Graphical Methods of Determining Predictor Importance and Effect

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Aaron Kjell Rendahl

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Sanford Weisberg, Adviser

August 2008

Acknowledgements

Many thanks to the faculty, staff, and fellow students of the School of Statistics. With your expertise, support, and friendship, I have learned more statistics than I ever imagined.

Special thanks to my adviser, Sandy Weisberg, with whom I have enjoyed exploring all kinds of plots and learning about model combining methods. I have especially valued your expertise in statistical consulting, both in doing statistics and in dealing with clients, and look forward to working with you in the Statistical Clinic.

Finally, this thesis would never have been completed without the steadfast support and encouragement of my wife, Tessa. I love you more than I can say.

Abstract

Many experiments and studies are designed to discover how a group of predictors affect a single response. For example, an agricultural scientist may perform an experiment to determine how rainfall, sunlight, and fertilizer affect plant growth. In situations like this, graphical methods to show how the various predictors affect the response and the relative importance of each predictor can be invaluable, not only in helping the researcher understand the results, but also in communicating the findings to non-specialists.

For settings where a simple statistical model can be used to fit the data, several graphical methods for showing the effect of individual predictors already exist. However, few methods are available for more complex settings that require more complex models. A framework for understanding the existing methods is developed using Cook's net-effect plots, and a criterion for evaluating and creating methods is proposed. This criterion states that for a plot to be most useful in showing how a given predictor affects the response, the conditional distribution of the vertical axis given the horizontal axis should be independent of the other predictors. That is, the plot should not hide any additional information gained by knowing the other predictors.

This proposed framework and criterion is used to develop graphical methods appropriate for use in more complex modeling algorithms. In particular, these plots have been explored in the context of model combining methods, and various versions compared and analyzed. Additionally, the weights from these model combining methods are used to modify existing methods of determining predictor importance values, resulting in improved values for spurious predictors.

Contents

List of Tables

List of Figures

Chapter 1

Overview of useful methods

Given a univariate response Y and a set of p predictors $X = (x_1, \ldots, x_p)$, a regression analysis studies the conditional distribution of $Y|X$ as X varies. There are at least two main reasons for this analysis: to predict values of Y from given values of X, and to understand how the predictors X affect the response Y. The goal of this thesis is to explore graphical methods that show the importance and effect of a subset of q predictors $X_2 = (x_{p-q+1}, \ldots, x_p)$ to aid in the understanding of the relationship between the predictors and the response.

This chapter will develop criteria to determine when a plot might be useful and review various types of useful plots in the context of multivariate normal (X, Y). After a framework for understanding these plots is developed, several specific model types and fitting methods will be explored.

A plot of x on the horizontal axis and y on the vertical axis will be designated with the notation $\{x, y\}$.

1.1 Scatterplot

Consider the case where $p = q = 1$, so there is one predictor X, and a response Y. The most common method of displaying their relationship is a *scatterplot*. With X

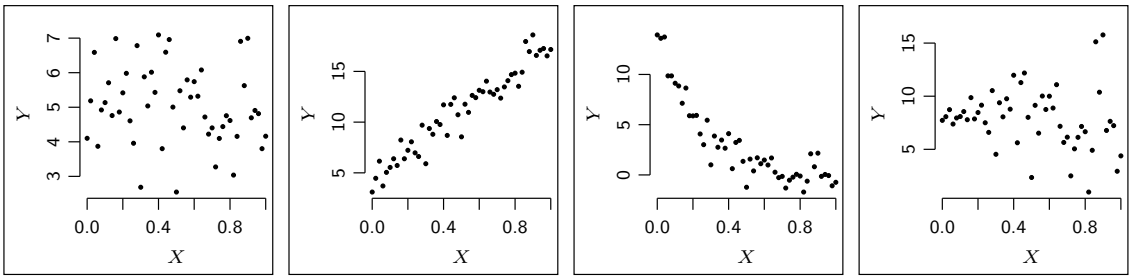
on the horizontal axis and Y on the vertical axis, each (X, Y) point is plotted, which shows the joint distribution of X and Y. The conditional distribution of $Y|X$ can be examined by visualizing narrow vertical slices of the data along the x-axis. The points in each slice correspond to possible values for $Y|(X \in \text{slice})$. Thus various features of the conditional distribution can be inspected, including the mean, the variance, and possibly the skewness and outliers, as well as how these features change as X is varied. Because there are no other predictors, this plot contains a full description of how the distribution of $Y|X$ behaves for all X. Thus in general a two-dimensional scatterplot provides visualization of the conditional distribution of $Y|X$ as X is varied. A scatterplot in this situation is called a *response plot* of Y versus X.

Figure 1.1 demonstrates how scatterplots look for several different conditional mean and variance functions. Figure 1.1a shows independent X and Y, so $E(Y|X)$ and $\text{Var}(Y|X)$ are both constant, and the distribution of points for each visualized slice along the x-axis is the same. Figure 1.1b shows a relationship where $E(Y|X)$ is a linear function of X and $\text{Var}(Y|X)$ is constant, so the distribution of points for each visualized slice along the x-axis changes only in location, and that change is linear. Figure 1.1c shows a similar relationship; $\text{Var}(Y|X)$ is still constant, so the distribution of points for each visualized slice along the x-axis changes only in location, but that change is non-linear, showing that $E(Y|X)$ is nonlinear. Figure 1.1d shows a relationship where $\text{Var}(Y|X)$ increases in X but $E(Y|X)$ is constant; we see that the spread of points in each visualized slice increases, although the average location remains constant.

In the bivariate normal context, suppose

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & s \\ s & 1 \end{pmatrix}\right). \tag{1.1}$$

If the means of X and Y are not zero, there is a location change in the plot, but not



(a) Independent $\{X, Y\}$.

(b) Linear mean function, constant variance.

(c) Non-linear mean function, constant variance.

(d) Constant mean function, increasing variance.

Figure 1.1: Scatterplots showing various conditional mean and variance functions.

a change in shape. Similarly, if the variances of X and Y are not one, there is only a scale change. Then

$$Y|X \sim N(sX, 1 - s^2), \tag{1.2}$$

so the mean function $E(Y|X) = sX$ is linear in X, and the variance has no X dependence.

In addition to the points on the scatterplot, a line representing this mean function may be added to increase visual perception of the mean. This is especially useful because visual perception of the mean from the points can be misleading. For example, in the bivariate normal context, contours of constant probability for the bivariate normal are g, so the scatter of points observed in the $\{X, Y\}$ plot will also be elliptical. However, the mean function of the conditional distribution does not line up with the principal axes of the ellipses (see Freedman et al., 1978, p. 147). Instead, the conditional mean function is visually estimated by estimating the mean of vertical slices. In Figure 1.2, ellipses of constant probability are shown for $s = 0.5$, along with the mean function $Y = sX$. The mean function does not line up with the principal axes of the ellipses, but it does bisect the vertical slice marked by the dotted line (for $x \in (2, 2.4)$).

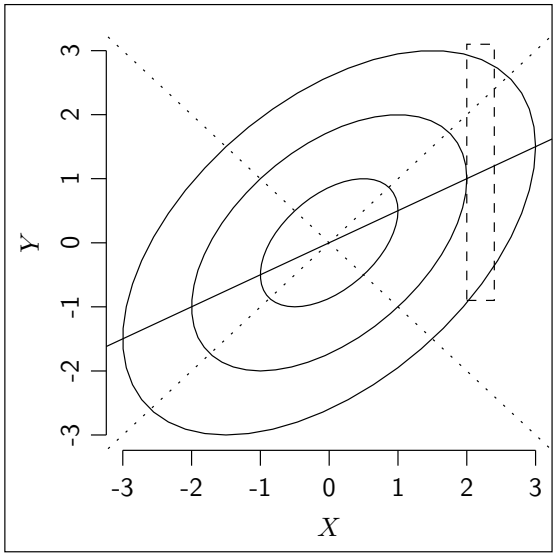
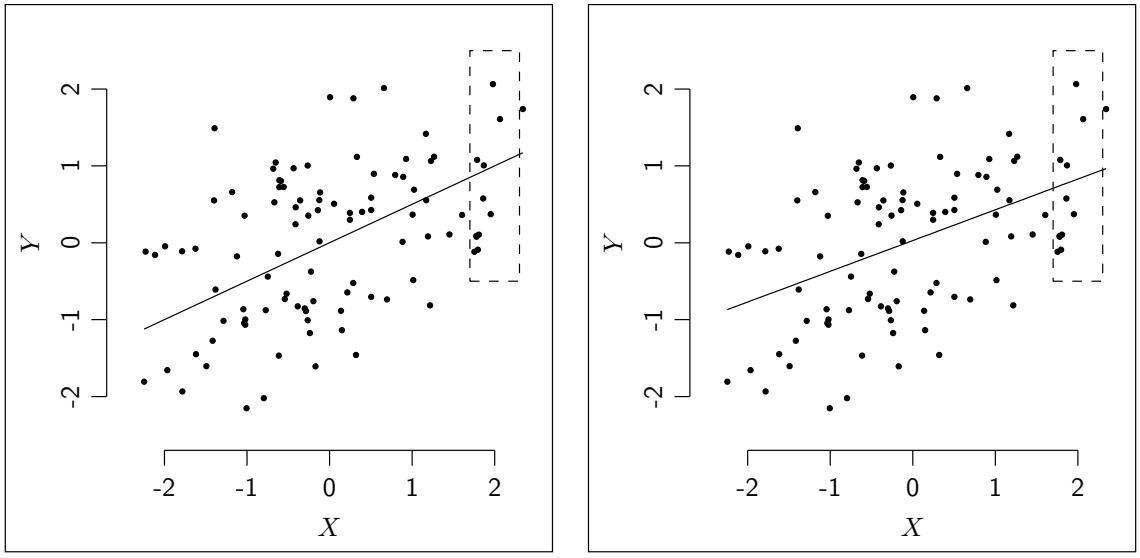


Figure 1.2: Elliptical contours of constant probability for a bivariate normal distribution. Dotted lines show the principal axes, and the solid line shows the conditional mean function.

A difficulty with adding the mean function to the plot is that the mean function is not usually known, but instead must be estimated from the data. In this investigation, examples will often be simulated and thus fully understood, so the plots can be drawn using either the known true values, or estimates of these values from the data. Plots based on known population values will be called *idealized plots*, and plots based on estimates from the data will be called *estimated plots*. This distinction will become more important in more complex plots where the plotted points themselves must be estimated along with the mean function.

Example 1.1 (Scatterplot of bivariate normal data with one predictor.) A sample of size 100 has been taken from the bivariate normal described in (1.1), with $s = 0.5$. Scatterplots of $\{X, Y\}$ are shown in Figure 1.3, showing the joint distribution of X and Y. Judging by the density of the points, the contours of equal probability appear elliptical. Additionally, for any given value of X, the conditional



(a) With idealized conditional mean. (b) With estimated conditional mean.

Figure 1.3: Scatterplots of the bivariate normal data from Example 1.1, with idealized and estimated conditional mean functions added.

distribution of $Y|X$ can be inspected by visualizing a narrow slice of the data around that value. Take, for instance, a narrow slice around $X = 2$ (shown in both figures). The Y values of the points in this slice are centered roughly around $Y = 1$; this is a rough estimate of the conditional mean for $X = 2$. Similarly, a rough estimate of the conditional mean at $X = 0$ is $Y = 0$. Using these points, a rough estimate for the slope is $(1 - 0)/(2 - 0) = 0.5$, which turns out to be exactly s. Figure 1.3a shows the idealized conditional mean line, with slope $s = 0.5$ and intercept 0, and Figure 1.3b shows the least squares estimate of the conditional mean line, with slope 0.4 and intercept 0.03. Finally, the variance seems constant for all values of X. □

1.2 Marginal response plot

Now consider the case where there are two predictors with one of interest, so $p = 2$ and $q = 1$, and the goal remains to study the dependence of Y on X_2. A first

method is to simply ignore the information in X_1, and plot the scatterplot $\{X_2, Y\}$. As in the case where $p = 1$, the conditional distribution of $Y|X_2$ can be inspected by visualizing vertical slices. As this scatterplot ignores X_1 and so only contains the marginal information of X_2 and Y, it is called a *marginal response plot* (Cook and Weisberg, 1999b).

In the multivariate normal context, let

$$\begin{pmatrix} Y \\ X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r & s \\ r & 1 & \rho \\ s & \rho & 1 \end{pmatrix}\right). \tag{1.3}$$

Ignoring X_1, or equivalently, integrating it out, the conditional distribution of $Y|X_2$ is

$$Y|X_2 \sim N(sX_2, 1 - s^2), \tag{1.4}$$

exactly as in the case where $p = 1$. This can also be found by the rule of iterated expectations,

$$f(y|x_2) = E\left(f(y|x_1, x_2)|x_2\right), \tag{1.5}$$

where $f(a)$ denotes the density of A. This relationship will be used again in Section 1.8 for the marginal model checking plot.

Example 1.2 (Marginal response plot for data with two predictors.)
A sample of size 100 has been taken from the multivariate normal of (1.3), with $r = 0.9$, $s = 0.8$ and $\rho = 0.95$. Figure 1.4 shows the marginal response plot for X_2, $\{X_2, Y\}$. The idealized conditional mean line, with slope 0.8, is drawn with a solid line, and the least squares estimate of the conditional mean line, with slope of 0.81

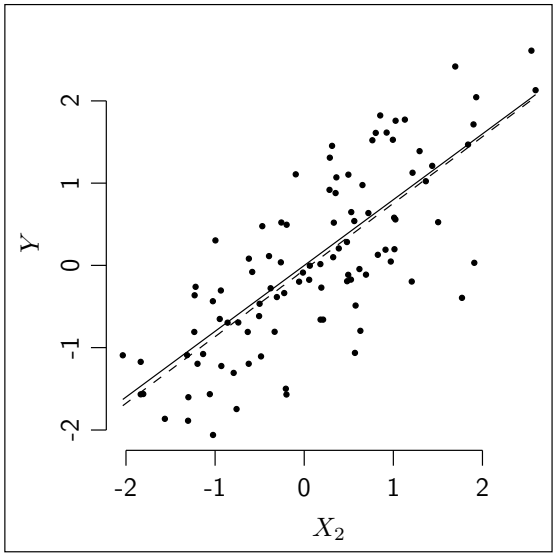


Figure 1.4: Marginal response plot for X_2 from Example 1.2, with idealized (solid line) and conditional (dashed line) mean functions added.

and intercept -0.05, is drawn with a dotted line. By visualizing vertical slices for various values of X_2, these conditional mean lines can be verified, and the variance confirmed to be constant over X_2. □

1.3 Local net-effect plots

While a marginal response plot like Figure 1.4 is useful in that it does show the relationship between Y and X_2, as the effects of X_1 have been ignored, it does not show that this relationship may depend on X_1. *Local net-effect plots*, introduced by Cook (1995), provide a general method for showing this dependence.

The idea is to consider the distribution of $(Y, X_2)|X_1 = x_1$ for the possible values of x_1. Since X_1 is held fixed, this shows how X_2 is related to Y after accounting for the effect of X_1, or, the "net-effect" of X_2 when $X_1 = x_1$. Again, this distribution is visualized with a scatterplot of $\{X_2, Y\}$, but this time only points where $X_1 = x_1$ are used. As in the case where $p = 1$, the conditional distribution can be inspected,

though this time it is the distributions of $Y|X_1 = x_1$ given $X_2|X_1 = x_1$ that are of interest. Then to understand the relationship between X_2 and Y given X_1, one can inspect both the form of this conditional distribution for various values of X_1, and how this distribution changes as X_1 changes.

In the multivariate context of (1.3), the conditional distributions are easily derived to be

$$Y|\left(X_1 = x_1, X_2\right) \sim N\left(\beta_1 x_1 + \beta_2 X_2, 1 - \frac{r^2 - 2\rho rs + s^2}{1 - \rho^2}\right), \tag{1.6}$$

where

$$\beta_1 = \frac{r - \rho s}{1 - \rho^2}, \text{ and } \beta_2 = \frac{s - \rho r}{1 - \rho^2}.$$

Like in the marginal response plot, the mean function

$$E(Y|X_1 = x_1, X_2) = c(x_1) + \beta_2 X_2 \tag{1.7}$$

is a linear function of X_2 and the variance function has no X_2 dependence for all values of x_1.

Thus the effect of X_2 on Y can look very different depending if X_1 is considered fixed or ignored. For every unit increase in X_2, Y increases s units when X_1 is ignored, but the increase is β_2 units when $X_1 = x_1$ is fixed. When $\rho r > s > 0$ or $\rho r < s < 0$, the effect when X_1 is fixed will even be in the opposite direction of the effect when X_1 is ignored.

This hidden variable bias is a well-known issue, especially when X_1 and X_2 are factors, where it is known as Simpson's paradox (Simpson, 1951). In fact, the local net-effect plot is a generalization of the interaction plot used with factors.

This approach has precedent and seems useful, but there are practical issues that arise when X_1 is continuous, for in practice a random sample from $(Y, X_2)|X_1 = x_1$

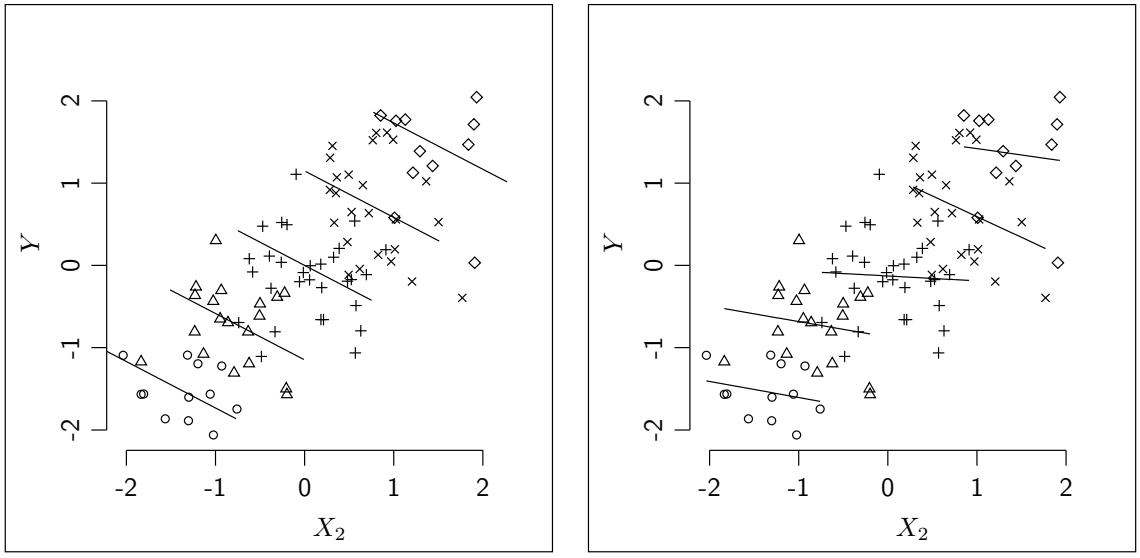
will not be accessible. Instead, apart from rounding, each x_1 value will be unique with probability one. One solution is to assume that the distribution of $(Y, X_2)|X_1$ is similar for nearby values of X_1, and so to group together points where $X_1 \in (a_i, a_{i+1}]$, where the values of $a_i, i = 1, ..., k$, create k slices on the recorded values of X_1. How many slices to choose, and how to choose the boundary points $\{a_i\}$ will depend on the particular data set and the usual bias/precision tradeoffs.

This process of slicing X_1 will still result in k distinct plots, of the k distributions of $(Y, X_2)|X_1 \in (a_i, a_{i+1}]$ for $i = 1, \ldots, k$. To better show how the distribution changes with X_1, these plots can be overlapped, using different symbols for each slice. Fitting a line to the points in each slice can help to visualize how the mean function changes with X_1. When the plots are not overlapped but instead placed side by side, this is also called a *conditioning plot*, or *co-plot* (Cleveland, 1993).

Example 1.3 (Local net-effect plots)

Using the multivariate normal data from Example 1.2, $s = 0.8$ and $\beta_2 = -0.56$, so the conditional mean lines should differ depending if X_1 is considered fixed or ignored. Since $\rho r = 0.95 \times 0.9 = 0.86 > 0.8 = s$, their slopes have opposite signs.

To demonstrate this, two net-effect plots for X_2 have been drawn in Figure 1.5. The points were sliced over X_1 into 5 slices, using $\{-2, -1.2, -0.4, 0.4, 1.2, 2\}$ as the boundaries, and marginal response plots drawn, with the points in each slice using a different symbol. Figure 1.5a shows the idealized conditional means, calculated for each slice by setting X_1 equal to the midpoint, and using the knowledge about the multivariate normal, so each has slope $\beta_2 = \frac{s-\rho r}{1-\rho^2} = -0.56$. Figure 1.5b shows the estimated conditional means for each slice, by fitting the data points in each slice with least squares. While the slope of these estimated lines vary from the idealized lines, both plots do demonstrate that the conditional means differ depending if X_1 is ignored or not. When the slices on X_1 are ignored, as in Figure 1.4, the slope is



(a) With idealized conditional means. (b) With estimated conditional means.

Figure 1.5: Local net-effect plots from Example 1.3, with idealized and conditional mean functions aded.

positive, but when they are taken into account, as they are here, the slope is negative. Additionally, the slope of the conditional mean function is independent of X_1, as it is the same for all slices. □

1.4 Combining local net-effect plots

While local net-effect plots in this form have the advantage of generally showing the distribution of $(Y, X_2)|X_1$, the result is a complex plot with several symbols and multiple levels of information, so it may be difficult to interpret. In many contexts, including the multivariate normal, the information in each slice is redundant, so the various local net-effect plots can be combined into one simpler plot.

In general, the goal when combining local net-effect plots is to create a new plot where the conditional distribution of the y-axis quantity, denoted by Y^*, given the

x-axis quantity, denoted by X^*, has no dependence on X_1, that is,

$$Y^*|(X^*, X_1) \sim Y^*|X^*. \tag{1.8}$$

A plot satisfying this condition, if available, will show something about the relationship between Y and X_2 because when (1.8) is true, the conditional relationship shown on the plot will be the same for all fixed values of the other variables. There is no information hidden in the plot that could be brought out by conditioning on X_1. The axes of the plot, X^* and Y^*, are allowed to depend on X_1, but this dependence is necessarily known in order to draw the plot, so the analysis can take this dependence into account. To look for plots of this type, define

$$Y^* = Y - g(X_1, X_2), \quad \text{and} \quad X^* = h(X_1, X_2) \tag{1.9}$$

and look for functions g and h that satisfy (1.8).

In the multivariate normal context, g and h can be limited to linear functions of the form

$$g(X_1, X_2) = aX_1 + bX_2$$
$$h(X_1, X_2) = cX_1 + dX_2$$

for appropriate constants a, b, c, and d because the relationships between X_1, X_2, and Y are all linear. Under this constraint, the distribution of $Y^*|X^*$will always be normal with constant variance, so to investigate the distribution of $Y^*|(X^*, X_1)$, only

the mean function need be investigated. Then because

$$\begin{aligned} E(Y|X_1, X_2) &= \beta_1 X_1 + \beta_2 X_2 \\ E\left(Y - (aX_1 + bX_2)|X_1, X_2\right) &= (\beta_1 - a)X_1 + (\beta_2 - b)X_2 \\ E(Y^*|cX_1 + dX_2, X_1) &= \tfrac{\beta_2 - b}{d}(cX_1 + dX_2) + \left(\beta_1 - a - c\left(\tfrac{\beta_2 - b}{d}\right)\right) X_1 \\ E(Y^*|X^*, X_1) &= \tfrac{\beta_2 - b}{d} X^* + \left(\beta_1 - a - \tfrac{c}{d}(\beta_2 - b)\right) X_1, \end{aligned}$$

any combination where

$$\beta_1 - a - \tfrac{c}{d}(\beta_2 - b) = 0 \tag{1.10}$$

will not have any conditional dependence on X_1.

There are infinitely many solutions to (1.10) and so infinitely many plots satisfy (1.8). Interpretability will be a guide to which plots to consider. Two helpful considerations will be the familiarity of the quantity on the x-axis and the choice of slope in the plot.

1.5 Component-plus-residual plot

One interpretable option for the x-axis is simply X_2, which is obtained when $c = 0$ and $d = 1$. Then by (1.10),

$$\beta_1 - a = 0 \times (\beta_2 - b) \tag{1.11}$$

so $a = \beta_1$ and b can be arbitrary. If the slope of the plot is restricted to be β_2, which is the slope of the mean function of the local net-effect plots for X_2 in the multivariate

normal context, then $b = 0$. The result is the plot of

$$\{X_2, Y - \beta_1 X_1\}. \tag{1.12}$$

This type of plot is known as a *component-plus-residual plot* or a *partial residual plot*, and has been explored in the literature extensively. First described by Ezekiel (1924), it was later rediscovered by Larsen and McCleary (1972), who called it a partial residual plot, and Wood (1973), who called it a component-plus-residual plot.

It is called a component-plus-residual plot because under the model $Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$,

$$Y - \beta_1 X_1 = \beta_2 X_2 + \epsilon, \tag{1.13}$$

so the plot of (1.12) can be written equivalently as

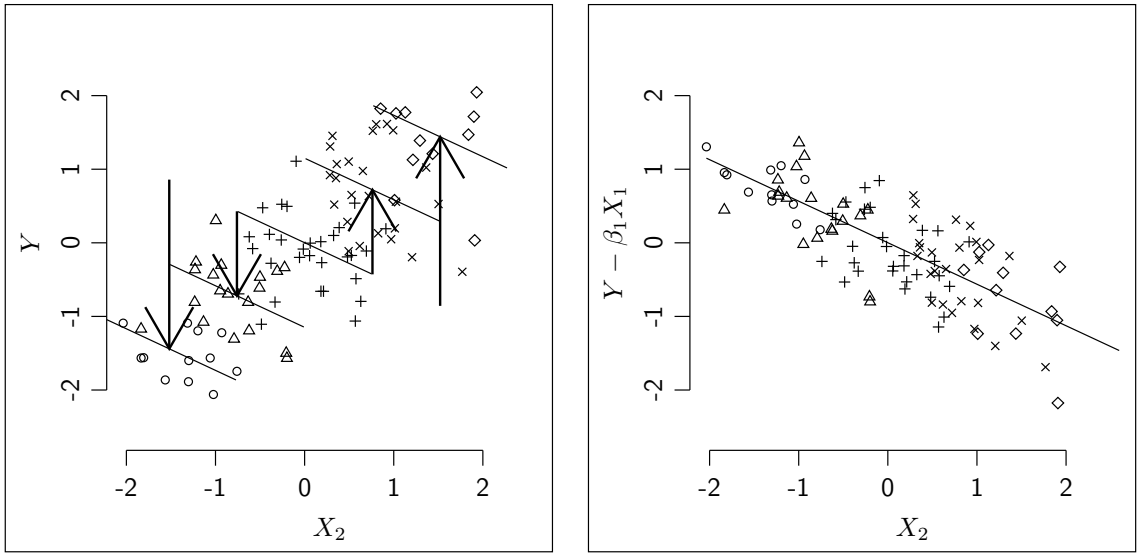
$$\{X_2, \beta_2 X_2 + \epsilon\}, \tag{1.14}$$

which has the X_2 component of the response due to X_2 plus the residuals on the y-axis. This will prove especially useful in additive models, where

$$Y = f_1(X_1) + f_2(X_2) + \epsilon, \tag{1.15}$$

because if f_1 can be estimated accurately, the form of f_2 can be plotted even if f_2 cannot be directly estimated.

These plots have been called partial residual plots because $Y - f_1(X_1)$ is called the partial residual for X_2. However, this name most often refers to plots where f_1 has been fitted linearly. Since a linear fit will not always result in a consistent estimate of $f_2(X_2) + \epsilon$, the term partial residual plot will be reserved for this specific case and



(a) Idealized local net-effect plots, with contributions from X_1, $\beta_1 X_1 = 1.44 X_1$, marked.

(b) Idealized component-plus-residual plot; the contributions from X_1 have been removed.

Figure 1.6: Graphical explanation of the component-plus-residual plot.

the term component-plus-residual plot will be used in general.

Graphically, the component-plus-residual plot combines the local net-effect plots by subtracting the part X_1 contributes to Y to form Y^*, while leaving the x-axis alone. Figure 1.6a shows the idealized local net-effect plot from Figure 1.5a, with arrows marking the effect due to X_1. In Figure 1.6b, these contributions have been removed. As a result, the conditional mean functions for each slice now line up perfectly, demonstrating that they no longer have X_1 dependence.

In practice β_1 is unknown and must be estimated. In the multivariate normal context, the ordinary least squares fit of Y on X_1 and X_2 yields consistent estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ for β_1 and β_2. Cook (1993) explored this plot in more complex situations, and showed that ordinary least squares consistently estimates β_1 only when the dependence of the response Y on the predictor X_2 given the other predictors is linear or when the relationship between the predictors is linear. That is, when either $E(Y|X_2)$

is a linear function of X_2 or $E(a'X|X)$ is linear in X for all a. Situations without this type of dependence will be explored more fully in Section 2.1.

1.5.1 Residual plot

Plots similar to the component-plus-residual plot can be drawn if the slope of the new plot is not restricted to be β_2, and b is allowed to take on other values. All of these plots will differ from the component-plus-residual plot only in the slope, as the only difference is that a multiple of the x-axis (X_2) is being subtracted from the y-axis. The slope for these plots will be $\beta_2 - b$.

One case of interest is $b = \beta_2$, which plots

$$\{X_2, Y - (\beta_1 X_1 + \beta_2 X_2)\}; \tag{1.16}$$

the residuals from the full model against X_2. As shown by Mansfield and Conerly (1987), this plot is a detrended component-plus-residual plot, so except for the slope, it contains the same information. However, Mansfield and Conerly (1987) show that despite not showing the slope, it can make certain non-linearities more apparent by increasing resolution in the plot by detrending.

1.5.2 Marginal-plus-residual plot

A second case that is apparently new and may be of interest is $b = -\rho\beta_1$, which plots

$$\{X_2, Y - (\beta_1(X_1 - \rho X_2))\}, \tag{1.17}$$

where $\rho X_2 = E(X_1|X_2)$, and shows the marginal relationship between Y and X_2, but with the residuals from the full model, so may be called a *marginal-plus-residual* plot.

