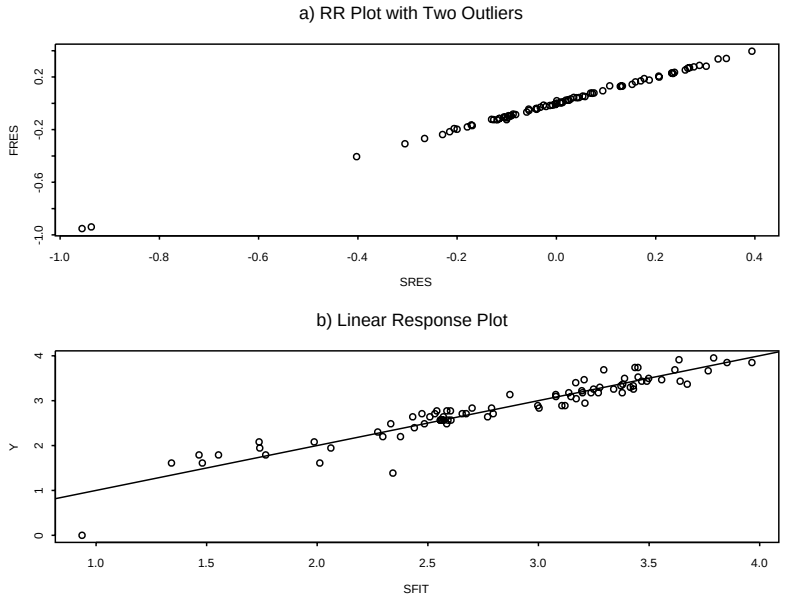

Figure 15.7: Mussel Data with log(Muscle Mass) as the Response

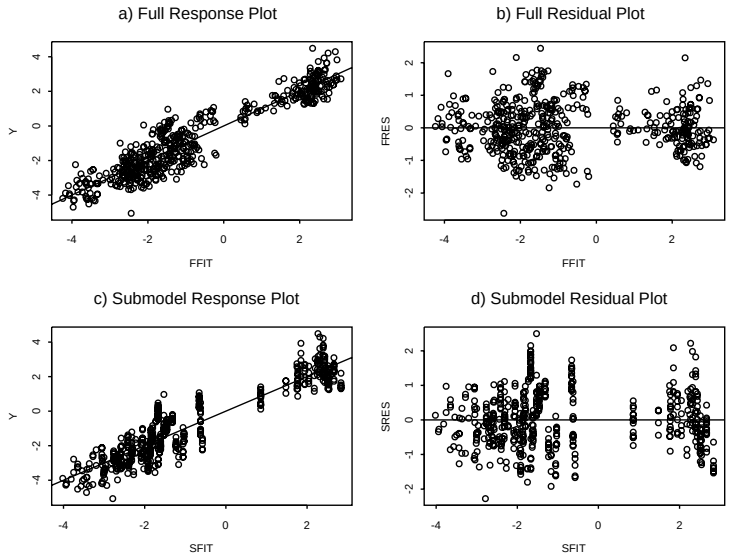

Figure 15.8: Response and Residual Plots for Boston Housing Data

The p–value for including log(height) in the model was 0.03, and making the FF and RR plots after deleting log(height) suggests that log(height) may not be needed in the model.

Example 15.6 According to Li (1997), the predictors in the Boston housing data of Harrison and Rubinfeld (1978) have a nonlinear quasi–helix relationship which can cause regression graphics methods to fail. Nevertheless, the graphical diagnostics can be used to gain interesting information from the data. The response $Y = \log(CRIM)$ where CRIM is the per capita crime rate by town. The predictors used were x_1 = proportion of residential land zoned for lots over 25,000 sq.ft., $\log(x_2)$ where x_2 is the proportion of non-retail business acres per town, x_3 = Charles River dummy variable (= 1 if tract bounds river; 0 otherwise), $x_4 = NOX$ = nitric oxides concentration (parts per 10 million), x_5 = average number of rooms per dwelling, x_6 = proportion of owner-occupied units built prior to 1940, $\log(x_7)$ where x_7 = weighted distances to five Boston employment centers, $x_8 = RAD$ = index of accessibility to radial highways, $\log(x_9)$ where x_9 = full-value property-tax rate per \$10,000, x_{10} = pupil-teacher ratio by town, $x_{11} = 1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town, $\log(x_{12})$ where $x_{12} = \%$ lower

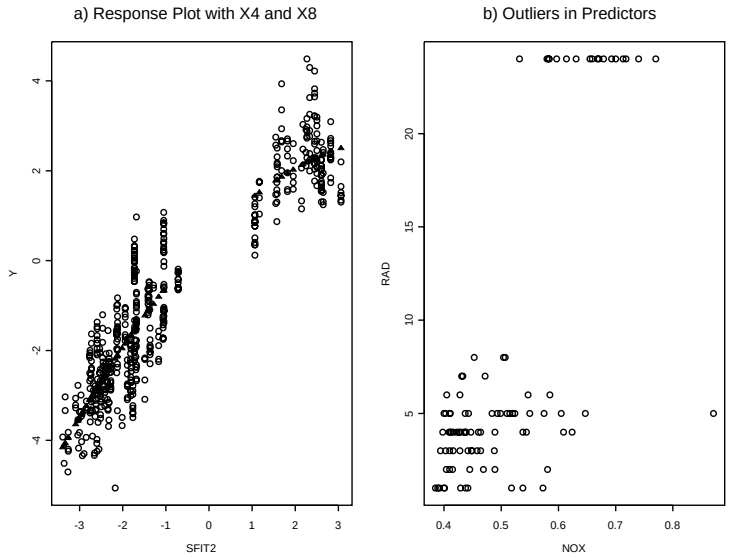

Figure 15.9: Relationships between NOX, RAD and Y = log(CRIM)

status of the population, and $\log(x_{13})$ where x_{13} = median value of owner-occupied homes in \$1000's. The full model has 506 cases and 13 nontrivial predictor variables.

Figure 15.8ab shows the response plot and residual plot for the full model. The residual plot suggests that there may be three or four groups of data, but a linear model does seem plausible. Backward elimination with C_p suggested the "min C_p submodel" with the variables $x_1, \log(x_2), NOX, x_6, \log(x_7), RAD, x_{10}, x_{11}$ and $\log(x_{13})$. The full model had $R^2 = 0.878$ and $\hat{\sigma} = 0.7642$. The C_p submodel had $C_p(I) = 6.576, R_I^2 = 0.878$, and $\hat{\sigma}_I = 0.762$. Deleting $\log(x_7)$ resulted in a model with $C_p = 8.483$ and the smallest coefficient p–value was 0.0095. The FF and RR plots for this model (not shown) looked like the identity line. Examining further submodels showed that NOX and RAD were the most important predictors. In particular, the OLS coefficients of x_1, x_6 and x_{11} were orders of magnitude smaller than those of NOX and RAD. The submodel including a constant, NOX, RAD and $\log(x_2)$ had $R^2 = 0.860$, $\hat{\sigma} = 0.811$ and $C_p = 67.368$. Figure 15.8cd shows the response plot and residual plot for this submodel.

