

# Chapter 1

## Introduction

*All models are wrong, but some are useful.*

Box (1979)

This chapter provides a preview of the book but is presented in a rather abstract setting and will be much easier to follow after the reading the rest of the book. The reader may omit this chapter on first reading and refer back to it as necessary.

In *data analysis*, an investigator is presented with a *problem* and *data* from some *population*. The population might be the collection of all possible outcomes from an experiment while the problem might be predicting a future value of the response variable  $Y$  or summarizing the relationship between  $Y$  and the  $p \times 1$  vector of predictor variables  $\mathbf{x}$ . A **statistical model** is used to provide a useful approximation to some of the important underlying characteristics of the population which generated the data. Many of the most used models for 1D regression, defined below, are families of conditional distributions  $Y|\mathbf{x} = \mathbf{x}_o$  indexed by  $\mathbf{x} = \mathbf{x}_o$ . A 1D regression model is a *parametric model* if the conditional distribution is completely specified except for a fixed finite number of parameters, otherwise, the 1D model is a *semiparametric model*.

**Definition 1.1.** *Regression* investigates how the response variable  $Y$  changes with the value of a  $p \times 1$  vector  $\mathbf{x}$  of nontrivial predictors. Often this *conditional distribution*  $Y|\mathbf{x}$  is described by a *1D regression model*, where  $Y$  is conditionally independent of  $\mathbf{x}$  given  $\beta^T \mathbf{x}$ , written

$$Y \perp\!\!\!\perp \mathbf{x} | \beta^T \mathbf{x} \quad \text{or} \quad Y \perp\!\!\!\perp \mathbf{x} | (\alpha + \beta^T \mathbf{x}). \quad (1.1)$$

This class of models is very rich. Generalized linear models (GLMs) are a special case of 1D regression, and an important class of parametric or semiparametric 1D regression models has the form

$$Y_i = g(\alpha + \boldsymbol{\beta}^T \mathbf{x}_i, e_i) \quad (1.2)$$

for  $i = 1, \dots, n$  where  $g$  is a bivariate function,  $\boldsymbol{\beta}$  is a  $p \times 1$  unknown vector of parameters, and  $e_i$  is a random error. Often the errors  $e_1, \dots, e_n$  are **iid** (independent and identically distributed) from a distribution that is known except for a scale parameter. For example, the  $e_i$ 's might be iid from a normal (Gaussian) distribution with *mean* 0 and unknown *standard deviation*  $\sigma$ . For this Gaussian model, estimation of  $\alpha$ ,  $\boldsymbol{\beta}$  and  $\sigma$  is important for inference and for predicting a new value of the response variable  $Y_f$  given a new vector of predictors  $\mathbf{x}_f$ .

**Notation.** Often the index  $i$  will be suppressed. For example, model (1.2) could be written as  $Y = g(\alpha + \boldsymbol{\beta}^T \mathbf{x}, e)$ . More accurately,  $Y|\mathbf{x} = g(\alpha + \boldsymbol{\beta}^T \mathbf{x}, e)$ , but the conditioning on  $\mathbf{x}$  will often be suppressed.

Many of the most used statistical models are 1D regression models. A *single index model* with additive error uses  $g(\alpha + \boldsymbol{\beta}^T \mathbf{x}, e) = m(\alpha + \boldsymbol{\beta}^T \mathbf{x}) + e$ , and thus

$$Y = m(\alpha + \boldsymbol{\beta}^T \mathbf{x}) + e. \quad (1.3)$$

An important special case is *multiple linear regression*

$$Y = \alpha + \boldsymbol{\beta}^T \mathbf{x} + e \quad (1.4)$$

where  $m$  is the identity function. The *response transformation model* uses

$$g(\alpha + \boldsymbol{\beta}^T \mathbf{x}, e) = t^{-1}(\alpha + \boldsymbol{\beta}^T \mathbf{x} + e) \quad (1.5)$$

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

$$t(Y) = \alpha + \boldsymbol{\beta}^T \mathbf{x} + e. \quad (1.6)$$

Several important *survival models* have this form. In a *1D binary regression model*, the  $Y|\mathbf{x}$  are independent Bernoulli $[\rho(\alpha + \boldsymbol{\beta}^T \mathbf{x})]$  random variables where

$$P(Y = 1|\mathbf{x}) \equiv \rho(\alpha + \boldsymbol{\beta}^T \mathbf{x}) = 1 - P(Y = 0|\mathbf{x}) \quad (1.7)$$

In particular, the *logistic regression model* uses

$$\rho(\alpha + \boldsymbol{\beta}^T \mathbf{x}) = \frac{\exp(\alpha + \boldsymbol{\beta}^T \mathbf{x})}{1 + \exp(\alpha + \boldsymbol{\beta}^T \mathbf{x})}.$$

In a *1D Poisson regression model*, the  $Y|\mathbf{x}$  are independent

$$\text{Poisson}[\mu(\alpha + \boldsymbol{\beta}^T \mathbf{x})]$$

random variables. In particular, the *loglinear regression model* uses

$$\mu(\alpha + \boldsymbol{\beta}^T \mathbf{x}) = \exp(\alpha + \boldsymbol{\beta}^T \mathbf{x}). \quad (1.8)$$

In the literature, the response variable is sometimes called the dependent variable while the predictor variables are sometimes called carriers, covariates, explanatory variables, or independent variables. The  $i$ th case  $(Y_i, \mathbf{x}_i^T)$  consists of the values of the response variable  $Y_i$  and the predictor variables  $\mathbf{x}_i^T = (x_{i,1}, \dots, x_{i,p})$  where  $p$  is the number of predictors and  $i = 1, \dots, n$ . The *sample size*  $n$  is the number of cases.

Box (1979) warns that “all models are wrong, but some are useful.” For example the function  $g$  or the error distribution could be misspecified. *Diagnostics* are used to check whether model assumptions such as the form of  $g$  and the proposed error distribution are reasonable. Often diagnostics use *residuals*  $r_i$ . If  $m$  is known, then the single index model (1.3) uses

$$r_i = Y_i - m(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)$$

where  $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$  is an estimate of  $(\alpha, \boldsymbol{\beta})$ .

*Exploratory data analysis* (EDA) can be used to find useful models when the form of the regression or multivariate model is unknown. For example, suppose  $g$  is a monotone function  $t^{-1}$  :

$$Y = t^{-1}(\alpha + \boldsymbol{\beta}^T \mathbf{x} + e). \quad (1.9)$$

Then the transformation

$$Z = t(Y) = \alpha + \boldsymbol{\beta}^T \mathbf{x} + e \quad (1.10)$$

follows a multiple linear regression model.

**Definition 1.2:** If the 1D model (1.1) holds, then  $Y \perp\!\!\!\perp \mathbf{x} | (a + c\boldsymbol{\beta}^T \mathbf{x})$  for any constants  $a$  and  $c \neq 0$ . The quantity  $a + c\boldsymbol{\beta}^T \mathbf{x}$  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  $\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \mathbf{x}$  where  $\tilde{\boldsymbol{\beta}}$  is an estimator of  $c\boldsymbol{\beta}$  for some nonzero constant  $c$ . An *estimated sufficient summary plot* (ESSP) or **response plot** is a plot of any ESP versus  $Y$ .