The slope given (1.3) is

$$\begin{aligned}\beta_2 + \beta_1\rho &= \frac{s-\rho r}{1-\rho^2} + \rho\frac{r-\rho s}{1-\rho^2} \\ &= \frac{s-\rho r+\rho r-\rho^2 s}{1-\rho^2} \\ &= s,\end{aligned}$$

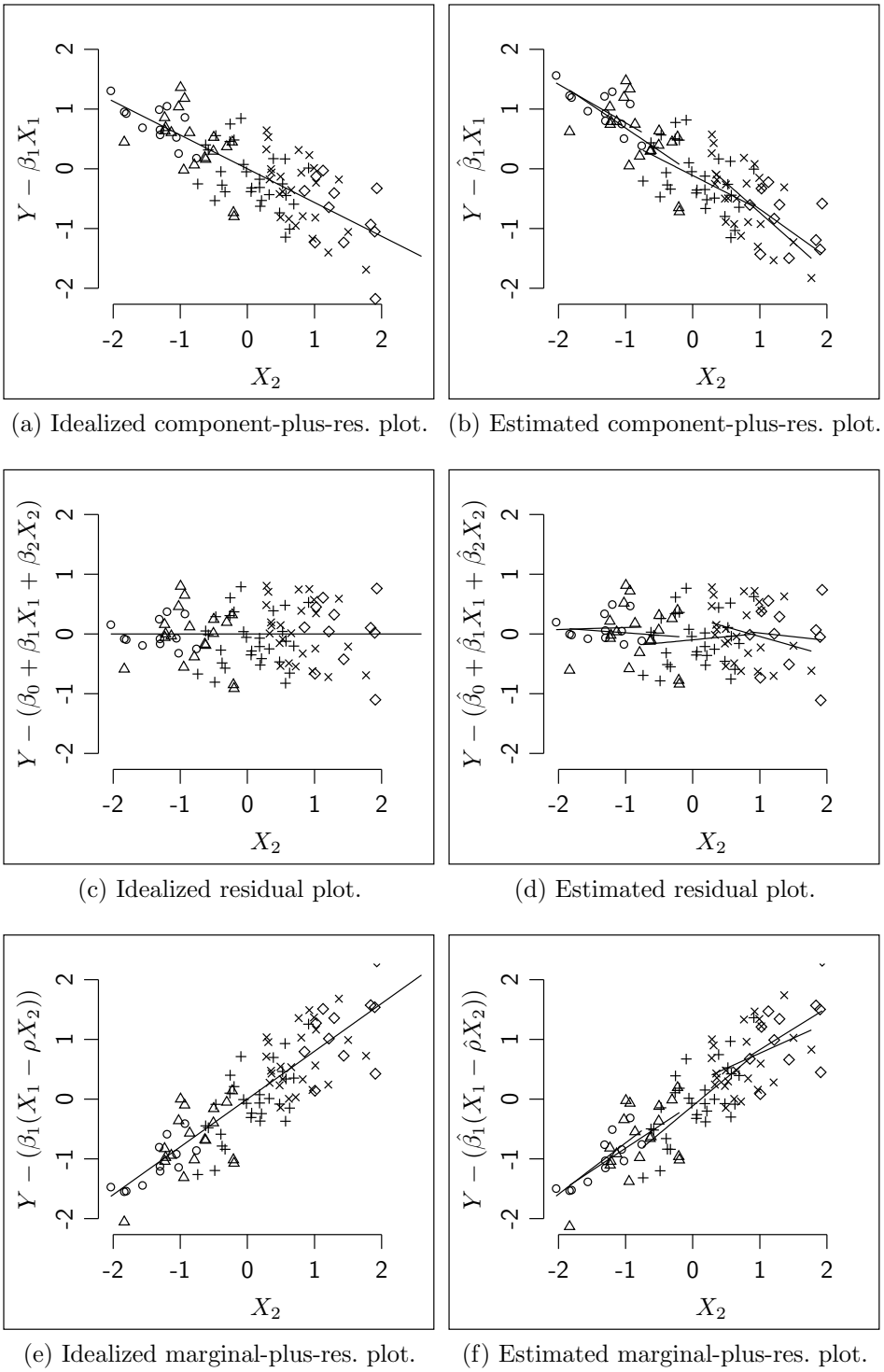
just as in the marginal response plot. The difference is that contributions to the response that are orthogonal to X_2 have been removed, as $X_1 - \rho X_2$ is the part of X_1 that is orthogonal to X_2. Removing these contributions can make the marginal mean relationship easier to see.

An alternate way of estimating this plot is to use the change in fitted values between a model including only X_2 and a model including both X_1 and X_2, as

$$\begin{aligned}E(Y|X_1,X_2) - E(Y|X_2) &= E(Y|X_1,X_2) - E_{X_1|X_2}(E(Y|X_1,X_2)) \\ &= (\beta_1 X_1 + \beta_2 X_2) - E_{X_1|X_2}(\beta_1 X_1 + \beta_2 X_2) \\ &= (\beta_1 X_1 + \beta_2 X_2) - (\beta_1 E(X_1|X_2) + \beta_2 X_2) \\ &= \beta_1(X_1 - \rho X_2).\end{aligned}$$

Example 1.4 (Component-plus-residual plots and related plots.)
Using the data from Example 1.2, idealized and estimated versions of the component-plus-residual plot, the residual plot, and the marginal-plus-residual plot, are each plotted in Figure 1.7. Using ordinary least squares, $\hat{\beta}_1 = 1.59$, $\hat{\beta}_2 = -0.68$, and the sample correlation is $\hat{\rho} = 0.94$.

To demonstrate that the conditional X_1 dependence really has been removed, each of the plots have been sliced over X_1, with the points in each slice drawn using a different symbol, and a fitted line drawn to each slice. In the idealized plots, these



(a) Idealized component-plus-res. plot. (b) Estimated component-plus-res. plot.

(c) Idealized residual plot. (d) Estimated residual plot.

(e) Idealized marginal-plus-res. plot. (f) Estimated marginal-plus-res. plot.

Figure 1.7: The three plots with X_2 on the x-axis, using data from Example 1.4.

lines exactly overlap because the slope is known. Although they are slightly different in the estimated plots because of random variation, they do share the same conditional mean.

While the conditional (vertical) dependence on X_1 has been removed, there is still dependence horizontally; that is, the x-axis (X_2) still depends on X_1. In all plots, the circles are associated with smaller X_2 values, and the diamonds with larger X_2 values. Since points that have the same symbols have similar X_1 values, this shows that X_2 and X_1 are associated. Additionally, the fitted lines in the estimated plots also demonstrate this dependence because each line only spans a specific section of the range of X_2. While this type of dependence is not required to be removed by condition (1.8), Section 1.6 will show that added-variable plots additionally remove this dependence.

The component-plus-residual plots have negative slopes, showing that the relationship between Y and X_2 is negative when X_1 is accounted for. In contrast, the marginal-plus-residual plots have positive slope, showing that marginally, this relationship is positive, again showing the hidden variable bias contained in this data.

Finally, as all three plots are the same except for the slope, the residual plots are detrended versions of the other two types of plots, and so allow certain details to be seen more clearly. For example, in the estimated plots, the difference in the slopes of the lines fitted for each slice is much easier to see in the residual plot than in the other two. □

1.6 Added-variable plot

Another easily interpretable variable to put on the x-axis is the part of X_2 that is orthogonal to X_1, or the new information in X_2. In the multivariate normal context, this is $X_2 - \rho X_1$, so $c = -\rho$ and $d = 1$. Because the x-axis is independent of X_1,

in these plots not only is the conditional distribution of the y-axis given the x-axis independent of X_1 as in (1.8), but the joint distribution will also be independent of X_1, that is,

$$(X^*, Y^*)|X_1 \sim (X^*, Y^*). \tag{1.18}$$

To remove X_1 dependence from these plots, a and b are chosen such that

$$\beta_1 - a = \rho(\beta_2 - b), \tag{1.19}$$

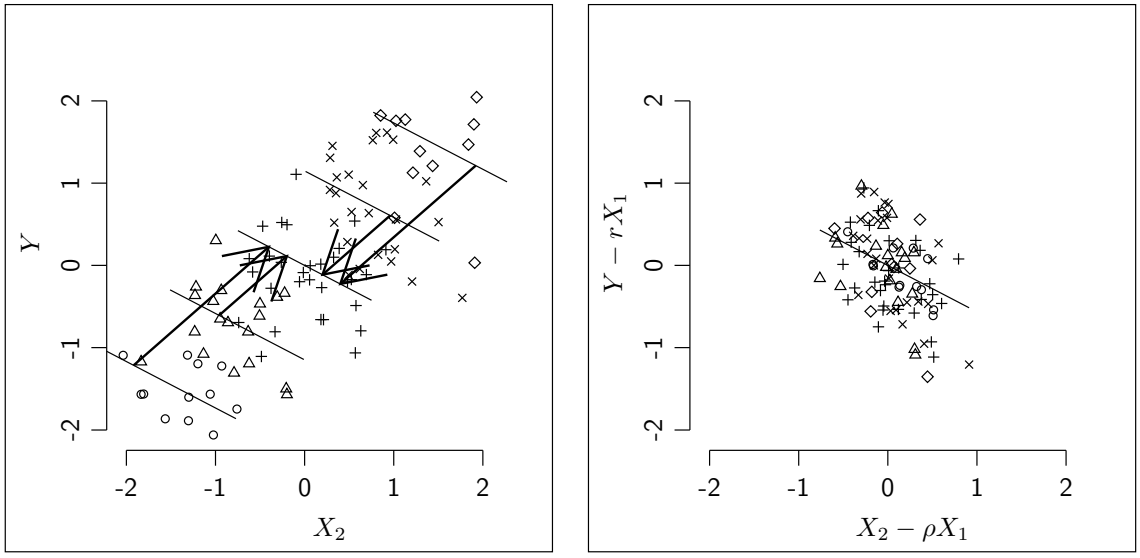
which again has multiple solutions.

First, the slope is restricted to β_2, so $b = 0$ and $a = \beta_1 + \rho\beta_2 = r$. The resulting plot is of

$$\{X_2 - \rho X_1, Y - rX_1\}. \tag{1.20}$$

Since $rX_1 = E(Y|X_1)$ and $\rho X_1 = E(X_2|X_1)$, this plot can be interpreted as showing the new information in X_2 that contributes to understanding the part of Y unexplained by X_1, or more simply, how adding the variable X_2 to the model improves the fit. Accordingly, Cook and Weisberg (1982) called it an *added-variable plot*. This type of plot first appeared in Cox (1958), and has been described by Mosteller and Tukey (1977), Belsley et al. (1980), Draper and Smith (1981), Atkinson (1985), Weisberg (1985), and others.

Graphically, the added-variable plot combines the local net-effect plots by centering around their conditional means, on both the x and the y axes, as seen in Figure 1.8, using the data from Example 1.2. This works because the structure of the distribution of $(X_2, Y)|X_1$ is identical for all values of X_1, and varies only in the mean. So an added-variable plot can combine these distributions by overlaying them



(a) Idealized local net-effect plots, with arrows showing how to center each slice around its mean.

(b) Idealized added-variable plot; the net-effect plots with each slice centered.

Figure 1.8: Graphical explanation of the added-variable plot.

on top of each other after centering without losing any information. This provides better information about the conditional distribution because the existing data is now pooled together.

1.6.1 Residual plot

As in the component-plus-residual plot, allowing different values for the slope results in additional useful plots. First, the plot can be detrended, as Cook and Weisberg (1991) suggest, to have zero slope by using $a = \beta_1$ and $b = \beta_2$, resulting in the plot of

$$\{X_2 - \rho X_1, Y - (\beta_1 X_1 + \beta_2 X_2)\}, \tag{1.21}$$

which puts the residuals of the full model on the y-axis.

1.6.2 General added-variable plot

Another useful plot is a general added-variable plot. Rescaling the x-axis by multiplying by β_2, the slope becomes 1, and $c = -\beta_2\rho$ and $d = \beta_2$, so the plot is

$$\{\beta_2(X_2 - \rho X_1), Y - rX_1\} \tag{1.22}$$

Then the x-axis is equal to the change in the fitted values between the full model and the model only including X_1:

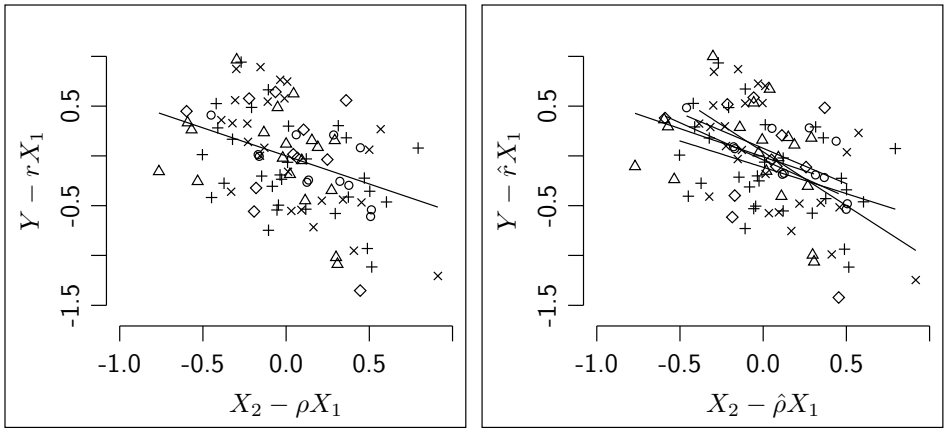
$$\begin{aligned} E(Y|X_1, X_2) - E(Y|X_1) &= (\beta_1 X_1 + \beta_2 X_2) - rX_1 \\ &= (\beta_1 - r)X_1 + \beta_2 X_2 \\ &= -\rho\beta_2 X_1 + \beta_2 X_2 \\ &= \beta_2(X_2 - \rho X_1). \end{aligned}$$

This plot has been discussed by Cook and Weisberg (1991), and is especially useful for seeing the effect of adding more than one variable at once, as it results in a two-dimensional plot even when $q > 1$.

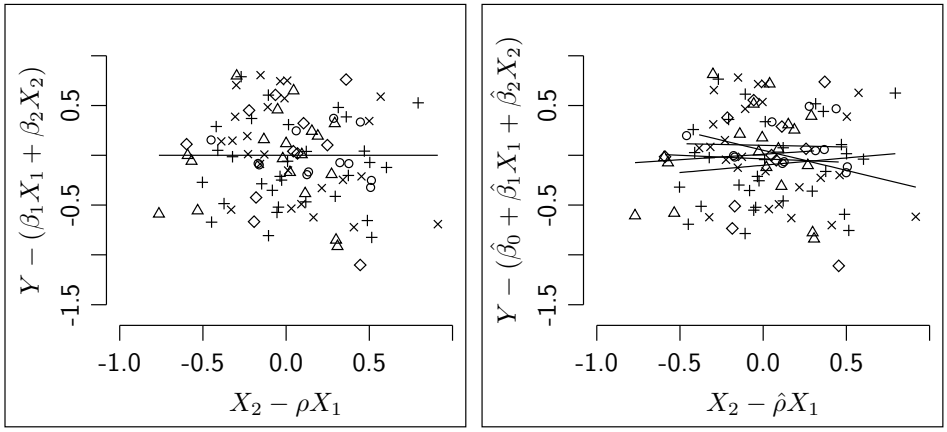
Example 1.5

Using the data from Example 1.2, idealized and estimated versions of the added-variable plot, the residual plot, and the general added-variable plot are plotted in Figure 1.9. Using ordinary least squares, $\hat{r} = 0.94$; the other estimated values are as in Example 1.4.

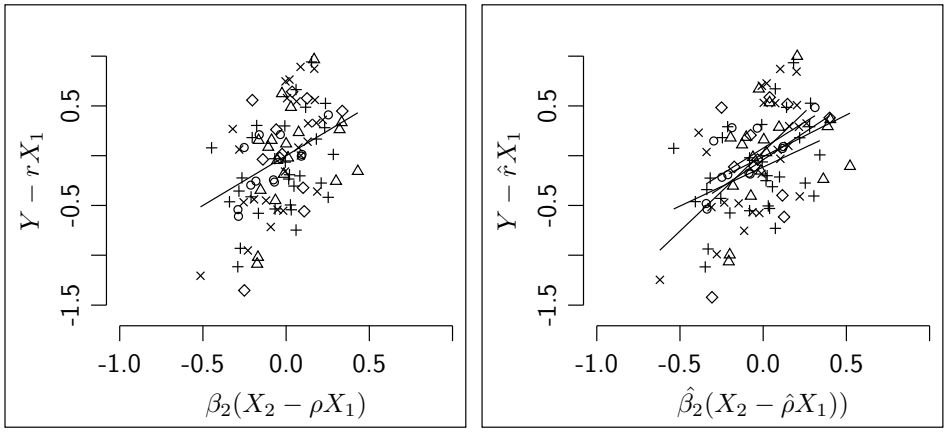
To demonstrate that the X_1 dependence really has been removed, each of the plots have been sliced over X_1, with the points in each slice drawn using a different symbol, and a fitted line drawn to each slice. In the idealized plots, these lines exactly overlap because the slope is known. Although they are slightly different in the estimated plots



(a) Idealized added-variable plot. (b) Estimated added-variable plot.



(c) Idealized residual plot. (d) Estimated residual plot.



(e) Idealized general added-var. plot. (f) Estimated general added-varplot.

Figure 1.9: The three types of added-variable plots, using data from Example 1.5.

because of random variation, they do share the same conditional mean. In contrast to the component-plus-residual plots, in addition to removing the conditional (vertical) dependence on X_1, these added-variable plots remove horizontal dependence; that is, the x-axis no longer depends on X_1. In all plots, the points with similar symbols are scattered evenly from left to right. Since points that have the same symbols have similar X_1 values, this shows that the x-axis is not associated with X_1. The fitted lines in the estimated plots also demonstrate this independence because each line spans the entire range of X_2. This horizontal independence is not required by condition (1.8), but is an additional feature of the added-variable plot.

In the estimated versions of the plots the points for each X_1 slice are also exactly overlaid on each other, as the fitted lines for each slice are similar both vertically and horizontally. It is possible to overlay these points without losing information about the distribution because the distribution for each slice is identical except for its location.

Finally, the only difference between the three plots is in the slopes. The standard added-variable plot has a slope of $\beta_2 = -0.56$; meaning that when X_2 is larger than X_1 would predict, Y is less than X_1 would predict. The slope of the other plots is defined by their construction; the residual plot has slope zero, and the general added-variable plot has slope one. □

1.7 ARES plot

So far, the plots discussed in this section have been simple two-dimensional plots exploring what a given model says about how a predictor is related to the response. The *ARES plot*, developed by Cook and Weisberg (1989), is different in two ways; first, instead of showing how a predictor is related to the response, it attempts to show the general importance of a predictor in the given model, and secondly, it does

so using an animated plot.

Consider the linear model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon. \tag{1.23}$$

The ARES plot shows how the plot of the fitted values against the residuals changes as the predictor X_2 is added. The idea is to start with the plot for the model $Y = \beta_1 X_1 + \epsilon$, and smoothly change to the plot for the model $Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$. By doing this, one can see both how the fit of the model improves when X_2 is added to the model.

Because this plot only uses fitted values and residuals, changing the parameterization of the predictors will not change the plot. So let us consider only adding the part of X_2 that is orthogonal to X_1, normalized to unit length:

$$\widetilde{X_2} = Q_1 X_2 / ||Q_1 X_2||,$$

so the model is

$$Y = Z\beta^* + \epsilon = X_1\beta_1^* + \widetilde{X_2}\beta_2^* + \epsilon. \tag{1.24}$$

The smooth transition is obtained by calculating the fitted values and residuals for $0 \leq \lambda \leq 1$,

$$\hat{\beta}_\lambda = \left(Z'Z + \frac{1-\lambda}{\lambda} bb' \right)^{-1} Z'Y \tag{1.25}$$

where b is a vector of length p, of all zeros except for a single 1 corresponding to X_2. Then because of the orthogonality of X_1 and $\widetilde{X_2}$, the fitted values $\widehat{Y}(\lambda) = Z\hat{\beta}_\lambda$ and residuals $e(\lambda) = Y - \widehat{Y}(\lambda)$ are simply

$$\widehat{Y}(\lambda) = \widehat{Y}_1 + \lambda(\widehat{Y} - \widehat{Y}_1) \tag{1.26}$$

and

$$e(\lambda) = e_{y|1} - \lambda(\widehat{Y} - \widehat{Y}_1), \tag{1.27}$$

where $\widehat{Y}_1$ and $e_{y|1}$ are the fitted values and residuals from the fit of Y on only X_1.

Example 1.6

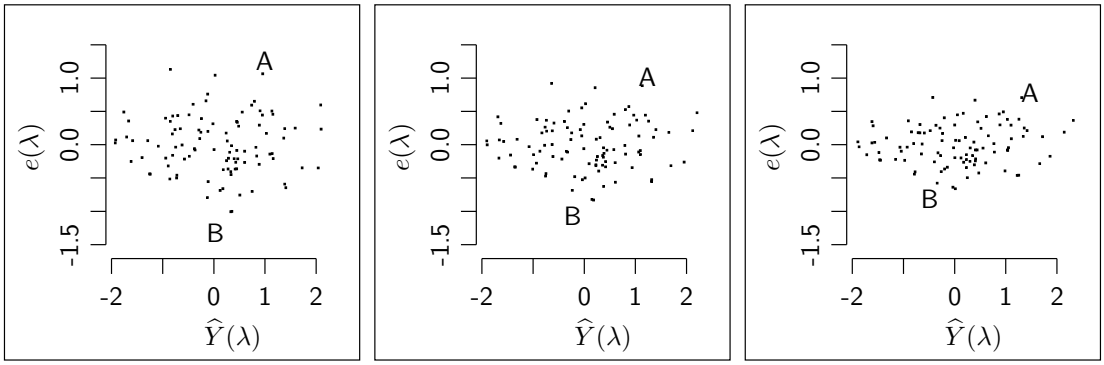
A sample of size 100 has been taken from the multivariate normal of (1.3), with $r = 0.9$, $s = 0.8$ and $\rho = 0.95$. Figure 1.10 shows six frames of an ARES plot for this data, as λ moves from 0 to 1. The residuals reduce significantly as λ increases, so X_2 is having an effect in the model. Also, the two points have the largest residuals when $\lambda = 0$ have been marked with A and B. As λ increases and X_2 is gradually added to the model, both the fitted values and the residuals for these points change, resulting in a better fit. For point A, the initial residual is positive, and as X_2 is added, the fitted value becomes larger, making the residual smaller. The opposite happens for point B. □

1.7.1 Connection to added-variable plot

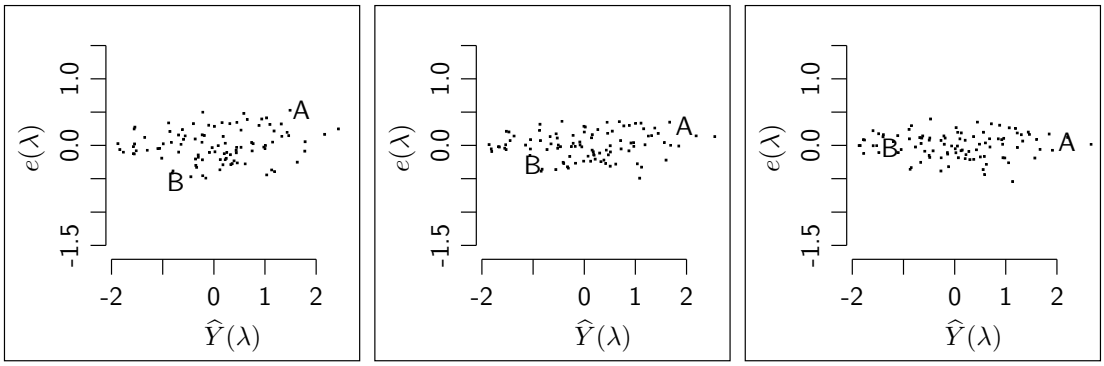
The ARES plot is connected to the general added-variable plot and the detrended general added-variable plot; the y axis of the ARES plot goes between the residuals from the full model and the residuals from the model without X_2, which are the y-axes for these two added-variable plots. Additionally, the x-axis for the general added-variable plot is the change in fitted values between these two models, which is the distance that the points travel in the x-direction in the ARES plot.

Example 1.7

Using the data from Example 1.6, Figure 1.11a shows the path of the points in the ARES plot. The end of the paths farther from zero are the residuals from the

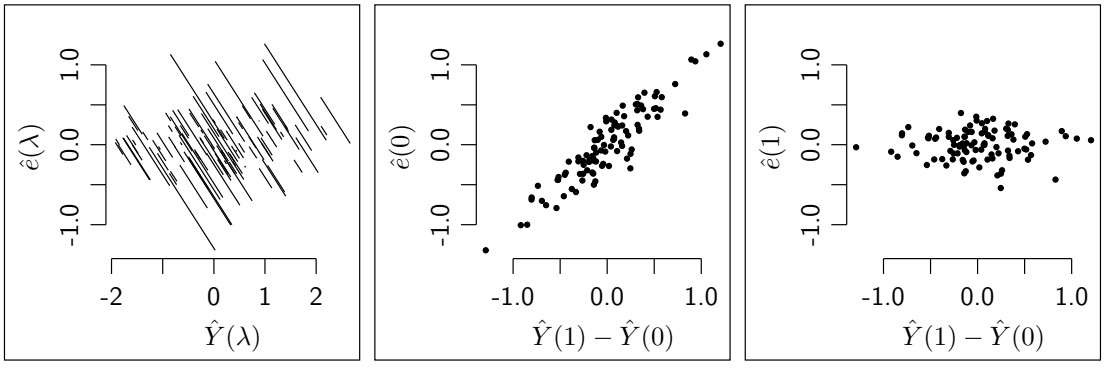


(a) $\lambda = 0$ (b) $\lambda = 1/5$ (c) $\lambda = 2/5$



(d) $\lambda = 3/5$ (e) $\lambda = 4/5$ (f) $\lambda = 1$

Figure 1.10: ARES plots for multivariate normal data, when adding X_2.



(a) The path of the points in the ARES plot.

(b) General added variable plot.

(c) Detrended general added variable plot.

Figure 1.11: Connection between the ARES plot and the added-variable plot.

submodel, and are equal to the y-values of the general added-variable plot of Figure 1.11b. The end of the paths closer to zero are the residuals from the full model, and are equal to the y-values in the detrended general added-variable plot of Figure 1.11c. The length of each line, which is the distance each point in the ARES plot travels, is proportional to its horizontal shift, which is equal to the x-axis in both added-variable plots. □

1.7.2 Derivation of ARES plot

The formulas for the ARES plot can be derived by gradually replacing the part of the desired variable orthogonal to the other predictors with random noise that is orthogonal to all predictors and the response. Only this part need be replaced because whatever can be explained by the remainder of the variable is also explained by the other variables, so when it is fully replaced, the chosen variable has no added predictive power.

Again, let the predictor of interest be X_2, and let the other predictors be X_1, and let $\tilde{X}_2 = Q_1 X_2$ be the part of X_2 that is orthogonal to X_1. Then let E be a vector such that $E'E = \tilde{X}_2{}'\tilde{X}_2$ and $E \perp\!\!\!\perp (X_1, \tilde{X}_2, Y, e)$, where $e = Q_X Y$. Without loss of generality, assume all vectors have mean zero.

To gradually remove $\tilde{X}_2$ from the model, let

$$X_2^* = w_1 \tilde{X}_2 + w_2 E,$$

where $w_1^2 + w_2^2 = 1$ so that $X_2^{*\prime} X_2^* = \tilde{X}_2{}'\tilde{X}_2$. Projecting Y onto $X^* = (X_1, X_2^*)$, by

orthogonality

$$\begin{aligned}
P_{X^*}Y &= P_{X_1}Y + P_{X_2^*}Y \\
&= P_{X_1}Y + X_2^*(X_2^{*\prime}X_2^*)^{-1}X_2^{*\prime}Y \\
&= P_{X_1}Y + (w_1\tilde{X}_2 + w_2E)(\tilde{X_2}'\tilde{X}_2)^{-1}(w_1\tilde{X}_2 + w_2E)'Y \\
&= P_{X_1}Y + w_1^2\tilde{X}_2(\tilde{X_2}'\tilde{X}_2)^{-1}\tilde{X_2}'Y + w_1w_2E(\tilde{X_2}'\tilde{X}_2)^{-1}\tilde{X_2}'Y \\
&= P_{X_1}Y + w_1^2P_{\tilde{X}_2}Y + w_1w_2E(\tilde{X_2}'\tilde{X}_2)^{-1}\tilde{X_2}'Y.
\end{aligned}$$

The final term is in the direction of E, which is assumed to be orthogonal to Y, so it is simply adding noise to the prediction. Removing it,

$$\hat{Y}^* = P_{X_1}Y + w_1^2P_{\tilde{X_2}}Y,$$

where the hat notation is used because this is no longer a projection, but only an estimate of Y. Letting $\lambda = w_1^2$, this is the result in (2.3) of Cook and Weisberg (1989), as

$$\begin{aligned}
\hat{Y}(\lambda) &= P_{X_1}Y + \lambda P_{\tilde{X_2}}Y \\
&= P_{X_1}Y + \lambda(P_XY - P_{X_1}Y) \\
&= \hat{Y}_1 + \lambda(\hat{Y} - \hat{Y}_1).
\end{aligned}$$

The coefficient vector $\hat{\beta}_\lambda$ for a given value of $\lambda = w_1^2$ can be then derived to be

$$\hat{\beta}_\lambda = \left(Z'Z + \frac{1-\lambda}{\lambda}bb'\right)Z'Y,$$

where $Z = \begin{pmatrix} X_1 & \tilde{X_2} \end{pmatrix}$. By orthogonality of X_1 and $\tilde{X}_2$,

$$\begin{aligned}
Z\hat{\beta}^* &= \hat{Y}^* \\
&= P_{X_1}Y + \lambda P_{\tilde{X_2}}Y \\
&= X_1(X_1'X_1)^{-1}X_1'Y + \lambda\tilde{X_2}(\tilde{X_2}'\tilde{X_2})^{-1}\tilde{X_2}'Y \\
&= \begin{pmatrix} X_1 & \tilde{X_2} \end{pmatrix} \begin{pmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & \lambda(\tilde{X_2}'\tilde{X_2})^{-1} \end{pmatrix} \begin{pmatrix} X_1 & \tilde{X_2} \end{pmatrix}' Y \\
&= \begin{pmatrix} X_1 & \tilde{X_2} \end{pmatrix} \begin{pmatrix} X_1'X_1 & 0 \\ 0 & \left(1 + \frac{1-\lambda}{\lambda}\right)\tilde{X_2}'\tilde{X_2} \end{pmatrix}^{-1} \begin{pmatrix} X_1 & \tilde{X_2} \end{pmatrix}' Y \\
&= Z\left(Z'Z + \frac{1-\lambda}{\lambda}bb'\right)Z'Y
\end{aligned}$$

where b is a vector of zeros except for a single 1 corresponding to $\tilde{X}_2$.

The ARES plot shows the residuals against the fitted values, dynamically changing λ between 0 and 1. When λ is 0, the fit is totally without X_2 and when λ is 1, the fit is with all variables.

It may also be of interest to note how the R^2 value changes as λ changes. Writing $R^2(\lambda) = ||\hat{Y}^*||^2/||Y||^2$, the numerator can be written as

$$\begin{aligned}
||\hat{Y}^*||^2 &= ||P_{X_1}Y + \lambda P_{\tilde{X_2}}Y||^2 \\
&= ||P_{X_1}Y||^2 + \lambda^2||P_{\tilde{X_2}}Y||^2 \\
&= ||P_{X_1}Y||^2 + \lambda^2\left(||P_ZY||^2 - ||P_{X_1}Y||^2\right)
\end{aligned}$$

so dividing by $||Y||^2$,

$$R^2(\lambda) = R_1^2 + \lambda^2(R^2 - R_1^2)$$

so the change in R^2 is linear in λ^2.

If instead the R^2 value is calculated using the projection onto X^*, so that it includes the term in the E direction, there is an added term in the numerator of

$$\begin{aligned}||w_1 w_2 E(\tilde{X_2}'\tilde{X_2})^{-1}\tilde{X_2}'Y||^2 &= w_1^2 w_2^2 \left(Y'\tilde{X_2}(\tilde{X_2}'\tilde{X_2})^{-1}E'E(\tilde{X_2}'\tilde{X_2})^{-1}\tilde{X_2}'Y\right) \\ &= \lambda(1-\lambda)\left(Y'\tilde{X_2}(\tilde{X_2}'\tilde{X_2})^{-1}\tilde{X_2}'Y\right) \\ &= (\lambda-\lambda^2)||P_{\tilde{X_2}}Y||^2\end{aligned}$$

so the R^2 for a given value of λ is instead

$$\begin{aligned}R^2(\lambda)^* &= R_1^2 + \lambda^2(R^2 - R_1^2) + (\lambda - \lambda^2)(R^2 - R_1^2) \\ &= R_1^2 + \lambda(R^2 - R_1^2)\end{aligned}$$

and so is linear in λ, instead of λ^2.