Although this submodel has nearly the same R^2 as the full model, the residuals show more variability than those of the full model. Nevertheless,

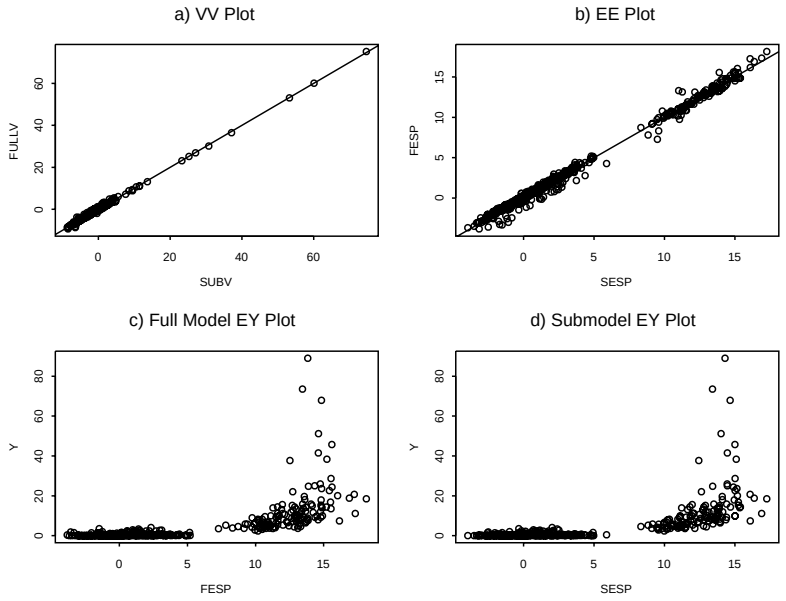

Figure 15.10: Boston Housing Data: Nonlinear 1D Regression Model

we can examine the effect of NOX and RAD on the response by deleting $\log(x_2)$. This submodel had $R^2 = 0.842$, $\hat{\sigma} = 0.861$ and $C_p = 138.727$. Figure 15.9a shows that the response plot for this model is no longer linear. The residual plot (not shown) also displays curvature. Figure 15.9a shows that there are two groups, one with high Y and one with low Y. There are three clusters of points in the plot of NOX versus RAD shown in Figure 15.9b (the single isolated point in the southeast corner of the plot actually corresponds to several cases). The two clusters of high NOX and high RAD points correspond to the cases with high per capita crime rate.

The tiny filled in triangles if Figure 15.9a represent the fitted values for a quadratic. We added NOX^2, RAD^2 and $NOX * RAD$ to the full model and again tried variable selection. Although the full quadratic in NOX and RAD had a linear response plot, the submodel with NOX, RAD and $\log(x_2)$ was very similar. For this data set, NOX and RAD seem to be the most important predictors, but other predictors are needed to make the model linear and to reduce residual variation.

Example 15.7. In the Boston housing data, now let $Y = CRIM$. Since

$\log(Y)$ has a linear relationship with the predictors, Y should follow a nonlinear 1D regression model. Consider the full model with predictors $\log(x_2)$, x_3, x_4, x_5, $\log(x_7)$, x_8, $\log(x_9)$ and $\log(x_{12})$. Regardless of whether Y or $\log(Y)$ is used as the response, the minimum C_p model from backward elimination used a constant, $\log(x_2)$, x_4, $\log(x_7)$, x_8 and $\log(x_{12})$ as predictors. If Y is the response, then the model is nonlinear and $C_p = 5.699$. Remark 15.5 suggests that if $C_p \leq 2k$, then the points in the VV plot should tightly cluster about the identity line even if a multiple linear regression model fails to hold. Figure 15.10 shows the VV and EE plots for the minimum C_p submodel. The response (EY) plots for the full model and submodel are also shown. Note that the clustering in the VV plot is indeed higher than the clustering in the EE plot. Note that the response plots are highly nonlinear but are nearly identical.

15.5 Inference

This section follows Chang and Olive (2010) closely. Inference can be performed for trimmed views if M is chosen without using the response, eg if the trimming is done with a DD plot, and the dimension reduction (DR) method such as OLS is performed on the data $(Y_{Mi}, \boldsymbol{x}_{Mi})$ that remains after trimming $M\%$ of the cases with ellipsoidal trimming based on the MBA or FCH estimator.

First we review some theoretical results for OLS as a DR method and give the main theoretical result for OLS. Let

$$\mathrm{Cov}(\boldsymbol{x}) = \mathrm{E}[(\boldsymbol{x} - \mathrm{E}(\boldsymbol{x}))(\boldsymbol{x} - \mathrm{E}(\boldsymbol{x}))^{\mathrm{T}}] = \boldsymbol{\Sigma}_{\boldsymbol{x}}$$

and $\mathrm{Cov}(\boldsymbol{x}, Y) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(Y - E(Y))] = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Let the OLS estimator be $(\hat{\alpha}_{OLS}, \hat{\boldsymbol{\beta}}_{OLS})$. Then the population coefficients from an OLS regression of Y on $\boldsymbol{x}$ are

$$\alpha_{OLS} = E(Y) - \boldsymbol{\beta}_{OLS}^T E(\boldsymbol{x}) \text{ and } \boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}\mathrm{Y}}. \tag{15.20}$$

Let the data be $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. Let the $p \times 1$ vector $\boldsymbol{\eta} = (\alpha, \boldsymbol{\beta}^T)^T$, let $\boldsymbol{X}$ be the $n \times p$ OLS design matrix with ith row $(1, \boldsymbol{x}_i^T)$, and let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$. Then the OLS estimator $\hat{\boldsymbol{\eta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$. The sample covariance of $\boldsymbol{x}$ is

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

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

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

The first result shows that $\hat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$.

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

$$\hat{\alpha}_{OLS} = \overline{Y} - \hat{\boldsymbol{\beta}}_{OLS}^T\overline{\boldsymbol{x}} \xrightarrow{D} \alpha_{OLS}$$

and

$$\hat{\boldsymbol{\beta}}_{OLS} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \xrightarrow{D} \boldsymbol{\beta}_{OLS} \text{ as } \mathrm{n} \to \infty.$$

The following OLS results need some notation. Many 1D regression models have an error e with

$$\sigma^2 = \mathrm{Var(e)} = \mathrm{E(e^2)}. \tag{15.21}$$

Let $\hat{e}$ be the error residual for e. Let the population OLS residual

$$v = Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T\boldsymbol{x} \tag{15.22}$$

with

$$\tau^2 = E[(Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T\boldsymbol{x})^2] = E(v^2), \tag{15.23}$$

and let the OLS residual be

$$r = Y - \hat{\alpha}_{OLS} - \hat{\boldsymbol{\beta}}_{OLS}^T\boldsymbol{x}. \tag{15.24}$$

Typically the OLS residual r is not estimating the error e and $\tau^2 \neq \sigma^2$, but the following results show that the OLS residual is of great interest for 1D regression models.

Assume that a 1D model holds, $Y \perp\!\!\!\perp \boldsymbol{x}|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$, which is equivalent to $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}$. Then under regularity conditions, results ii) – iv) below hold.

ii) Li and Duan (1989): $\boldsymbol{\beta}_{OLS} = c\boldsymbol{\beta}$ for some constant c.

iii) Li and Duan (1989) and Chen and Li (1998):

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - c\boldsymbol{\beta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \boldsymbol{C}_{OLS}) \tag{15.25}$$

where

$$\boldsymbol{C}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}E[(Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T\boldsymbol{x})^2(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T]\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}. \tag{15.26}$$

iv) Chen and Li (1998): Let $\boldsymbol{A}$ be a known full rank constant $k \times (p-1)$ matrix. If the null hypothesis Ho: $\boldsymbol{A\beta} = \boldsymbol{0}$ is true, then

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS} - c\boldsymbol{A\beta}) = \sqrt{n}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS} \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{C}_{OLS}\boldsymbol{A}^T)$$

and

$$\boldsymbol{A}\boldsymbol{C}_{OLS}\boldsymbol{A}^T = \tau^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{A}^T. \tag{15.27}$$

Notice that $\boldsymbol{C}_{OLS} = \tau^2\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ if $v = Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{x}$ or if the MLR model holds. If the MLR model holds, $\tau^2 = \sigma^2$.