Assume that the data has been collected and that a 1D regression model (1.1) has been fitted. Suppose that the *sufficient predictor*

$$SP = \alpha + \boldsymbol{\beta}^T \mathbf{x} = \alpha + \boldsymbol{\beta}_R^T \mathbf{x}_R + \boldsymbol{\beta}_O^T \mathbf{x}_O \quad (1.11)$$

where the  $r \times 1$  vector  $\mathbf{x}_R$  consists of the nontrivial predictors in the *reduced model*. Then the investigator will often want to check whether the model is useful and to perform inference. Several things to consider are listed below.

i) Use the response plot (and/or the sufficient summary plot) to explain the 1D regression model to consulting clients, students or researchers.

ii) Goodness of fit: use the response plot to show that the model provides a simple, useful approximation for the relationship between the response variable  $Y$  and the nontrivial predictors  $\mathbf{x}$ . The response plot is used to visualize the conditional distribution of  $Y | (\alpha + \boldsymbol{\beta}^T \mathbf{x})$  when the 1D regression model holds.

iii) Check for lack of fit of the model (eg with a residual plot of the ESP versus the residuals).

iv) Check whether  $Y$  is independent of  $\mathbf{x}$  by testing  $H_o : \boldsymbol{\beta} = \mathbf{0}$ , that is, check whether the nontrivial predictors  $\mathbf{x}$  are needed in the model.

v) Test  $H_o : \boldsymbol{\beta}_O = \mathbf{0}$ , that is, check whether the reduced model can be used instead of the full model.

vi) Use variable selection to find a good submodel.

vii) Estimate the mean function  $E(Y_i | \mathbf{x}_i) = \mu(\mathbf{x}_i) = d_i \tau(\mathbf{x}_i)$  or estimate  $\tau(\mathbf{x}_i)$  where the  $d_i$  are known constants.

viii) Predict  $Y_i$  given  $\mathbf{x}_i$ .

The field of statistics known as *regression graphics* gives useful results for examining the 1D regression model (1.1) even when it is unknown or

misspecified. The following sections show that the sufficient summary plot is useful for explaining the given 1D model while the response plot can often be used to visualize the conditional distribution of  $Y|(\alpha + \boldsymbol{\beta}^T \mathbf{x})$ . 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  $\boldsymbol{\beta}$  is unknown and only a response plot can be made. In Definition 1.2, since  $\tilde{\alpha}$  can be any constant,  $\tilde{\alpha} = 0$  is often used.

## 1.1 Multiple Linear Regression

Suppose that the response variable  $Y$  is quantitative and that at least one predictor variable  $x_i$  is quantitative. Then the multiple linear regression (MLR) model is often a very useful model. For the MLR model,

$$Y_i = \alpha + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \alpha + \mathbf{x}_i^T \boldsymbol{\beta} + e_i = \alpha + \boldsymbol{\beta}^T \mathbf{x}_i + e_i \quad (1.12)$$

for  $i = 1, \dots, n$ . Here  $Y_i$  is the response variable,  $\mathbf{x}_i$  is a  $p \times 1$  vector of nontrivial predictors,  $\alpha$  is an unknown constant,  $\boldsymbol{\beta}$  is a  $p \times 1$  vector of unknown coefficients, and  $e_i$  is a random variable called the error.

The Gaussian or normal MLR model makes the additional assumption that the errors  $e_i$  are iid  $N(0, \sigma^2)$  random variables. This model can also be written as  $Y = \alpha + \boldsymbol{\beta}^T \mathbf{x} + e$  where  $e \sim N(0, \sigma^2)$ , or  $Y|\mathbf{x} \sim N(\alpha + \boldsymbol{\beta}^T \mathbf{x}, \sigma^2)$  or  $Y|\mathbf{x} \sim N(SP, \sigma^2)$ . The normal MLR model is a parametric model since, given  $\mathbf{x}$ , the family of conditional distributions is completely specified by the parameters  $\alpha$ ,  $\boldsymbol{\beta}$  and  $\sigma^2$ . Since  $Y|SP \sim N(SP, \sigma^2)$ , the conditional mean function  $E(Y|SP) \equiv M(SP) = \mu(SP) = SP = \alpha + \boldsymbol{\beta}^T \mathbf{x}$ . The MLR model is discussed in detail in Chapters 2, 3 and 4.

A sufficient summary plot (SSP) of the sufficient predictor  $SP = \alpha + \boldsymbol{\beta}^T \mathbf{x}_i$  versus the response variable  $Y_i$  with the mean function added as a visual aid can be useful for describing the multiple linear regression model. This plot can not be used for real data since  $\alpha$  and  $\boldsymbol{\beta}$  are unknown. To make Figure 1.1, the artificial data used  $n = 100$  cases with  $k = 5$  nontrivial predictors. The data used  $\alpha = -1$ ,  $\boldsymbol{\beta} = (1, 2, 3, 0, 0)^T$ ,  $e_i \sim N(0, 1)$  and  $\mathbf{x}$  from a multivariate normal distribution  $\mathbf{x} \sim N_5(\mathbf{0}, \mathbf{I})$ .

In Figure 1.1, notice that the *identity line* with unit slope and zero intercept corresponds to the mean function since the identity line is the line  $Y = SP = \alpha + \boldsymbol{\beta}^T \mathbf{x} = \mu(SP) = E(Y|SP)$ . The vertical deviation of  $Y_i$