1.7.3 Derivation without orthogonality

It is also of interest to see how this changes when the entire variable is replaced with random noise, not just the orthogonal part. Let $X_2^* = w_1 X_2 + w_2 E$. Now projecting Y onto $X^* = (X_1, X_2^*)$ is more complex because X_1 and X_2^* are not necessarily orthogonal. Projecting Y onto X^* by using the orthogonal basis of X_1 and $Q_{X_1}X_2^*$,

$$P_{X^*}Y = P_{X_1}Y + P_{Q_{X_1}X_2^*}Y.$$

When $w_1 = 1$,

$$P_X Y = P_{X_1}Y + P_{Q_{X_1}X_2}Y,$$

so only the second term will be investigated, with the goal of putting $P_{X^*}Y$ in terms of P_XY. First

$$\begin{aligned} P_{Q_{X_1}X_2^*}Y &= \frac{(Q_{X_1}X_2^*)(Q_{X_1}X_2^*)'Y}{||Q_{X_1}X_2^*||^2} \\ &= \frac{(w_1Q_{X_1}X_2 + w_2E)\,(w_1Q_{X_1}X_2 + w_2E)'\,Y}{||Q_{X_1}X_2^*||^2} \\ &= \frac{w_1^2(Q_{X_1}X_2)(Q_{X_1}X_2)'Y + w_1w_2E(Q_{X_1}X_2'Y)}{||Q_{X_1}X_2^*||^2}. \end{aligned}$$

Additionally, as

$$\begin{aligned} Q_{X_1}X_2^* &= X_2^* - P_{X_1}X_2^* \\ &= w_1X_1 + w_2E - P_{X_1}(w_1X_1 + w_2E) \\ &= w_1X_1 + w_2E - P_{X_1}(w_1X_1) \\ &= w_1Q_{X_1}X_2 + w_2E \end{aligned}$$

the denominator can be written as

$$\begin{aligned} ||Q_{X_1}X_2^*||^2 &= ||w_1Q_{X_1}X_2 + w_2E||^2 \\ &= w_1^2||Q_{X_1}X_2||^2 + w_2^2||E||^2 \\ &= ||Q_{X_1}X_2||^2\left(w_1^2 + w_2^2\frac{||X_2||^2}{||Q_{X_1}X_2||^2}\right) \\ &= ||Q_{X_1}X_2||^2\left(w_1^2 + w_2^2\frac{||X_2||^2}{||X_2||^2 - ||P_{X_1}X_2||^2}\right) \\ &= ||Q_{X_1}X_2||^2\left(\lambda + (1-\lambda)\left(\frac{1}{1-R_X^2}\right)\right) \end{aligned}$$

where $R_X^2 = ||P_{X_1}X_2||^2/||X_2||^2$.

Then removing the term in the E direction from $P_{Q_{X_1}X_2^*}Y$ as it is known to be orthogonal to Y,

$$\begin{aligned}\hat{Y}^* &= P_{X_1}Y + \frac{w_1^2(Q_{X_1}X_2)(Q_{X_1}X_2)'Y}{||Q_{X_1}X_2||^2\left(w_1^2 + w_2^2\left(\frac{1}{1-R_X^2}\right)\right)} \\ &= P_{X_1}Y + \frac{\lambda}{\lambda + (1-\lambda)\left(\frac{1}{1-R_X^2}\right)}P_{Q_{X_1}X_2}Y. \\ &= \hat{Y}_1 + g(\lambda)\left(\hat{Y} - \hat{Y}_1\right)\end{aligned}$$

where

$$g(\lambda) = \frac{\lambda}{\lambda + (1-\lambda)\left(\frac{1}{1-R_X^2}\right)}.$$

For $\lambda = 0$, it is as if X_2 is not in the model at all and for $\lambda = 1$, it as if X_2 is fully included. However, the change is not linear in λ but is instead curved. The change in R^2 is nonlinear in the same way (except squared),

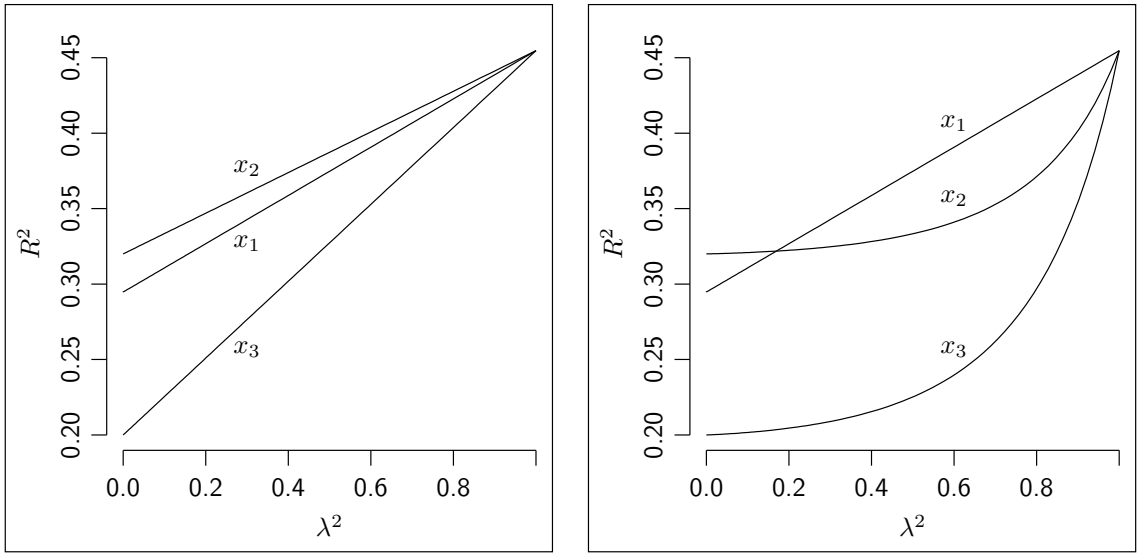
$$R^2(\lambda)^* = R_1^2 + (g(\lambda))^2\left(R^2 - R_1^2\right).$$

So even if fully removing two variables results in the same change in R^2, adding the same amount of noise to each may result in different changes in R^2.

Example 1.8

For this example, consider three predictors and a single response that are multivariate normal with covariance matrix

$$\text{Cov}\begin{pmatrix} y \\ x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1.00 & 0.40 & 0.20 & 0.40 \\ 0.40 & 1.00 & 0.00 & 0.00 \\ 0.20 & 0.00 & 1.00 & 0.90 \\ 0.40 & 0.00 & 0.90 & 1.00 \end{pmatrix}.$$



(a) When removing only the orthogonal part.

(b) When removing the entire variable.

Figure 1.12: The change in R^2 when removing each of the three variables of Example 1.8 gradually.

The changes in R^2 as λ changes were calculated directly from this covariance matrix, and plotted in Figure 1.12, both when only the orthogonal part of the variable was replaced with noise and when the entire variable was replaced with noise.

The variable x_1 is completely orthogonal to x_2 and x_3, so the change for x_1 will be linear in λ^2 using both methods. However, x_2 and x_3 are not orthogonal, so there is instead a non-linear relationship, as shown in Figure 1.12. Additionally, the values in the covariance matrix have been chosen so that when only part of the entire variable is removed, the change in R^2 for x_2 is larger than the change for x_1, even though when the entire variable is removed, the change for x_1 is larger. That is, the lines for x_1 and x_2 cross. □

1.8 Marginal model checking plot

To model the general regression context, where the object of study is the conditional distribution of a response Y given the value of predictors X, let

$$f_T(y|x)$$

represent the true, but unknown, conditional density of $Y|X$, and let

$$\hat{f}_M(y|x)$$

be the estimate of f_T from the sample data when assuming a model M is true. If the method of finding this estimate is Fisher-consistent, let

$$\hat{f}_M(y|x) \to f_M(y|x).$$

Then if M is true, then the estimated density approaches the true density as the sample size increases, and $f_M = f_T$.

One purpose of plots in this situation is diagnostic, with the goal of discovering if $M = T$. There are many different kinds of plots that do this, including residual plots, quantile plots, and many others. Which to use often depends on the type of model. One particular method, the *marginal model checking plot*, is a general method for testing if $M = T$. It will be discussed in detail because it can also aid in understanding the relationship between a given predictor X_2 and the response Y as estimated under model M, and whether that estimated relationship matches the data.

This plot was developed by Cook and Weisberg (1997) with the goal of providing plots that check the adequacy of a model M by comparing the estimated distribution under the assumption that M is true, $\hat{f}_M(y|x)$, with the true but unknown distribution, $f_T(y|x)$. If the distributions are similar, the model can then be declared adequate. Pardoe (2001) explored Bayesian methods of formally testing whether or not the distributions are similar.

A straightforward comparison of the conditional distributions $\hat{f}_M(y|x)$ and $f_T(y|x)$ is difficult because they are multivariate functions that depend on $p+1$ variables, p from the p predictors in x, and one from the response y. The comparison is greatly simplified by reducing the distributions to univariate functions so that they can easily be compared on two-dimensional plots. As Cook and Weisberg (1997) show, for two conditional distributions $F(y|x)$ and $G(y|x)$, $F(y|x) = G(y|x)$ for all x in the sample space Ω_x if and only if $F(y|a'x) = G(y|a'x)$ for all values of $a'x \in \{a'x|a \in \mathcal{R}^p, ||a|| = 1, x \in \Omega_x\}$. So instead of comparing the multivariate functions $\hat{f}_M(y|x)$ and $f_T(y|x)$, the bivariate functions $\hat{f}_M(y|a'x)$ and $f_T(y|a'x)$ can be compared over several values of $a'x$. Checking more combinations of $a'x$ provides more support for the model, as the model holds only if all marginal models are true. Cook and Weisberg (1997) suggest several choices for a, including setting $a'x$ equal to each predictor in turn.

This reduces the distributions to functions that depend on two variables; the response y and a linear combination of the predictors $a'x$. To visualize in two dimensional plots, certain functions of these distributions must be chosen to be compared. Usually the mean, $E(y|a'x)$, and the variance, $\text{Var}(y|a'x)$, are chosen, although quantile regression (see Koenker, 2005) provides the possibility of comparing the distributions in other ways. For f_T, these functions can be estimated by smoothing plots of $\{a'x, y\}$, as these plot shows the conditional distributions of $y|a'x$.

For $\hat{f}_M$, applying the rules of iterated expectations also leads to smoothing, as will be demonstrated in Section 1.8.1. Smoothing is now done on plots of $\{a'x, \hat{y}\}$, where $\hat{y}$ refers to the fitted values of $\hat{f}_M$. Functions of $\hat{f}_M$ may also be computed directly from the model if the relationship between the predictors can be well understood.

A *marginal model plot* is created by plotting $\{a'x, y\}$ and adding lines for the estimated functions of $\hat{f}_M$ and f_T; $E(y|x) \pm SD(y|x)$ is a usual choice. The adequacy of the model M is supported if the lines describing $\hat{f}_M$ and f_T are similar. Again, checking more values of $a'x$ provides more support for the model, as the model holds

only if all marginal models are true.

1.8.1 With a single predictor on the x-axis

Choosing $a'x$ equal to the predictor X_2 is also useful as an explanatory plot. Since this is a marginal response plot of $\{X_2, Y\}$ with the desired functions of $\hat{f}_M(Y|X_2)$ and $f_T(Y|X_2)$ added, it shows both the true relationship between X_2 and Y when ignoring the other variables, and how this relationship is estimated by the model M.

The estimated true mean of $Y|X_2$, $\widehat{E}_T(Y|X_2)$, can be found by smoothing the plot of $\{X_2, Y\}$, and the estimated true variance by smoothing $\left\{X_2, \left(Y - \widehat{E}_T\left(Y|X_2\right)\right)^2\right\}$.

For the estimated mean under model M, the rule of iterated expectations with estimated expectations is

$$\begin{aligned}\widehat{E}_M(Y|X_2) &= \widehat{E}\left(\widehat{E}_M(Y|X_1, X_2)|X_2\right) \\ &\approx \widehat{E}\left(\widehat{Y}|X_2\right) \end{aligned} \tag{1.28}$$

where $\widehat{E}_M(Y|X_1, X_2)$ are the fitted values from the fit of M, or $\widehat{Y}$. In most cases, it is simplest to estimate these values by smoothing. The goal is to estimate $E\left(\widehat{Y}|X_2\right)$ by finding the mean value of $\widehat{Y}$ over the possible values of X_1 for a fixed X_2. Although with continuous data there is at most one value of X_1 for a fixed X_2, by assuming a smooth relationship between X_1 and X_2, points with X_2 near the fixed X_2 can be used to help in the estimation by using a non-parametric smoother. So the estimated mean under model M, $\widehat{E}_M(Y|X_2)$, is found by smoothing $\left\{X_2, \widehat{Y}\right\}$.

If $\mathrm{Var}_M(Y|X) = \sigma^2$ is a constant, then for the estimated variance under model

M, the rule of iterated expectation for variance using estimated expectations yields

$$\begin{aligned}\widehat{\mathrm{Var}}_M(Y|X_2) &= \widehat{\mathrm{Var}}\left(\widehat{E}_M(Y|X_1,X_2)|X_2\right) + \widehat{E}\left(\widehat{\mathrm{Var}}_M(Y|X_1,X_2)|X_2\right)\\ &\approx \widehat{\mathrm{Var}}\left(\widehat{Y}|X_2\right) + \widehat{E}(\hat{\sigma}^2|X_2)\\ &= \widehat{\mathrm{Var}}\left(\widehat{Y}|X_2\right) + \hat{\sigma}^2, \end{aligned} \tag{1.29}$$

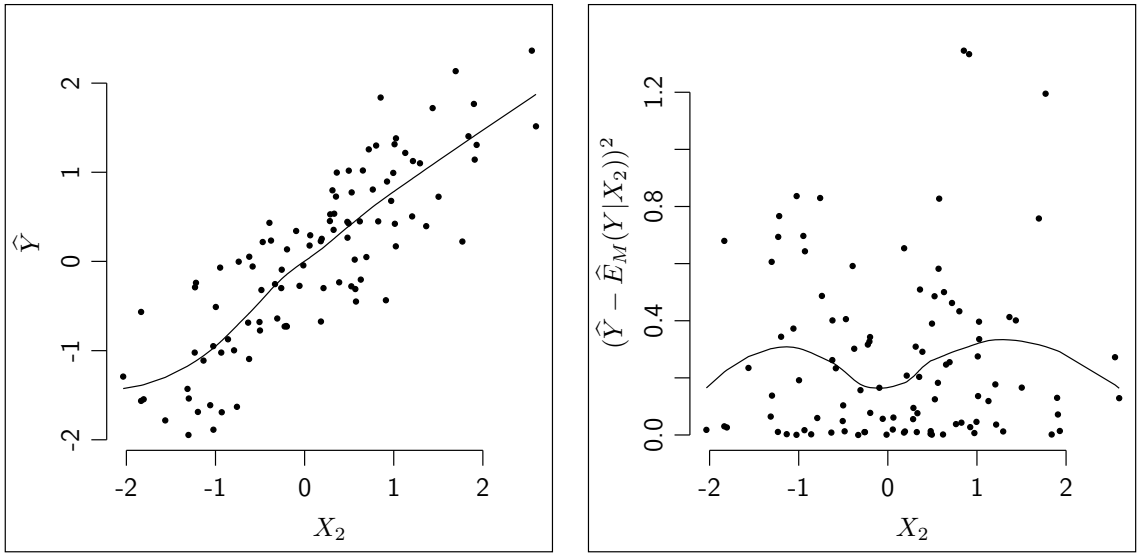
where $\widehat{\mathrm{Var}}\left(\widehat{Y}|X_2\right)$ can be estimated by smoothing $\left\{X_2, \left(\widehat{Y} - \widehat{E}_M\left(Y|X_2\right)\right)^2\right\}$.

When the relationship between X_1 and X_2 can be easily estimated, the conditional mean and variance under model M can also be directly estimated. For example, in the context of the multivariate normal of (1.3), OLS regression can be used to estimate $\widehat{E}(X_1|X_2) = \widehat{X}_{1|2}$ and $\widehat{\mathrm{Var}}(X_1|X_2) = \hat{\sigma}^2_{1|2}$. Under the full model, let the estimated mean $\widehat{E}_M(Y|X_1,X_2)$ be $\hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_2X_2$ and the estimated variance $\widehat{\mathrm{Var}}_M(Y|X_1,X_2)$ be $\hat{\sigma}^2$. Then using estimated expectations,

$$\begin{aligned}\widehat{E}_M(Y|X_2) &= \widehat{E}\left(\widehat{E}_M(Y|X_1,X_2)|X_2\right)\\ &= \widehat{E}(\hat{\beta}_0 + \hat{\beta}_1X_1 + \hat{\beta}_2X_2|X_2)\\ &= \hat{\beta}_0 + \hat{\beta}_1\widehat{E}(X_1|X_2) + \hat{\beta}_2X_2\\ &= \hat{\beta}_0 + \hat{\beta}_1\widehat{X}_{1|2} + \hat{\beta}_2X_2. \end{aligned} \tag{1.30}$$

And by the rule of iterated expectations for variance,

$$\begin{aligned}\widehat{\mathrm{Var}}_M(Y|X_2) &= \widehat{\mathrm{Var}}\left(\widehat{E}_M(Y|X_1,X_2)|X_2\right) + \widehat{E}\left(\widehat{\mathrm{Var}}_M(Y|X_1,X_2)|X_2\right)\\ &= \widehat{\mathrm{Var}}(\hat{\beta}_0 + \hat{\beta}_1X_1 + \hat{\beta}_2X_2|X_2) + \hat{\sigma}^2\\ &= \hat{\beta}_1{}^2\widehat{\mathrm{Var}}(X_1|X_2) + \hat{\sigma}^2\\ &= \hat{\beta}_1{}^2\hat{\sigma}^2_{1|2} + \hat{\sigma}^2. \end{aligned} \tag{1.31}$$



(a) Smoothing to get $\widehat{E}_M(Y|X_2)$.

(b) Smoothing to get $\widehat{\text{Var}}(\widehat{Y}|X_2)$.

Figure 1.13: Steps in using the smoothing method to find the model-based conditional mean and variance to build a marginal model checking plot, in Example 1.9.

Example 1.9 (Marginal model checking plot.)

An ordinary least squares model M was fit to the data of Example 1.2, and the conditional model mean and variance were estimated both by smoothing and by using a linear model. These estimates were plotted on Figure 1.14 over a scatterplot of $\{X_2, Y\}$, the marginal response plot of Section 1.2. Remember that this plot shows the relationship of Y and X_2 without taking X_1 into account, so it may look quite different from the component-plus-residual plot or the added-variable plot, but instead has a similar form to the marginal-plus-residual plot.

The necessary plots and smooths for the smoothing method for $\hat{f}_M$ are shown in Figure 1.13. These smooths were used to find the expected value and the standard deviation as a function of X_2, as shown by the dashed line in Figure 1.14.

To perform the direct estimation using a linear model, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_1$, and $\hat{\sigma}^2$ were estimated by a least squares fit of Y on X_1 and X_2. Additionally, $\widehat{X}_{1|2}$ and $\hat{\sigma}^2_{1|2}$ were

estimated from the least squares fit of X_2 on X_1. For each value of X_2, (1.30) yields

$$\begin{aligned}\widehat{E}_M(Y|X_2) &= \hat{\beta}_0 + \hat{\beta}_1 \widehat{X}_{1|2} + \hat{\beta}_2 X_2 \\ &= -0.01 + 1.59(-0.03 + 0.93X_2) + 1.59X_2\end{aligned}$$

and (1.31) yields

$$\begin{aligned}\widehat{\mathrm{Var}}_M(Y|X_2) &= \hat{\beta}_1{}^2 \hat{\sigma}^2_{1|2} + \hat{\sigma}^2 \\ &= 1.59^2 \cdot 0.33^2 + 0.45^2.\end{aligned}$$

□

The mean and standard deviation are shown by the thinner solid line in Figure 1.14.

Finally, estimates of $E_T(Y|X_2)$ and $\mathrm{Var}_T(Y|X_2)$ were found by smoothing plots of $\{X_2, Y\}$ and $\{X_2, (Y - \widehat{E}_T(Y|X_2))^2\}$, and the mean and standard deviation shown on Figure 1.14 using the thick line. Both of the model-based estimates match up well with this model-free estimate. Although there is some disagreement at the edge of the plot, this is probably due to edge effects in the particular smoother that was used. This agreement provides some evidence that the model is a good fit. To be more certain, the process would have to be repeated for several different linear combinations of X.

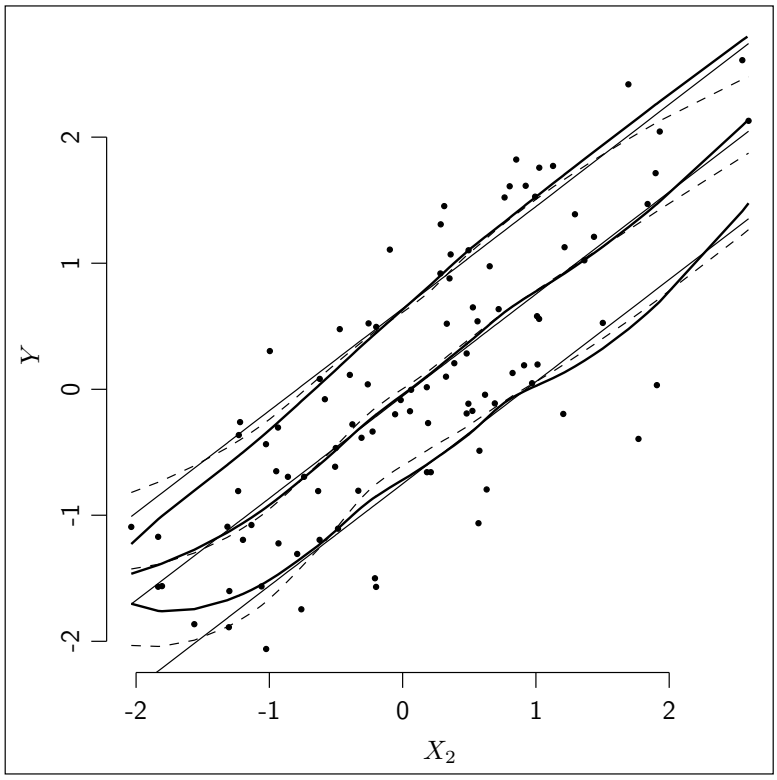


Figure 1.14: Marginal model checking plot, from Example 1.9.

1.9 Example: Island deforestation

When Europeans first reached the islands of the South Pacific, some islands had already been deforested, and the societies that had previously existed on the islands had collapsed. On other islands, the original trees had survived or been replaced with introduced species and the societies had survived. Rolett and Diamond (2004) studied the environmental conditions of 81 sites on 69 islands in the South Pacific to determine if there were environmental explanations for these differences, as opposed to differences in the societies themselves.

The predictors they considered were rainfall, latitude, island age, amount of volcanic ash fallout, called *tephra*, dust fallout, elevation, area, and distance from the nearest island with at least 25% of the area of the current island, called isolation. They additionally studied the terrain type, specifically, how much of the island was covered in makatea, a terrain in which plants and trees have difficulty growing. Since only a few islands had any makatea, the data will be analyzed without that predictor. Any data with missing values for any of these eight predictors will also be removed, leaving only 75 sites. Additionally, rainfall, elevation, area, and isolation were log-transformed.

Two response variables were studied, how much of the island was deforested, called deforestation, how much of the island had non-native species that had replaced the original ones, called replacement. Both were measured on a five point scale that they treated as a continuous ordinal scale. Only the deforestation response will be considered here. Data was provided by the authors at `http://www.nature.com/nature/journal/v431/n7007/suppinfo/nature02801.html`, last accessed on July 1, 2008.

The main goal of Rolett and Diamond was to determine which of these predictors affected deforestation, what the possible effects looked like, and which of the predictors

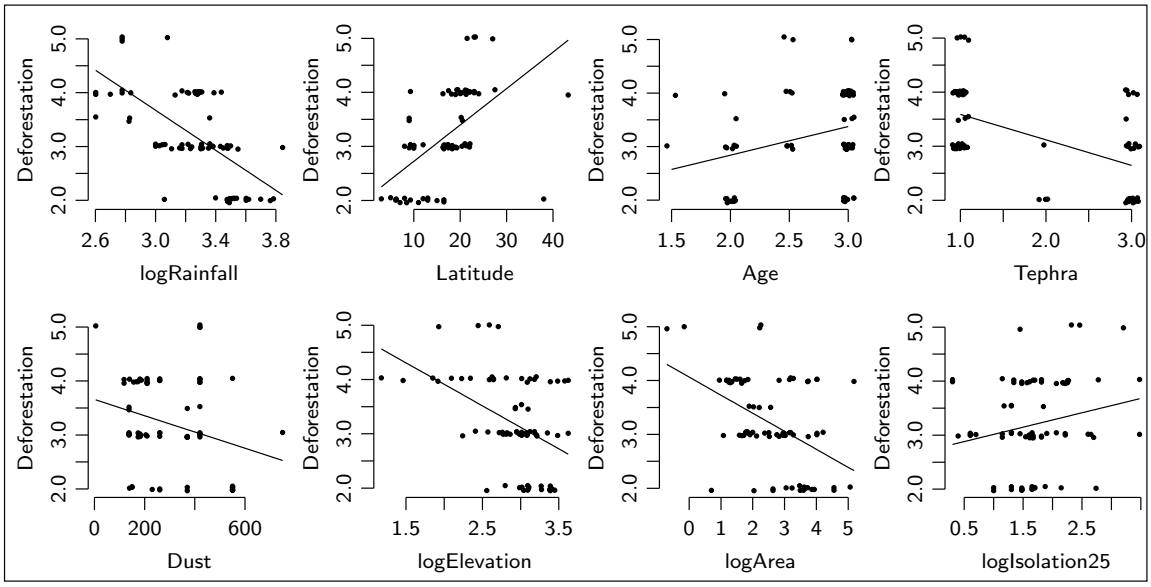


Figure 1.15: Marginal response plots of the eight predictors in the island deforestation data set.

were most important. This example will revisit these questions using the methods developed so far, using linear models to fit the data. Later chapters will revisit this example with other modeling methods and algorithms, which may prove to be more appropriate.

1.9.1 Marginal response plots

Figure 1.15 shows marginal response plots of the eight predictors, with linear fits added. Because the response is only integers, the y-axis has been jittered to separate points and make the patterns clearer. Additionally, the x-axis has been jittered on the Age and Tephra plots. To complement these plots, the coefficient and p-value of the univariate regressions are shown in Table 1.1.

Marginally, all eight variables are statistically significant with p-values less than 0.02, except for isolation. As the p-values are mostly very small, it is difficult to know from the table what the relative effects of each are. The response plots help make

	Estimate	p-value
logRainfall	−1.860	0.000
Latitude	0.067	0.000
Age	0.533	0.011
Tephra	−0.471	0.000
Dust	−0.002	0.017
logElevation	−0.792	0.000
logArea	−0.335	0.000
logIsolation25	0.266	0.078

Table 1.1: Univariate regression coefficients of the island deforestation data.

these effects clearer, for instance, that rainfall seems to have a stronger effect than dust.

From both the plots and the regression, more rainfall, smaller latitudes, younger ages, more tephra, more dust, higher elevation, and more area are associated with less deforestation. According to Rolett and Diamond, these environmental explanations are all sensible.

However, all of these plots and regressions are all done marginally, and do not take into account any possible correlation between the predictors which might cause the conditional effect to be different than the marginal effect.

1.9.2 Net-effect plots

For instance, logRainfall and Latitude are correlated ($\rho = -0.42$), as shown in the scatterplot in Figure 1.16. Additionally, this plot displays the Deforestation for each site by sizing each point proportionally, so large points have higher deforestation values. The points get larger in the upper-left corner, and smaller in the lower-right corner, but it is unclear if this due to the change in Latitude, the change in Rainfall, or some of both.

Figure 1.17 shows these relationships using net-effect plots. Deforestation is plot-

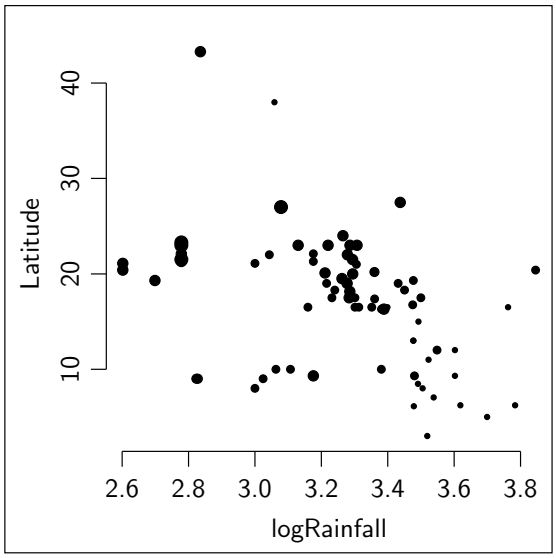
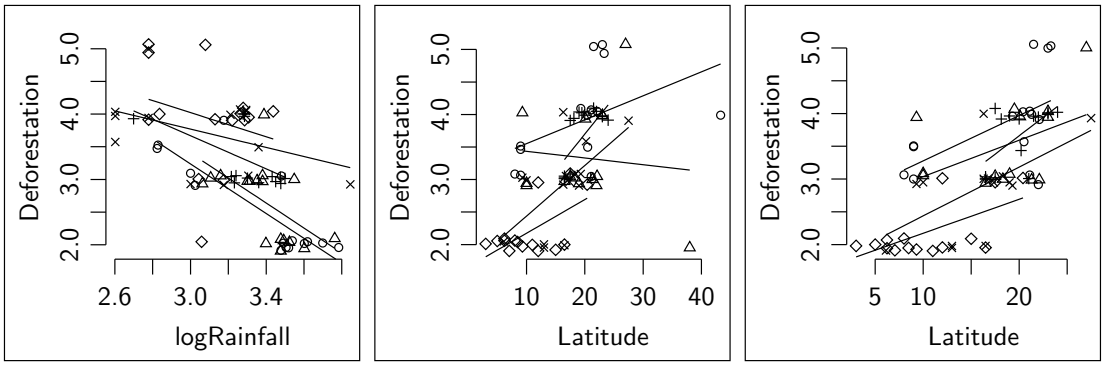


Figure 1.16: Scatterplot of logRainfall and Latitude, with point diameter proportional to Deforestation.



(a) Deforestation against logRainfall, sliced by Latitude.

(b) Deforestation against Latitude, sliced by logRainfall.

(c) Deforestation against Latitude, sliced by logRainfall, without the two sites with highest Latitude values.

Figure 1.17: Sliced net-effect plots of Deforestation against logRainfall and Latitude.

ted against one of logRainfall and Latitude, with slices created using the other. Lines are shown for each slice, fit using least squares. These plots show that there is some effect from both predictors. In Figure 1.17a, the lines show the effect of logRainfall for sites with similar latitudes; the effect of logRainfall is negative, regardless of latitude. Conversely, in Figure 1.17b, the lines show the effect of Latitude for sites with similar rainfall levels. However, the effect of latitude does not seem to be the same for all rainfall levels. If so, this would be evidence of an interaction between rainfall and latitude, as the effect of latitude would be different for different rainfalls. But in fact, only the two sites with the largest latitude values are causing this effect. In Figure 1.17c these sites have been removed, and there is no longer visible evidence of the interaction. However, this doesn't necessarily mean that these sites should still be considered outliers after all eight predictors are accounted for; perhaps one or more of the other predictors explain the large difference seen here.

This example also illustrates how difficult it would be to view net-effect plots for each predictor by accounting for every other predictor. The above process could be repeated for all pairs of predictors, but that would result in a prohibitively large number of plots. Additionally, it would not account for any possible interactions in the other predictors. Cook (1995) states that one option is to slice over a function of all the other predictors, but in this case, it is unclear what function might be appropriate. This example demonstrates the need for global net-effect plots, such as the component-plus-residual plot or the added-variable plot, which gather the information from many plots together in appropriate ways.

1.9.3 Component-plus-residual plot

To build the component-plus-residual plots, the necessary coefficients must first be obtained by performing a multiple regression. These are shown in Table 1.2, along with the corresponding p-values. Some of these coefficients are different than the

marginal coefficients; for example, the coefficient for logArea changes from -0.33 to -0.088 and is now not statistically significant.