To create test statistics, the estimator

$$\hat{\tau}^2 = \text{MSE} = \frac{1}{\text{n} - \text{p}} \sum_{\text{i}=1}^{\text{n}} \text{r}_\text{i}^2 = \frac{1}{\text{n} - \text{p}} \sum_{\text{i}=1}^{\text{n}} (\text{Y}_\text{i} - \hat{\alpha}_\text{OLS} - \hat{\boldsymbol{\beta}}_\text{OLS}^\text{T}\boldsymbol{x}_\text{i})^2$$

will be useful. The estimator $\hat{\boldsymbol{C}}_{OLS} =$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} [(Y_i - \hat{\alpha}_{OLS} - \hat{\boldsymbol{\beta}}_{OLS}^T\boldsymbol{x}_i)^2 (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T] \right] \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \tag{15.28}$$

can also be useful. Notice that for general 1D regression models, the OLS MSE estimates τ^2 rather than the error variance σ^2.

v) Result iv) suggests that a test statistic for $Ho : \boldsymbol{A\beta} = \boldsymbol{0}$ is

$$W_{OLS} = n\hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{A}^T [\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\boldsymbol{A}^T]^{-1} \boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS}/\hat{\tau}^2 \xrightarrow{D} \chi_k^2, \tag{15.29}$$

the chi–square distribution with k degrees of freedom.

Before presenting the main theoretical result, some results from OLS MLR theory are needed. Let the $p \times 1$ vector $\boldsymbol{\eta} = (\alpha, \boldsymbol{\beta}^T)^T$, the known $k \times p$ constant matrix $\tilde{\boldsymbol{A}} = [\boldsymbol{a}\ \boldsymbol{A}]$ where $\boldsymbol{a}$ is a $k \times 1$ vector, and let $\boldsymbol{c}$ be a known $k \times 1$ constant vector. Following Seber and Lee (2003, p. 99–106), the usual F statistic for testing $Ho : \tilde{A}\boldsymbol{\eta} = \boldsymbol{c}$ is

$$F_0 = \frac{(SSE(Ho) - SSE)/k}{SSE/(n-p)} = \tag{15.30}$$

$$(\tilde{\boldsymbol{A}}\hat{\boldsymbol{\eta}} - \boldsymbol{c})^T [\tilde{\boldsymbol{A}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\tilde{\boldsymbol{A}}^T]^{-1} (\tilde{\boldsymbol{A}}\hat{\boldsymbol{\eta}} - \boldsymbol{c})/(k\hat{\tau}^2)$$

where $MSE = \hat{\tau}^2 = SSE/(n-p)$, $SSE = \sum_{i=1}^{n} r_i^2$ and

$$SSE(Ho) = \sum_{i=1}^{n} r_i^2(Ho)$$

is the minimum sum of squared residuals subject to the constraint $\tilde{\boldsymbol{A}}\boldsymbol{\eta} = \boldsymbol{c}$. Recall that if Ho is true, the MLR model holds and the errors e_i are iid $N(0, \sigma^2)$, then $F_o \sim F_{k,n-p}$, the F distribution with k and $n-p$ degrees of freedom. Also recall that if $Z_n \sim F_{k,n-p}$, then

$$Z_n \xrightarrow{D} \chi_k^2/k \tag{15.31}$$

as $n \to \infty$.

The main theoretical result of this section is Theorem 15.4 below. This theorem and (15.31) suggest that OLS output, originally meant for testing with the MLR model, can also be used for testing with many 1D regression data sets. Without loss of generality, let the 1D model $Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$ be written as

$$Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O)$$

where the reduced model is $Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha_R + \boldsymbol{\beta}_R^T \boldsymbol{x}_R)$ and $\boldsymbol{x}_O$ denotes the terms outside of the reduced model. Notice that OLS ANOVA F test corresponds to Ho: $\boldsymbol{\beta} = \boldsymbol{0}$ and uses $\boldsymbol{A} = \boldsymbol{I}_{p-1}$. The tests for Ho: $\beta_i = 0$ use $\boldsymbol{A} = (0, ..., 0, 1, 0, ..., 0)$ where the 1 is in the ith position and are equivalent to the OLS t tests. The test Ho: $\boldsymbol{\beta}_O = \boldsymbol{0}$ uses $\boldsymbol{A} = [\boldsymbol{0}\ \boldsymbol{I}_j]$ if $\boldsymbol{\beta}_O$ is a $j \times 1$ vector, and the test statistic (15.30) can be computed by running OLS on the full model to obtain SSE and on the reduced model to obtain $SSE(R) \equiv SSE(Ho)$.

In the theorem below, it is crucial that Ho: $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$. Tests for Ho: $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{1}$, say, may not be valid even if the sample size n is large. Also, confidence intervals corresponding to the t tests are for $c\boldsymbol{\beta}_i$, and are usually not very useful when c is unknown.

Theorem 15.4. Assume that a 1D regression model (15.1) holds and that Equation (15.29) holds when $Ho : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$ is true. Then the test statistic (15.30) satisfies

$$F_0 = \frac{n-1}{kn} W_{OLS} \xrightarrow{D} \chi_k^2/k$$

as $n \to \infty$.

Proof. Notice that by (15.29), the result follows if $F_0 = (n-1)W_{OLS}/(kn)$. Let $\tilde{\boldsymbol{A}} = [\boldsymbol{0}\ \boldsymbol{A}]$ so that Ho:$\tilde{\boldsymbol{A}}\boldsymbol{\eta} = \boldsymbol{0}$ is equivalent to Ho:$\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$. Following Seber and Lee (2003, p. 106),

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1} \end{pmatrix} \tag{15.32}$$

where the $(p-1) \times (p-1)$ matrix

$$\boldsymbol{D}^{-1} = [(n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}]^{-1} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}/(n-1). \tag{15.33}$$

Using $\tilde{\boldsymbol{A}}$ and (15.32) in (15.30) shows that $F_0 =$

$$(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS})^T \left[[\boldsymbol{0}\ \boldsymbol{A}] \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{0}^T \\ \boldsymbol{A}^T \end{pmatrix} \right]^{-1} \boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS}/(k\hat{\tau}^2),$$

and the result follows from (15.33) after algebra. QED

Ellipsoidal trimming can be used to create outlier resistant 1D methods that can give useful results when the assumption of linearly related predictors (15.6) is violated. To perform ellipsoidal trimming, a robust estimator of multivariate location and dispersion $(T, \boldsymbol{C})$ is computed and used to create the Mahalanobis distances $D_i(T, \boldsymbol{C})$. The ith case $(Y_i, \boldsymbol{x}_i)$ is trimmed if $D_i > D_{(j)}$. For example, if $j \approx 0.9n$, then about M% = 10% of the cases are trimmed, and OLS can be computed from the cases that remain.

For theory and outlier resistance, the choice of $(T, \boldsymbol{C})$ and M are important. The MBA estimator $(T_{MBA}, \boldsymbol{C}_{MBA})$ will be used for $(T, \boldsymbol{C})$ (although the FCH estimator may be a better choice because of its combination of speed, robustness and theory). The classical Mahalanobis distance uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})$. Denote the robust distances by RD_i and the classical distances by MD_i. Then the DD plot of the MD_i versus the RD_i can be used to choose M. The plotted points in the DD plot will follow the identity line with zero intercept and unit slope if the predictor distribution is multivariate normal (MVN), and will follow a line with zero intercept but non–unit slope if the distribution is elliptically contoured with nonsingular covariance matrix but not MVN. Delete M% of the cases with the largest MBA distances so that the remaining cases follow the identity line (or some line through the origin) closely. Let $(Y_{Mi}, \boldsymbol{x}_{Mi})$ denote the data that was not trimmed where $i = 1, ..., n_M$. Then apply OLS on these n_M cases.