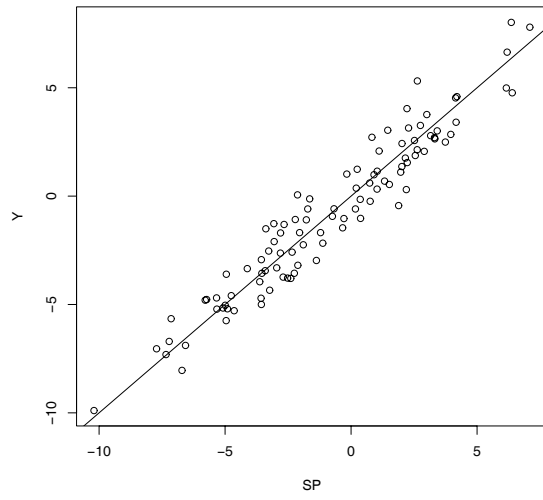


Figure 1.1: SSP for MLR Data

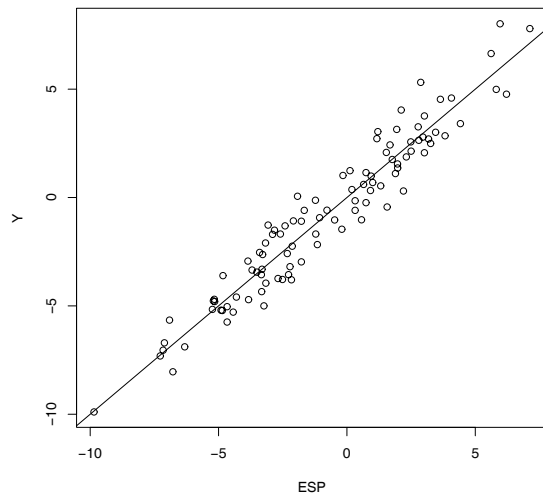


Figure 1.2: ESSP = Response Plot for MLR Data

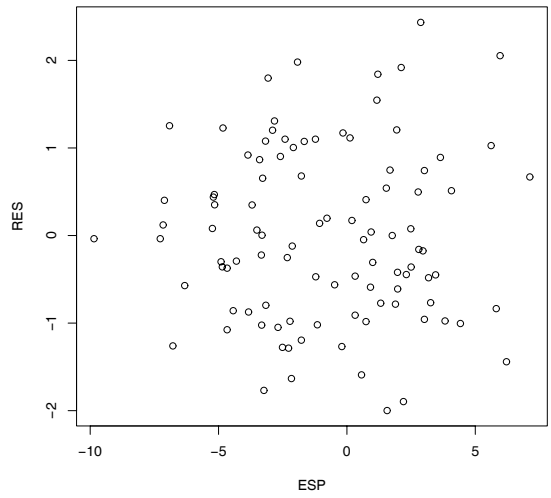


Figure 1.3: Residual Plot for MLR Data

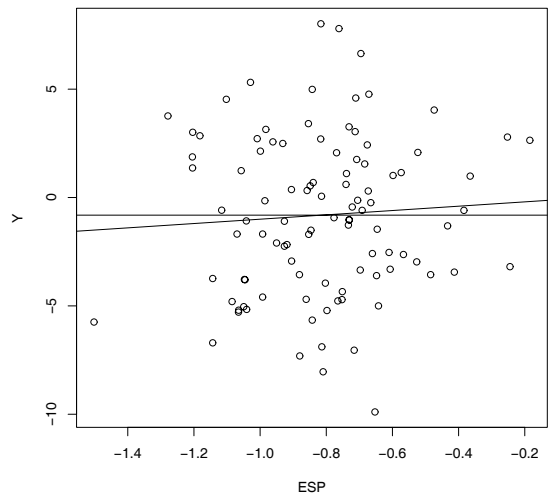


Figure 1.4: Response Plot when  $Y$  is Independent of the Predictors

from the line is equal to  $e_i = Y_i - (\alpha + \boldsymbol{\beta}^T \mathbf{x}_i)$ . For a given value of  $SP$ ,  $Y_i \sim N(SP, \sigma^2)$ . For the artificial data,  $\sigma^2 = 1$ . Hence if  $SP = 0$  then  $Y_i \sim N(0, 1)$ , and if  $SP = 5$  then  $Y_i \sim N(5, 1)$ . Imagine superimposing the  $N(SP, \sigma^2)$  curve at various values of  $SP$ . If all of the curves were shown, then the plot would resemble a road through a tunnel. For the artificial data, each  $Y_i$  is a sample of size 1 from the normal curve with mean  $\alpha + \boldsymbol{\beta}^T \mathbf{x}_i$ .

The estimated sufficient summary plot (ESSP) is a plot of  $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i$  versus  $Y_i$  with the identity line added as a visual aid. For MLR, the ESP =  $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}$  and the estimated conditional mean function is  $\hat{\mu}(ESP) = ESP$ . The estimated or fitted value of  $Y_i$  is equal to  $\hat{Y}_i = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}$ . Now the vertical deviation of  $Y_i$  from the identity line is equal to the residual  $r_i = Y_i - (\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)$ . The interpretation of the ESSP is almost the same as that of the SSP, but now the mean SP is estimated by the estimated sufficient predictor (ESP). This plot is also called the **response plot** and is used as a goodness of fit diagnostic. The residual plot is a plot of the ESP versus  $r_i$  and is used as a lack of fit diagnostic. These two plots should be made immediately after fitting the MLR model and before performing inference. Figures 1.2 and 1.3 show the response plot and residual plot for the artificial data.

The response plot is also a useful visual aid for describing the ANOVA F test (see § 2.4) which tests whether  $\boldsymbol{\beta} = \mathbf{0}$ , that is, whether the nontrivial predictors  $\mathbf{x}$  are needed in the model. If the predictors are not needed in the model, then  $Y_i$  and  $E(Y_i|\mathbf{x}_i)$  should be estimated by the sample mean  $\bar{Y}$ . If the predictors are needed, then  $Y_i$  and  $E(Y_i|\mathbf{x}_i)$  should be estimated by the ESP  $\hat{Y}_i = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i$ . If the identity line clearly fits the data better than the horizontal line  $Y = \bar{Y}$ , then the ANOVA F test should have a small pvalue and reject the null hypothesis  $H_o$  that the predictors  $\mathbf{x}$  are not needed in the MLR model. Figure 1.2 shows that the identity line fits the data better than any horizontal line. Figure 1.4 shows the response plot for the artificial data when only  $X_4$  and  $X_5$  are used as predictors with the identity line and the line  $Y = \bar{Y}$  added as visual aids. In this plot the horizontal line fits the data about as well as the identity line which was expected since  $Y$  is independent of  $X_4$  and  $X_5$ .

It is easy to find data sets where the response plot looks like Figure 1.4, but the pvalue for the ANOVA F test is very small. In this case, the MLR

model is statistically significant, but the investigator needs to decide whether the MLR model is practically significant.

## 1.2 Logistic Regression

Multiple linear regression is used when the response variable is quantitative, but for many data sets the response variable is categorical and takes on two values: 0 or 1. The *binary regression model* states that  $Y_1, \dots, Y_n$  are independent random variables with

$$Y_i \equiv Y_i | \mathbf{x}_i \sim \text{binomial}(1, \rho(\mathbf{x}_i)).$$

The *binary logistic regression model* is the special case where

$$P(Y = 1 | \mathbf{x}_i) = 1 - P(Y = 0 | \mathbf{x}_i) = \rho(\mathbf{x}_i) = \frac{\exp(\alpha + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha + \boldsymbol{\beta}^T \mathbf{x}_i)}. \quad (1.13)$$

The artificial data set used in the following discussion used  $\alpha = -1.5$  and  $\boldsymbol{\beta} = (1, 1, 1, 0, 0)^T$ . Let  $N_i$  be the number of cases where  $Y = i$  for  $i = 0, 1$ . For the artificial data,  $N_0 = N_1 = 100$ , and hence the total sample size  $n = N_1 + N_0 = 200$ .

Again a sufficient summary plot (SSP) of the sufficient predictor  $SP = \alpha + \boldsymbol{\beta}^T \mathbf{x}_i$  versus the response variable  $Y_i$  with the mean function added as a visual aid can be useful for describing the logistic regression (LR) model. The artificial data described above was used because the plot can not be used for real data since  $\alpha$  and  $\boldsymbol{\beta}$  are unknown.

Unlike the SSP for multiple linear regression where the mean function is always the identity line, the mean function in the SSP for LR can take a variety of shapes depending on the range of the SP. For the LR SSP,  $Y | SP \sim \text{binomial}(1, \rho(SP))$  where the mean function is

$$\rho(SP) = \frac{\exp(SP)}{1 + \exp(SP)}.$$

If the  $SP = 0$  then  $Y | SP \sim \text{binomial}(1, 0.5)$ . If the  $SP = -5$ , then  $Y | SP \sim \text{binomial}(1, \rho \approx 0.007)$  while if the  $SP = 5$ , then  $Y | SP \sim \text{binomial}(1, \rho \approx 0.993)$ . Hence if the range of the SP is in the interval  $(-\infty, -5)$ , then the

mean function is flat and  $\rho(SP) \approx 0$ . If the range of the SP is in the interval  $(5, \infty)$ , then the mean function is again flat but  $\rho(SP) \approx 1$ . If  $-5 < SP < 0$  then the mean function looks like a slide. If  $-1 < SP < 1$  then the mean function looks linear. If  $0 < SP < 5$  then the mean function first increases rapidly and then less and less rapidly. Finally, if  $-5 < SP < 5$  then the mean function has the characteristic “ESS” shape shown in Figure 1.5.

The estimated sufficient summary plot (ESSP or ESS plot or response plot) is a plot of  $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i$  versus  $Y_i$  with the estimated mean function

$$\hat{\rho}(ESP) = \frac{\exp(ESP)}{1 + \exp(ESP)}$$

added as a visual aid. The interpretation of the ESS plot is almost the same as that of the SSP, but now the SP is estimated by the estimated sufficient predictor (ESP).

The response plot is very useful as a goodness of fit diagnostic. Divide the ESP into  $J$  “slices” each containing approximately  $n/J$  cases. Compute the sample mean = sample proportion of the  $Y$ ’s in each slice and add the resulting step function to the response plot. This is done in Figure 1.6 with  $J = 10$  slices. This step function is a simple nonparametric estimator of the mean function  $\rho(SP)$ . If the step function follows the estimated LR mean function (the logistic curve) closely, then the LR model fits the data well. The plot of these two curves is a graphical approximation of the goodness of fit tests described in Hosmer and Lemeshow (2000, p. 147–156).