	Estimate	Pr(>|t|)
logRainfall	−1.222	0.000
Latitude	0.038	0.000
Age	0.229	0.116
Tephra	−0.092	0.277
Dust	0.000	0.686
logElevation	−0.394	0.056
logArea	−0.088	0.337
logIsolation25	0.231	0.014

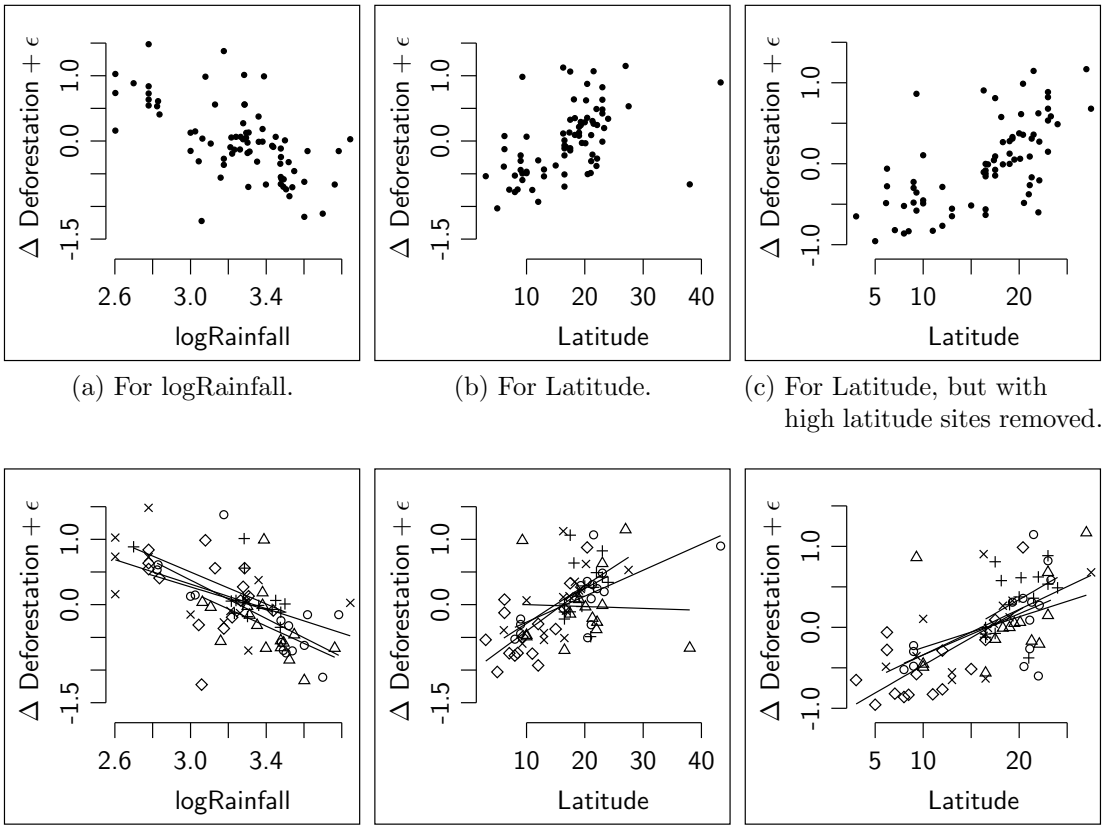
Table 1.2: Multivariate regression coefficients of the island deforestation data.

First, the two predictors investigated earlier, logRainfall and Latitude, will be revisited. Component-plus-residual plots for these variables are shown in Figure 1.18a and Figure 1.18b. The effect seen in the net-effect plots when only slicing for one variable holds up here after all the other variables are accounted for; an increase in logRainfall or a decrease in Latitude when all other variables are held constant is associated with a smaller Deforestation score.

Figure 1.18d and Figure 1.18e show these same plots, except sliced over the other variable. The dependence on the other variable has been properly removed, except again in the Latitude plot, because of the points with high Latitude. Figure 1.18c and Figure 1.18f show these plots recreated from the model without these two sites, and again the dependence has been properly removed.

Figure 1.19a shows the component-plus-residual plots for all eight variables, calculated from the full model. These show how each predictor affects the response, and give a sense of the size of the effect. As in the marginal plots, the points in the Age and Tephra plots have been jittered horizontally as those variables are discrete.

Researchers occasionally try to get a sense of the size of the effect of each predictor



(a) For logRainfall.

(b) For Latitude.

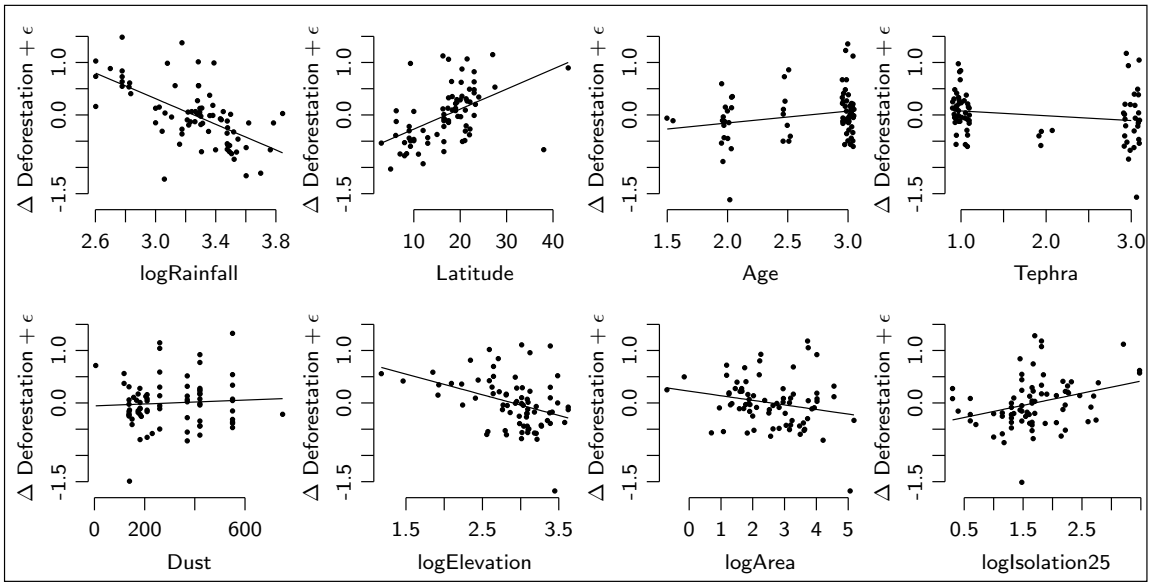
(c) For Latitude, but with high latitude sites removed.

(d) For logRainfall, sliced over Latitude.

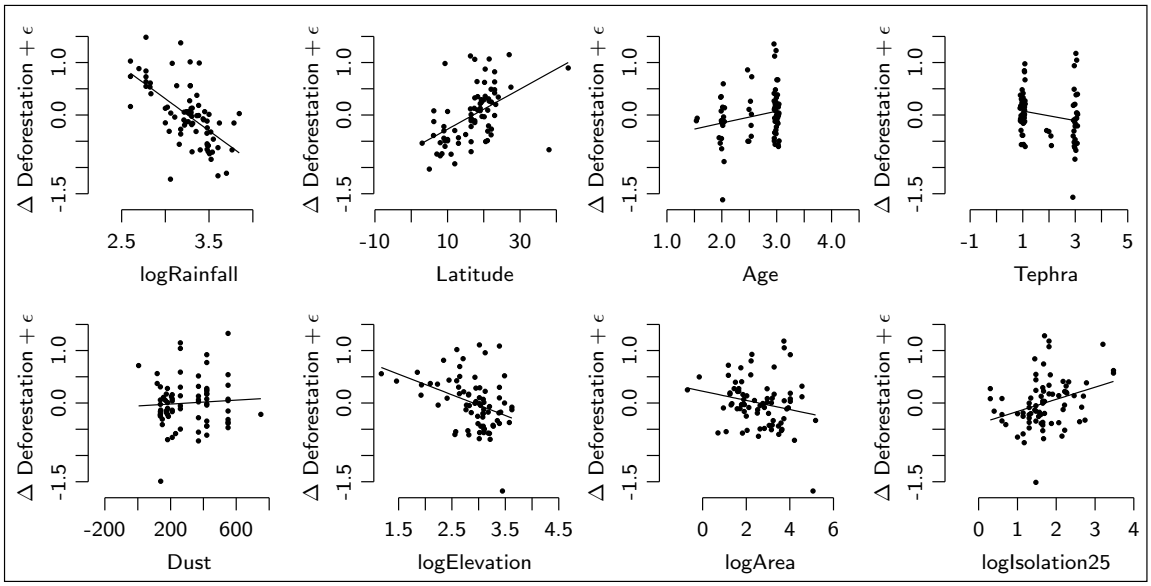
(e) For Latitude, sliced over logRainfall.

(f) For Latitude, but with high latitude sites removed, sliced over logRainfall.

Figure 1.18: Component-plus-residual plots for Latitude and logRainfall, using the coefficients from the multiple linear regression.



(a) With default x-axis scaling.



(b) With standardized x-axis scaling.

Figure 1.19: Component-plus-residual plots for all eight variables in the islands deforestation data set, using linear fits.

by standardizing the predictors to have equal standard deviations and then fitting the data; in this way the units of the coefficients are more directly comparable; for each variable the coefficient is the change in the response when that variable is increased by one standard deviation but all other variables are held constant. However, as Weisberg (1985) explains, these coefficients then depend on the standard deviation of each variable, so different experiments will result in different values if the experiments test the variables over different ranges. This occurs even if the unstandardized coefficients are identical. Nevertheless, this type of comparison can be achieved by viewing component-plus-residual plots where the horizontal axes are standardized to each include the same range of standard deviations, as in Figure 1.19b. The slopes of the fitted lines in these plots are directly comparable in the same way that standardized coefficients are. However, an advantage of using this method is that the standardization is hidden; the scale on the x-axis of each plot is still on the scale of the original predictor, it simply extends farther than would be necessary in order to make the slopes comparable. Comparing the coefficients in this way instead of standardizing them also makes comparisons between experiments clearer, as the x-axes of the plots from two different experiments can be directly compared.

Finally, it should be remembered that all these plots rely on the assumption that each of the predictors does enter the model linearly. As will be seen in Section 2.1, when this assumption does not hold, these plots do not properly show the component due to each variable and so do not properly remove the dependence on the other variables.

1.9.4 Added-variable plots

Figure 1.20a shows added-variable plots for each of the eight predictors. Unlike the component-plus-residual plot, these plots do not necessarily show the practical significance of each predictor, but the statistical significance. That is, the plot shows how

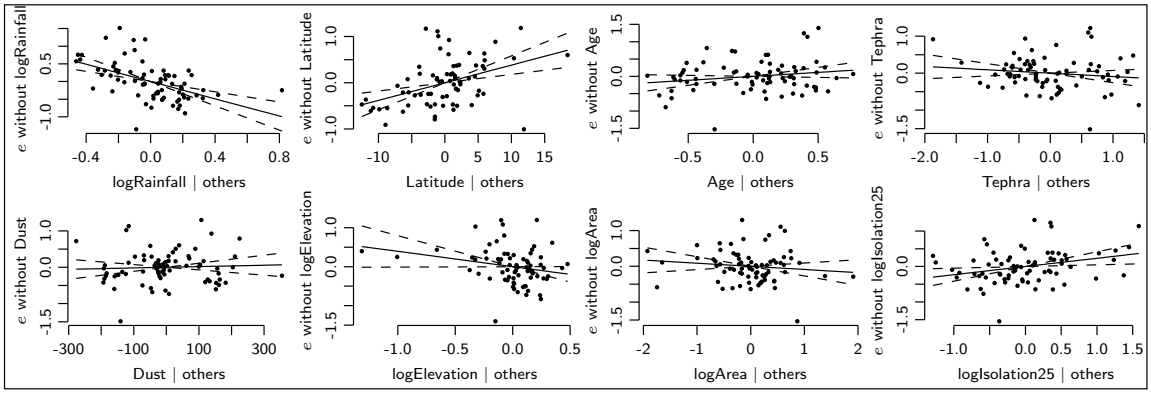
much of the information left unexplained by the other predictor can be explained by this particular predictor.

In these particular plots, lines have been drawn to show both the fitted slope and a 95% confidence interval for the slope. Thus a nonsignificant slope at the 5% level corresponds to bounds that include the line with slope 0, a helpful visual clue to which predictors are significant and which are not.

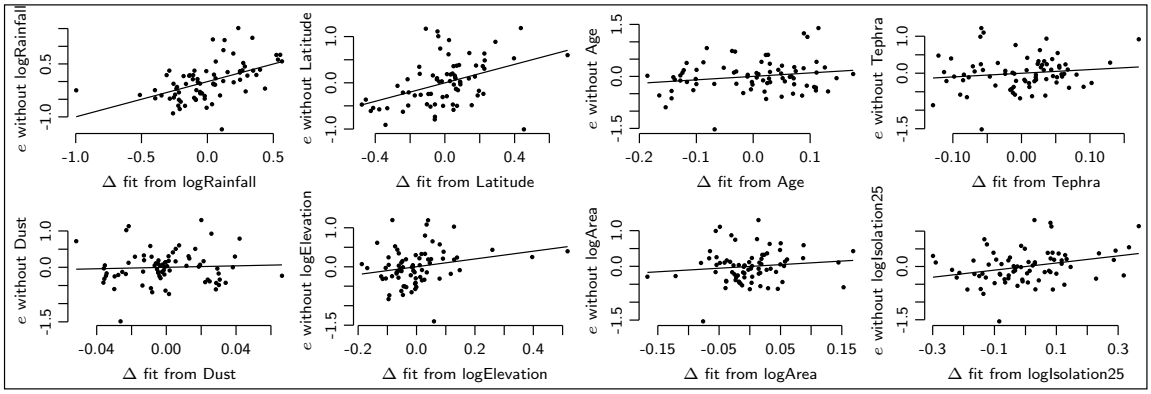
These plots can also help identify points that are either outliers or that have high leverage, after accounting for all the other variables. For example, in the logRainfall plot, there is one site that gets substantially more rainfall than might be predicted from the other variables; this point has a higher leverage and is more influential in the fit. Additionally, many of the plots show an outlier with substantially lower deforestation. This outlier is one of the sites with high latitude that was discussed earlier when investigating only logRainfall and Latitude.

The general added-variable plots are shown in Figure 1.20b. These instead have the change in fitted values when the given predictor is removed from the model on the x-axis. The points in these plots are all centered around a line with a slope of one, which has been added to each plot. These plots can also help identify any points that are outliers or have more influence, but since the slopes are all the same, it can be hard to use them to visually compare the effect of the variables.

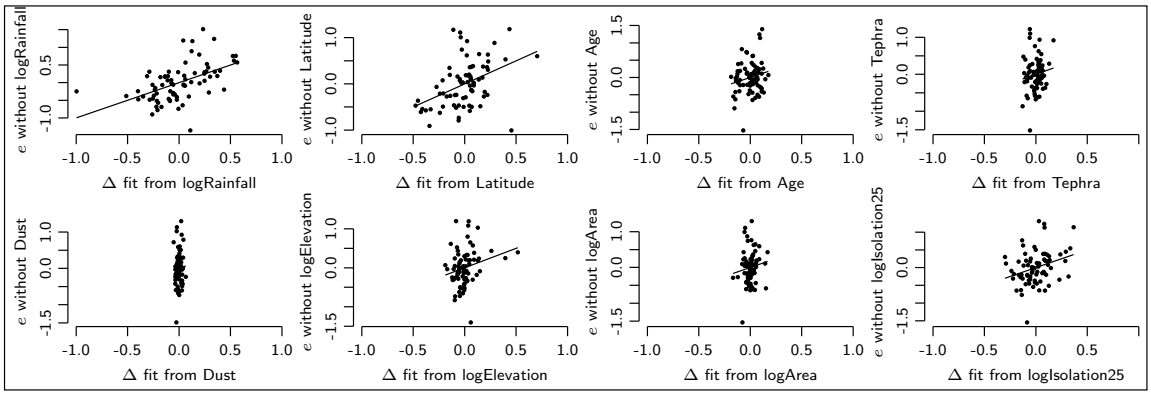
Standardizing the x-axis, as in Figure 1.20c, can aid in this. Now the horizontal spread of the plots, which show how much the fit changed when each predictor was added, can be compared. The plots for logRainfall and Latitude have the largest spreads, and so in this sense, are the most important. Conversely, Dust has very little spread, so is much less important.



(a) Standard added-variable plots, with default x-axis.



(b) General added-variable plots, with different x-axis scales.



(c) General added-variable plots, with a common x-axis scale.

Figure 1.20: Added-variable plots for all eight predictors in the island deforestation data set, using linear fits.

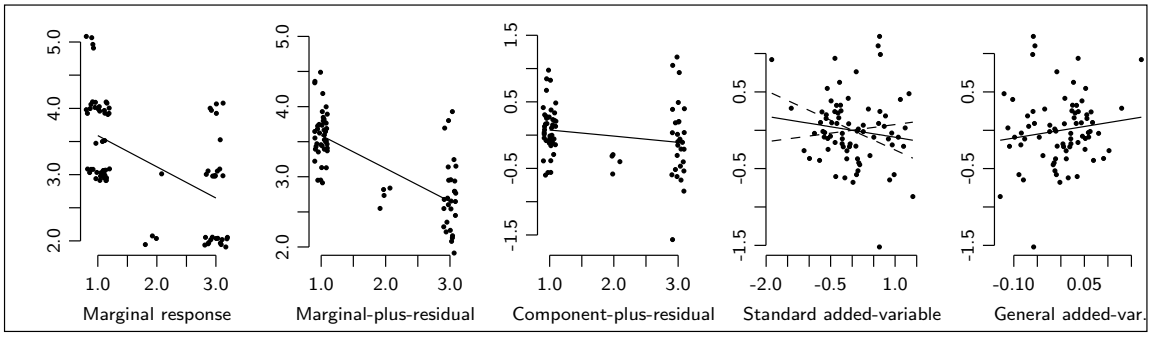
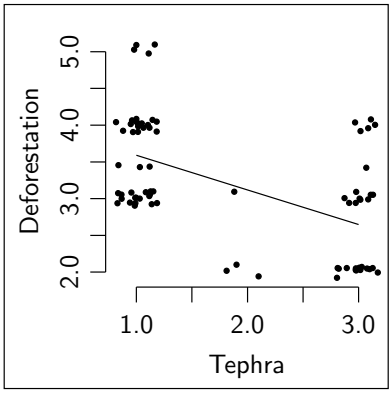


Figure 1.21: The marginal response, marginal-plus-residual, component-plus-residual, and standard and general added-variable plots for Tephra.

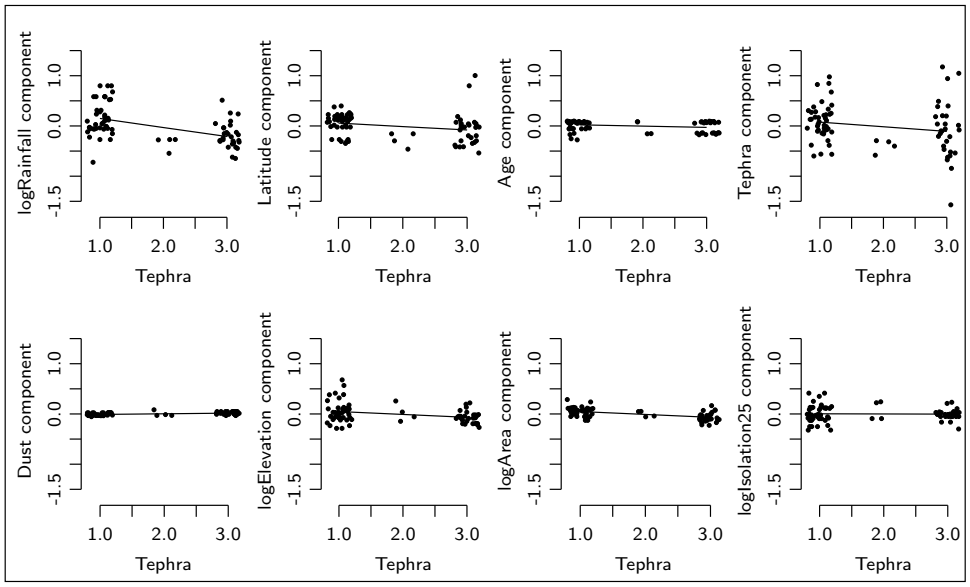
1.9.5 Comparing various plots

To compare these plots with each other, five plots for Tephra, including the marginal-plus-residual plot of Section 1.5.2, are shown in Figure 1.21. In this way, the marginal and conditional effects of a given predictor can be compared. Marginally, an increase in Tephra is associated with a decrease in Deforestation, but as most of this decrease can be accounted for the other variables, the conditional relationship is much weaker.

This comparison can also made by plotting the component due to each of the other variables, that is, $\hat{\beta}_k X_k$, where $\hat{\beta}_k$ is the coefficient and X_k is the variable, against Tephra, as in Figure 1.22. This shows how the contribution to Deforestation from each of the other predictors changes as Tephra changes. When Tephra is uncorrelated with a predictor, the contribution will be the same across the range of Tephra, as with logIsolation ($\rho = -0.02$). But when it is correlated, this contribution may change, as with logRainfall ($\rho = 0.49$). Thus the marginal plot of Figure 1.22a shows that Tephra is marginally related to Deforestation. But the component plots show that a large part of this relationship is due to logRainfall, and that an increase in Tephra is associated with an increase in logRainfall, which is associated with the decrease in Deforestation.



(a) Marginal plot for Tephra; both axes are jittered slightly.



(b) Component plots for each other variable plotted against Tephra. The plot for Tephra also contains the residuals, so is a component-plus-residual plot. Plots for variables with only a few values are jittered on the x-axis.

Figure 1.22: Marginal plot for Tephra, compared with component plots for each other variable plotted against Tephra.

Chapter 2

Additive models

When the relationship between the predictors and the response, or between the predictors themselves, are not linear, the plots explored in the multivariate normal context must be adapted. Consider an additive model with two predictors, X_1 and X_2, where one, X_2, is of interest, a response Y, and independent normal errors, so the true model can be expressed as

$$Y = f_1(X_1) + f_2(X_2) + \epsilon, \tag{2.1}$$

where $\epsilon \sim N(0, 1)$. In this model f_1 and f_2 may be nonlinear, as well as the relationship between X_1 and X_2, or between $f_1(X_1)$ and $f_2(X_2)$. Several nonparametric algorithms have been proposed to fit models like this, including backfitting and penalized spline smoothing. These have been implemented in the `gam` and `mgcv` libraries in `R`, respectively.

The backfitting method, described in detail by Hastie and Tibshirani (1999), estimates each function by iteratively smoothing the partial residuals until the fit doesn't change, where the partial residuals for each predictor are calculated by subtracting the current estimates of functions of the other predictors from the response. Buja et al. (1989) proved that this process converges to a unique solution for a large of class of smoothers, including polynomial regression, smoothing splines, and regres-

sion splines. One disadvantage of backfitting is that the degree of smoothness is difficult to estimate and usually must be specified beforehand.

In contrast, when using penalized spline smoothing to fit each of the smooths simultaneously, as in Wood (2006), estimation of the degree of smoothness can be integrated into the algorithm. However, more data points are needed, as the effective dimension of the system is greater when smoothing all of the predictors simultaneously than when smoothing each individually.

Regardless of the fitting method used, marginal response plots and local net-effect plots can be constructed exactly as in the multivariate normal context, and interpreted in exactly the same way. However, the ways that the local net-effect plots were combined depend on the linearity of the predictors and the response, so these methods must be modified to be useful in this context.

The goal will remain the same: to construct plots $\{X^*, Y^*\}$ where

$$Y^*|(X^*, X_1) \sim Y^*|X^* \tag{2.2}$$

so the plot has no conditional dependence on X_1. As in Chapter 1, let

$$Y^* = Y - g(X_1, X_2), \quad \text{and} \quad X^* = h(X_1, X_2).$$

Assuming the errors are additive and independent of the predictors, the conditional distribution of $Y^*|X^*$ will again change only in the mean, so determining sufficient conditions for (2.2) to hold may be found by investigating only $E(Y^*|X^*)$. As

$$\begin{aligned} E(Y^*|X_1, X_2) &= E(Y - g(X_1, X_2)|X_1, X_2) \\ &= f_1(X_1) + f_2(X_2) - g(X_1, X_2), \end{aligned} \tag{2.3}$$

to satisfy (2.2), only functions of $X^* = h(X_1, X_2)$ must be on the right side. There are again many choices, so interpretability will be a guide to the choice of X^*. There

are three main classes of plots that are easily interpretable, corresponding to the component-plus-residal plot, the marginal-plus-residual plot, and the added-variable plot.

2.1 Component-plus-residual plot

One interpretable option for the x-axis is simply $X^* = X_2$. Then there are several ways to make (2.3) a function of only X_2; the simplest option is to set $g(X_1, X_2) = f_1(X_1)$ so (2.3) becomes simply $f_2(X_2)$. This results in the idealized plot of

$$\{X_2, Y - f_1(X_1)\}. \tag{2.4}$$

Because $Y - f_1(X_1) = f_2(X_2) + \epsilon$, this plot may also be written as

$$\{X_2, f_2(X_2) + \epsilon\}, \tag{2.5}$$

and is now recognizable as a component-plus-residual plot, as in Section 1.5. A plot of this type allows the form of f_2 to be inspected, and shows how X_2 is related to Y for fixed X_1.

A version of this plot using an estimate $\hat{f}_2$ of f_2 and without the residuals,

$$\{X_2, \hat{f}_2(X_2)\}, \tag{2.6}$$

is also commonly used; it one of the default plots in the `R` implementation of both the backfitting (`gam`) and the spline smoothing (`mgcv`) methods of fitting additive models. This plot makes the characteristics of $\hat{f}_2$ easier to see because it is uncluttered by the underlying data. However, this clarity comes at a price, as the variability of the data around this estimate is no longer shown. Including the residuals allows one to see if

the estimated form of f_2 is significant in relationship to the residuals and can help in determining if $\hat{f}_2$ is influenced by any individual points.

When the method used for fitting the model results in an estimate of f_2, as in backfitting and spline smoothing, estimated versions of these component plots, either with or without the residuals, are especially easy to construct. An advantage of using one of these methods to fit the model and create this plot is these methods do not rely on linearity in f_1, f_2, or between the predictors, so can be used for any additive model. Section 2.1.1 will show that when certain of these relationships are linear, other methods exist, including partial residual plots, augmented partial residual plots, and CERES plots. All of these methods estimate each component only up to an arbitrary additive constant.

As in the multivariate normal context, the component-plus-residual plot can be modified and still satisfy (2.2). One useful option is to detrend the plot, which according to Mansfield and Conerly (1987) can allow one to better see the form of f_2. Another is to simply plot the residuals against X_2.

2.1.1 Linear f_1

When f_1 is linear, say

$$f_1(X_1) = \beta_1 X_1, \tag{2.7}$$

several methods exist to create the component-plus-residual plot. These methods do not estimate f_2 directly, but instead estimate the linear function $f_1(X_1) = \beta_1 X_1$ and use the equality

$$Y - f_1(X_1) = f_2(X_2) + \epsilon. \tag{2.8}$$

to estimate the component-plus-residual, $f_2(X_2)+\epsilon$. Several methods exist, depending on the form of $E(X_1|X_2)$.

When $E(X_1|X_2)$ is linear in X_2 or when X_1 and X_2 are independent, a consistent estimate of β_1 can be found by fitting the model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon. \tag{2.9}$$

Although this model is known to be incorrect because $f_2(X_2) \neq \beta_2 X_2$, Cook (1993) proved that it does provide a consistent estimate of β_1 that can be used in constructing the component-plus-residual plot. Additionally, this model will consistently estimate β_1 when f_2 is linear, no matter what the form of $E(X_1|X_2)$, because then the model is correct. Plots constructed using this model are called *partial residual plots*. Earlier references include Larsen and McCleary (1972) and Wood (1973).

When $E(X_1|X_2)$ is quadratic in X_2, the model (2.9) no longer consistently estimates β_1. Instead, the model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2 + \epsilon \tag{2.10}$$

should be used, even though this model may again be incorrect. This model also is appropriate when f_2 is quadratic, no matter the form of $E(X_1|X_2)$. These plots are called *augmented partial residual plots*, and were proposed by Mallows (1986).

A general method for consistently estimating β_1, no matter the form of $E(X_1|X_2)$, was introduced by Cook (1993), who proved that the model

$$Y = \beta_1 X_1 + \beta_2 m(X_2) + \epsilon, \tag{2.11}$$

where $m(X_2) = E(X_1|X_2) - E(X_1)$, will always consistently estimate β_1. Plots using the estimate of β_1 from this model are called *CERES plots*, which stands for

Combining Expectations and RESiduals. If $E(X_1|X_2)$ is unknown, a nonparametric estimate of $E(X_1|X_2)$ can be used.

Proof 2.1 (Fisher consistency of the CERES plot)
Consider the true model

$$Y = \alpha_0 + \alpha_1 X_1 + g(X_2) + \epsilon$$

where $\epsilon \sim N(0, I)$. It is often of interest to plot the form of g using a component-plus-residual plot

$$\{X_2, g(X_2) + \epsilon\},$$

which is equivalent (up to a constant) to the partial residual plot

$$\{X_2, Y - \alpha_1 X_1\}.$$

To estimate this plot, a Fisher-consistent estimate of α_1 is needed. An early idea was to fit the model

$$\hat{Y}_1 = b_0 + b_1 X_1 + b_2 X_2$$

by minimizing the loss $L(Y - \hat{Y}_1)$ for some convex loss function. However, Cook showed that this only results in Fisher-consistent estimates when $E(X_1|X_2)$ is linear or $g(X_2)$ is linear. Instead, the model

$$\widehat{Y} = b_0 + b_1 X_1 + b_2 m(X_2)$$

where $m(X_2) = E(X_1|X_2) - E(X_1)$ should be used.

To show that b_1 is Fisher-consistent for α_1, let $\beta = (\beta_0, \beta_1, \beta_2)$ be the values that minimize the expected loss (the risk) of

$$R(b_0, b_1, b_2) = EL\left[(\alpha_0 + \alpha_1 X_1 + g(X_2) + \epsilon) - (b_0 + b_1 X_1 + b_2 m(X_2))\right],$$

so by definition, $b = (b_0, b_1, b_2)$ is Fisher consistent for β. Assuming β is unique, if α_1 also minimizes this expected loss, then $\alpha_1 = \beta_1$, and b_1 is Fisher consistent for α_1.

For any values of (b_0, b_1, b_2),

$$\begin{aligned}
R(b_0, b_1, b_2) &= EL\left[((\alpha_0 + \alpha_1 X_1 + g(X_2) + \epsilon) - (b_0 + b_1 X_1 + b_2 m(X_2))\right] \\
&\leq E_{X_1,\epsilon} L\left[E_{X_2}\left(\alpha_0 - b_0 + (\alpha_1 - b_1)X_1 + g(X_2) - b_2 m(X_2) + \epsilon\right)\right] \\
&= E_{X_1,\epsilon} L\left[\alpha_0 - b_0 + (\alpha_1 - b_1)E(X_1|X_2) + g(X_2) - b_2 m(X_2) + \epsilon\right] \\
&= EL[\alpha_0 - b_0 - (b_2 - (\alpha_1 - b_1))m(X_2) \\
&\qquad + g(X_2) + \epsilon + (\alpha_1 - b_1)E(X_1) \\
&= EL\left[(\alpha_0 + \alpha_1 X_1 + g(X_2) + \epsilon) - (b_0^* + \alpha_1 X_1 + b_2^* m(X_2))\right] \\
&= R(b_0^*, \alpha_1, b_2^*)
\end{aligned}$$

where $b_0^* = b_0 - (\alpha_1 - b_1)E(X_1)$, and $b_2^* = (b_2 - (\alpha_1 - b_1))$. So the risk for $b_1 = \alpha_1$ is no larger than the risk for any other value of b_1. If this lower bound is actually met, then α_1 is the minimizer; it is actually met because at $b_1 = \alpha_1$, $b_0^* = b_0$ and $b_2^* = b_2$, so

$$R(b_0, \alpha_1, b_2) \leq R(b_0^*, \alpha_1, b_2^*) = R(b_0, \alpha_1, b_2) \quad \text{when} \quad b_1 = \alpha_1.$$

Thus α_1 minimizes the risk and $\alpha_1 = \beta_1$, so b_1 is a Fisher consistent estimate for α_1.