As long as M is chosen only using the predictors, OLS theory will apply if the data $(Y_M, \boldsymbol{x}_M)$ satisfies the regularity conditions. For example, if the MLR model is valid and the errors are iid $N(0, \sigma^2)$, then the OLS estimator

$$\hat{\boldsymbol{\eta}}_M = (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^T \boldsymbol{Y}_M \sim N_p(\boldsymbol{\eta}, \sigma^2 (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1}).$$

More generally, let $\phi_M = \lim_{n\to\infty} n/n_M$, let c_M be a constant and let $\hat{\boldsymbol{\beta}}_M$ denote the OLS estimator applied to $(Y_{Mi}, \boldsymbol{x}_{Mi})$ with

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_M - c_M \boldsymbol{\beta}) = \frac{\sqrt{n}}{\sqrt{n_M}} \sqrt{n_M}(\hat{\boldsymbol{\beta}}_M - c_M \boldsymbol{\beta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \phi_M \boldsymbol{C}_M). \quad (15.34)$$

If $Ho : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$ is true and $\hat{\boldsymbol{C}}_M$ is a consistent estimator of $\boldsymbol{C}_M$, then

$$W_M = n_M \hat{\boldsymbol{\beta}}_M^T \boldsymbol{A}^T [\boldsymbol{A} \hat{\boldsymbol{C}}_M \boldsymbol{A}^T]^{-1} \boldsymbol{A} \hat{\boldsymbol{\beta}}_M / \hat{\tau}_M^2 \xrightarrow{D} \chi_k^2.$$

Notice that $M = 0$ corresponds to the full data set and $n_0 = n$.

A tradeoff is that low amounts of trimming may not work while large amounts of trimming may be inefficient if low amounts of trimming work since $n/n_M \geq 1$ and the diagonal elements of $\boldsymbol{C}_M$ typically become larger with M.

Trimmed views can also be used to select $M \equiv M_{TV}$. If the MLR model holds and OLS is used, then the resulting trimmed views estimator $\hat{\boldsymbol{\beta}}_{M,TV}$ is $\sqrt{n}$ consistent, but need not be asymptotically normal.

Adaptive trimming can be used to obtain an asymptotically normal estimator that may avoid large efficiency losses. First, choose an initial amount of trimming M_I by using, eg, $M_I = 50$ or the DD plot. Let $\hat{\boldsymbol{\beta}}$ denote the first direction of the DR method. Next compute $|corr(\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}, \hat{\boldsymbol{\beta}}_{M_I}^T \boldsymbol{x})|$ for $M = 0, 10, ..., 90$ and find the smallest value $M_A \leq M_I$ such that the absolute correlation is greater than 0.95. If no such value exists, then use $M_A = M_I$. The resulting adaptive trimming estimator is asymptotically equivalent to the estimator that uses 0% trimming if $\hat{\boldsymbol{\beta}}_0$ is a consistent estimator of $c_0 \boldsymbol{\beta}$ and if $\hat{\boldsymbol{\beta}}_{M_I}$ is a consistent estimator of $c_{M_I} \boldsymbol{\beta}$.

The following example and Tables 15.1 and 15.2 show that ellipsoidal trimming can be useful for 1D regression when $\boldsymbol{x}$ is not EC. There is a myth that transforming predictors is free, but using a log transformation for the example below will destroy the 1D structure.

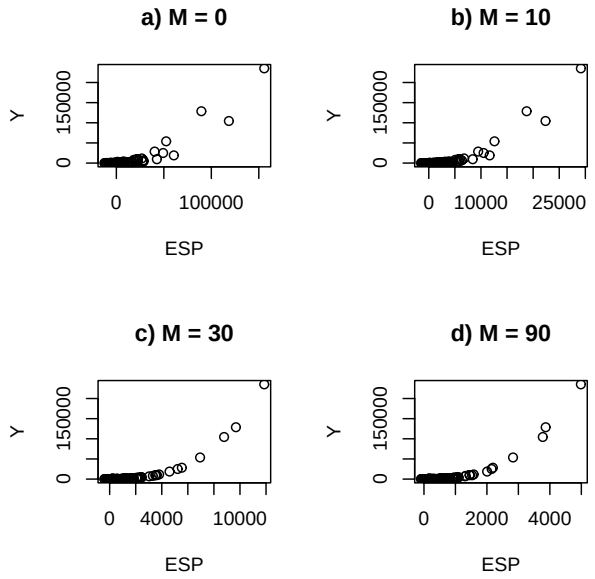

Figure 15.11: Trimmed Views

Example 15.8. An artificial data set was generated with $Y = (\alpha + \boldsymbol{\beta}^T \boldsymbol{x})^3 + e$ where $n = 100$, $\alpha = 0, \boldsymbol{\beta} = (1, 2, 3)^T, e \sim N(0, 1)$ and $x_i \sim lognormal(0, 1)$ for $i = 1, 2, 3$ where the x_i are iid. Figure 15.11 shows the trimmed views for M = 0, 10, 30 and 90. Table 15.1 shows the values of $\hat{\boldsymbol{\beta}}_M$. Notice that the 30% and 90% trimmed views capture the cubic function much better then the OLS = 0% trimmed view. Notice that $\hat{\boldsymbol{\beta}}_{30} \approx 205\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}_{90} \approx 86\boldsymbol{\beta}$.

Table 15.1: Trimming with Non-EC Predictors, $\boldsymbol{\beta} = c(1, 2, 3)^T$

M	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
0	346.034	3394.260	9000.226
10	292.575	731.751	1616.625
30	191.516	421.577	616.201
90	86.024	160.877	258.987

Table 15.2: Trimming with Outlier Percentage = γ, $\boldsymbol{\beta} = c(1,0,0,0)^T$

γ	M	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
0	0	5.974	.0083	$-.0221$	.0008
0	50	4.098	.0166	.0017	$-.0016$
49	0	2.269	$-.7509$	$-.7390$	$-.7625$
49	50	5.647	.0305	.0011	.0053

In a small simulation, the clean data $Y = (\alpha + \boldsymbol{\beta}^T\boldsymbol{x})^3 + e$ where $n = 1000$, $\alpha = 1$, $\boldsymbol{\beta} = (1,0,0,0)^T$, $e \sim N(0,1)$ and $\boldsymbol{x} \sim N_4(\mathbf{0}, \boldsymbol{I}_4)$. The outlier percentage γ was either 0% or 49%. The 2 clusters of outliers were about the same size with $Y \sim N(0,1)$ and $\boldsymbol{x} \sim N_4(\pm 10(1,1,1,1)^T, \boldsymbol{I}_4)$. Table 15.2 records the averages of $\hat{\beta}_i$ over 100 runs where OLS used $M = 0$ or $M = 50\%$ trimming. When outliers were present, the average of $\hat{\boldsymbol{\beta}}_{50} \approx c(1,0,0,0)^T$.