The deviance test described in Chapter 10 is used to test whether  $\boldsymbol{\beta} = \mathbf{0}$ , and is the analog of the ANOVA F test for multiple linear regression. If the LR model is a good approximation to the data but  $\boldsymbol{\beta} = \mathbf{0}$ , then the predictors  $\mathbf{x}$  are not needed in the model and  $\hat{\rho}(\mathbf{x}_i) \equiv \hat{\rho} = \bar{Y}$  (the usual univariate estimator of the success proportion) should be used instead of the LR estimator

$$\hat{\rho}(\mathbf{x}_i) = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}.$$

If the logistic curve clearly fits the step function better than the line  $Y = \bar{Y}$ , then  $H_o$  will be rejected, but if the line  $Y = \bar{Y}$  fits the step function about as well as the logistic curve (which should only happen if the logistic curve is linear with a small slope), then  $Y$  may be independent of the predictors.

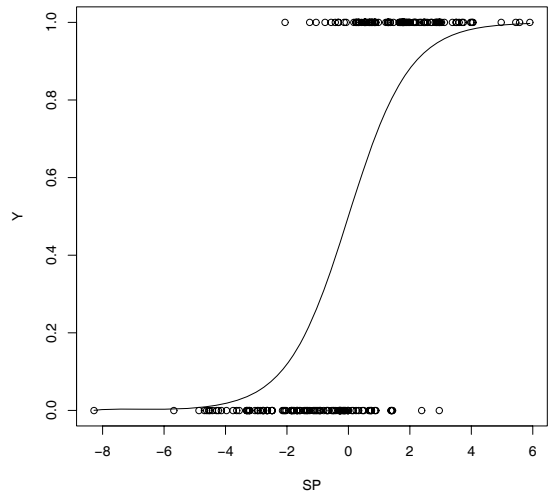


Figure 1.5: SSP for LR Data

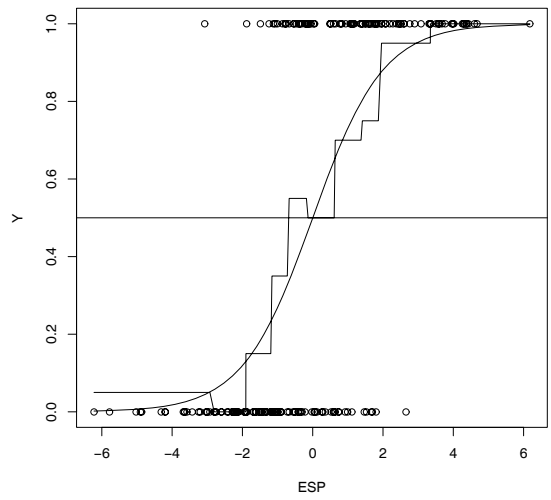


Figure 1.6: Response Plot for LR Data

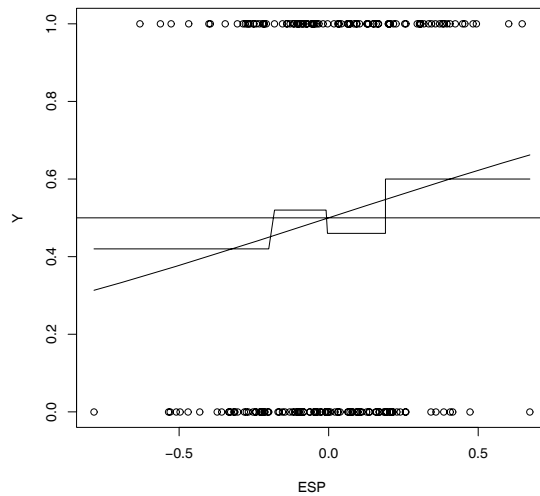


Figure 1.7: Response Plot When  $Y$  Is Independent Of The Predictors

Figure 1.7 shows the response plot when only  $X_4$  and  $X_5$  are used as predictors for the artificial data, and  $Y$  is independent of these two predictors by construction. It is possible to find data sets that look like Figure 1.7 where the pvalue for the deviance test is very small. Then the LR relationship is statistically significant, but the investigator needs to decide whether the relationship is practically significant.

For binary data the  $Y_i$  only take two values, 0 and 1, and the residuals do not behave very well. Thus the response plot is both a goodness of fit plot and a lack of fit plot. For binomial regression, described in Chapter 10, the  $Y_i$  take on values 0, 1, ...,  $m_i$ , and residual plots may be useful if  $m_i \geq 5$  for some of the cases.

### 1.3 Poisson Regression

If the response variable  $Y$  is a count, then the *Poisson regression model* is often useful. This model states that  $Y_1, \dots, Y_n$  are independent random variables with

$$Y_i \equiv Y_i | \mathbf{x}_i \sim \text{Poisson}(\mu(\mathbf{x}_i)).$$

The *loglinear regression model* is the special case where

$$\mu(\mathbf{x}_i) = \exp(\alpha + \boldsymbol{\beta}^T \mathbf{x}_i). \quad (1.14)$$

A sufficient summary plot (SSP) of the sufficient predictor  $SP = \alpha + \boldsymbol{\beta}^T \mathbf{x}_i$  versus the response variable  $Y_i$  with the mean function added as a visual aid can be useful for describing the loglinear regression (LLR) model. Artificial data needs to be used because the plot can not be used for real data since  $\alpha$  and  $\boldsymbol{\beta}$  are unknown. The data used in the discussion below had  $n = 100$ ,  $\mathbf{x} \sim N_5(\mathbf{1}, \mathbf{I}/4)$  and

$$Y_i \sim \text{Poisson}(\exp(\alpha + \boldsymbol{\beta}^T \mathbf{x}_i))$$

where  $\alpha = -2.5$  and  $\boldsymbol{\beta} = (1, 1, 1, 0, 0)^T$ .

The shape of the mean function  $\mu(SP) = \exp(SP)$  for loglinear regression depends strongly on the range of the SP. The variety of shapes occurs because the plotting software attempts to fill the vertical axis. If the range of the SP is narrow, then the exponential function will be rather flat. If the range of the SP is wide, then the exponential curve will look flat in the left of the plot but will increase sharply in the right of the plot. Figure 1.8 shows the SSP for the artificial data. Notice that  $Y|SP = 0 \sim \text{Poisson}(1)$ . In general,  $Y|SP \sim \text{Poisson}(\exp(SP))$ .

The estimated sufficient summary plot (ESSP or response plot) is a plot of the  $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i$  versus  $Y_i$  with the estimated mean function

$$\hat{\mu}(ESP) = \exp(ESP)$$

added as a visual aid. The interpretation of the response plot is almost the same as that of the SSP, but now the SP is estimated by the estimated sufficient predictor (ESP).

The response plot is very useful as a goodness of fit diagnostic. The lowess curve is a nonparametric estimator of the mean function called a “scatterplot smoother.” The lowess curve is represented as a jagged curve to distinguish it from the estimated LLR mean function (the exponential curve) in Figure 1.9. If the lowess curve follows the exponential curve closely (except possibly for the largest values of the ESP), then the LLR model fits the data well. A *useful lack of fit plot* is a plot of the ESP versus the *deviance residuals* that