Additionally, b_0 is a Fisher consistent estimate of $\alpha_0 - E(g(X_2))$ because

$$\begin{aligned}
EL(b_0, b_1, b_2) &\geq E\left[((\alpha_0 + \alpha_1 X_1 + g(X_2) + \epsilon) - (b_0 + b_1 X_1 + b_2 m(X_2))\right] \\
&= \alpha_0 + E(g(X_2)) + (\alpha_1 - b_1)X_1 + b_2 E(m(X_2)) + b_0 \\
&= \alpha_0 + E(g(X_2)) + b_0 \quad \text{when} \quad b_1 = \alpha_1,
\end{aligned}$$

which equals zero when $b_0 = \alpha_0 + E(g(X_2))$. □

A dynamic version of the CERES plot was proposed by Seo (1999) to show the

change in the form of f_2 when one of the other predictors is added or removed. For example, suppose the full model has three predictors where the third may be nonlinear, so

$$Y = \beta_1 x_1 + \beta_2 x_2 + f_3(x_3) + \epsilon.$$

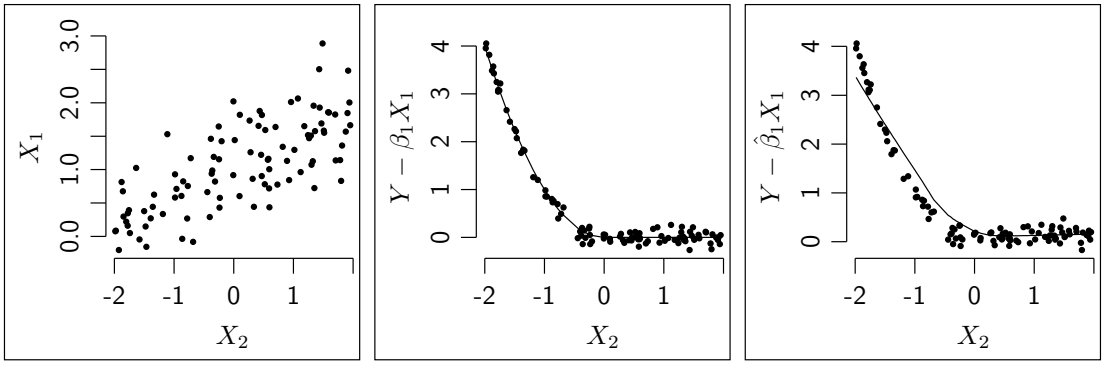
To see how f_3 changes when x_1 is removed from the model, this plot morphs between a CERES plot for x_3 calculated using the full model, and a CERES plot for x_3 calculated using the submodel without x_1 included.

As discussed earlier, an alternative method in any of these cases is to use nonparametric techniques to estimate f_2 directly. When X_1 is assumed to enter the model linearly, this method can be used by restricting the smoother to only consider linear terms of X_1. This is called AMONE by Berk and Booth (1995), who cite the discussion of Breiman and Friedman (1985) for inspiration, and find it to be comparable to the CERES plot.

Example 2.1 (Partial residual plot)
For this example, $f_1(X_1) = 2X_1$, so β_1=2, and $f_2(X_2) = I(X_2 < 0)X_2^2$. 100 data points were constructed as follows: $X_2 \sim \text{Unif}(-2, 2)$, $X_1 = (0.5X_2 + 1) + 0.5\epsilon_1$, and $Y = f_1(X_1) + f_2(X_2) + 0.1\epsilon$, where ϵ and ϵ_1 are a independent standard normal random variables.

Figure 2.1a shows the relationship between the predictors, and Figure 2.1b shows the idealized component-plus-residual plot.

Because $f_1(X_1)$ and $E(X_1|X_2)$ are both linear, it is appropriate to use the linear model of (2.9) to estimate β_1. Although this model is known to be wrong, the least squares estimate is $\hat{\beta}_1 = 1.91$, which is very close to the true value of $\beta_1 = 2$. Figure 2.1c shows the corresponding partial residual plot, which agrees well with the idealized component-plus-residual plot.



(a) The relationship between the predictors.

(b) Idealized component-plus-residual plot for X_2, with true component.

(c) Estimated partial residual plot for X_2, with lowess fit.

Figure 2.1: Component-plus-residual plots for Example 2.1, when $E(X_1|X_2)$ is linear.

Example 2.2 (Augmented partial residual plot)

When the relationship between the predictors is not linear but quadratic, then an augmented partial residual plot should be used. This example uses the same setup as Example 2.1, except that the relationship between the predictors is changed, so

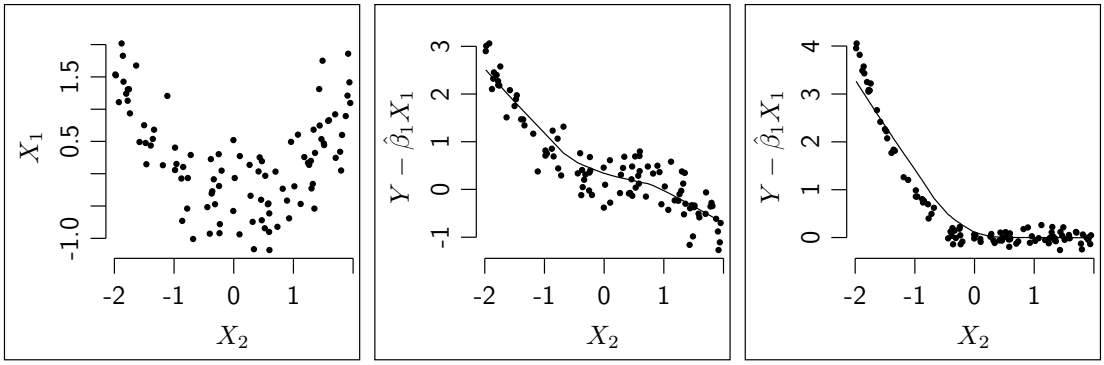
$$X_1 = \tfrac{1}{2}(X_2^2 + 1) + 0.5\epsilon_1,$$

as shown in Figure 2.2a. Since the component for X_2 and the residuals did not change, the idealized component-plus-residual plot is exactly as shown in Figure 2.1b.

Both the linear model (2.9) and the augmented model (2.10) were then used to estimate β_1. Using the linear model, $\hat{\beta}_1 = 2.68$, which is not as close to the true value of $\beta_1 = 2$ as in the last example. This discrepancy is visible in the partial residual plot (Figure 2.2b), which no longer shows the form of the true component as shown in Figure 2.1b. However, using the augmented model, $\hat{\beta}_1 = 2.00$, which is much closer to the true value of β_1, and so the augmented partial residual plot (Figure 2.2c) does show the true form of f_2. □

Example 2.3 (CERES plot)

As in Example 2.2, this example uses the same setup as Example 2.1 except for the



(a) The relationship between the predictors.

(b) Estimated partial residual plot for X_2, with lowess fit.

(c) Estimated augmented partial residual plot for X_2, with lowess fit.

Figure 2.2: Component-plus-residual plots for Example 2.2, when $E(X_1|X_2)$ is quadratic.

relationship between X_1 and X_2. In this case,

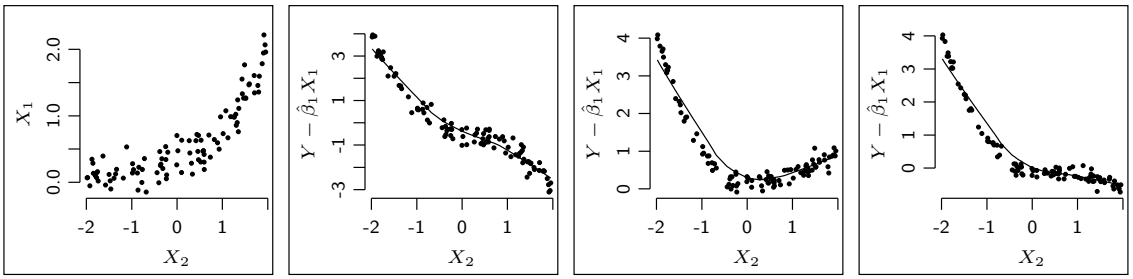
$$X_1 = \exp(X_2 + 2)/25 + 0.1\epsilon_1,$$

as seen in Figure 2.3a. The component from X_2 remains the same, so the idealized component-plus-residual plot for this example is still Figure 2.1b.

Because $E(X_1|X_2)$ is not linear or quadratic, neither the linear model or the augmented model estimate β_1 well; the estimates are 3.4 and 1.51, respectively. But using the known $E(X_1|X_2)$, the CERES plot estimates β_1 to be 2.28, close to the true $\beta_1 = 2$. Figure 2.3 shows that the linear and augmented partial residual plots do not accurately show the form of f_2, but the CERES plot does. □

2.1.2 Nonlinear f_1

Although CERES plots are designed for the situation where X_1 enters the model linearly, Cook and Weisberg (1999a) do suggest guidelines for their use when several of the predictors appear to be non-linear. Since the nonlinear effects of other predictors can leak into the CERES plot for X_2, the plot can show a curve even when X_2 is



(a) The relationship between the predictors.

(b) Estimated partial residual plot, with lowess fit.

(c) Estimated augmented partial residual plot, with lowess fit.

(d) Estimated CERES plot, with lowess fit.

Figure 2.3: Component-plus-residual plots for Example 2.3, when $E(X_1|X_2)$ has degree > 2.

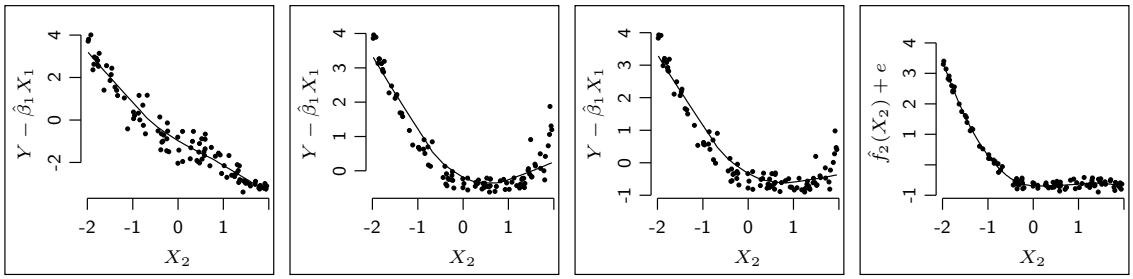
linear. They suggest transforming the predictor that shows the strongest visual fit in its CERES plot first.

An alternate method is to use a full nonparametric fit, called AMALL by Berk and Booth (1995), to estimate the full model, and then plot the component and the residuals.

Example 2.4 (CERES vs. AMALL)

This example uses the same setup as Example 2.3 except that $f_1(X_1)$ is not linear, but equals X_1^2. As before, f_2 does not change, so the idealized component-plus-residual for this example is as before, and is shown in Figure 2.1b.

But with a nonlinear f_1, even the CERES plot does not accurately show the form of f_2, as the curve in X_1 affects the plot. The estimates of β_1 using the linear, augmented, and CERES models are 3.57, 1.37, 1.78, respectively. None are sufficiently close to the true value of 2, and the corresponding plots in Figure 2.4 do not accurately show the form of f_2. But the component-plus-residual plot from the model found by spline smoothing, using the `mgcv` library in `R`, does accurately show the form of f_2, as shown in Figure 2.4d. □



(a) Estimated partial residual plot, with lowess fit.

(b) Estimated augmented partial residual plot, with lowess fit.

(c) Estimated CERES plot, with lowess fit.

(d) Estimated component-plus-residual plot, using `mgcv`.

Figure 2.4: Component-plus-residual plots for Example 2.4, when $E(X_1|X_2)$ is non-quadratic and f_1 is non-linear.

2.2 Marginal-plus-residual plot

Setting $g(X_1, X_2) = f_1(X_1)$ is not the only way for $E(Y^*|X_1, X_2)$ to be a function of only X_2. From (2.3),

$$E(Y^*|X_1, X_2) = f_1(X_1) + f_2(X_2) - g(X_1, X_2). \tag{2.12}$$

By the rule of iterated expectations, $E(Y|X_2) = E(f_1(X_1)|X_2) + f_2(X_2)$. Solving for $f_2(X_2)$ and substituting into (2.12),

$$E(Y^*|X_1, X_2) = f_1(X_1) + E(Y|X_2) - E(f_1(X_1)|X_2) - g(X_1, X_2). \tag{2.13}$$

Now there are several options for g that result in $E(Y^*|X_1, X_2)$ being a function of X_2. One that results in an interpretable plot is to set

$$g(X_1, X_2) = f_1(X_1) - E(f_1(X_1)|X_2) \tag{2.14}$$

so $E(Y^*|X_1, X_2) = E(Y|X_2)$, which is indeed a function of only X_2. The resulting plot is

$$\{X_2, Y - [f_1(X_1) - E(f_1(X_1)|X_2)]\}, \tag{2.15}$$

which is a marginal-plus-residual plot, like that in Section 1.5.2. This plot is apparently new. It makes the marginal conditional mean relationship between X_2 and Y easier to see by removing variability that can be explained by the other variables.

Unless f_1 is linear, the idealized form of g will be difficult to compute. However, it is not difficult to estimate, since it is simply the change in fitted values between a model with only X_2 and a model with both X_1 and X_2, as

$$\begin{aligned} E(Y|X_1, X_2) - E(Y|X_2) &= E(Y|X_1, X_2) - E(E(Y|X_1, X_2)|X_2) \\ &= \big[f_1(X_1) + f_2(X_2)\big] - \big[E(f_1(X_1)|X_2) + f_2(X_2)\big] \\ &= f_1(X_1) - E(f_1(X_1)|X_2) \\ &= g(X_1, X_2), \end{aligned}$$

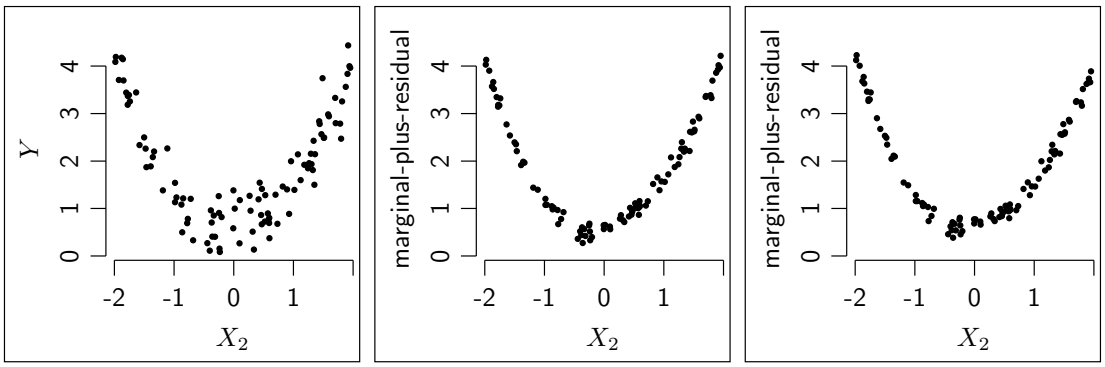
so an equivalent way to write (2.15) is

$$\{X_2, Y - [E(Y|X_1, X_2) - E(Y|X_2)]\}. \tag{2.16}$$

Example 2.5 (Marginal-plus-residual plot)
This example uses the simulated data from Example 2.3, where $f_1(X_1) = \beta_1 X_1$, $f_2(X_2) = I(X_2 < 0)X_2^2$, and $X_1 = \exp(X_2 + 2)/25 + 0.1\epsilon$.

A first view of the marginal conditional mean of $Y|X_2$ is shown in a marginal response plot, Figure 2.5a. The relationship appears close to quadratic, though there is enough variation that further details could be masked.

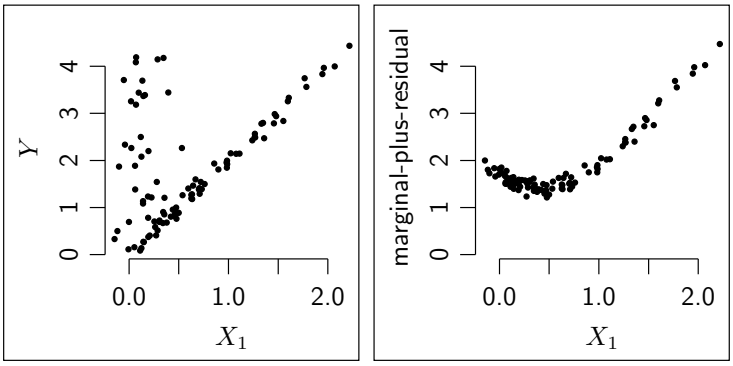
The marginal-plus-residual plot removes variation that can be explained by the



(a) Marginal response plot for X_2.

(b) Idealized marginal-plus-residual plot for X_2.

(c) Estimated marginal-plus-residual plot for X_2.



(d) Marginal response plot for X_1.

(e) Estimated marginal-plus-residual plot for X_1.

Figure 2.5: Marginal plots from Example 2.5.

other variables, so may allow a closer look at the conditional mean function. In this example, the linearity of f_1 allows the idealized marginal-plus-residual plot for X_2 to be calculated. From (2.14),

$$\begin{aligned} g(X_1, X_2) &= f_1(X_1) - E(f_1(X_1)|X_2 \\ &= \beta_1 X_1 - E(\beta_1 X_1|X_2) \\ &= \beta_1 X_1 - \beta_1 E(X_1|X_2) \\ &= \beta_1 X_1 - \beta_1 \exp(X_2 + 2)/25. \end{aligned}$$

Figure 2.5b shows the idealized marginal-plus-residual plot created using this g.

An estimated version of the marginal-plus-residual plot, shown in Figure 2.5c, was calculated using the second form of g, as used in (2.16), where

$$g(X_1, X_2) = E(Y|X_1, X_2) - E(Y|X_2).$$

These conditional means were estimated using the lowess smoother, and the fitted values used to create the estimated marginal-plus-residual plot. This estimated version matches the idealized version well, demonstrating that the two forms of g available are equivalent.

The role that a marginal-plus-residual plot can play is clearer when X_1 is considered to be the variable of interest. Again, the marginal response plot, shown in Figure 2.5d, is a good first view of the marginal relationship. But here, there is significant variation for small X_1. Figure 2.5e shows an estimated marginal-plus-residual plot for X_1, calculated in the same way as the estimated plot for X_2. In this plot, the extra variation around small X_1 has been removed, as it was explainable by the other variables, allowing the shape of the marginal mean function to be clearer than in the marginal response plot. □

2.3 Added-variable plot

Returning to the original goal, g and h must be chosen to make $E(Y^*|X_1, X_2)$ a function of only $X^* = h(X_1, X_2)$, where, as in (2.3),

$$E(Y^*|X_1, X_2) = f_1(X_1) + f_2(X_2) - g(X_1, X_2). \tag{2.17}$$

Similar to calculations done for the marginal-plus-residual plot, by the rule of iterated expectations, $E(Y|X_1) = f_1(X_1) + E(f_2(X_2)|X_1)$. Solving for f_1 and substituting into (2.17),

$$E(Y^*|X_1, X_2) = E(Y|X_1) - E(f_2(X_2)|X_1) + f_2(X_2) - g(X_1, X_2). \tag{2.18}$$

Then setting $g(X_1, X_2) = E(Y|X_1)$, and $h(X_1, X_2) = f_2(X_2) - E(f_2(X_2)|X_1)$ satisfies the desired condition, as $E(Y^*|X^*) = X^*$, and results in the plot of

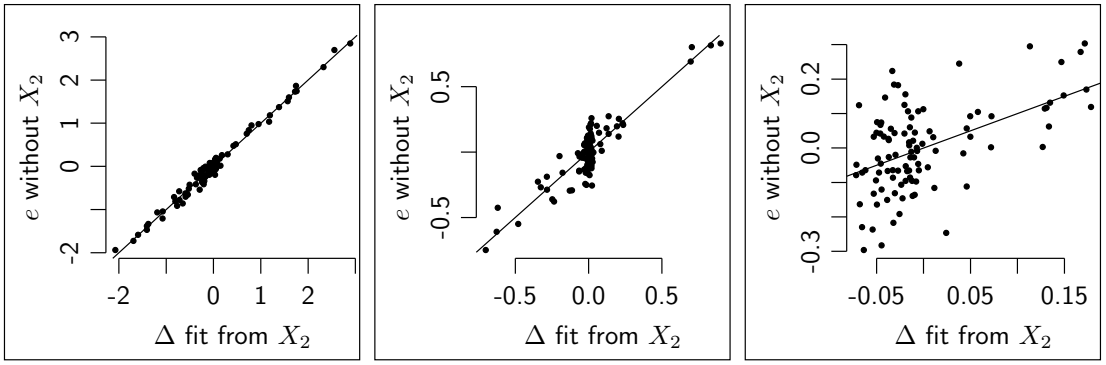
$$\{f_2(X_2) - E(f_2(X_2)|X_1), Y - E(Y|X_1)\}. \tag{2.19}$$

As shown for the marginal-plus-residual plot, $f_2(X_2) - E(f_2(X_2)|X_1)$ is equal to the change in fitted values between a model with only X_1 and a model with both X_1 and X_2, so this plot can equivalently be written as

$$\{E(Y|X_1, X_2) - E(Y|X_1), Y - E(Y|X_1)\}. \tag{2.20}$$

The x-axis can be interpreted as the change in fitted values when adding X_2 to the model, and the y-axis can be interpreted as the amount unexplained by X_1, so this plot is a direct generalization of the general added-variable plot of Section 1.6.2.

Because $E(Y^*|X^*) = X^*$, the plot will always have a slope of 1, so the relative importance of the predictor X_2 cannot be determined by the slope. Instead, the x



(a) For X_2, when $\hat{\rho} = 0.73$.

(b) For X_2, when $\hat{\rho} = 0.99$.

(c) For X_2, when f_2 is reduced by 95%.

Figure 2.6: General added-variable plots for X_2, from Example 2.6, showing how correlated data and magnitude of f_2 affect the plot.

and y range of the plot holds clues to how important a predictor is, as in Example 2.6.

Example 2.6 (General added-variable plot)

This example uses the simulated data from Example 2.1, where $f_1(X_1) = \beta_1 X_1$, $f_2(X_2) = I(X_2 < 0)X_2^2$, and $X_1 = (0.5X_2 + 1) + 0.5\epsilon$. Figure 2.6a shows an estimated general added-variable plot for X_2, where $E(Y|X_1)$ and $E(Y|X_1, X_2)$ were estimated with a nonparametric smoother. The points in this plot are all very close to a line of slope 1. So for instance, if the amount unexplained by X_1 for a certain point is +2, then the amount that X_2 changes the fitted value by is close to +2, so adding X_2 removes most of the variability remaining in the model.

To demonstrate further how this added-variable plot can show the importance of adding a term to the model, the example has been tweaked in two ways, and the added-variable plot reconstructed.

The first change is to let $X_1 = (0.5X_2 + 1) + 0.1\epsilon$, which increases the sample correlation between X_1 and X_2 to 0.99, X_2 will have less new information about Y. The added-variable plot is shown in Figure 2.6b. The points are still near the line of slope 1, but not as close as in Figure 2.6a, so adding X_2 removes less of the variability

that remains in the model. Also both the horizontal and vertical ranges of the plot are smaller, showing that there is less variability unexplained by X_1 and also less explained by adding X_2.

The second change is to reduce the effect of X_2 on Y by 95% by multiplying f_2 by 0.05, so adding X_2 will have a smaller effect. The added-variable plot is shown in Figure 2.6b. There are two changes in the plot; first, the vertical range is substantially smaller, showing that without the added variation from X_2, X_1 can explain a larger part of the variation in Y. Secondly, the points do not fall as clearly on a line with slope 1, so adding X_2 is not as able to explain the variation that does remain in the model. □

2.3.1 Linear f_2

When f_2 is linear, say $f_2(X_2) = \beta_2 X_2$,

$$\begin{aligned} f_2(X_2) - E(f_2(X_2)|X_1) &= \beta_2 X_2 - E(\beta_2 X_2|X_1 \\ &= \beta_2(X_2 - E(X_2|X_1). \end{aligned}$$

So in this case, $X_2 - E(X_2|X_1)$ can be put on the x-axis, as in the plot

$$\{X_2 - E(X_2|X_1), Y - E(Y|X_1)\}, \tag{2.21}$$

which will have slope β_2. This version of the plot is a generalization of the standard added-variable plot introduced in Section 1.6, and was suggested by Cook (1995) as a more general way to combine local net-effect plots. It has the same interpretation as the multivariate normal version; it shows how the information left unexplained in the response after fitting X_1 is related to new information in X_2 beyond that in X_1.

Estimating $E(X_2|X_1)$ or $E(Y|X_1)$ accurately can be important, especially when

they are nonlinear. This can happen even when the true model is linear,

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon, \tag{2.22}$$

as this model does not specify anything about the relationship between the predictors or about $E(Y|X_1)$. When these expectations are estimated using a linear fit when they are not really linear, there will be more information about Y left unexplained by X_1 and more new information in X_2 beyond X_1, so the added-variable plot will show X_2 to be more important than it really is (Cook, 1995, p. 267).

A plot using linear fits in this context will satisfy (2.2), and remove the conditional dependence on X_1. But the interpretation is different than the usual added-variable plot, as it only shows the effect of adding the new variable X_2 to a linear fit on X_1. If the linear fit on X_1 is not appropriate, this may have little meaning.

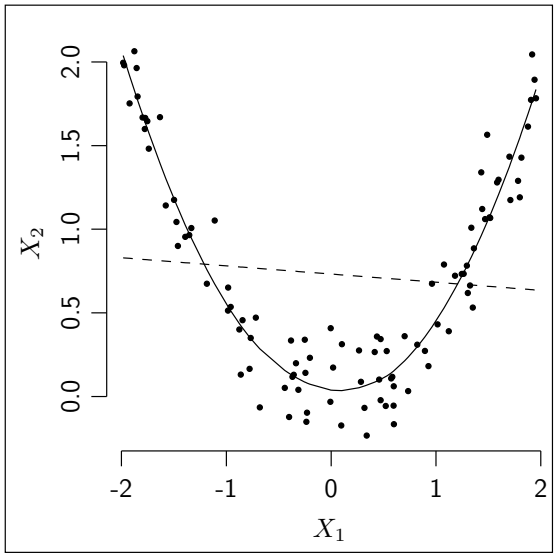
Example 2.7

For this example, 100 data points were simulated from the linear model of (2.22), with $\beta_1 = \beta_2 = 1$, and a nonlinear relationship between X_1 and X_2, where $X_1 \sim \text{Unif}(-2, 2)$, and

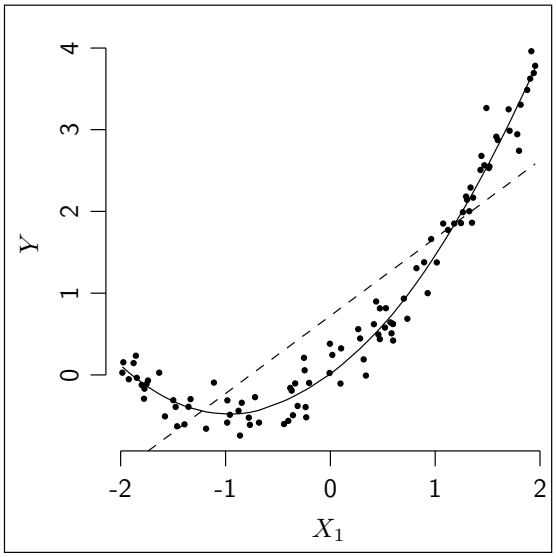
$$X_2 = X_1^2 + 0.2\epsilon.$$

Then to show that simply using linear estimates is not as precise, added-variable plots for X_2 were estimated using both nonparametric and linear fits. Figure 2.7 shows the fits that were used, as well as the two added-variable plots. The added-variable plots were sliced over X_1, and individual fitted lines drawn for each slice. The lines overlap well for each plot, showing that the dependence on X_1 has been removed for each plot and both plots meet the required criteria.

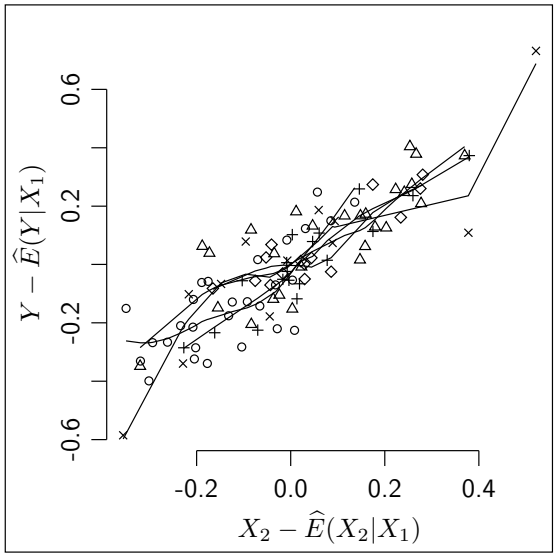
Additionally, the added-variable plot using linear fits shows adding X_2 to be significantly more valuable than the added-variable plot using the nonlinear fits. This



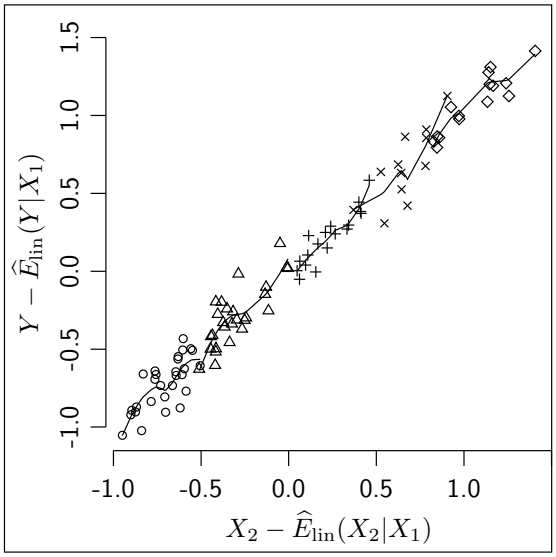
(a) Linear and nonparametric fits of X_1 on X_2.



(b) Linear and nonparametric fits of Y on X_1.



(c) Added-variable plot for X_2, using a nonparametric fit.



(d) Added-variable plot for X_2, using a linear fit.

Figure 2.7: Marginal fits and added-variable plots of the data in Example 2.7, using linear and non-parametric fits. The added-variable plot using the linear fits overstates the importance of X_2. In the added-variable plots, points with similar X_1 values were given similar point shapes, and fitted lines were drawn through these points using a loess smoother, showing that both added-variable plots do remove conditional X_1 dependence.

is true; it is more important when adding it to only a linear fit. But a linear fit is a poor model, so it overstates the true effect of adding X_2 to the analysis. □

2.3.2 Nonlinear f_2

When f_2 is nonlinear and there is a curve in the variable of interest, the plot

$$\{X_2 - E(X_2|X_1), Y - E(Y|X_1)\} \tag{2.23}$$

is not as useful, especially for finding the form of f_2. This has been analyzed by many, including Cook (1996), Berk and Booth (1995), Landwehr (1986), Johnson and McCulloch (1987). Landwehr and Pregibon (1993) describe the problem as a "jumbling" the x-axis, so the values of X_2 are no longer in order. That is, because the x-axis consists of $X_2 - E(X_2|X_1)$ instead of X_2, it will not accurately show a curve in terms of X_2.

Similar reasoning is provided by Landwehr (1986) (as cited by Berk and Booth, 1995), who proved when using linear fits, an added-variable plot for X_2 is equivalent to a partial residual plot for $X_2 - E(X_2|X_1)$. While under certain conditions partial residual plots do accurately show the shape of the curve, this shows that in general this version of the added-variable plot shows the curve of $X_2 - E(X_2|X_1)$, not the curve of X_2.

Cook (1996) formalized this, determining that the amount of collinearity between the predictors determines how useful the added-variable plot is for showing the curve. When there is no collinearity and the predictors are independent, $X_2 - E(X_2|X_1) = X_2$, so X_2 is on the x-axis. But as the collinearity increases and $X_2 - E(X_2|X_1)$ and X_2 become more different, X_2 is no longer represented on the x-axis.

As the problem seems to be on the x-axis, several authors have suggested simply

using X_2 on the x-axis,

$$\{X_2, Y - E(Y|X_1)\} \tag{2.24}$$

including Draper and Smith (1981), Mosteller and Tukey (1977) and Atkinson (1985). According to Berk and Booth (1995), Landwehr (1986) showed that this modified added-variable plot does not show the correct slope when X_2 is linear.

This analysis adds another reason that these two plots are not useful for showing the form of f_2, which is that both of them do not correctly remove the conditional dependence on X_1. From (2.3),

$$f_1(X_1) + f_2(X_2) - g(X_1, X_2) \tag{2.25}$$

must be a function only of X^*, the value on the x-axis. For the regular added-variable plot of (2.23), $g(X_1, X_2) = E(Y|X_1)$, and $X^* = X_2 - E(X_2|X_1)$. Then

$$\begin{aligned} f_1(X_1) + f_2(X_2) - g(X_1, X_2) &= f_1(X_1) + f_2(X_2) - E(Y|X_1) \\ &= f_1(X_1) + f_2(X_2) - [f_1(X_1) + E(f_2(X_2)|X_1)] \\ &= f_2(X_2) - E(f_2(X_2)|X_1) \end{aligned}$$

which can only be written as a function of $X_2 - E(X_2|X_1)$ if f_2 is linear. For the modified added-variable plot of (2.24), $X^* = X_2$, so this should be a function of X_2; this is only true if $E(f_2(X_2)|X_1)$ is zero, which may not be true even when f_2 is linear.