The following simulation study is extracted from Chang (2006) who used eight types of predictor distributions: d1) $\boldsymbol{x} \sim N_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1})$, d2) $\boldsymbol{x} \sim 0.6N_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1}) + 0.4N_{p-1}(\mathbf{0}, 25\boldsymbol{I}_{p-1})$, d3) $\boldsymbol{x} \sim 0.4N_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1}) + 0.6N_{p-1}(\mathbf{0}, 25\boldsymbol{I}_{p-1})$, d4) $\boldsymbol{x} \sim 0.9N_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1}) + 0.1N_{p-1}(\mathbf{0}, 25\boldsymbol{I}_{p-1})$, d5) $\boldsymbol{x} \sim LN(\mathbf{0}, \boldsymbol{I})$ where the marginals are iid lognormal(0,1), d6) $\boldsymbol{x} \sim MVT_{p-1}(3)$, d7) $\boldsymbol{x} \sim MVT_{p-1}(5)$ and d8) $\boldsymbol{x} \sim MVT_{p-1}(19)$. Here $\boldsymbol{x}$ has a multivariate t distribution $\boldsymbol{x}_i \sim MVT_{p-1}(\nu)$ if $\boldsymbol{x}_i = \boldsymbol{z}_i/\sqrt{W_i/\nu}$ where $\boldsymbol{z}_i \sim N_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1})$ is independent of the chi–square random variable $W_i \sim \chi^2_\nu$. Of the eight distributions, only d5) is not elliptically contoured. The MVT distribution gets closer to the MVN distribution d1) as $\nu \to \infty$. The MVT distribution has first moments for $\nu \geq 3$ and second moments for $\nu \geq 5$. See Johnson and Kotz (1972, pp. 134-135). All simulations used 1000 runs.

The simulations for single index models used $\alpha = 1$. Let the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. Then the seven models considered were m1) $Y = SP + e$, m2) $Y = (SP)^2 + e$, m3) $Y = \exp(SP) + e$, m4) $Y = (SP)^3 + e$, m5) $Y = \sin(SP)/SP + 0.01e$, m6) $Y = SP + \sin(SP) + 0.1e$ and m7) $Y = \sqrt{|SP|} + 0.1e$ where $e \sim N(0,1)$. Models m2), m3) and m4) can result in large $|Y|$ values which can cause numerical difficulties for OLS if $\boldsymbol{x}$ is heavy tailed.

For single index models with EC $\boldsymbol{x}$, OLS can fail if m is symmetric about the median θ of the distribution of $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. If m is symmetric about a, then OLS may become effective as $|\theta - a|$ gets large. This fact is

often overlooked in the literature and is demonstrated by models m7), m5) and m2) where $Y = (SP)^2 + e$ with $\theta = \alpha = 1$. OLS has trouble with $Y = (SP - a)^2 + e$ as a gets close to $\theta = 1$ for the EC distributions. The type of symmetry where OLS fails is easily simulated, but may not occur often in practice.

First, coefficient estimation was examined with $\boldsymbol{\beta} = (1, 1, 1, 1)^T$, and for OLS the sample standard deviation (SD) of each entry $\hat{\beta}_{Mi,j}$ of $\hat{\boldsymbol{\beta}}_{M,j}$ was computed for $i = 1, 2, 3, 4$ with $j = 1, ..., 1000$. For each of the 1000 runs, the formula

$$SE_{cl}(\hat{\boldsymbol{\beta}}_{Mi}) = \sqrt{n_M^{-1}(\hat{\boldsymbol{C}}_M)_{ii}}$$

was computed where

$$\hat{\boldsymbol{C}}_M = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}_M}^{-1} \left[\frac{1}{n_M} \sum_{i=1}^{n_M} [(Y_{Mi} - \hat{\alpha}_M - \hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_{Mi})^2 (\boldsymbol{x}_{Mi} - \overline{\boldsymbol{x}}_M)(\boldsymbol{x}_{Mi} - \overline{\boldsymbol{x}}_M)^T] \right] \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}_M}^{-1}$$

is the estimate (15.28) applied to $(Y_M, \boldsymbol{x}_M)$. The average of $\hat{\boldsymbol{\beta}}_M$ and of $\sqrt{n} SE_{cl}$ were recorded as well as $\sqrt{n} SD$ of $\hat{\boldsymbol{\beta}}_{Mi,j}$ under the labels $\overline{\beta}_M$, $\sqrt{n}\ \overline{SE}_{cl}$ and $\sqrt{n} SD$. Under regularity,

$$\sqrt{n}\ \overline{SE}_{cl} \approx \sqrt{n} SD \approx \sqrt{\frac{1}{1 - \frac{M}{100}} \operatorname{diag}(\boldsymbol{C}_{\mathrm{M}})}$$

where $\boldsymbol{C}_M$ is (15.26) applied to $(Y_M, \boldsymbol{x}_M)$.

For MVN $\boldsymbol{x}$, MLR and 0% trimming, all three recorded quantities were near (1,1,1,1) for $n = 60, 500$, and 1000. For 90% trimming and $n = 1000$, the results were $\overline{\beta}_{90} = (1.00, 1.00, 1.01, 0.99)$, $\sqrt{n}\ \overline{SE}_{cl} = (7.56, 7.61, 7.60, 7.54)$ and $\sqrt{n} SD = (7.81, 8.02, 7.76, 7.59)$, suggesting that $\hat{\boldsymbol{\beta}}_{90}$ is asymptotically normal but inefficient.

For other distributions, results for 0 and 10% trimming were recorded as well as a "good" trimming value M_B. Results are "good" if all of the entries of both $\overline{\beta}_{M_B}$ and $\sqrt{n}\ \overline{SE}_{cl}$ were approximately equal, and if the theoretical $\sqrt{n}\ \overline{SE}_{cl}$ was close to the simulated $\sqrt{n} SD$. The results were good for MVN $\boldsymbol{x}$ and all seven models, and the results were similar for $n = 500$ and $n = 1000$. The results were good for models m1 and m5 for all eight distributions. Model m6 was good for 0% trimming except for distribution d5 and model m7 was good for 0% trimming except for distributions d5, d6 and d7. Trimming

Table 15.3: OLS Coefficient Estimation with Trimming

m	$\boldsymbol{x}$	M	$\overline{\boldsymbol{\beta}}_M$	$\sqrt{n}\,\overline{SE}_{cl}$	$\sqrt{n}\,SD$
m2	d1	0	2.00,2.01,2.00,2.00	7.81,7.79,7.76,7.80	7.87,8.00,8.02,7.88
m5	d4	0	$-.03, -.03, -.03, -.03$	.30,.30,.30,.30	.31,.32,.33,.31
m6	d5	0	1.04,1.04,1.04,1.04	.36,.36,.37,.37	.41,.42,.42,.40
m7	d6	10	.11,.11,.11,.11	.58,.57,.57,.57	.60,.58,.62,.61

usually helped for models m2, m3 and m4 for distributions d5 – d8. For $n = 500$, Table 15.3 shows that $\hat{\boldsymbol{\beta}}_M$ estimates $c_M\boldsymbol{\beta}$ and the average of the Chen and Li (1998) SE is often close to the simulated SD.

Next testing was considered. Let F_M denote the OLS statistic (15.30) applied to the n_M cases $(Y_M, \boldsymbol{x}_M)$ that remained after trimming. Ho was rejected for OLS if $F_M > F_{k,n_M-p}(0.95)$. Let $\hat{p}$ be the proportion of runs where H_0 was rejected. Since 1000 runs were used, the count $1000\hat{p} \sim$ binomial(1000, $1 - \delta_n$) where $1 - \delta_n$ converges to the true large sample level $1 - \delta$. The standard error for the proportion is $\sqrt{\hat{p}(1-\hat{p})/1000} \approx 0.0069$ for $p = 0.05$. An observed coverage $\hat{p} \in (0.03, 0.07)$ suggests that there is no reason to doubt that the true level is 0.05.