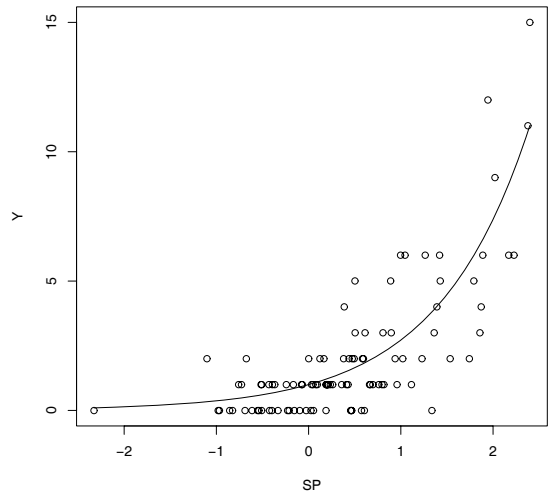


Figure 1.8: SSP for Poisson Regression

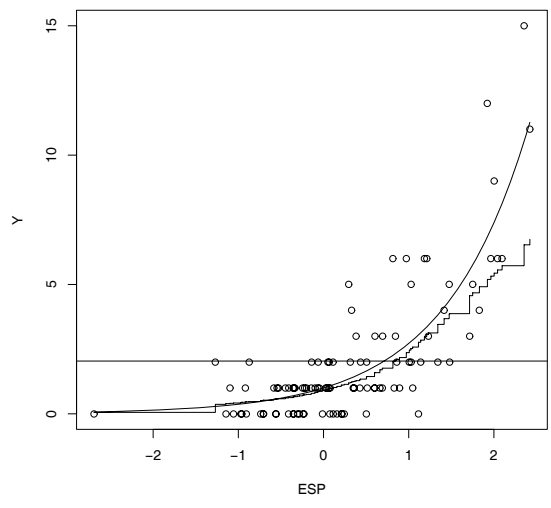


Figure 1.9: Response Plot for Poisson Regression

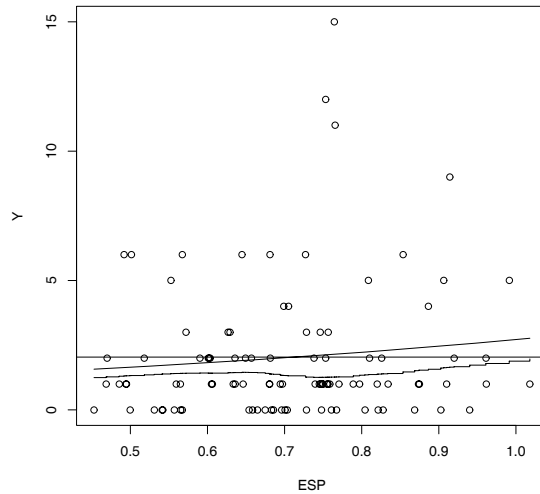


Figure 1.10: Response Plot when  $Y$  is Independent of the Predictors

are often available from the software. Additional plots are given in Chapter 11.

The deviance test described in Chapter 11 is used to test whether  $\beta = \mathbf{0}$ , and is the analog of the ANOVA F test for multiple linear regression. If the LLR model is a good approximation to the data but  $\beta = \mathbf{0}$ , then the predictors  $\mathbf{x}$  are not needed in the model and  $\hat{\mu}(\mathbf{x}_i) \equiv \hat{\mu} = \bar{Y}$  (the sample mean) should be used instead of the LLR estimator

$$\hat{\mu}(\mathbf{x}_i) = \exp(\hat{\alpha} + \hat{\beta}^T \mathbf{x}_i).$$

If the exponential curve clearly fits the lowest curve better than the line  $Y = \bar{Y}$ , then  $H_o$  should be rejected, but if the line  $Y = \bar{Y}$  fits the lowest curve about as well as the exponential curve (which should only happen if the exponential curve is approximately linear with a small slope), then  $Y$  may be independent of the predictors. Figure 1.10 shows the ESSP when only  $X_4$  and  $X_5$  are used as predictors for the artificial data, and  $Y$  is independent of these two predictors by construction. It is possible to find data sets that look like Figure 1.10 where the pvalue for the deviance test is very small. Then the LLR relationship is statistically significant, but the investigator needs to decide whether the relationship is practically significant.

## 1.4 Single Index Models

The *single index model* with additive error

$$Y = m(\alpha + \boldsymbol{\beta}^T \mathbf{x}) + e = m(SP) + e \quad (1.15)$$

includes the multiple linear regression model as a special case. In the sufficient summary plot of  $SP = \alpha + \boldsymbol{\beta}^T \mathbf{x}$  versus  $Y$ , the plotted points fall about the curve  $m(SP)$ . The vertical deviation from the curve is  $Y - m(SP) = e$ . If the  $e_i$  are iid  $N(0, 1)$  random variables, then  $Y|SP \sim N(m(SP), \sigma^2)$ . Often  $m$  and/or the distribution of  $e$  is unknown, and then the single index model is a *semiparametric model*. See Chapter 15.

The response plot of the ESP versus  $Y$  can be used to visualize the conditional distribution  $Y|SP$  and to visualize the conditional mean function  $E(Y|SP) \equiv M(SP) = m(SP)$ . The response plot can also be used to check the goodness of fit of the single index model. If  $m$  is known, add the estimated mean function  $\hat{M}(\mathbf{x}) = m(ESP)$  to the plot. If  $m$  is unknown, add a nonparametric estimator of the mean function  $\hat{M}(\mathbf{x}) = \hat{m}(ESP)$  such as lowess to the response plot. If the data randomly scatters about the estimated mean function, then the single index model may be a useful approximation to the data. The residual plot of the ESP versus the residuals  $r = Y - \hat{m}(ESP)$  should scatter about the horizontal line  $r = 0$  if the errors are iid with mean zero and constant variance  $\sigma^2$ . The response plot can also be used as a diagnostic for  $H_o : \boldsymbol{\beta} = \mathbf{0}$ . If the estimated mean function  $\hat{m}(ESP)$  fits the data better than any horizontal line, then  $H_o$  should be rejected.

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

$$h_{a,c}(w) = m\left(\frac{w - a}{c}\right)$$

for  $c \neq 0$ . In other words, if  $m$  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  $Y_i = m(\alpha + \boldsymbol{\beta}^T \mathbf{x}_i) + e_i$  if the errors  $e_i$  are small. Ordinary least squares (OLS) often provides a useful estimator of  $c\boldsymbol{\beta}$  where  $c \neq 0$ , but OLS can result in  $c = 0$  if  $m$  is symmetric about the median of  $\alpha + \boldsymbol{\beta}^T \mathbf{x}$ .

The software packages *Splus* (MathSoft 1999ab) and *R*, the free version of *Splus* available from ([www.r-project.org/](http://www.r-project.org/)), can be used to generate artificial single index model data sets. The *R/Splus* commands

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

were used to generate 100 trivariate Gaussian predictors  $\mathbf{x} \sim N_3(\mathbf{0}, \mathbf{I}_3)$  and the response  $Y = (\boldsymbol{\beta}^T \mathbf{x})^3 + e = (x_1 + 2x_2 + 3x_3)^3 + e$  where  $e \sim N(0, 1)$ . This is a single index model where  $m$  is the cubic function,  $\boldsymbol{\beta} = (1, 2, 3)^T$  and  $\alpha = 0$ . Figure 1.11 shows the sufficient summary plot of  $\boldsymbol{\beta}^T \mathbf{x}$  versus  $Y$ , and Figure 1.12 shows the sufficient summary plot of  $-\boldsymbol{\beta}^T \mathbf{x}$  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.

```
plot(SP,Y)
plot(-SP,Y)
```

An amazing result is that the unknown function  $m$  can often be visualized by the response plot called the “OLS view,” a plot of the OLS ESP (the OLS fit, possibly ignoring the constant) versus  $Y$  generated by the following commands.

```
bols <- lsfit(X,Y)$coef[-1]
plot(X %*% bols, Y)
```

The OLS view, shown in Figure 1.13, can be used to visualize  $m$  and for prediction. Note that  $Y$  appears to be a cubic function of the OLS ESP and that if the OLS ESP = 0, then the graph suggests using  $\hat{Y} = 0$  as the predicted value for  $Y$ . Since the plotted points cluster about a smooth curve better than any horizontal line, the OLS view suggests that a single index model is appropriate and that  $\boldsymbol{\beta} \neq \mathbf{0}$ .