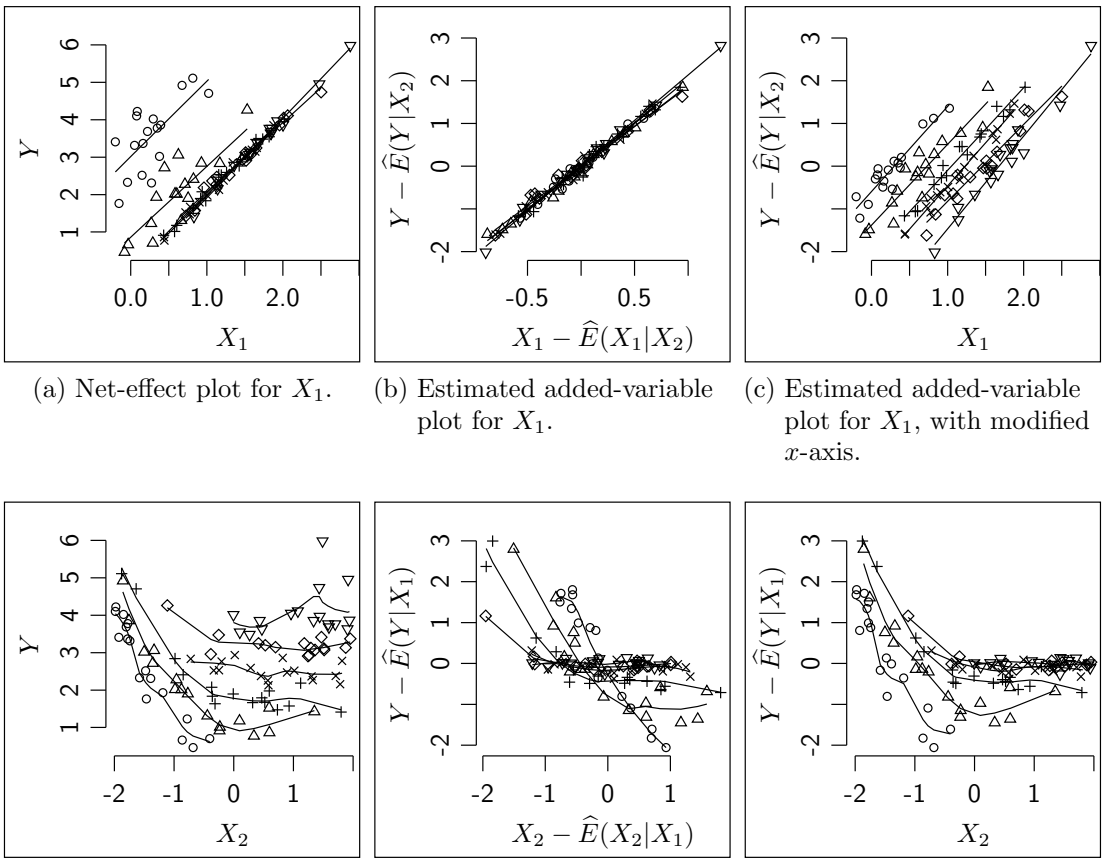
Example 2.8 (Added-variable plots and nonlinear f_2)
This example uses the data from Example 2.3, where $E(X_2|X_1) = \exp(2+X_2)/25$ and $E(Y|X_1, X_2) = 2X_1 + I(X_2 < 0)X_2^2$. X_1 enters the model linearly, so added-variable plots for X_1 should successfully remove conditional dependence on X_2, but X_2 enters

the model nonlinearly, so the conditional dependence on X_1 should not be removed.

Figure 2.8 shows net-effect plots, added-variable plots, and added-variable plots with modified x-axis for both X_1 and X_2. All expected values were estimated using the lowess nonparametric smoother.

The net-effect plots for X_1 shows that the conditional mean for each slice of X_2 is linear, with nearly the same slope. So the added-variable plot is able to combine the slices by centering around their means in both directions and thus remove conditional dependence on X_2. But the modified added-variable plot, which only centers the slices around their mean in the vertical direction, fails to remove this dependence.

In the net-effect plots for X_2, the shape of the conditional mean function depends on X_2, so the structure is not identical except for the mean. Thus the added-variable plot, which centers each around the mean, still has conditional X_1 dependence. Finally, although the modified added-variable plot is better, the conditional mean of each slice is different. This is because $E(f_2(X_2)|X_1)$, which equals $E(I(X_2 < 0)X_2^2|X_1)$ in this example, is non-zero. □



(a) Net-effect plot for X_1.

(b) Estimated added-variable plot for X_1.

(c) Estimated added-variable plot for X_1, with modified x-axis.

(d) Net-effect plot for X_2.

(e) Estimated added-variable plot for X_2.

(f) Estimated added-variable plot for X_2, with modified x-axis.

Figure 2.8: Various added-variable plots for data with nonlinear f_2, from Example 2.8.

2.4 Alternate forms

The three main plots in this chapter, the ordinary added-variable plot, the component-plus-residual plot, and the marginal-plus-residual plot, can each be written in an alternate form, based on the following more general form. Remember that the goal is to define X^* and Y^* so

$$Y^*|(X^*, X_1) \sim Y^*|X^*. \tag{2.26}$$

Then choosing

$$X^* = X_2 - g(X_1) \quad \text{and} \tag{2.27}$$

$$Y^* = Y - E\left(Y|X_1, X_2 = g(X_1)\right) \tag{2.28}$$

will satisfy this requirement when either X_2 enters the model linearly, or g is a constant and the model is additive. This form will be used directly to derive the added-variable plot and the component-plus-residual plot; a similar idea will be used for the marginal-plus-residual plot.

2.4.1 Added-variable plot

When X_2 enters the model linearly, $E(Y|X_1, X_2) = f_1(X_1) + \beta_2 X_2$, so this becomes

$$\begin{aligned}
E(Y^*|X_1, X^*) &= E\left[Y - E\left(Y|X_1, X_2 = g(X_1)\right)|X_1, X^* = X_2 - g(X_1)\right] \\
&= E\left(Y|X_1, X_2 = g(X_1) + X^*\right) - E\left(Y|X_1, X_2 = g(X_1)\right) \\
&= f_1(X_1) + \beta_2\left(g(X_1) + X^*\right) - f_1(X_1) - \beta_2 g(X_1) \\
&= \beta_2 X^*,
\end{aligned}$$

and dependence on X_1 is removed. The ordinary added-variable plot results when $g(X_1) = E(X_2|X_1)$.

The general added-variable plot can also be made using this alternate form, using

$$\{E(Y|X) - E(Y|X_1, X_2 = g(X_1)), Y - E(Y|X_1, X_2 = g(X_1))\},$$

which has the change in predicted values using the actual values of X_2 and $X_2 = g(X_1)$ on the x-axis, and the residuals from the predictions using $X_2 = g(X_1)$ on the y-axis. With $g(X_1) = E(X_2|X_1)$, this equals the standard general added-variable plot when X_2 enters the model linearly.

2.4.2 Component-plus-residual plot

Consider an additive model $E(Y|X_1, X_2) = f_1(X_1) + f_2(X_2)$, and constant $g(X_1) = k$. Then $X^* = X_2 - k$ and $Y^* = Y - E(Y|X_1, X_2 = k)$. Then

$$\begin{aligned} E(Y^*|X_1, X^*) &= E(Y|X_1, X_2 = X^* + k) - E(Y|X_1, X_2 = k) \\ &= f_1(X_1) + f_2(X^* + k) - f_1(X_1) - f_2(k) \\ &= f_2(X^* + k) - f_2(k), \end{aligned}$$

which can be written instead as

$$E(Y^*|X_2) = f_2(X_2) + C$$

which not only removes the dependence on X_1, but is the standard component-plus-residual plot. Because in an additive model,

$$f_2(X_2) + \epsilon = Y - f_1(X_1),$$

the above method can also directly retrieve the component for X_1, for

$$E(Y|X_1, X_2 = k) = f_1(X_1) + f_2(k)$$

which can be written as $f_1(X_1) + C$ and so is the component for X_1, with no residuals.

2.4.3 Marginal-plus-residual plot

A version of the marginal-plus-residual plot can be written in a similar manner. The original marginal-plus-residual plot for X_2 has X_2 on the x-axis and $E(Y|X_2) + \epsilon$ on the y-axis. Using the framework of building plots using

$$E(Y|X_1, X_2 = g(X_1)),$$

set $g(X_1) = E(X_2|X_1)$ and instead put

$$E\left(Y|X_1, X_2 = E(X_2|X_1)\right) + \epsilon$$

on the y-axis. The ϵ term could also be removed to simply plot the line given by the expected value.

When the model is additive,

$$E\left(Y|X_1, X_2\right) = f_1(X_1) + f_2(X_2);$$

this alternate marginal expectation is

$$E\left(Y|X_1, X_2 = E(X_2|X_1)\right) = f_1(X_1) + f_2(E(X_2|X_1)). \tag{2.29}$$

This is not equal to the actual marginal expectation

$$E(Y|X_1) = f_1(X_1) + E(f_2(X_2)|X_1), \tag{2.30}$$

so this plot does not show the actual marginal relationship between X_1 and Y, but instead an alternate marginal relationship: the mean function of the response against X_1 when the other variables equal their predictions given X_1 alone. Other options for $g(X_1)$ could also be considered, such as the median or other quantiles of interest. Only the behavior for $g(X_1) = E(X_2|X_1)$ will be investigated here.

This alternate marginal mean relationship may be of interest because it shows how the response Y depends on the variable X_1, using only the data from X_1, but including any knowledge about how the other variables behave given X_1 and the fitted model reflected through $\widehat{E}(Y|X_1, X_2)$. As will be seen, at times this will be equal to the actual marginal function; at times it will be equal to the component from X_1, and at times it will be some compromise between them. Additionally, this function is straightforward to calculate because it depends only on one-dimensional smooths of each other variable given X_1, and a model that can produce predicted values from the resulting data.

When X_2 enters the model linearly, so $f_2(X_2) = \beta_2 X_2$, the actual and alternate marginal relationships are equal, as

$$f_2(E(X_2|X_1)) = \beta_2 E(X_2|X_1) = E(\beta_2 X_2|X_1) = E(f_2(X_2)|X_1).$$

But when f_2 is not linear, they are not equal. Their relationship can be investigated further by using the Taylor approximation, called the delta method (Oehlert, 1992). For general θ,

$$f_2(X) = f_2(\theta) + f_2'(\theta)(X - \theta) + f_2''(\theta)(X - \theta)^2/2 + \cdots .$$

Letting $\theta = EX$ and taking the expectation of both sides,

$$Ef_2(X) = f_2(EX) + f_2''(EX)\operatorname{Var}(X)/2 + \cdots .$$

With X_2 as X and $\theta = E(X_2|X_1)$, and taking the expectation with respect to X_1,

$$E\left(f_2(X_2)|X_1\right) = f_2(E(X_2|X_1)) + f_2''(E(X_2|X_1))\operatorname{Var}(X_2|X_1)/2 + \cdots .$$

If the remainder of the terms are small, the difference between $E(f_2(X_2)|X_1)$ and $f_2(E(X_2|X_1))$ is approximately equal to

$$f_2''(E(X_2|X_1))\operatorname{Var}(X_2|X_1)/2$$

and depends only on $f_2''(E(X_2|X_1))$ and $\text{Var}(X_2|X_1)$. This is true even if X_2 is multivariate. A few examples will illustrate this dependence.

Example 2.9 (Linear f_2)

Let X_1 be a sequence of 1000 points evenly spaced between -1 and 1, and consider three possibilities for X_2. First, with constant conditional mean and variance: $X_2 \sim \text{Unif}(-1, 1)$. Second, with changing conditional mean, but constant variance: $X_2 \sim \text{Unif}(X_1 - 2, X_1 + 2)$. Third, with constant conditional mean, but changing variance: $X_2 \sim \text{Unif}(-X_1 - 1, X_1 + 1)$. Pictures of these three relationships are shown in Figure 2.9a. Let the response variable be

$$Y = -X_1 + 2X_2.$$

For clarity, no error will be added. Here $f_2(X_2) = 2X_2$, so f_2 is linear and $f_2''(\cdot) = 0$, so there is no difference between the actual and alternate marginal relationship. This is true whether or not $E(X_2|X_1)$ and $\text{Var}(X_2|X_1)$ are constant. Marginal response plots for X_1 for each of the three sets of predictors are shown in Figure 2.9b, with a solid line added to show both the alternate and actual marginal relationships. □

Example 2.10 (Quadratic f_2)

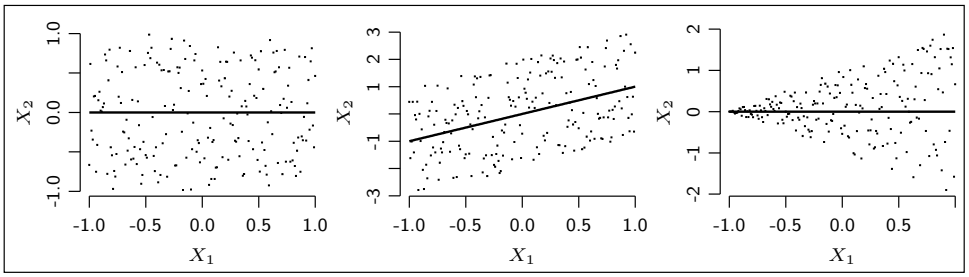
Now consider a response variable

$$Y = -X_1 + X_2^2.$$

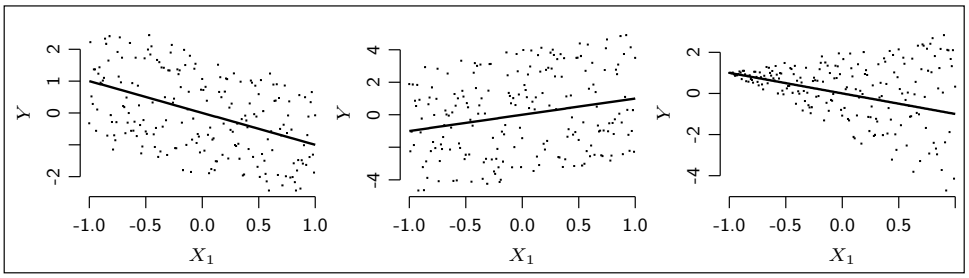
Again for clarity, no error will be added. Here $f_2(X_2) = X_2^2$, so f_2 is quadratic and $f_2''(\cdot) = 2$. Here the difference between the true relationships is exactly

$$f_2''(E(X_2|X_1))\,\text{Var}(X_2|X_1)/2;$$

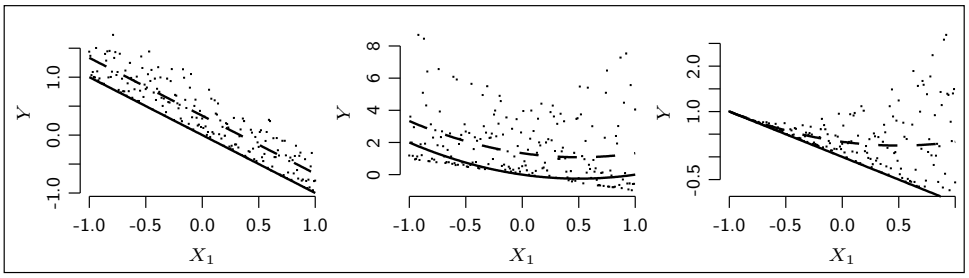
the additional terms are zero because the additional derivatives of f_2 are zero. So for quadratic f_2, a changing $E(X_2|X_1)$ does not affect the difference, but a changing



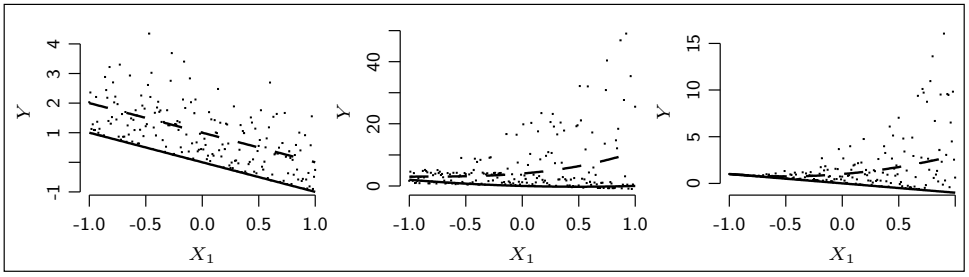
(a) Three predictor relationships: constant conditional mean and variance, changing mean but constant variance, and constant mean but changing variance.



(b) Marginal relationship for linear f_2.



(c) Marginal relationship for quadratic f_2.



(d) Marginal relationship for cubic f_2.

Figure 2.9: Marginal relationships for linear, quadratic, and cubic f_2, for three possible predictor relationships. The dotted line shows the actual marginal relationship; the solid line shows the alternate marginal relationship.

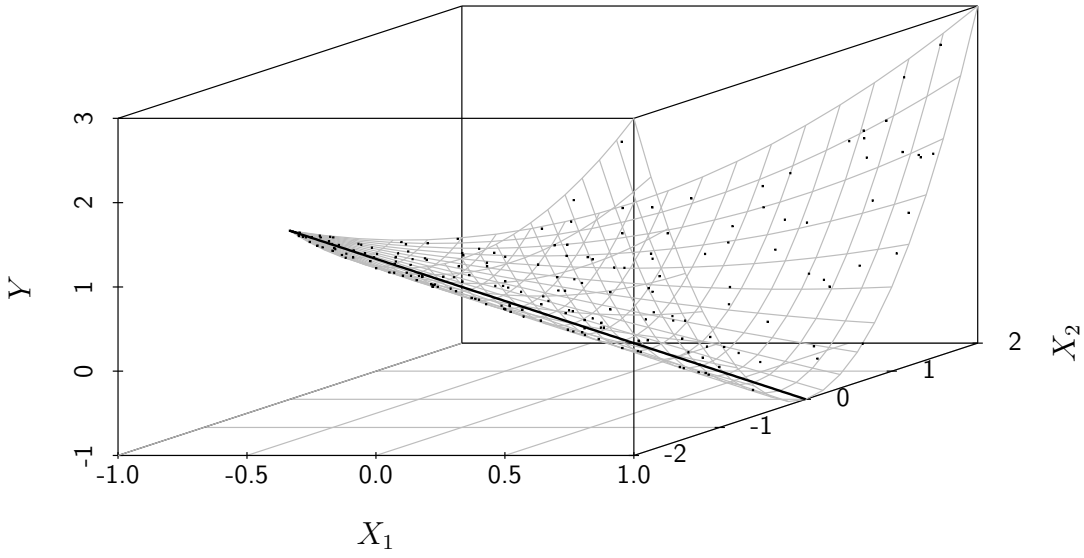


Figure 2.10: Plot of the data set with nonconstant variance and quadratic f_2, from Example 2.10, with the alternate marginal function added.

variance does. A constant variance results in a constant difference, while a changing variance will change the difference.

Plots of the marginal relationship between X_1 and Y for each of the three sets of predictors given in Example 2.9 are shown in Figure 2.9c, with lines added to show the alternate (solid line) and actual (dotted line) marginal relationships. The difference between them is constant in the first two plots, where the variance is constant, but changes in the third plot, where the variance is not constant.

The cause of the difference in the third plot is perhaps clearer in a three dimensional plot, shown in Figure 2.10. The line at $X_2 = E(X_2|X_1) = 0$ is always at the bottom of the quadratic curve, so it goes down as X_1 increases, even as the curve causes the true marginal relationship to increase as X_1 increases. □

Example 2.11 (Cubic f_2)

Finally, consider a response variable

$$Y = -X_1 + 3X_2^2 + X_2^3.$$

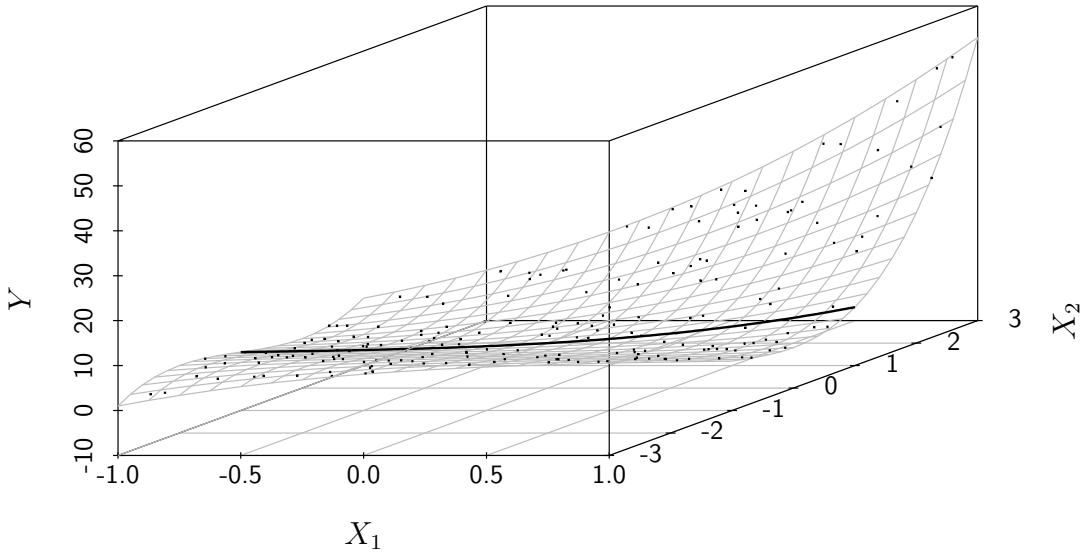


Figure 2.11: Plot of the data set with nonconstant variance and quadratic f_2, from Example 2.11, with the alternate marginal function added.

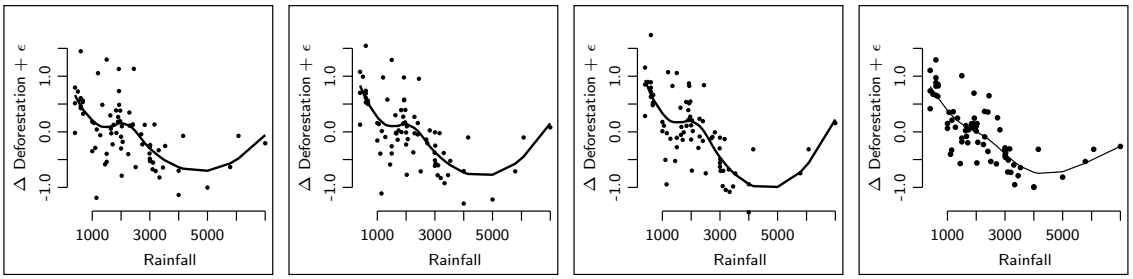
Again for clarity, no error will be added. Here f_2 is cubic and $f_2''(\cdot) = 6X_2 + 6$. The difference between the true relationships is again exactly

$$f_2''(E(X_2|X_1))\,\mathrm{Var}(X_2|X_1)/2.$$

The next term contains the third central moment, which is zero because the distribution of $X_2|X_1$ is symmetric, and the additional terms are zero because additional derivatives of f_2 are zero. So in this case, a changing $E(X_2|X_1)$ does affect the difference, as well as a changing variance.

Marginal response plots for X_1 for each of the three sets of predictors given in Example 2.9 are shown in Figure 2.9d, with lines added to show the alternate (solid line) and actual (dotted line) marginal relationships The difference between them is constant only in the first plot, where the conditional mean and variance are both constant. It changes in the both the second plot, where the mean is not constant, and the third plot, where the variance is not constant.

The cause of the difference in the second plot is investigated again in a three-



(a) Using linear fit. (b) Using quadratic fit. (c) Using CERES. (d) Using `gam`.

Figure 2.12: Component-plus-residual plots for Rainfall, using four methods.

dimensional plot, shown in Figure 2.11. The alternate mean function is the line shown, at $X_2 = E(X_2|X_1)$. It does increase with X_1, but because the function is cubic, the actual marginal mean function increases at a greater rate. □

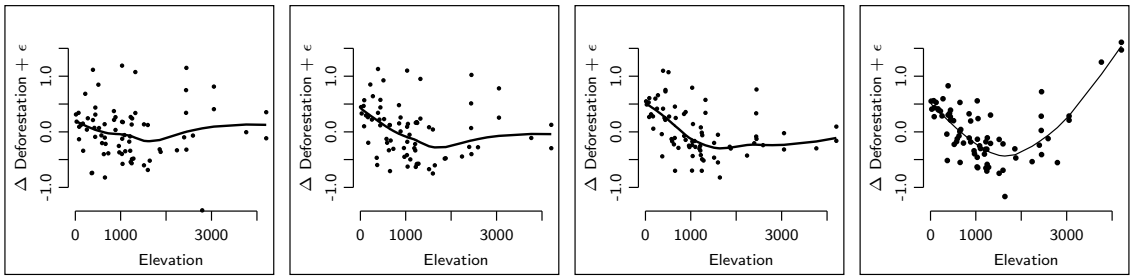
Finally, these examples only explore the case where X_1 and X_2 are additive. If there is any interaction, it can be more complex.

2.5 Example: Island deforestation

Previously, plots for the island deforestation data were made assuming that each variable entered the model linearly. While several of the variables were log transformed to better meet this assumption, no systematic analysis was made, so it is possible that those plots are not as accurate as they could be, so using the methodology in this chapter, these plots will now be revisited using non-parametric additive methods. As the relationships are no longer required to be linear, Rainfall and Elevation will no longer be log-transformed.

2.5.1 Component-plus-residual plots

Figure 2.12 and Figure 2.13 show component-plus-residual plots for Rainfall and Elevation, respectively, estimated using four different methods; a partial residual plot, an



(a) Using linear fit. (b) Using CERES. (c) Using CERES after Rainfall fit. (d) Using `gam`.

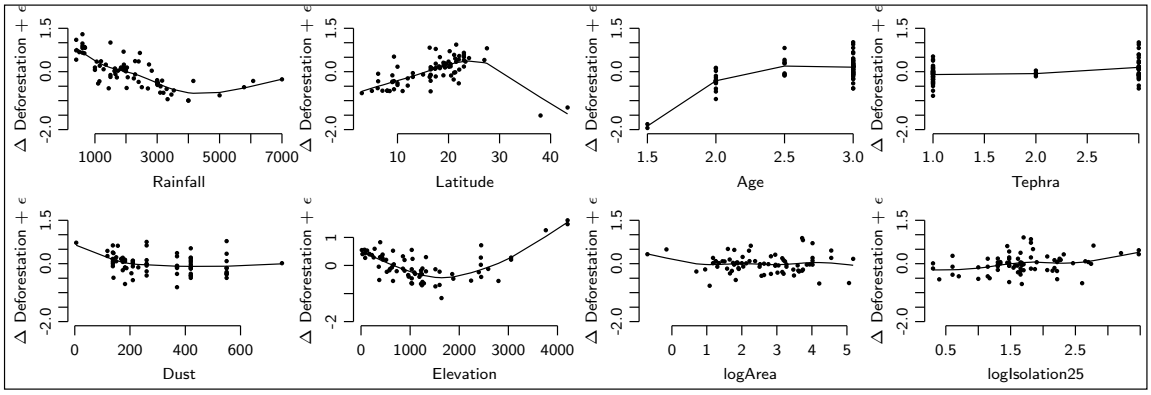
Figure 2.13: Component-plus-residual plots for Elevation, using several methods.

augmented partial residual plot, a CERES plot, and a plot using the `gam` estimates. For Rainfall, each of the plots is roughly the same and show a nonlinearity.

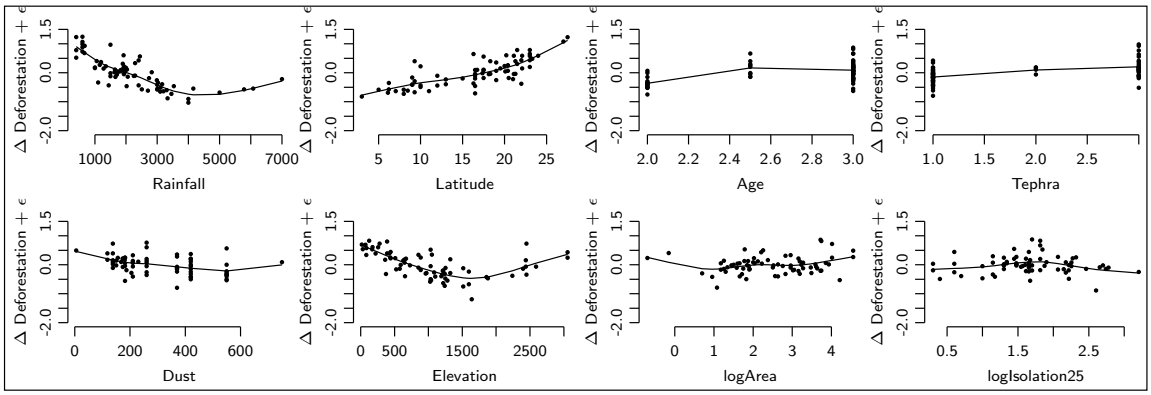
However, for Elevation, the plots for the `gam` estimate are distinctly different. The first three methods all depend on only one term being nonlinear, while the `gam` fitting method allows multiple terms to be nonlinear. A idea that sometimes works with CERES plots when multiple predictors are thought to be nonlinear is to fit CERES plots for all the terms, and then refit the model after subtracting the effect of the term with the largest curve from the response. In this case, Rainfall had the largest curve; the CERES plot after removing its effect is shown in Figure 2.13c. It may have slightly more curve than the first CERES plot, but is still quite unlike the plot from `gam`.

Figure 2.14a shows component-plus-residual plots for all variables from the `gam` fit, allowing the effect of each variable to be visually compared. Rainfall, Latitude, Age, and Elevation seem to have more of an effect than the other predictors.

However, it also seems that several points may have especially large influences; two points with high latitudes, two points with Age equal to 1.5 and three points with the high elevations. These points represent only four sites; two on Hawaii and two on New Zealand. Specifically, the two Hawaii points have both low Age and high Elevation; and according to the fit, the Age causes a large decrease, while Elevation



(a) For all variables from the `gam` fit.



(b) For all variables from the `gam` fit, but with four outliers removed.

Figure 2.14: Component-plus-residual plots for all variables in the island deforestation data set, using an additive model.

causes a large increase. It is likely that the model is overfitting to these points; component-plus-residual plots without them are shown in Figure 2.14b. The fit does not significantly change, but the visual assessment of the comparative effect does; Age and Elevation now seem to be less significant than they did in the previous plots, as the vertical change over the displayed range of the predictor is smaller. Whether the fit for Age and Elevation is correct or not, this example does show how the measured range of each predictor can affect the visual assessment of importance.

2.5.2 Marginal-plus-residual plots

In addition to the component-plus-residual plots, which show the conditional effect of each variable, it can be helpful to show the marginal and the marginal-plus-residual plots, which show the marginal effects.

The regular marginal plots, shown in Figure 2.15a, simply plot the response against each variable in turn. Lines showing the marginal `gam` fit and alternate marginal mean functions for the linear and `gam` fits has also been added; they all show similar behavior here.

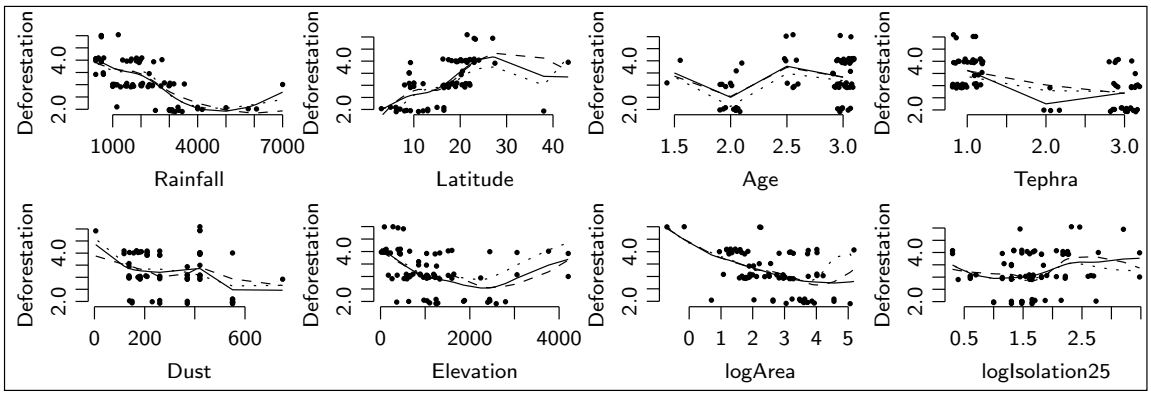
The marginal-plus-residual plots, shown in Figure 2.15b, have the same marginal line, but the scatter of points around the lines correspond to the residuals from the full model. This shows how large the marginal effect is compared to the residuals. For example, the marginal effects of Rainfall and Latitude seem strong, but logIsolation25 does not.