Suppose a 1D model holds but $Y \perp\!\!\!\perp \boldsymbol{x}$. Then the Y_i are iid and the model reduces to $Y = E(Y) + e = c_\alpha + e$ where $e = Y - E(Y)$. As a special case, if $Y = m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) + e$ and if $Y \perp\!\!\!\perp \boldsymbol{x}$, then $Y = m(\alpha) + e$. For the corresponding test $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$, the OLS F statistic (15.30) is invariant with respect to a constant. This test is interesting since if Ho holds, then the results do not depend on the 1D model (15.1), but only on the distribution of $\boldsymbol{x}$ and the distribution of e. Since $\boldsymbol{\beta}_{OLS} = c\boldsymbol{\beta}$, power can be good if $c \neq 0$. The OLS test is equivalent to the ANOVA F test from MLR of Y on $\boldsymbol{x}$. Under H_0, the test should perform well provided that the design matrix is nonsingular and the error distribution and sample size are such that the central limit theorem holds. For the simulated data with $\boldsymbol{\beta} = \mathbf{0}$, the model is linear and normal, and the exact OLS level is 0.05 for $n > p$. Table 15.4 illustrates this claim for $n = 100$ and $n = 500$.

Next the test $Ho : \beta_2 = 0$ was considered. The OLS test is equivalent

Table 15.4: Rejection Proportions for H_0: $\boldsymbol{\beta} = \mathbf{0}$

$\boldsymbol{x}$	n	F	n	F
d1	100	0.041	500	0.050
d2	100	0.050	500	0.045
d3	100	0.047	500	0.050
d4	100	0.045	500	0.048
d5	100	0.055	500	0.061
d6	100	0.042	500	0.036
d7	100	0.054	500	0.047
d8	100	0.044	500	0.060

Table 15.5: Rejection Proportions for Ho: $\beta_2 = 0$

m	$\boldsymbol{x}$	70	60	50	40	30	20	10	0	ADAP
1	1	.061	.056	.062	.051	.046	.050	.044	.043	.043
5	1	.019	.023	.019	.019	.020	.022	.027	.037	.029
2	2	.023	.024	.026	.070	.183	.182	.142	.166	.040
4	3	.027	.058	.096	.081	.071	.057	.062	.123	.120
6	4	.026	.024	.030	.032	.028	.044	.051	.088	.088
7	5	.058	.058	.053	.054	.046	.044	.051	.037	.037
3	6	.021	.024	.019	.025	.025	.034	.080	.374	.036
6	7	.027	.032	.023	.041	.047	.053	.052	.055	.055

to the t test from MLR of Y on $\boldsymbol{x}$. The true model used $\alpha = 1$ and $\boldsymbol{\beta} = (1, 0, 1, 1)^T$. To simulate adaptive trimming, $|corr(\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}, \boldsymbol{\beta}^T \boldsymbol{x})|$ was computed for $M = 0, 10, ..., 90$ and the initial trimming proportion M_I maximized this correlation. This process should be similar to choosing the best trimmed view by examining 10 plots. The rejection proportions were recorded for $M = 0, ..., 90$ and for adaptive trimming. The seven models, eight distributions and sample sizes $n = 60, 150$, and 500 were used.

The test that used adaptive trimming had proportions ≤ 0.072 except for model m4 with distributions d2, d3, d4, d6, d7 and d8; m2 with d4, d6 and d7 for n = 500 and d6 with n = 150; m6 with d4 and n = 60, 150; m5 with d7 and n = 500 and m7 with d7 and n = 500. With the exception of m4, when the adaptive $\hat{p} > 0.072$, then 0% trimming had a rejection proportion near 0.1. Occasionally adaptive trimming was conservative with $\hat{p} < 0.03$. The 0% trimming worked well for m1 and m6 for all eight distributions and for d1 and d5 for all seven models. Models m2 and m3 usually benefited from adaptive trimming. For distribution d1, the adaptive and 0% trimming methods had identical $\hat{p}$ for $n = 500$ except for m3 where the values were 0.038 and 0.042. Table 15.5 used $n = 150$ and supports the claim that the adaptive trimming estimator can be asymptotically equivalent to OLS (0% trimming) and that trimming can greatly improve the type I error.

15.6 Complements

For 1D regression models, suppose that $|corr(\hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x}, \hat{\boldsymbol{\beta}}^T \boldsymbol{x})| \geq 0.95$ where $\hat{\boldsymbol{\beta}}$ is a good estimator of $d\boldsymbol{\beta}$ for $d \neq 0$, or that the 1D regression can be visualized with the OLS response plot. For example, the plotted points cluster tightly about the mean function m. Then OLS should be a useful 1D estimator and output originally meant for MLR is also often useful for 1D regression (1DR) data. In particular, i) $\hat{\boldsymbol{\beta}}_{OLS}$ estimates $\boldsymbol{\beta}$ for MLR and $c\boldsymbol{\beta}$ for 1DR. ii) The F test statistics tend to have a χ_k^2/k limiting distribution for MLR, and the $F_{k,n-p}$ cutoffs tend to be useful for exploratory purposes for 1DR. iii) Variable selection with the C_p statistic is effective. iv) The MSE estimates σ^2 for MLR and τ^2 for 1DR. v) The OLS response plot is a very effective tool for visualizing the regression and outlier detection. The estimated mean function for MLR is the unit slope line through the origin, but tends to be nonlinear for 1DR. vi) Resistant $\sqrt{n}$ consistent estimators based on OLS and ellipsoidal trimming exist for both MLR and 1DR. vii) Cook's distance is a

useful influence diagnostic.

To see vii) for 1DR, notice that the ith Cook's distance

$$\mathrm{CD}_i = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^T(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{p\widehat{\sigma}^2} = \frac{\|ESP(i) - ESP\|^2}{(p+1)MSE}$$

where $ESP(i) = \boldsymbol{X}^T\hat{\boldsymbol{\eta}}_{(i)}$ and $\hat{\boldsymbol{\eta}}_{(i)}$ is computed without the ith case, and the estimated sufficient predictor $ESP = \boldsymbol{X}^T\hat{\boldsymbol{\eta}}$ estimates $\alpha_{OLS} + c\,\boldsymbol{\beta}^T\boldsymbol{x}_j$ for some constant c and $j = 1, ..., n$. Thus Cook's distances give useful information on cases that influence the OLS ESP.

Fast exploratory analysis with OLS can be used to complement alternative 1D methods, especially if tests and variable selection for the 1D method are slow or unavailable from the software.

An excellent introduction to 1D regression and regression graphics is Cook and Weisberg (1999a, ch. 18, 19, and 20) and Cook and Weisberg (1999b). More advanced treatments are Cook (1998a) and Li (2000). Important papers include Brillinger (1977, 1983), Li and Duan (1989) and Stoker (1986). Xia, Tong, Li and Zhu (2002) provides a method for single index models (and multi–index models) that does not need the linearity condition.

The response plot is crucial for checking the goodness of fit of the model. Also see Stute and Zhu (2005) and Xia, Li, Tong and Zhang (2004). One goal for future research is to develop better methods for visualizing 1D regression. Trimmed views seem to become less effective as the number of predictors $k = p - 1$ increases. Consider the sufficient predictor $\mathrm{SP} = x_1 + \cdots + x_k$. With the sin(SP)/SP data, several trimming proportions gave good views with $k = 3$, but only one of the ten trimming proportions gave a good view with $k = 10$. In addition to problems with dimension, it is not clear which regression estimator and which multivariate location and dispersion (MLD) estimator should be used. We suggest using the FCH = `covfch` MLD estimator or classical MLD estimator with OLS as the regression estimator. See Olive (2009a, $\oint$ 10.7).