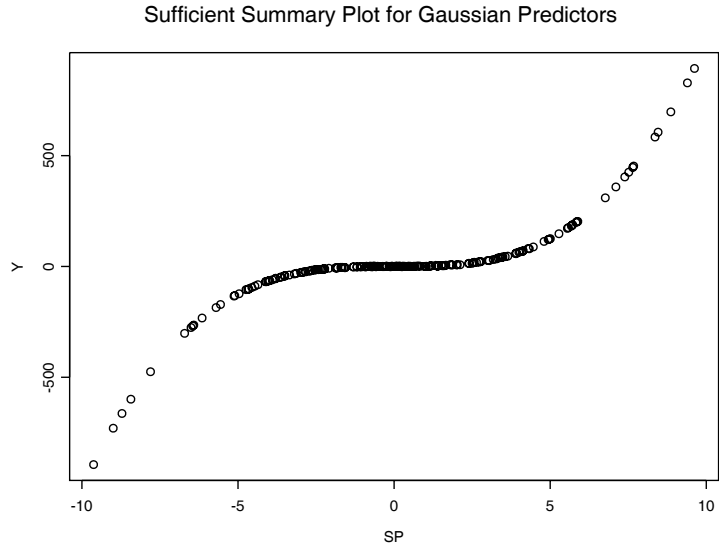


Figure 1.11: SSP for  $m(u) = u^3$

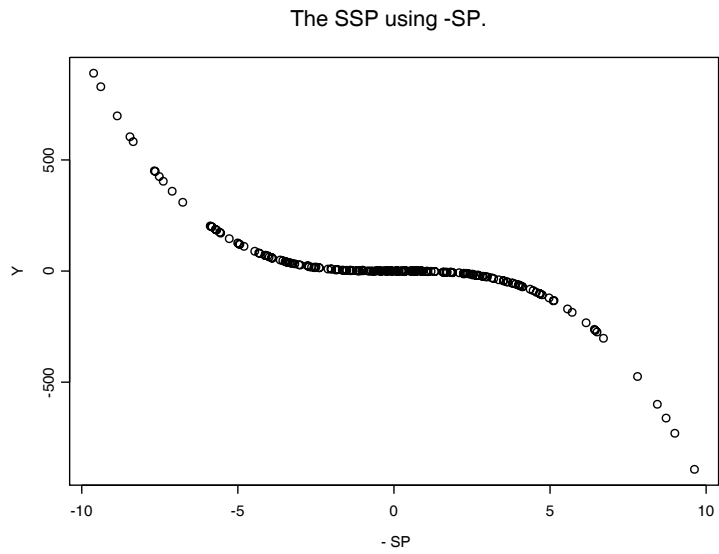


Figure 1.12: Another SSP for  $m(u) = u^3$

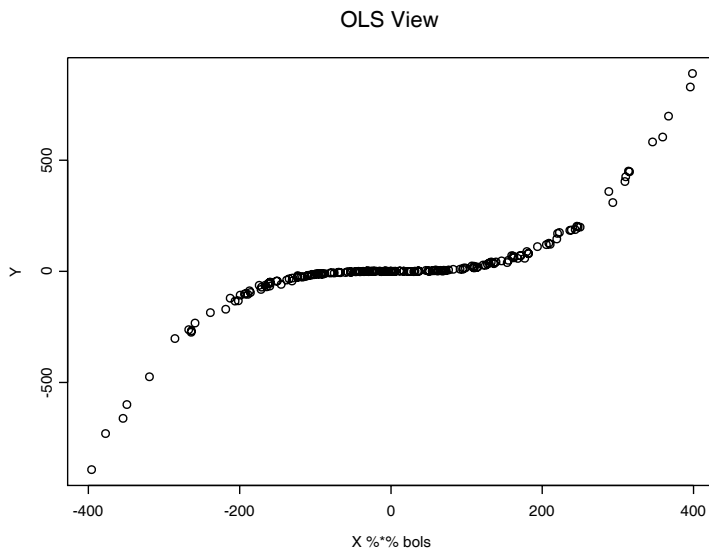


Figure 1.13: OLS View for  $m(u) = u^3$

## 1.5 Survival Regression Models

The most important survival regression models are 1D models, and are described in detail in Chapter 16. For these models, the conditional survival function  $S_{Y|SP}(t) = P(Y > t | \boldsymbol{\beta}^T \mathbf{x}) = P(Y > t | SP)$  and the conditional hazard function  $h_{Y|SP}(t)$  are of great interest. Hence the response plot is no longer of great interest. Instead, the slice survival plot is used to visualize  $S_{Y|SP}(t)$ .

The *Cox proportional hazards* regression model (Cox 1972) is a semiparametric model with  $SP = \boldsymbol{\beta}_C^T \mathbf{x}$  and

$$h_{\mathbf{x}}(t) \equiv h_{Y|SP}(t) = \exp(\boldsymbol{\beta}_C^T \mathbf{x}) h_0(t) = \exp(SP) h_0(t)$$

where the baseline hazard function  $h_0(t)$  is left unspecified. The survival function is

$$S_{\mathbf{x}}(t) \equiv S_{Y|SP}(t) = [S_0(t)]^{\exp(\boldsymbol{\beta}_C^T \mathbf{x})} = [S_0(t)]^{\exp(SP)}$$

where  $S_0(t)$  is the unspecified baseline survival function.

For *parametric proportional hazards* regression models, the baseline function is parametric and the parameters are estimated via maximum likelihood.

Then as a 1D regression model,  $SP = \boldsymbol{\beta}_P^T \mathbf{x}$ , and

$$h_{Y|SP}(t) \equiv h_{\mathbf{x}}(t) = \exp(\boldsymbol{\beta}_P^T \mathbf{x}) h_{0,P}(t) = \exp(SP) h_{0,P}(t)$$

where the parametric baseline function depends on  $k$  unknown parameters but does not depend on the predictors  $\mathbf{x}$ . The survival function is

$$S_{\mathbf{x}}(t) \equiv S_{Y|SP}(t) = [S_{0,P}(t)]^{\exp(\boldsymbol{\beta}_P^T \mathbf{x})} = [S_{0,P}(t)]^{\exp(SP)},$$

and

$$\hat{S}_{\mathbf{x}}(t) = [\hat{S}_{0,P}(t)]^{\exp(\hat{\boldsymbol{\beta}}_P^T \mathbf{x})} = [\hat{S}_{0,P}(t)]^{\exp(ESP)}.$$

The Weibull regression model is an important special case.

For a parametric *accelerated failure time* model,

$$\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \mathbf{x}_i + \sigma e_i$$

where the  $e_i$  are iid from a location scale family. Let  $SP = \boldsymbol{\beta}_A^T \mathbf{x}$ . Then as a 1D regression model,  $\log(Y)|SP = \alpha + SP + e$ . The parameters are again estimated by maximum likelihood and the survival function is

$$S_{\mathbf{x}}(t) \equiv S_{Y|\mathbf{x}}(t) = S_0 \left( \frac{t}{\exp(\boldsymbol{\beta}_A^T \mathbf{x})} \right),$$

and

$$\hat{S}_{\mathbf{x}}(t) = \hat{S}_0 \left( \frac{t}{\exp(\hat{\boldsymbol{\beta}}_A^T \mathbf{x})} \right)$$

where  $\hat{S}_0(t)$  depends on  $\hat{\alpha}$  and  $\hat{\sigma}$ .

## 1.6 Variable Selection

A standard problem in 1D regression is variable selection, also called subset or model selection. Assume that  $Y \perp\!\!\!\perp \mathbf{x} | (\alpha + \boldsymbol{\beta}^T \mathbf{x})$ , that a constant is always included, that  $\mathbf{x} = (x_1, \dots, x_{p-1})^T$  are the  $p - 1$  nontrivial predictors and that  $(1, \mathbf{x})^T$  has full rank. Then *variable selection* is a search for a subset of predictor variables that can be deleted without important loss of information.