2.5.3 Added-variable plots

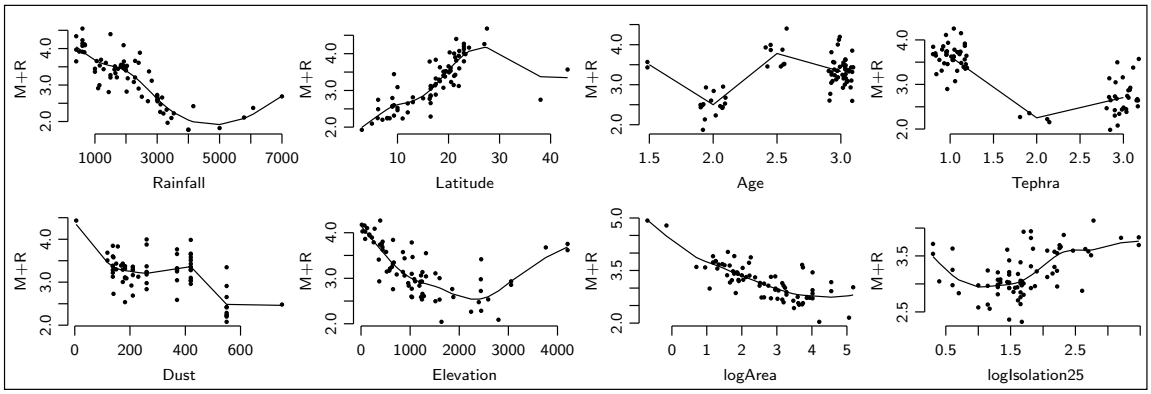
The added-variable plot is another way to show the importance of the various predictors, though instead of showing how much the response changes, it shows how much new information about the response is included in the predictor.

When the variables enter the model linearly, the standard added-variable plot for a given predictor plots the residuals from fitting the response against the other predictors against the residuals from fitting the given predictor against the others, and shows both the slope of the fit and the residuals. But when the variables do not enter linearly, this plot is not as useful, and instead the general added-variable plot should be used, which has the change in fitted values when the given predictor is removed from the model on the x-axis.

Figure 2.16a shows these added-variable plots for all eight variables, where all fits

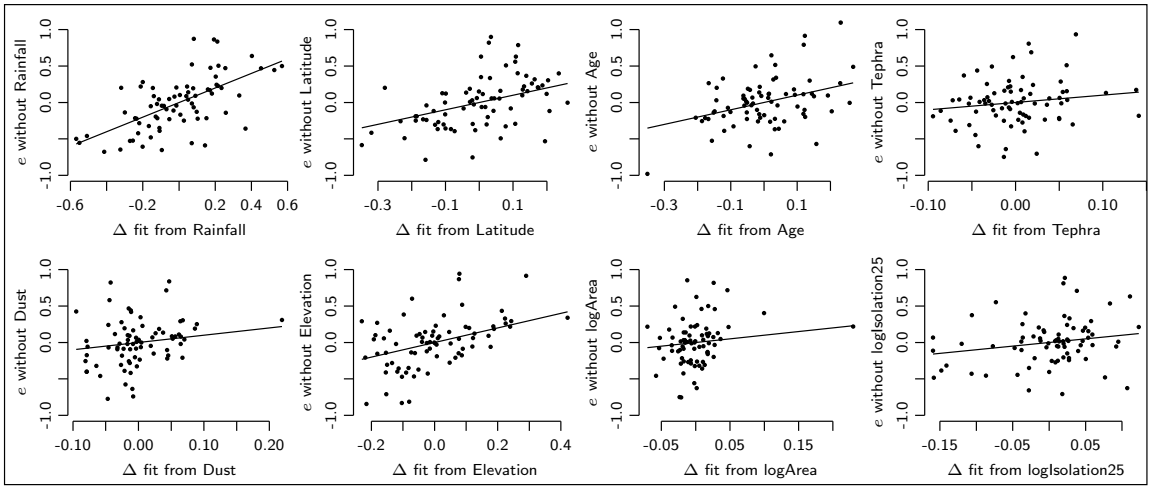


(a) Marginal plots, with marginal `gam` fits (solid lines) and alternate marginal mean functions for the linear fit (dashed lines) and additive fit (dotted lines) added.

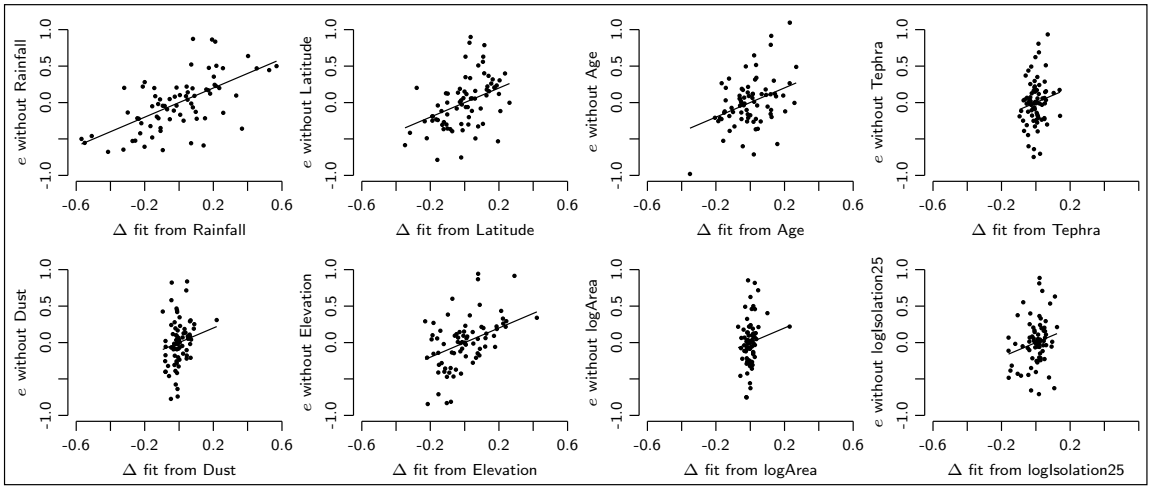


(b) Marginal-plus-residual plots.

Figure 2.15: Marginal and marginal-plus-residual plots for all variables in the island deforestation data set, using an additive model.



(a) General added-variable plots with different x-axis scales.



(b) General added-variable plots with a common x-axis scale.

Figure 2.16: Added-variable plots using the additive model fit with the `gam` library.

are performed using the `gam` library. The points in these plots are all centered around a line with a slope of one, which has been added to each plot. These plots help to show any points that may be outliers or have more influence; an example is in the plot for logArea, where there is one point farther to the right than the other points. This point corresponds to the Necker island of the Hawaii chain, which had the smallest area of the islands in the data set. For the other islands the change in fitted values due to logArea was usually no more than ± 0.1, but for this island it was over 0.2. It also falls almost precisely on the line with slope 1; this shows that the residual for this point without logArea is almost exactly equal to the change in fitted values when logArea is added. It is likely that he model is overfitting to this point.

A disadvantage of these plots as in Figure 2.16a is that it can be difficult to visually compare the effect of the eight predictors. Standardizing the scale of the x-axis, as in Figure 2.16b, can aid in this. Now the horizontal spread of the plots, which show how much the fit changed when each predictor was added, can be compared. Rainfall has the largest spread, and so in this sense, it is the most important. Conversely, logArea has very little spread, so is much less important.

Chapter 3

Model selection and combining

According to the criteria of (1.8), a good plot for showing how a predictor X_k is related to the response Y is $\{X^*, Y^*\}$, where

$$Y^*|(X^*, X_{\backslash k}) \sim Y^*|X^*,$$

so the conditional distribution of Y^* given X^* is independent of the other variables. The added-variable plot and the component-plus-residual plot both meet this criteria.

These two plots are especially useful because they show both the conditional relationship between X_k and Y and the residuals of the model. In making these plots, a distinction was made between *idealized* plots, which were constructed using the true, but unknown, population parameters, and *estimated* plots, which were constructed using estimates of these parameters. However, the effect of the method of estimation was not investigated, except for the specific case of CERES plots, but even then, the specific estimation technique was unspecified. Partly, this is because these plots work for any sensible estimation method, if the estimation method results in a model that meets the requirements of additivity and/or linearity, depending on the desired plot. Of course, the better the estimated values, the better the plot will be.

3.1 Correlated predictors

One of the causes of poor estimated values, especially values regarding the effect of a single predictor, can be correlation between the predictors. When this correlation exists, it creates uncertainty about which of the predictors should be associated with the effect.

Additionally, error in estimation with correlated predictors can significantly change the way these plots look. When predictors are uncorrelated and the effects are additive, the marginal effect of a given predictor X_k is equal to the conditional effect. That is, when

$$E(Y|X_k, X_{\setminus k}) = f_k(X_k) + g(X_{\setminus k}),$$

the conditional effect of X_k is described by f_k, the marginal effect of X_k is

$$E(Y|X_k) = f_k(X_k) + E(g(X_{\setminus k})|X_k).$$

When X_k and $X_{\setminus k}$ are uncorrelated, $E(g(X_{\setminus k})|X_k)$ is constant so the effect is still described by f_k. But when they are correlated in some way, this may nonconstant, and the marginal effect different from the conditional effect.

Example 3.1

Consider two data sets, each with 2 predictors and 100 data points. In the first, the two predictors are completely uncorrelated; in the second their correlation is 0.9. In both cases, the response Y equals $X_1 + 0.1X_2 + \epsilon$, where $\epsilon \sim N(0, 1)$, and the variance of the predictors is equal to one. The estimate of the coefficient for X_1 is similar in the two data sets (0.65 vs. 0.55), but the standard error is more than twice as large when the correlation is equal to 0.9 (0.3 vs. 0.69). Figure 3.1 shows confidence regions for these two coefficients for both models.

Figure 3.2 shows marginal and component-plus residual plots for X_2, for both the data set with uncorrelated predictors and the data set with correlated predictors.

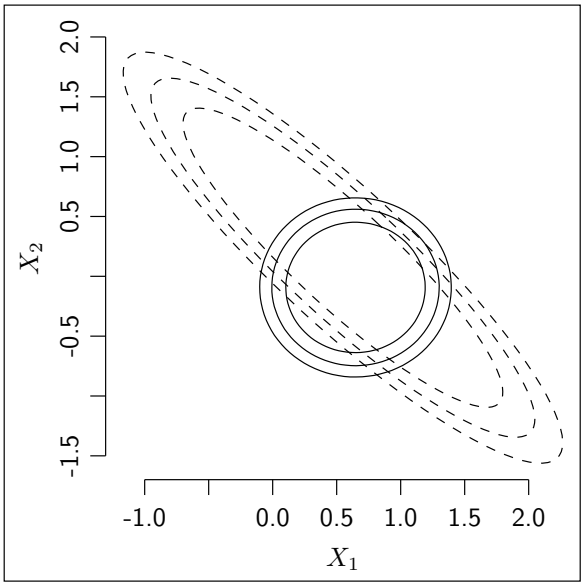
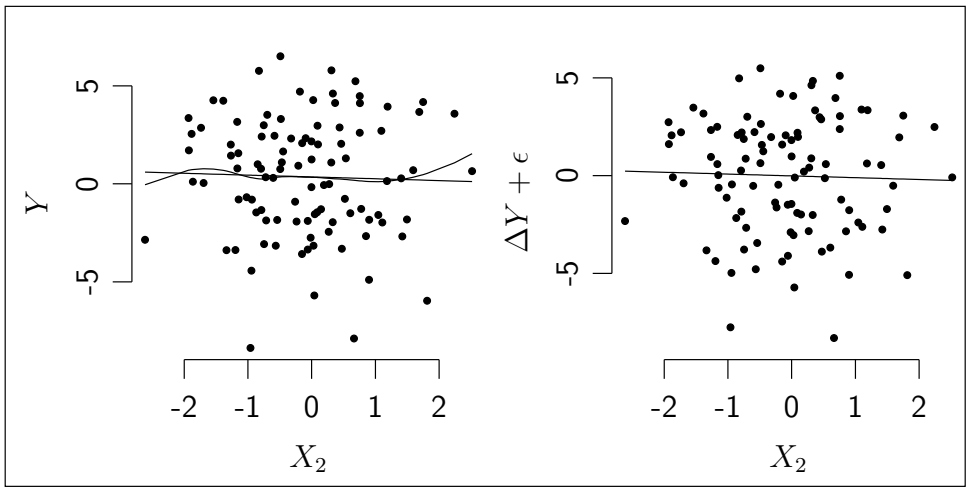


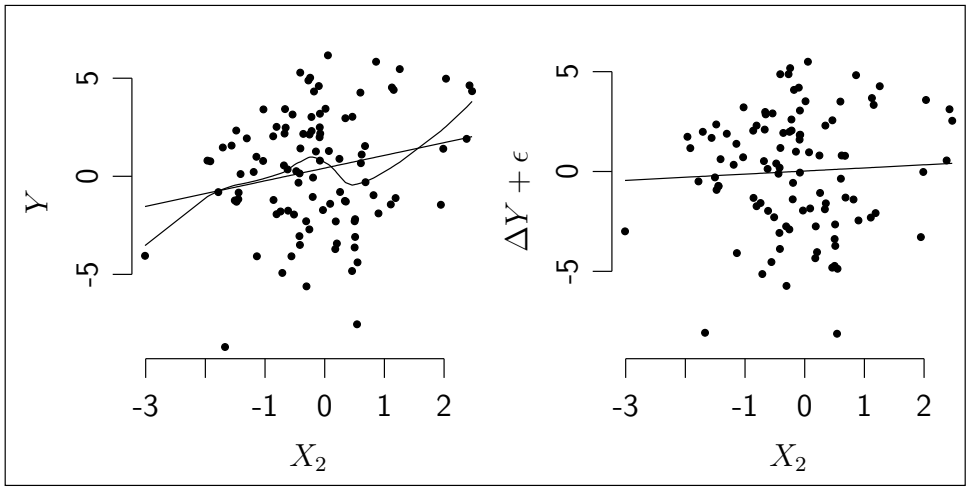
Figure 3.1: Confidence regions for $\alpha = 0.2$, 0.1, and 0.05, for coefficients from linear models using uncorrelated predictors (solid lines) and correlated predictors (dashed lines).

The plots for the data set with uncorrelated predictors have identical slopes, but the plots for the data set with correlated predictors have different slopes, as the marginal plot for X_2 includes not only the effect from X_2, but also part of the effect from X_1, because they are correlated. □

Although the following sections apply whether or not the predictors are correlated, the focus will be on cases with correlated predictors, as the possibility of misleading plots is increased. In particular, the focus will be on cases where there are many predictors, some of which probably have little to no effect on the response, but are correlated with those that do.



(a) For the data with uncorrelated predictors.



(b) For the data with correlated predictors.

Figure 3.2: Marginal and component-plus-residual plots for data with uncorrelated and correlated predictors, from Example 3.1.

3.2 Variable selection

In data sets with a large number of variables, often a better model can be obtained by leaving some of the variables out of the model altogether. The process of choosing which to leave out and which to leave in is called variable selection, and is a special case of a more general process called model selection in which one model is selected among many candidate models, some of which may be of a completely different functional form. There are several competing variable selection criteria, including AIC and BIC, and algorithms, such as forward and backward selection. These methods are described in Cook and Weisberg (1994) and other standard texts. These algorithms all result in a set of variables to include in the model, denoted by X_A, and a set to leave out, denoted by $X_{\setminus A}$. For example, in the case of a linear model with mean function

$$E(Y|X) = \sum_i \beta_i X_i,$$

if a variable selection procedure includes only variables in the set X_A, the estimates of the coefficients β_i are zero for all $i \notin A$, so

$$\widehat{Y} = \sum_{i \in A} \hat{\beta}_i X_i + \sum_{i \notin A} 0 X_i.$$

Remember that the criteria for a good plot about X_k is

$$Y^*|(X^*, X_{\setminus k}) \sim Y^*|X^*, \tag{3.1}$$

where $X_{\setminus k}$ denotes all predictors except X_k. However, now that only the terms X_A remain the model, an alternate criteria is possible; that

$$Y^*|(X^*, X_{A\setminus k}) \sim Y^*|X^*. \tag{3.2}$$

This condition only requires that the axes of the plot are conditionally independent of variables that remain in the model.

These criteria will be compared for the component-plus-residual plot and the added-variable plot.

3.2.1 Component-plus-residual plot

When the relationship between the response and the predictors is of the form

$$Y = f_k(X_k) + g(X_{\setminus k}) + \epsilon,$$

the idealized component-plus-residual plot for X_k is defined as

$$\{X_k, f_k(X_k) + \epsilon\},$$

or equivalently,

$$\{X_k, Y - g(X_{\setminus k})\}.$$

This plot meets (3.1) and is conditionally independent of all of the other variables. Now the plot that only requires conditional independence of $X_{A\setminus k}$ can be written as

$$\{X_k, Y - g(X_{A\setminus k})\}.$$

However, since only the set of variables X_A are needed in this model, $g(X_{\setminus k}) = g(X_{A\setminus k})$, so this plot is equivalent to the earlier plot. It therefore satisfies both the full criteria (3.1) and the reduced criteria (3.2).

Incidentally, this holds true for X_k with both $k \in A$ and $k \notin A$; if $k \notin A$ the plot reduces to the residual plot

$$\{X_k, \epsilon\},$$

as f_k is zero.

As described in the earlier sections, there are several ways to estimate this plot. If f and g are both nonlinear, methods such as backfitting or spline smoothing are appropriate. And if g is linear, the CERES plot (Cook, 1993) is a Fisher-consistent way to estimate the coefficients of $X_{\setminus k}$. As AIC and BIC are Fisher-consistent methods, they may be used in the process of estimating the CERES plot.

3.2.2 Added-variable plot

When the relationship between the response and the predictors is of the form

$$Y = \beta_k X_k + g(X_{\setminus k}) + \epsilon,$$

the idealized added-variable plot for X_k is defined as

$$\{X_k - E(X_k|X_{\setminus k}), Y - E(Y|X_{\setminus k})\}.$$

and, as written here, is conditionally independent of $X_{\setminus k}$ as in (3.1).

However, if only conditional independence of $X_{A\setminus k}$ is required, as in (3.2), the variables $X_{\setminus A}$ are ignored, and the plot is

$$\{X_k - E(X_k|X_{A\setminus k}), Y - E(Y|X_{A\setminus k})\}, \tag{3.3}$$

which can look noticably different. Interestingly, though, because the variables in $X_{\setminus A}$ are not needed to compute $g(X_{\setminus k})$, this second plot satisfies the full condition of (3.1) as well.

This can be shown with the following form of (4.14). Any plot $\{X^*, Y^*\}$ where

$$\begin{aligned} X^* &= X_k - h(X_{\setminus k}) \text{ and} \\ Y^* &= Y - E\Big(Y\Big|X_k = h(X_{\setminus k}), X_{\setminus k}\Big) \end{aligned}$$

will satisfy the desired criterion of

$$E\left(Y^*|X^*, X_{\setminus k}\right) = E\left(Y^*|X^*\right).$$

Working with the full conditional expected value,

$$\begin{aligned} E\left(Y^*|X^*, X_{\backslash k}\right) &= E\Big(Y - E\left(Y|X_k = h(X_{\backslash k}), X_{\backslash k}\right)\Big|X^* = X_k - h(X_{\backslash k}), X_{\backslash k}\Big) \\ &= E\Big(Y\Big|X_k = X^* + h(X_{\backslash k}), X_{\backslash k}\Big) - E\Big(Y\Big|X_k = h(X_{\backslash k}), X_{\backslash k}\Big) \\ &= \beta_k\left(X^* + h(X_{\backslash k})\right) + g(X_{\backslash k}) - \beta_k h(X_{\backslash k}) - g(X_{\backslash k}) \\ &= \beta_k X^*, \end{aligned}$$

which depends only on X^* and is thus conditionally independent of $X_{\backslash k}$.

In the case under consideration, $h(X_{\backslash k}) = E(X_k|X_{A\backslash k})$, so to have full conditional independence of $X_{\backslash k}$,

$$\begin{aligned} Y^* &= Y - E\left(Y|X_k = E(X_k|X_{A\backslash k}), X_{\backslash k}\right) \\ &= Y - \left(\beta_k E(X_k|X_{A\backslash k}) + g(X_{\backslash k})\right) \end{aligned}$$

is needed. In this case, $g(X_{\backslash k})$ only depends on X_A, so $g(X_{\backslash k}) = g(X_{A\backslash k})$ and

$$\begin{aligned} Y^* &= Y - E(\beta_k X_k + g(X_{A\backslash k})|X_{A\backslash k}) \\ &= Y - E(Y|X_{A\backslash k}) \end{aligned}$$

which is the same as in (3.3), the plot derived initially only from conditional independence of $X_{A\backslash k}$.

Thus both options actually fulfill the strongest requirement for a good plot, so the question of which plot to use becomes a matter of interpretation, and will be investigated in the following example.

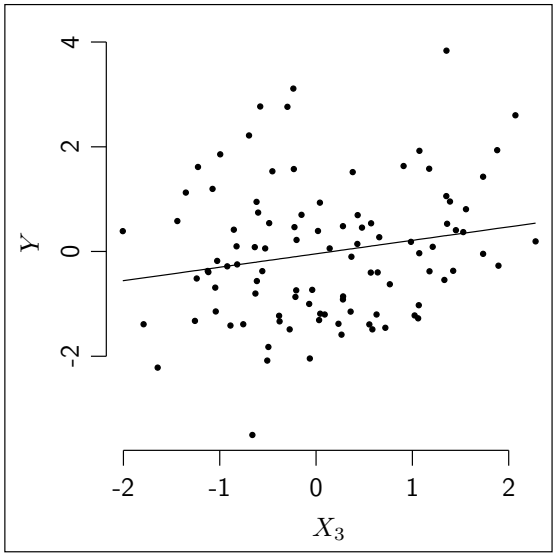


Figure 3.3: Marginal response plot of X_3 from Example 3.2, with least squares line added; X_3 and Y are not independent marginally, even though they are independent given X_1 and X_2.

Example 3.2

Let $N = 50$ data points be constructed using

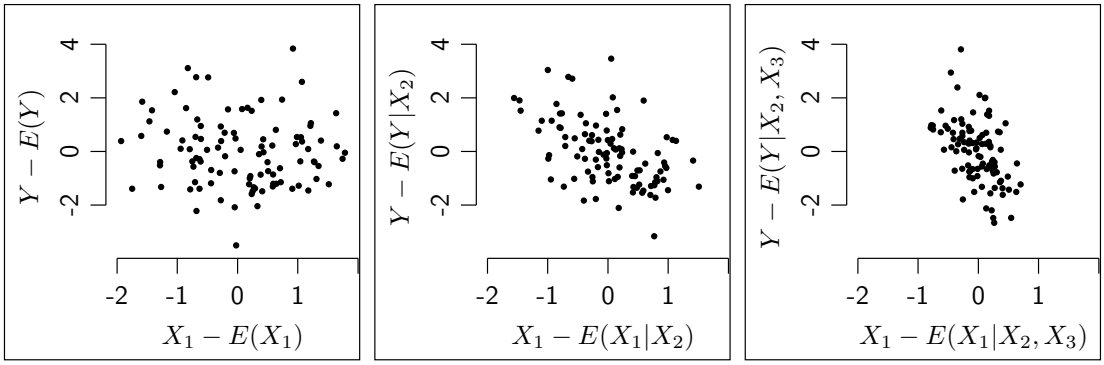
$$X \sim N\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}1.00 & 0.70 & 0.90\\0.70 & 1.00 & 0.90\\0.90 & 0.90 & 1.00\end{pmatrix}\right),$$

and

$$Y = -1X_1 + 1X_2 + 0X_3 + \epsilon,$$

where $\epsilon \sim N(0, I_N)$. In this example, X_3 is not in the full model. Should the added-variable plot for X_1 be constructed by taking X_3 into account, or by ignoring it?

In this example, although X_3 does not have any additional information about Y after X_1 and X_2 are accounted for, when viewed marginally (Figure 3.3), there is a positive association with Y. The question is whether this association should be



(a) Centered marginal plot. (b) Added-variable plot, ignoring X_3. (c) Added-variable plot, accounting for X_3.

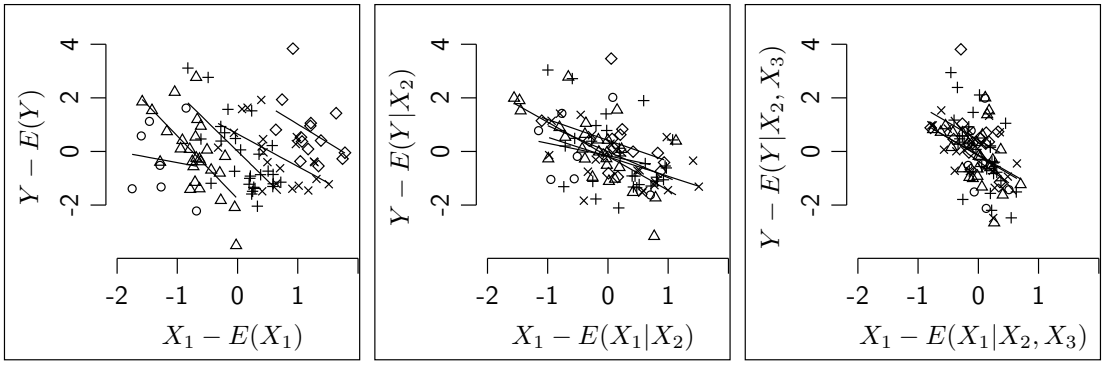
Figure 3.4: Idealized added-variable plots showing the relationship between X_1 and Y, marginally, accounting for only X_2, and accounting for both X_2 and X_3.

accounted for before constructing the added-variable plot for X_1. Figure 3.4 shows both options, as well as the centered marginal plot of X_1 and Y.

If predictors that are left out of the model when constructing the added-variable plot are ignored, the fact that those variables may indeed have some information about the response is lost, as in this example. Indeed, the added-variable plot without X_3 (Figure 3.4b) overstates the importance of X_1 compared to the plot with X_3 (Figure 3.4c).

However, if a variable really should be left out of the model, the argument can also be made that it should also be left out of the added-variable plot. For example, there may be many variables that are associated with the variable of interest that have no effect on the response. If all of them are included, the added-variable plot will show that the variable of interest is of little importance after accounting for the others. While this is true, it can also be considered misleading, as they probably shouldn't be accounted for in the first place. From this perspective, the full added-variable plot of Figure 3.4c understates the importance of X_1.

It is also instructive to slice these plots along X_3, as in Figure 3.5, to see if they are indeed conditionally independent of X_3. In these plots, the lines associated with each



(a) Centered marginal plot. (b) Added-variable plot, ignoring X_3. (c) Added-variable plot, accounting for X_3.

Figure 3.5: Idealized added-variable plots, as in Figure 3.4, but now sliced along X_3 to show any X_3 dependence.

slice are the least squares lines through those points, so because of random variation, they do not line up exactly. In the marginal plot, there is still X_3 dependence because when ignoring X_2, Y does depend on both X_1 and X_3 because of the correlation between X_2 and X_3. In the first added-variable plot, the axes are conditionally independent of X_3 even though it does not take X_3 into account, because conditional on X_1 and X_2, X_3 is independent of Y. Finally, the last added-variable plot does account for X_3, and its axes are conditionally independent of X_3. □

3.2.3 Estimated versions

When it is clear which variables to include in the model, the above methods work well, as the estimated values are consistent for the true population values. Which variable to include is not always clear, however, and different choices may yield very different results.

Example 3.3

Consider a sample data set of size 50, constructed in the following way: $X_1 \sim \text{Unif}(-1, 1)$, $X_2 = X_1^2 + 0.1\epsilon_1$, and $Y = 1X_1 + 1X_2 + 0.4\epsilon$, where ϵ and ϵ_1 are

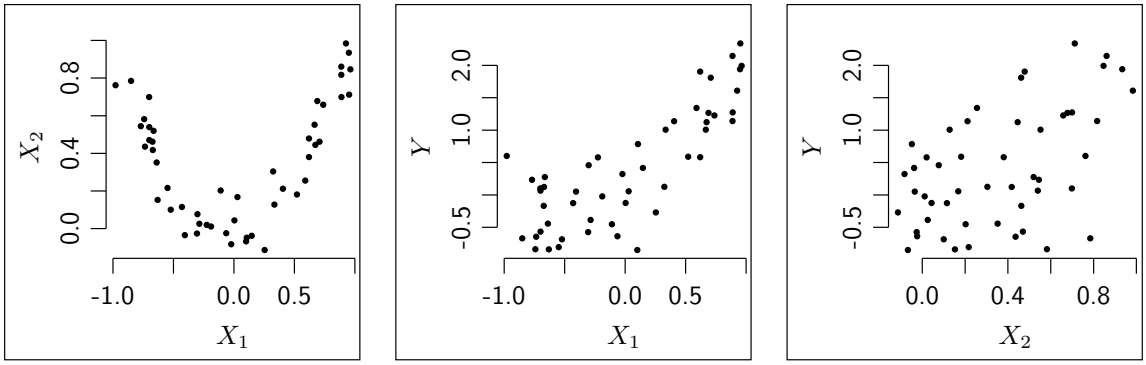
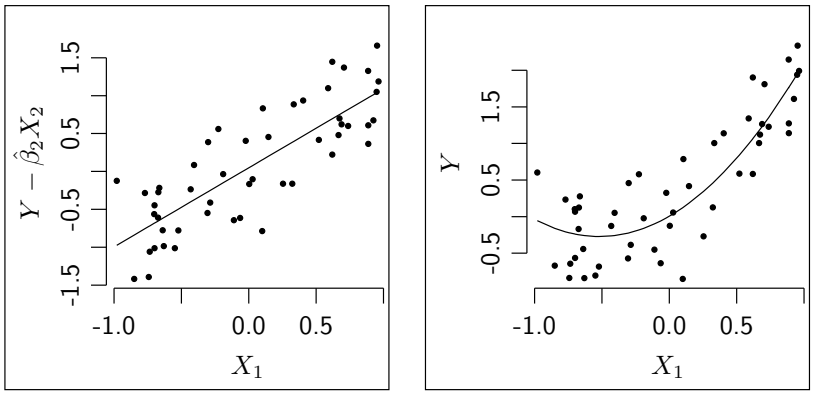


Figure 3.6: Scatterplots of the marginal relationships for the variables in Example 3.3.



(a) For the model with both X_1 and X_2.

(b) For the model with only X_1.

Figure 3.7: Component-plus-residual plots for X_1 in Example 3.3.

independent standard normals. What is the effect of X_1 on the response Y?

Scatterplots showing the relationships between the variables are shown in Figure 3.6. Because of the strong relationship between X_1 and X_2, it maybe difficult to decide which of these terms, or both, should be in the model. This choice will affect the apparent effect of X_1.

First, consider the model where Y depends on both X_1 and X_2. This model is actually the true model, but this information would not be known when analyzing a real data set. A linear least squares fit for this model result in fitted values of $\widehat{Y} = 0.05 + 1.04X_1 + 0.95X_2$, an AIC value of 69.9, and an R-squared value of

0.75. The component-plus-residual plot for X_1 based on these estimates is shown in Figure 3.7a, along with the fitted component line. Because a linear model was used, it is linear.

Next, consider the model where Y depends only on X_1. This relationship is quadratic, so a quadratic least squares fit was used, resulting in fitted values of $\widehat{Y} = 0.01 + 1.07X_1 + 1.03X_1^2$, an AIC value of 68.7, and an R-squared value of 0.75. The component-plus-residual plot for X_1 based on these estimates is shown in Figure 3.7b, along with the fitted component line. As expected from the form of the model used, it is quadratic.