There are many ways to estimate 1D models, including maximum likelihood for parametric models. The literature for estimating $c\boldsymbol{\beta}$ when model (15.1) holds is growing, and OLS frequently performs well if there are no strong nonlinearities present in the predictors. In addition to OLS, specialized methods for 1D models with an unknown inverse link function (eg models

(15.2) and (15.3)) have been developed, and often the focus is on developing asymptotically efficient methods. See the references in Cavanagh and Sherman (1998), Delecroix, Härdle and Hristache (2003), Härdle, Hall and Ichimura (1993), Horowitz (1998), Hristache, Juditsky, Polzehl, and Spokoiny (2001), Stoker (1986), Weisberg and Welsh (1994), Xia (2006) and Xia, Tong, Li and Zhu (2002).

Some of these methods standardize $\hat{\boldsymbol{\beta}}$ so $\hat{\beta}_1 = 1$. This standardization may cause problems for testing $\boldsymbol{\beta} = \mathbf{0}$ and $\beta_1 = 0$.

Several papers have suggested that outliers and strong nonlinearities need to be removed from the predictors. See Brillinger (1991), Cook (1998a, p. 152), Cook and Nachtsheim (1994) and Li and Duan (1989, p. 1011, 1041, 1042). Trimmed views were introduced by Olive (2002, 2004b). Li, Cook and Nachtsheim (2004) find clusters, fit OLS to each cluster and then pool the OLS estimators into a final estimator. This method uses all n cases while trimmed views gives $M\%$ of the cases weight zero. The trimmed views estimator will often work well when outliers and influential cases are present.

Section 15.4 follows Olive and Hawkins (2005) closely. The literature on numerical methods for variable selection in the OLS multiple linear regression model is enormous, and the literature for other given 1D regression models is also growing. Li, Cook and Nachtsheim (2005) give an alternative method for variable selection that can work without specifying the model. Also see, for example, Claeskins and Hjort (2003), Efron, Hastie, Johnstone and Tibshirani (2004), Fan and Li (2001, 2002), Hastie (1987), Kong and Xia (2007), Lawless and Singhai (1978), Leeb and Pötscher (2006), Naik and Tsai (2001), Nordberg (1982) and Tibshirani (1996). For generalized linear models, forward selection and backward elimination based on the AIC criterion are often used. See Chapters 11, 12 and 13, Agresti (2002, p. 211-217), Cook and Weisberg (1999a, p. 485, 536-538). Again, if the variable selection techniques in these papers are successful, then the estimated sufficient predictors from the full and candidate model should be highly correlated, and the EE, VV and response plots will be useful. Survival regression models also use AIC. See Chapter 16.

The variable selection model with $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ and $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T\boldsymbol{x}_S$ is not the only variable selection model. Burnham and Anderson (2004) note that for many data sets, the variables can be ordered in decreasing

importance from x_1 to x_{p-1}. The "tapering effects" are such that if $n >> p$, then all of the predictors should be used, but for moderate n it is better to delete some of the least important predictors.

Section 15.5 followed Chang and Olive (2010) closely. More examples and simulations are in Chang (2006). Severini (1998) discusses when OLS output is relevant for the Gaussian additive error single index model. Li and Duan (1989) and Li (1997) suggest that OLS F tests are asymptotically valid if $\boldsymbol{x}$ is multivariate normal and if $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \mathbf{0}$. Freedman (1981), Brillinger (1983) and Chen and Li (1998) also discuss $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{OLS})$. Formal testing procedures for the single index model are given by Simonoff and Tsai (2002) and Gao and Liang (1997). Chang and Olive (2007) shows how to apply ellipsoidal trimming to general 1D methods, including OLS.

The mussel data set is included as the file *mussel.lsp* in the *Arc* software and can be obtained from the web site (http://www.stat.umn.edu/arc/). The Boston housing data can be obtained from the text website or from the STATLIB website (http://lib.stat.cmu.edu/datasets/boston).

15.7 Problems

15.1. Refer to Definition 15.3 for the Cox and Snell (1968) definition for residuals, but replace $\boldsymbol{\eta}$ by $\boldsymbol{\beta}$.

a) Find $\hat{e}_i$ if $Y_i = \mu + e_i$ and $T(Y)$ is used to estimate μ.

b) Find $\hat{e}_i$ if $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$.

c) Find $\hat{e}_i$ if $Y_i = \beta_1 \exp[\beta_2(x_i - \bar{x})]e_i$ where the e_i are iid exponential(1) random variables and $\bar{x}$ is the sample mean of the $x_i's$.

d) Find $\hat{e}_i$ if $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i/\sqrt{w_i}$.

15.2*. (Aldrin, Bϕlviken, and Schweder 1993). Suppose

$$Y = m(\boldsymbol{\beta}^T\boldsymbol{x}) + e \tag{15.35}$$

where m is a possibly unknown function and the zero mean errors e are independent of the predictors. Let $z = \boldsymbol{\beta}^T\boldsymbol{x}$ and let $\boldsymbol{w} = \boldsymbol{x} - E(\boldsymbol{x})$. Let $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \mathrm{Cov}(\boldsymbol{x}, Y)$, and let $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \mathrm{Cov}(\boldsymbol{x}) = \mathrm{Cov}(\boldsymbol{w})$. Let $\boldsymbol{r} = \boldsymbol{w} - (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w}$.

a) Recall that $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{Y}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$ and show that $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = E(\boldsymbol{w}Y)$.

b) Show that $E(\boldsymbol{w}Y) = \boldsymbol{\Sigma}_{\boldsymbol{x},Y} = E[(\boldsymbol{r} + (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})\ m(z)] =$

$$E[m(z)\boldsymbol{r}] + E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}.$$

c) Using $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$, show that $\boldsymbol{\beta}_{OLS} = c(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{u}(\boldsymbol{x})$ where the constant

$$c(\boldsymbol{x}) = E[\boldsymbol{\beta}^T(\boldsymbol{x} - E(\boldsymbol{x}))m(\boldsymbol{\beta}^T\boldsymbol{x})]$$

and the bias vector $\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}E[m(\boldsymbol{\beta}^T\boldsymbol{x})\boldsymbol{r}]$.

d) Show that $E(\boldsymbol{w}z) = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}$. (Hint: Use $E(\boldsymbol{w}z) = E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{x}^T\boldsymbol{\beta}] = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x}^T - E(\boldsymbol{x}^T) + E(\boldsymbol{x}^T))\boldsymbol{\beta}]$.)

e) Assume $m(z) = z$. Using d), show that $c(\boldsymbol{x}) = 1$ if $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$.

f) Assume that $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$. Show that $E(z\boldsymbol{r}) = E(\boldsymbol{r}z) = \boldsymbol{0}$. (Hint: Find $E(\boldsymbol{r}z)$ and use d).)

g) Suppose that $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$ and that the distribution of $\boldsymbol{x}$ is multivariate normal. Then the joint distribution of z and $\boldsymbol{r}$ is multivariate normal. Using the fact that $E(z\boldsymbol{r}) = \boldsymbol{0}$, show $\text{Cov}(\boldsymbol{r}, z) = 0$ so that z and $\boldsymbol{r}$ are independent. Then show that $\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{0}$.

(Note: the assumption $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$ can be made without loss of generality since if $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = d^2 > 0$ (assuming $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is positive definite), then $y = m(d(\boldsymbol{\beta}/d)^T\boldsymbol{x}) + e \equiv m_d(\boldsymbol{\eta}^T\boldsymbol{x}) + e$ where $m_d(u) = m(du)$, $\boldsymbol{\eta} = \boldsymbol{\beta}/d$ and $\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\eta} = 1$.)