To clarify ideas, assume that there exists a subset  $S$  of predictor variables such that if  $\mathbf{x}_S$  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  $\mathbf{x} = (\mathbf{x}_S^T, \mathbf{x}_E^T)^T$ . Then

$$SP = \alpha + \boldsymbol{\beta}^T \mathbf{x} = \alpha + \boldsymbol{\beta}_S^T \mathbf{x}_S + \boldsymbol{\beta}_E^T \mathbf{x}_E = \alpha + \boldsymbol{\beta}_S^T \mathbf{x}_S. \quad (1.16)$$

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

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

where  $\mathbf{x}_{I/S}$  denotes the predictors in  $I$  that are not in  $S$ . Since this is true regardless of the values of the predictors,  $\boldsymbol{\beta}_O = \mathbf{0}$  if  $S \subseteq I$ . Hence for any subset  $I$  that includes all relevant predictors, the population correlation

$$\text{corr}(\alpha + \boldsymbol{\beta}^T \mathbf{x}_i, \alpha + \boldsymbol{\beta}_I^T \mathbf{x}_{I,i}) = 1. \quad (1.17)$$

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  $j = 1, 2, \dots, p - 1$  nontrivial predictors, keep track of subsets  $I$  that provide the largest values of  $\text{corr}(\text{ESP}, \text{ESP}(I))$ . 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  $\text{ESP}(I)$  satisfies

$$\text{corr}(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \mathbf{x}_i, \tilde{\alpha}_I + \tilde{\boldsymbol{\beta}}_I^T \mathbf{x}_{I,i}) = \text{corr}(\tilde{\boldsymbol{\beta}}^T \mathbf{x}_i, \tilde{\boldsymbol{\beta}}_I^T \mathbf{x}_{I,i}) \geq 0.95. \quad (1.18)$$

The difficulty with this approach is that fitting large numbers of possible submodels involves substantial computation. Fortunately, OLS frequently gives a useful ESP and methods originally meant for multiple linear regression using the Mallows'  $C_p$  criterion (see Jones 1946 and Mallows 1973) also work for more general 1D regression models. As a rule of thumb, the OLS ESP is useful if  $|\text{corr}(\text{OLS ESP}, \text{ESP})| \geq 0.95$  where ESP is the standard ESP (eg, for generalized linear models, the ESP is  $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}$  where  $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$  is the maximum likelihood estimator of  $(\alpha, \boldsymbol{\beta})$ ), or if the OLS response plot suggests that the

OLS ESP is good. Variable selection will be discussed in much greater detail in Chapters 3, 10, 11, 12, 15 and 16, but the following methods are useful for a large class of 1D regression models.

Perhaps the simplest method of variable selection is the *t directed search* (see Daniel and Wood 1980, p. 100–101). Let  $k$  be the number of predictors in the model, including the constant. Hence  $k = p$  for the full model. Let  $X_1, \dots, X_{p-1}$  denote the nontrivial predictor variables and let  $W_1, W_2, \dots, W_{p-1}$  be the predictor variables in decreasing order of importance. Use theory if possible, but if no theory is available then fit the full model using OLS and let  $t_i$  denote the t statistic for testing  $H_o : \beta_i = 0$ . Let  $|t|_{(1)} \leq |t|_{(2)} \leq \dots \leq |t|_{(p-1)}$ . Then  $W_i$  corresponds to the  $X_j$  with  $|t|_{(p-i)}$  for  $i = 1, 2, \dots, p - 1$ . That is,  $W_1$  has the largest t statistic,  $W_2$  the next largest, etc. Then use OLS to compute  $C_p(I_j)$  for the  $p - 1$  models  $I_j$  where  $I_j$  contains  $W_1, \dots, W_j$  and a constant for  $j = 1, \dots, p - 1$ .

**Forward selection** starts with a constant =  $W_0$ .

Step 1)  $k = 2$ : compute  $C_p$  for all models containing the constant and a single predictor  $X_i$ . Keep the predictor  $W_1 = X_j$ , say, that corresponds to the model with the smallest value of  $C_p$ .

Step 2)  $k = 3$ : Fit all models with  $k = 3$  that contain  $W_0$  and  $W_1$ . Keep the predictor  $W_2$  that minimizes  $C_p$ .

Step j)  $k = j + 1$ : Fit all models with  $k = j + 1$  that contains  $W_0, W_1, \dots, W_j$ . Keep the predictor  $W_{j+1}$  that minimizes  $C_p$ .

Step  $p - 1$ )  $k = p$ : Fit the full model.

**Backward elimination:** starts with the full model. All models contain a constant =  $U_0$ . Hence the full model contains  $U_0, X_1, \dots, X_{p-1}$ . We will also say that the full model contains  $U_0, U_1, \dots, U_{p-1}$  where  $U_i$  need not equal  $X_i$  for  $i \geq 1$ .

Step 1)  $k = p - 1$ : fit each model with  $p - 1$  predictors including a constant. Delete the predictor  $U_{p-1}$ , say, that corresponds to the model with the smallest  $C_p$ . Keep  $U_0, \dots, U_{p-2}$ .

Step 2)  $k = p - 2$ : fit each model with  $p - 2$  predictors including the constant. Delete the predictor  $U_{p-2}$  that corresponds to the smallest  $C_p$ . Keep  $U_0, U_1, \dots, U_{p-3}$ .

Step j)  $k = p - j$ : fit each model with  $p - j$  predictors and a constant. Delete the predictor  $U_{p-j}$  that corresponds to the smallest  $C_p$ . Keep  $U_0, U_1, \dots, U_{p-j-1}$ .

Step  $p - 2$ )  $k = 2$ : The current model contains  $U_0, U_1$  and  $U_2$ . Fit the model

$U_0, U_1$  and the model  $U_0, U_2$ . Assume that model  $U_0, U_1$  minimizes  $C_p$ . Then delete  $U_2$  and keep  $U_0$  and  $U_1$ . (Step  $p - 1$ ) which finds  $C_p$  for the model that only contains the constant  $U_0$  is often omitted.)

**All subsets variable selection** examines all subsets and keeps track of several (up to three, say) subsets with the smallest  $C_p(I)$  for each group of submodels containing  $k$  predictors including a constant. This method can be used for  $p \leq 30$  by using the efficient “leaps and bounds” algorithms when OLS and  $C_p$  is used (see Furnival and Wilson 1974).

**Rule of thumb for variable selection** (assuming that the cost of each predictor is the same): find the submodel  $I_m$  with the minimum  $C_p$ . If  $I_m$  uses  $k_m$  predictors including a constant, do not use any submodel that has more than  $k_m$  predictors. Since the minimum  $C_p$  submodel **often has too many predictors**, also look at the submodel  $I_o$  with the smallest value of  $k$ , say  $k_o$ , such that  $C_p \leq 2k$ . This submodel **may have too few predictors**. So look at the predictors in  $I_m$  but not in  $I_o$  and see if they can be deleted or not. (If  $I_m = I_o$ , then it is a good candidate for the best submodel.)

Variable selection with the  $C_p$  criterion is closely related to the partial  $F$  test for testing whether a reduced model should be used instead of the full model. *The following results are properties of OLS and hold even if the data does not follow a 1D model.* If the candidate model of  $\mathbf{x}_I$  has  $k$  terms (including the constant), then the partial F test for reduced model  $I$  uses test statistic

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

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

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p - k)(F_I - 1) + k \quad (1.19)$$

where MSE is the residual mean square for the full model. Let  $ESP(I) = \hat{\alpha}_I + \hat{\beta}_I^T \mathbf{x}$  be the ESP for the submodel and let  $V_I = Y - ESP(I)$  so that  $V_{I,i} = Y_i - \hat{\alpha}_I + \hat{\beta}_I^T \mathbf{x}_i$ . Let ESP and  $V$  denote the corresponding quantities for the full model. Then Olive and Hawkins (2005) show that  $\text{corr}(V_I, V) \rightarrow 1$

forces  $\text{corr}(\text{OLS ESP}, \text{OLS ESP}(I)) \rightarrow 1$  and that

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

Also  $C_p(I) \leq 2k$  corresponds to  $\text{corr}(V_I, V) \geq d_n$  where

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

Notice that the submodel  $I_k$  that minimizes  $C_p(I)$  also maximizes  $\text{corr}(V, V_I)$  among all submodels  $I$  with  $k$  predictors including a constant. If  $C_p(I) \leq 2k$  and  $n \geq 10p$ , then  $0.948 \leq \text{corr}(V, V(I))$ , and both  $\text{corr}(V, V(I)) \rightarrow 1.0$  and  $\text{corr}(\text{OLS ESP}, \text{OLS ESP}(I)) \rightarrow 1.0$  as  $n \rightarrow \infty$ .