15.3. Suppose that you have a statistical model where both fitted values and residuals can be obtained. For example this is true for time series and for nonparametric regression models such as $Y = f(x_1, ..., x_p) + e$ where $\hat{y} = \hat{f}(x_1, ..., x_p)$ and the residual $\hat{e} = Y - \hat{f}(x_1, ..., x_p)$. Suggest graphs for variable selection for such models.

Output for Problem 15.4.
BEST SUBSET REGRESSION MODELS FOR CRIM
(A)LogX2 (B)X3 (C)X4 (D)X5 (E)LogX7 (F)X8 (G)LogX9 (H)LogX12
3 "BEST" MODELS FROM EACH SUBSET SIZE LISTED.

k	CP	ADJUSTED R SQUARE	R SQUARE	RESID SS	MODEL VARIABLES
1	379.8	0.0000	0.0000	37363.2	INTERCEPT ONLY
2	36.0	0.3900	0.3913	22744.6	F
2	113.2	0.3025	0.3039	26007.8	G
2	191.3	0.2140	0.2155	29310.8	E
3	21.3	0.4078	0.4101	22039.9	E F
3	25.0	0.4036	0.4059	22196.7	F H
3	30.8	0.3970	0.3994	22442.0	D F
4	17.5	0.4132	0.4167	21794.9	C E F
4	18.1	0.4125	0.4160	21821.0	E F H
4	18.8	0.4117	0.4152	21850.4	A E F
5	10.2	0.4226	0.4272	21402.3	A E F H
5	10.8	0.4219	0.4265	21427.7	C E F H
5	12.0	0.4206	0.4252	21476.6	A D E F
6	5.7	0.4289	0.4346	21125.8	A C E F H
6	9.3	0.4248	0.4305	21279.1	A C D E F
6	10.3	0.4237	0.4294	21319.9	A B E F H
7	6.3	0.4294	0.4362	21065.0	A B C E F H
7	6.3	0.4294	0.4362	21066.3	A C D E F H
7	7.7	0.4278	0.4346	21124.3	A C E F G H
8	7.0	0.4297	0.4376	21011.8	A B C D E F H
8	8.3	0.4283	0.4362	21064.9	A B C E F G H
8	8.3	0.4283	0.4362	21065.8	A C D E F G H
9	9.0	0.4286	0.4376	21011.8	A B C D E F G H

15.4. The output above is for the Boston housing data from software that does all subsets variable selection. The full model is a 1D transformation model with response variable Y = CRIM and uses a constant and variables A, B, C, D, E, F, G and H. (Using log(CRIM) as the response would give an MLR model.) From this output, what is the best submodel? Explain briefly.

15.5*. a) Show that $C_p(I) \leq 2k$ if and only if $F_I \leq p/(p-k)$.

b) Using (15.19), find E(C_p) and Var(C_p) assuming that an MLR model is appropriate and that Ho (the reduced model I can be used) is true.

c) Using (15.19), $C_p(I_{full}) = p$ and the notation in Section 15.4, show that

$$C_p(I_i) = C_p(I_{full}) + (t_i^2 - 2).$$

R/Splus Problems

Warning: Use the command *source("A:/regpack.txt")* **to download the programs. See Preface or Section 17.2.** Typing the name of the `regpack` function, eg *trviews*, will display the code for the function. Use the `args` command, eg *args(trviews)*, to display the needed arguments for the function.

15.6. Use the following *R/Splus* commands to make 100 $N_3(\mathbf{0}, I_3)$ cases and 100 trivariate non-EC cases.

```
n3x <- matrix(rnorm(300),nrow=100,ncol=3)
ln3x <- exp(n3x)
```

In *R*, type the command *library(MASS).*

a) Using the commands *pairs(n3x)* and *pairs(ln3x)* and include both scatterplot matrices in *Word.* (Click on the plot and hit *Ctrl* and *c* at the same time. Then go to *file* in the *Word* menu and select *paste.*) Are strong nonlinearities present among the MVN predictors? How about the non-EC predictors? (Hint: a box or ball shaped plot is linear.)

b) Make a single index model and the sufficient summary plot with the following commands

```
ncy <- (n3x%*%1:3)^3 + 0.1*rnorm(100)
plot(n3x%*%(1:3),ncy)
```

and include the plot in *Word.*

c) The command *trviews(n3x, ncy)* will produce ten plots. To advance the plots, click on the *rightmost mouse button* (and in *R* select *stop*) to advance to the next plot. The last plot is the OLS view. Include this plot in *Word.*

d) After all 10 plots have been looked at the output will show 10 estimated predictors. The last estimate is the OLS (least squares) view and might look like

```
Intercept        X1        X2        X3
 4.417988 22.468779 61.242178 75.284664
```

If the OLS view is a good estimated sufficient summary plot, then the plot created from the command (leave out the intercept)

```
plot(n3x%*%c(22.469,61.242,75.285),n3x%*%1:3)
```

should cluster tightly about some line. Your linear combination will be different than the one used above. Using your OLS view, include the plot using the command above (but with your linear combination) in *Word.* Was this plot linear? Did some of the other trimmed views seem to be better that the OLS view, that is, did one of the trimmed views seem to have a smooth mean function with a smaller variance function than the OLS view?

e) Now type the *R/Splus* command

```
lncy <- (ln3x%*%1:3)^3 + 0.1*rnorm(100).
```

Use the command *trviews(ln3x,lncy)* to find the best view with a smooth mean function and the smallest variance function. This view should not be the OLS view. Include your best view in *Word.*

f) Get the linear combination from your view, say $(94.848, 216.719, 328.444)^T$, and obtain a plot with the command

```
plot(ln3x%*%c(94.848,216.719,328.444),ln3x%*%1:3).
```

Include the plot in *Word.* If the plot is linear with high correlation, then your response plot in e) should be good.

15.7. (At the beginning of your *R/Splus* session, use the *source("A:/regpack.txt")* command (and *library(MASS)* in *R.*))

a) Perform the commands

```
> nx <- matrix(rnorm(300),nrow=100,ncol=3)
> lnx <- exp(nx)
> SP <- lnx%*%1:3
> lnsincy <- sin(SP)/SP + 0.01*rnorm(100)
```

For parts b), c) and d) below, to make the best trimmed view with `trviews`, `ctrviews` or `lmsviews`, you may need to use the function twice. The first view trims 90% of the data, the next view trims 80%, etc. The last view trims 0% and is the OLS view (or `lmsreg` view). Remember to advance the view with the rightmost mouse button (and in *R*, highlight "stop"). Then click on the plot and next simultaneously hit *Ctrl* and *c*. This makes a copy of the plot. Then in *Word*, use the menu commands "Copy>paste."

b) Find the best trimmed view with `OLS` and `covfch` with the following commands and include the view in *Word.*

```
> trviews(lnx,lnsincy)
```

(With `trviews`, suppose that 40% trimming gave the best view. Then instead of using the procedure above b), you can use the command

```
> essp(lnx,lnsincy,M=40)
```

to make the best trimmed view. Then click on the plot and next simultaneously hit *Ctrl* and *c*. This makes a copy of the plot. Then in *Word*, use the menu commands "Copy>paste". Click the rightmost mouse button (and in *R*, highlight "stop") to return the command prompt.)

c) Find the best trimmed view with `OLS` and $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ using the following commands and include the view in *Word.* See the paragraph above b).

```
> ctrviews(lnx,lnsincy)
```

d) Find the best trimmed view with `lmsreg` and `cov.mcd` using the following commands and include the view in *Word.* See the paragraph above b).

```
> lmsviews(lnx,lnsincy)
```

e) Which method or methods gave the best response plot? Explain briefly.