If 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 graphical and numerical checks on this assumption should be made. Also assume that the OLS ESP is useful. This assumption can be checked by making an OLS response plot or by verifying that  $|\text{corr}(\text{OLS ESP}, \text{ESP})| \geq 0.95$ . Then we suggest that submodels  $I$  are “interesting” if  $C_p(I) \leq \min(2k, p)$ .

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 pvalues 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 the model has  $k = p - 1$  predictors including the constant, and the test statistic is  $t_i$  where

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

Using (1.19) and  $C_p(I_{full}) = p$ , notice that

$$C_p(I_i) = (p - (p - 1))(t_i^2 - 1) + (p - 1) = t_i^2 - 1 + C_p(I_{full}) - 1,$$

or

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

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, for the partial  $F$  test, notice that by (1.19),  $C_p(I) \leq 2k$  iff  $(p - k)F_I - p + 2k \leq 2k$  iff  $(p - k)F_i \leq p$  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  $k \rightarrow 0$ , the change in SS  $F$  test will reject  $H_0$  (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 partial  $F$  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.

## 1.7 Other Issues

The 1D regression models offer a unifying framework for many of the most used regression models. By writing the model in terms of the sufficient predictor  $SP = \alpha + \boldsymbol{\beta}^T \mathbf{x}$ , many important topics valid for all 1D regression models can be explained compactly. For example, the previous section presented variable selection, and equation (1.19) can be used to motivate the test for whether the reduced model can be used instead of the full model. Similarly, the sufficient predictor can be used to to unify the interpretation of coefficients and to explain models that contain interactions and factors.

### Interpretation of Coefficients

One interpretation of the coefficients in a 1D model is that  $\beta_i$  is the rate of change in the SP associated with a unit increase in  $x_i$  when all other predictor variables  $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p$  are held fixed. Denote a model by  $SP = \alpha + \boldsymbol{\beta}^T \mathbf{x} = \alpha + \beta_1 x_1 + \dots + \beta_p x_p$ . Then

$$\beta_i = \frac{\partial SP}{\partial x_i} \text{ for } i = 1, \dots, p.$$

Of course, holding all other variables fixed while changing  $x_i$  may not be possible. For example, if  $x_1 = x$ ,  $x_2 = x^2$  and  $SP = \alpha + \beta_1 x + \beta_2 x^2$ , then  $x_2$  can not be held fixed when  $x_1$  increases by one unit, but

$$\frac{d SP}{dx} = \beta_1 + 2\beta_2 x.$$

The interpretation of  $\beta_i$  changes with the model in two ways. First, the interpretation changes as terms are added and deleted from the SP. Hence the interpretation of  $\beta_1$  differs for models  $SP = \alpha + \beta_1 x_1$  and  $SP = \alpha + \beta_1 x_1 + \beta_2 x_2$ . Secondly, the interpretation changes as the parametric or semiparametric form of the model changes. For multiple linear regression,  $E(Y|SP) = SP$  and an increase in one unit of  $x_i$  increases the conditional expectation by  $\beta_i$ . For binary logistic regression,

$$E(Y|SP) = \rho(SP) = \frac{\exp(SP)}{1 + \exp(SP)},$$

and the change in the conditional expectation associated with a one unit increase in  $x_i$  is more complex.

### Factors for Qualitative Variables

The interpretation of the coefficients also changes if interactions and factors are present. Suppose a factor  $W$  is a qualitative random variable that takes on  $c$  categories  $a_1, \dots, a_c$ . Then the 1D model will use  $c - 1$  indicator variables  $W_i = 1$  if  $W = a_i$  and  $W_i = 0$  otherwise, where one of the levels  $a_i$  is omitted, eg, use  $i = 1, \dots, c - 1$ .

### Interactions

Suppose  $X_1$  is quantitative and  $X_2$  is qualitative with 2 levels and  $X_2 = 1$  for level  $a_2$  and  $X_2 = 0$  for level  $a_1$ . Then a first order model with interaction is  $SP = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ . This model yields two unrelated lines in the sufficient predictor depending on the value of  $x_2$ :  $SP = \alpha + \beta_2 + (\beta_1 + \beta_3)x_1$  if  $x_2 = 1$  and  $SP = \alpha + \beta_1 x_1$  if  $x_2 = 0$ . If  $\beta_3 = 0$ , then there are two parallel lines:  $SP = \alpha + \beta_2 + \beta_1 x_1$  if  $x_2 = 1$  and  $SP = \alpha + \beta_1 x_1$  if  $x_2 = 0$ . If  $\beta_2 = \beta_3 = 0$ , then the two lines are coincident:  $SP = \alpha + \beta_1 x_1$  for both values of  $x_2$ . If  $\beta_2 = 0$ , then the two lines have the same intercept:  $SP = \alpha + (\beta_1 + \beta_3)x_1$  if  $x_2 = 1$  and  $SP = \alpha + \beta_1 x_1$  if  $x_2 = 0$ . In general, as factors have more levels and interactions have more terms, eg  $x_1 x_2 x_3 x_4$ , the interpretation of the model rapidly becomes very complex.

## 1.8 Complements

To help explain the given 1D model, use the sufficient summary plot (SSP) of  $SP = \alpha + \boldsymbol{\beta}^T \mathbf{x}_i$  versus  $Y_i$  with the mean function added as a visual aid. If  $p = 1$ , then  $Y \perp\!\!\!\perp x|x$  and the plot of  $x_i$  versus  $Y_i$  is a SSP and has been widely used to explain the simple linear regression (SLR) model and the logistic regression model with one predictor. See Agresti (2002, cover illustration and p. 169) and Collett (1999, p. 74). Replacing  $x$  by  $SP$  has two major advantages. First, the plot can be made for  $k \geq 1$  and secondly, the possible shapes that the plot can take is greatly reduced. For example, in a plot of  $x_i$  versus  $Y_i$ , the plotted points will fall about some line with slope  $\beta$  and intercept  $\alpha$  if the SLR model holds, but in a plot of  $SP = \alpha + \boldsymbol{\beta}^T \mathbf{x}_i$  versus  $Y_i$ , the plotted points will fall about the identity line with unit slope and zero intercept if the multiple linear regression model holds.

Important theoretical results for the single index model were given by Brillinger (1977, 1983) and Aldrin, Bølviken and Schweder (1993). Li and Duan (1989) extended these results to models of the form

$$Y = g(\alpha + \boldsymbol{\beta}^T \mathbf{x}, e) \tag{1.20}$$

where  $g$  is a bivariate inverse link function. Olive and Hawkins (2005) discuss variable selection while Chang (2006) and Chang and Olive (2009) discuss OLS tests. Severini (1998) discusses when OLS output is relevant for the Gaussian additive error single index model.

## 1.9 Problems

**1.1.** Explain why the model  $Y = g(\alpha + \boldsymbol{\beta}^T \mathbf{x}, e)$  can also be written as  $Y = g(\alpha + \mathbf{x}^T \boldsymbol{\beta}, e)$ .