These two models fit the data almost equally well, and in fact, by AIC, the true model is slightly worse. But the effect due to X_1 as shown in the component-plus-residual plots, is distinctly different.

Finally, consider the model where Y depends only on X_2. An linear least squares fit for this model results in fitted values of $\widehat{Y} = -0.13 + 1.52X_2$, an AIC value of 120.1, and an R-squared value of 0.28. This model is not as good as the other two, so plots showing these results will not be considered. □

Chatfield (1995) discusses this effect. He says that traditionally, statisticians have performed inference assuming that the chosen model is correct. He then shows that in several simple cases, the variability from model selection can outweigh the variance included in any particular model, and encourages statisticians to consider these sorts of issues when doing an analysis.

Indeed, different models can result in different plots, with different inferential conclusions about the effect of a single predictor on the response. To determine which model to use, a model selection procedure is traditionally used. But even in cases where two models have similar selection criteria, different plots and conclusions can result. One possible solution is to combine the most plausible models together.

3.3 Model combining

In multiple linear regression settings with many predictors, the final model used often does not include all the possible predictor variables. This can avoid overfitting the model to the data, especially when some variables are truly not important, and result in a model that has more predictive power than the model with all variables. As discussed previously, several model selection criteria, such as AIC and BIC, have been proposed, and there are various procedures for using them to choose a model, including forward and backward selection.

However, in many cases it can be difficult to choose which particular model to use. Two models may have very similar AIC values but include entirely different sets of predictors. It is also well known that ad hoc procedures like forward and backward selection can result in two different models. In cases like these, it can be difficult to decide which model to use. And as Chatfield (1995) describes, the uncertainty in the model selection process is usually ignored once the model has been chosen. But because each of the possible models could give different conclusions, choosing one while ignoring the other possibilities can lead to incomplete or erroneous conclusions. This can be especially true for models that are sensitive to the particular data set used, such as tree-based models.

In response to these difficulties, several model combining methods have been proposed. The basic idea of these methods is to not choose only one model, but to choose several likely models and combine them in a given way. This can have the advantages of avoiding overfitting and decreasing bias while avoiding the uncertainty involved in choosing only one model.

One method, stacking, finds a weighted linear combination of the possible models that minimizes the mean squared error by using a type of cross-validation. Initially proposed by Stone (1974) and called "model-mix," it was later redeveloped indepen-

dently by Wolpert (1992) in the neural network literature and applied to regression by Brieman (1995). Another method, bootstrap aggregation, or bagging, was developed by Brieman (1996). For this method, a set of N possible models are obtained by applying a given model selection criteria to N independent bootstrap samples. The average of the N fitted models is the final estimate. Several variations and generalizations of this type of method have been proposed in the machine learning under the name of learning ensembles; Friedman and Popescu (2003) provide a general framework.

A method called adaptive regression by mixing (ARM) has been developed by Yang (2001). This method also uses a weighted average of the possible models, where each of the possible models are found by choosing a subset of the data at random and applying a given model selection technique to the subset. The weights are proportional to the error, calculated using the other half of the data set. Yuan and Yang (2005) adds a screening step to this method, and also provides a measure of the instability of selecting a given model for a particular data set to help in deciding whether to combine possible models or to select just one. This enhanced method is called adaptive regression by mixing with screening (ARMS).

A Bayesian approach to model combining has also been studied. Called Bayesian model averaging (BMA), the basic idea is to put a prior on the possible models as well as on the parameters in those models and then calculate the posterior distribution of the possible models. The expected value of the quantity of interest can be found by computing a weighted average of the estimates from each model, using the posterior probabilities of each model as weights. An overview of techniques is provided by Hoeting et al. (1999).

While many of the graphical procedures presented earlier may be applied to other methods, in this the focus will be on applications using ARMS. The basic idea of ARMS is as follows: The data is divided into two parts. The various models being

considered are fit using the first part, and the top m models under both AIC and BIC are kept. Then the accuracy of each model is measured using the second part, and each model still under consideration is weighted accordingly. Finally, the process is repeated many times and the weights from each repetition are averaged.

While ARMS is not necessarily restricted to linear models, the focus here will be on that case because it is easiest to apply there. In this case, all the models under consideration are linear models, so the fitted values for model i can be written as

$$\widehat{Y_i} = \sum_k \hat{\beta}_{ki} X_k,$$

and since the final model is a linear combination of these linear models, the fitted model can also be written as a linear model, namely,

$$\widehat{Y} = \sum_i w_i \widehat{Y_i} = \sum_i w_i \sum_k \hat{\beta}_{ki} X_k = \sum_k \left(\sum_i w_i \hat{\beta}_{ki} \right) X_k \doteq \sum_k \hat{\beta}_k X_k.$$

3.3.1 ARMS and predictor correlation

Yuan and Yang (2005) did a simulation study on ARMS to determine when it performed better than model selection using AIC and BIC, and BMA. These simulations were performed using predictors that were independent and uniformly distributed. However, because graphics on data sets with independent predictors are not as complex as when the predictors are dependent, how ARMS does with correlated predictors was first investigated, using predictors chosen from a multivariate normal distribution.

With uncorrelated predictors

To better compare these results with the uncorrelated results in Yuan and Yang (2005), their simulations were rerun with independent normal predictors instead of independent uniform predictors.

The simulated scenarios are as follows. There are ten independent candidate predictors; in Yuan and Yang (2005) they were uniformly distributed on $[-1, 1]$, here they are normally distributed with mean 0 and variance 1/3, so the variance is the same in the two simulations. The sample size is one hundred. The response variable is $X\beta + \epsilon$, where ϵ is independent of the predictors and normally distributed with mean 0 and variance σ^2. Simulations were run with four different predictor coefficients (β), as described below, and five different variances ($\sigma^2 = 0.1$, 0.5, 1.0, 2.25, and 4.0). In these simulations, the true intercept is always zero.

The first model, *Small Coefficients*, has four small and four large coefficients, and two coefficients of zero:

$$Y = 1.5X_1 + 1.6X_2 + 1.7X_3 + 1.5X_4 + 0.4X_5 + 0.3X_6 + 0.2X_7 + 0.1X_8 + \epsilon.$$

The second model, *Large Coefficients*, has five large coefficients and five coefficients of zero:

$$Y = 1.0X_1 + 1.0X_2 + 1.0X_3 + 1.0X_4 + 1.0X_5 + \epsilon.$$

The third model, *Large Model*, has all nonzero coefficients of varying sizes:

$$Y = 1.8X_1 + 1.9X_2 + 2.0X_3 + 1.2X_4 + 1.5X_5 + 0.9X_6 + 0.8X_7 + 0.4X_8 + 0.3X_9 + 0.1X_{10} + \epsilon.$$

The fourth model, *Small Model*, has only two nonzero coefficients:

$$Y = 0.8X_1 + 0.9X_2 + \epsilon.$$

One hundred data sets were created from each of these models for each variance, and fitted using ARMS, AIC, and BIC. The risk for each model was then found by simulating 1000 new predictor values and calculating the average squared difference between the true function and the predicted value. Results are shown in Table 3.1,

with standard errors for each estimate in parentheses. The risk reduction is calculated as the percent improvement of the risk using the ARMS model compared with the smaller of the risks using AIC and BIC.

The results using normally distributed predictors have the same general pattern as those using uniformly distributed predictors. For most models, ARMS does better as the variance increases; this is because as the variance increases, model uncertainty also increases. In the models *Small Coefficients* and *Small Model*, this is especially true because AIC and BIC will occasionally remove entirely the predictors with small or zero coefficients; in contrast, ARMS will almost always give some weight to the correct model and models similar to it. In the model *Large Coefficients*, AIC and BIC usually can choose the correct model because it is clearer which terms are important and which are not. And in the model *Large Model*, AIC does well because all the terms are important, so there is no opportunity for it to overfit the model.

With correlated predictors

When predictors are uncorrelated, ARMS can improve the fit for certain models when variance is large. However, graphics in this case are fairly simple, because the relationship between the predictor and the response is the same whether or not one takes into account the effects of the other variables. So before investigating graphics for ARMS, whether or not ARMS is useful with correlated predictors will be investigated

First, ARMS is not invariant under transformations of the predictors. As it chooses potential models by leaving out various predictors, if the predictors change, the potential models will change, and the final model will change. Thus it is necessary to give ARMS the actual correlated predictors, rather than predictors that have been transformed to be uncorrelated.

Simulations were run using the same models and variances as in the uncorrelated

Small Coefficients

	0.1	0.5	1	2.25	4
ARMS	0.0113	0.0566	0.107	0.214	0.383
	(0.00062)	(0.0028)	(0.0051)	(0.012)	(0.025)
BIC	0.013	0.0726	0.135	0.246	0.442
	(0.00072)	(0.0033)	(0.0053)	(0.011)	(0.04)
AIC	0.0114	0.0594	0.123	0.268	0.446
	(0.00062)	(0.0031)	(0.0061)	(0.013)	(0.026)
RiskReduction	0%	5%	13%	13%	13%

Large Coefficients

	0.1	0.5	1	2.25	4
ARMS	0.00839	0.042	0.086	0.231	0.477
	(0.00047)	(0.0023)	(0.0053)	(0.014)	(0.025)
BIC	0.00739	0.037	0.0758	0.251	0.674
	(0.0005)	(0.0025)	(0.0058)	(0.023)	(0.042)
AIC	0.00998	0.0499	0.101	0.236	0.475
	(0.00052)	(0.0026)	(0.0059)	(0.014)	(0.028)
RiskReduction	−13%	−14%	−13%	2%	−1%

Large Model

	0.1	0.5	1	2.25	4
ARMS	0.0118	0.0608	0.121	0.27	0.496
	(0.00055)	(0.0028)	(0.0055)	(0.013)	(0.023)
BIC	0.0137	0.0771	0.145	0.322	0.657
	(0.00056)	(0.0032)	(0.0063)	(0.015)	(0.034)
AIC	0.0123	0.0616	0.13	0.279	0.492
	(0.00056)	(0.0029)	(0.0061)	(0.013)	(0.024)
RiskReduction	4%	1%	7%	3%	−1%

Small Model

	0.1	0.5	1	2.25	4
ARMS	0.00616	0.0309	0.0642	0.162	0.302
	(0.0005)	(0.0025)	(0.0054)	(0.013)	(0.022)
BIC	0.0054	0.027	0.0569	0.166	0.346
	(0.00055)	(0.0028)	(0.0062)	(0.016)	(0.026)
AIC	0.00877	0.0439	0.0877	0.207	0.39
	(0.00059)	(0.0029)	(0.0059)	(0.014)	(0.025)
RiskReduction	−14%	−14%	−13%	2%	13%

Table 3.1: Estimated mean squared error for prediction for ARMS, AIC, and BIC, using uncorrelated predictors. Risk reduction is the improvement in ARMS compared to the better of AIC and BIC. Negative values suggest ARMS is worse. Columns denote different values of σ^2.

case, but now with correlation $\rho = 0.9$ between all ten predictors. They all still have mean 0 and variance 1/3. Results for this simulation are shown in Table 3.2. The risk reduction is significantly greater in these simulations than in the uncorrelated simulations, because now predictors with small and nonzero coefficients are correlated with those with large coefficients, so it is more likely for AIC or BIC to select a predictor with a small or nonzero coefficient in place of a predictor with a large coefficient.

These two sets of results are directly compared in Figure 3.8, where the bars show the risk reduction calculated from the mean risk for ARMS and the mean risk for the better of AIC and BIC. Since the mean risk is an estimate, error bars were calculated using a parametric bootstrap to show the quality of the risk reduction estimate. For each estimate, 1000 ARMS risks and 1000 AIC/BIC risks were simulated from normal distributions with the estimated means and standard errors. The risk reduction was then calculated, and a 90% confidence interval plotted.

Although the confidence intervals on the risk reduction are large, ARMS does significantly better than AIC/BIC with correlated predictors than with uncorrelated predictors, particularly for the *Large Coefficients* model. As $n \to \infty$, the results with correlated predictors are expected to resemble the results from uncorrelated predictors.

Small Coefficients

	0.1	0.5	1	2.25	4
ARMS	0.0121	0.0515	0.112	0.242	0.381
	(0.00051)	(0.0027)	(0.0057)	(0.011)	(0.017)
BIC	0.0152	0.0575	0.131	0.343	0.58
	(0.00061)	(0.0035)	(0.0085)	(0.014)	(0.021)
AIC	0.0137	0.0581	0.125	0.302	0.505
	(0.00056)	(0.0026)	(0.0063)	(0.015)	(0.022)
RiskReduction	11%	10%	10%	20%	25%

Large Coefficients

	0.1	0.5	1	2.25	4
ARMS	0.00879	0.0603	0.115	0.215	0.318
	(0.00045)	(0.0026)	(0.0044)	(0.0088)	(0.014)
BIC	0.00776	0.0792	0.164	0.327	0.463
	(0.00044)	(0.004)	(0.0061)	(0.011)	(0.019)
AIC	0.0102	0.0635	0.133	0.277	0.46
	(0.00052)	(0.0033)	(0.0057)	(0.011)	(0.02)
RiskReduction	−13%	5%	13%	22%	31%

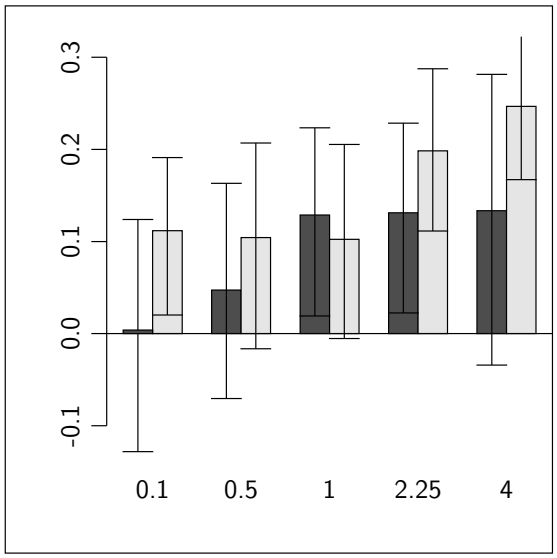
Large Model

	0.1	0.5	1	2.25	4
ARMS	0.0133	0.0706	0.146	0.311	0.492
	(0.0006)	(0.0034)	(0.0067)	(0.014)	(0.024)
BIC	0.0159	0.0896	0.198	0.472	0.762
	(0.00063)	(0.0046)	(0.0093)	(0.019)	(0.028)
AIC	0.0139	0.0721	0.145	0.341	0.584
	(0.00063)	(0.0032)	(0.007)	(0.016)	(0.026)
RiskReduction	4%	2%	0%	9%	16%

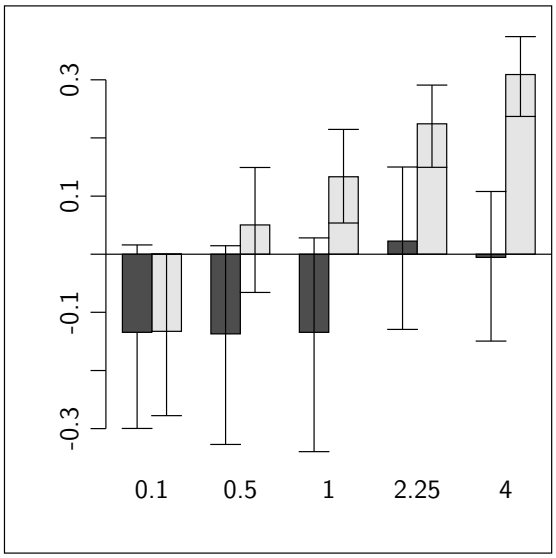
Small Model

	0.1	0.5	1	2.25	4	6.25
ARMS	0.00647	0.0393	0.0728	0.142	0.226	0.331
	(0.00055)	(0.0029)	(0.0052)	(0.011)	(0.018)	(0.028)
BIC	0.00581	0.0466	0.0927	0.17	0.257	0.365
	(0.00063)	(0.0037)	(0.0059)	(0.013)	(0.022)	(0.035)
AIC	0.00904	0.0507	0.104	0.224	0.358	0.546
	(0.00066)	(0.0037)	(0.0066)	(0.013)	(0.024)	(0.038)
RiskReduction	−11%	16%	21%	16%	12%	9%

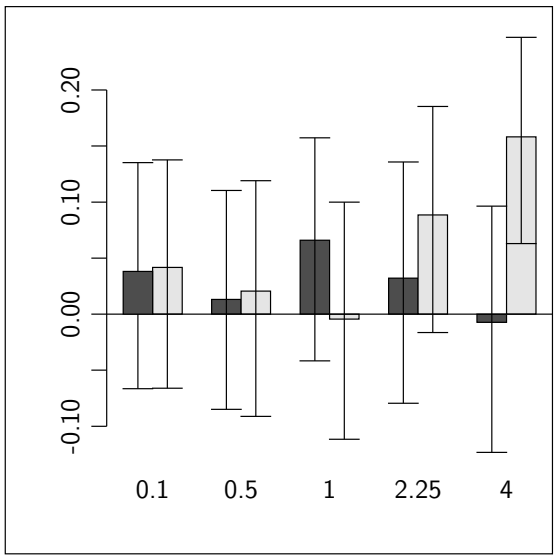
Table 3.2: Estimated mean squared error for prediction for ARMS, AIC, and BIC, using correlated predictors. Risk reduction is the improvement in ARMS compared to the better of AIC and BIC. Negative values suggest ARMS is worse. Columns denote different values of σ^2.



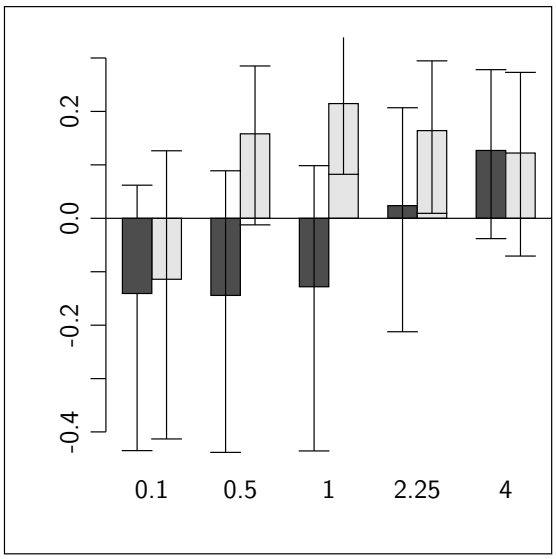
(a) Small Coefficients



(b) Large Coefficients



(c) Large Model



(d) Small Model

Figure 3.8: Comparison of Risk Reduction between ARMS and the better of AIC and BIC, for uncorrelated predictors (the dark bars) and correlated predictors (the light bars), for several values of σ^2.

3.3.2 Component-plus-residual plot

When the true model is

$$Y = f_k(X_k) + g(X_{\setminus k}) + \epsilon,$$

the idealized component-plus-residual plot for X_k is

$$\{X_k, f_k(X_k) + \epsilon\}, \text{ or } \{X_k, Y - g(X_{\setminus k})\}.$$

This plot shows the contribution to Y from X_k when all other variables are held constant. When g is linear, $g(X_{\setminus k}) = \sum_{i \neq k} \beta_i X_i$, and this can be written as

$$\{X_k, Y - \sum_{i \neq k} \beta_i X_i\}.$$

Since a model combining method such as ARMS can be thought of as a way of estimating β, using the ARMS estimates $\tilde{\beta}$ in the plot

$$\{X_k, Y - \sum_{i \neq k} \tilde{\beta}_i X_i\},$$

is the proper component-plus-residual plot; if the true value β of the estimate $\tilde{\beta}$, is substituted, the true component-plus-residual plot would result. Additionally, as $\tilde{\beta}$ is consistent for β, this estimated plot will be consistent for the true plot.

Fisher-consistency of ARMS and applicability to CERES plots

As described in the earlier sections, when g is linear and there is a nonlinear relationship between X_k and $X_{\setminus k}$, a CERES plot will obtain Fisher-consistent estimates of this plot. Since ARMS can be shown to be Fisher-consistent, as follows, it can be appropriate to use ARMS in estimating the coefficients for use in a CERES plot.

The ARMS method of model combining is Fisher-consistent for the true coefficient parameters when the composite methods are Fisher-consistent and at least one model including all predictors with non-zero coefficients is considered. For a method to be

Fisher-consistent, the method must result in the true parameters when applied to the entire population.

As described in Yuan and Yang (2005), ARMS first splits the data into two parts; given that the entire population is being considered, population calculations will be done no matter which part is being used. Then, the fitted values $\hat{f}_j$ are estimated for each model j using least squares, using the first half of the data. For models that include all predictors with nonzero coefficients, and perhaps also some unneeded predictors as well, the fitted values and the coefficients will be the true values, assuming the method used is Fisher consistent. For models that do not include all predictors with non-zero coefficients, the fitted values and coefficients will not be equal to the true values. ARMS then screens these models by only considering models with the lowest k AIC or BIC values. Given the population, the true model will have the lowest AIC and BIC, so that model is sure to be considered. AIC is defined as

$$AIC = n \log(\hat{\sigma}^2) + 2k$$

where the mean has k parameters. For models that leave off important terms, $\hat{\sigma}^2$ is too big, and eventually the first term will dominate. For models that include unimportant terms, $\hat{\sigma}^2$ will converge to the correct value, but the penalty $2k$ will be too big. Hence the true model is always selected, and it does not matter which other models are also considered.

An overall measure of discrepancy,

$$D_j = \sum_i \left(Y_i - \hat{f}_j(X_i) \right)^2,$$

is then computed for each model being considered, using the second half of the data. In the population, this is proportional to the error variance σ_j^2. So the proportional weight for model j

$$(\sigma_j)^{-n/2} \exp(-\sigma_j^{-2} D_j/2)$$

is in turn proportional to $(\sigma_j)^{-n/2}$. Thus models with smaller variances have larger weights, and the relative weight increases as n increases. Since this calculation is done in terms of the population, the model(s) with the smallest variance will have weight one and all other models will have weight zero.

Now, all models including all necessary predictors will have the same error variance, and all models missing one or more of these predictors will have a larger error variance, as the variance explained by the missing predictor will be included in the error variance.

So given the whole population, the models that do not include all necessary terms will have zero weight, and the models that have all necessary terms will have equal weight. Since these models all result in the same estimates (in the population), assuming they are Fisher-consistent, the ARMS method is also Fisher-consistent.

3.3.3 Added-variable plot

When using a fitting method such as ARMS or BMA that uses a weighted average of linear models, the method may not fully throw out certain predictors, but instead may give models with those predictors small weights. Let the fit for the ith submodel be

$$\widehat{Y}_i = \hat{\beta}_{ki} X_k + \hat{g}_i \left(X_{\setminus k} \right),$$

with residuals $e_i = Y - \widehat{Y}_i$. Because of the fitting method, each $\hat{g}_i$ will include a different subset of $X_{\setminus k}$. Then let the overall fit be

$$\begin{aligned} \widehat{Y} &= \textstyle\sum_i w_i \widehat{Y}_i \\ &= \left(\textstyle\sum_i w_i \hat{\beta}_{ki} \right) X_k + \textstyle\sum_i w_i \hat{g}_i \left(X_{\setminus k} \right) \\ &\doteq \hat{\beta}_k X_k + \hat{g} \left(X_{\setminus k} \right), \end{aligned}$$

with residuals $e = Y - \widehat{Y}$. By this definition, $\sum_i w_i \hat{\beta}_{ki} = \hat{\beta}_k$ and $\sum_i w_i e_i = e$.

Then to construct an added-variable plot for X_k, there are two options, as discussed in Section 3.2; either to construct the axes by accounting for all the other variables, or by only accounting for the variables that are selected by the various submodels.

Accounting for all the variables

When the true model is

$$Y = \beta_k X_k + g(X_{\setminus k}) + \epsilon,$$

one option to construct the added-variable plot for X_k is

$$\{X_k - E\left(X_k | X_{\setminus k}\right), Y - E\left(Y | X_{\setminus k}\right)\},$$

which takes into account all the other variables. There are several ways to estimate these expected values; one way is to estimate them independently using whatever methods are appropriate; ARMS, an additive model, or whatever. However, this plot will not be based at all on the full ARMS model, which is believed to be the best fit, nor will it show exactly the slope and residuals from that full model.

Remember that in an idealized added-variable plot, the residuals around a line with slope β_k will be equal to the errors ϵ, for

$$\begin{aligned} E\left(Y | X_{\setminus k}\right) &= E\left(E(Y|X) | X_{\setminus k}\right) \\ &= E\left(\beta_k X_k + g(X_{\setminus k}) | X_{\setminus k}\right) \\ &= \beta_k E\left(X_k | X_{\setminus k}\right) + g\left(X_{\setminus k}\right) \end{aligned} \tag{3.4}$$

so

$$\begin{aligned} Y^* &= Y - E(Y|X_{\setminus k}) \\ &= \left[\beta_k X_k + g\left(X_{\setminus k}\right) + \epsilon\right] - \left[\beta_k\, E\left(X_k|X_{\setminus k}\right) + g\left(X_{\setminus k}\right)\right] \\ &= \beta_k \left(X_k - E\left(X_k|X_{\setminus k}\right)\right) + \epsilon \qquad (3.5) \\ &= \beta_k X^* + \epsilon. \qquad (3.6) \end{aligned}$$

In these idealized versions of these plots, the slope and the residuals are exactly equal to the true values, as the equality in (3.4) holds exactly. But when these plots are estimated, the slope and the residuals may not be exactly equal to the values estimated by the full model. For this to be true, the equality in Equation (3.4) must hold when those values are estimated. That is, if

$$\begin{aligned} &\widehat{E}\left(Y|X\right) \doteq \hat{\beta}_k + \hat{g}\left(X_{\setminus k}\right) \quad \text{and} \\ &\widehat{E}\left(Y|X_{\setminus k}\right) = \hat{\beta}_k\, \widehat{E}\left(X_k|X_{\setminus k}\right) + \hat{g}\left(X_{\setminus k}\right), \qquad (3.7) \end{aligned}$$

then

$$Y - \widehat{E}(Y|X_{\setminus k}) = \hat{\beta}_k \left(X_k - \widehat{E}\left(X_k|X_{\setminus k}\right)\right) + e, \qquad (3.8)$$

as in (3.5). Now if the methods used to estimate these expectations are consistent, (3.5) will be true asymptotically. But for small samples, it will only hold in special cases, such as when ordinary least squares are used for all fits. In general, for the equality in (3.8) to hold exactly, only two of the three expected values in (3.7) should be estimated, and the third computed from the first two. Since the full model fit uses all of the available information, a usual choice is to use $\hat{\beta}_k$ and $\hat{g}$ from the estimate of $E(Y|X)$. Then to meet the equality in (3.8) exactly, either $E(Y|X_{\setminus k})$ can be

estimated and $E\left(X_k|X_{\setminus k}\right)$ computed, or vice versa.

In many cases, the plots are very similar, though this is not necessarily so. For example, when β_k is small, it is difficult to calculate $E\left(X_k|X_{\setminus k}\right)$ from $E(Y|X_{\setminus k})$ because it requires division by β_k, as from (3.4),

$$E\left(X_k|X_{\setminus k}\right) = \frac{1}{\beta_k}\Big(E\left(Y|X_{\setminus k}\right) - g\left(X_{\setminus k}\right)\Big).$$

This plot may nevertheless be helpful because it can be written in terms of a change in fitted values between the model including X_k and the model without X_k, as when computed this way,

$$\begin{aligned} X^* &= X_k - E\left(X_k|X_{\setminus k}\right) \\ &= X_k - \frac{1}{\beta_k}\Big(E\left(Y|X_{\setminus k}\right) - g\left(X_{\setminus k}\right)\Big) \\ &= \frac{1}{\beta_k}\Big(\beta_k X_k + g\left(X_{\setminus k}\right) - E\left(Y|X_{\setminus k}\right)\Big) \\ &= \frac{1}{\beta_k}\Big(E\left(Y|X\right) - E\left(Y|X_{\setminus k}\right)\Big) \end{aligned}$$

When β_k is large, this should not cause any problems, but when β_k is small, it may be more appropriate to use only the change in fitted values as X^*, in the form of the general added-variable plot, described in Section 1.6.2.

As for the other way, it might seem that there are two ways to compute $E(Y|X_{\setminus k})$ from an estimate of $E\left(X_k|X_{\setminus k}\right)$; either by using the full model estimate of β_k and g directly, or by using the estimates for each submodel β_{ki} and g_i and then weighting, but these turn out to be equivalent. Letting

$$\widehat{E}_i\left(Y|X\right) = \hat{\beta}_{ki}X_k + \hat{g}_i\left(X_{\setminus k}\right)$$

be the fit from the ith submodel, then the calculated version of $E(Y|X_{\setminus k})$ based on this submodel would be

$$\widehat{E}_i\left(Y|X_{\setminus k}\right) = \hat{\beta}_{ki}E\left(X_k|X_{\setminus k}\right) + \hat{g}_i\left(X_{\setminus k}\right)$$

so setting

$$\begin{aligned}\widehat{E}\left(Y|X_{\backslash k}\right) &\doteq \sum_i w_i \widehat{E}_i\left(Y|X_{\backslash k}\right) \\ &= \sum_i w_i \hat{\beta}_{ki} E\left(X_k|X_{\backslash k}\right) + \hat{g}_i\left(X_{\backslash k}\right) \\ &= \left(\sum_i w_i \hat{\beta}_{ki}\right) E\left(X_k|X_{\backslash k}\right) + \sum_i w_i \hat{g}_i\left(X_{\backslash k}\right) \\ &= \hat{\beta}_k E\left(X_k|X_{\backslash k}\right) + \hat{g}\left(X_{\backslash k}\right),\end{aligned}$$

which is the same result as when it is computed using the full model estimate.

Accounting for only the variables selected by the submodels

Alternatively, only the predictors suggested by each submodel could be used, weighted appropriately. Let A_i be the indices of the predictors chosen by submodel i, so according to submodel i, $E(Y|X_{A_i}) = E(Y|X)$. Then to construct an added-variable plot for X_k using this submodel only the set of predictors $X_{A_i\backslash k}$ would be used. Then

$$\begin{aligned}X_i^* &= X_k - E(X_k|X_{A_i\backslash k}) \quad \text{and} \\ Y_i^* &= Y - E(Y|X_{A_i\backslash k}).\end{aligned}$$

A weighted average of these plots gives a horizontal axis of $\sum_i w_i X_i^*$ and a vertical axis of $\sum_i w_i Y_i^*$.

This will not meet criteria (3.6), unless it gives full weight to the submodel that correctly includes the actual predictors. But because which submodel is actually true is unknown, this plot may be a better representation of the effect of X_k than just taking one of the plots generated by a submodel.

The axes in this plot are now a compromise between retaining and removing the

borderline variables. Although it is consistent for the true added-variable plot, the axes do not correspond to the correct expected values, as each of the submodels estimates different expected values. If it is believed that the correct added-variable plot should only account for the variables selected by the correct submodel, this plot is more appropriate than the version that accounts for all variables.

Accounting for only the variables selected by the submodels; an alternate method

A nice feature of added-variable plots is that they show both the proper slope and the residuals, so ideally,

$$Y^* = \hat{\beta}_k X^* + e. \tag{3.9}$$

For a weighted model to satisfy this,

$$\sum_i w_i Y_i^* = \hat{\beta}_k \sum_i w_i X_i^* + e. \tag{3.10}$$

However, for each submodel,

$$Y_i^* = \hat{\beta}_{ki} X_i^* + e_i,$$

so by taking the proper weighted average, instead

$$\begin{aligned}
\sum_i w_i Y_i^* &= \sum_i w_i \left(\hat{\beta}_{ki} X_i^* + e_i \right) \\
&= \sum_i w_i \hat{\beta}_{ki} X_i^* + e \\
&= \hat{\beta}_k \sum_i \left(\frac{w_i \hat{\beta}_{ki}}{\hat{\beta}_k} \right) X_i^* + e.
\end{aligned}$$

So using the full model weights to take the weighted average of each of the axes does not satisfy (3.10); on the x-axis weights equal to $w_i\hat{\beta}_{ki}/\hat{\beta}_k$ should instead be used, so the axes are

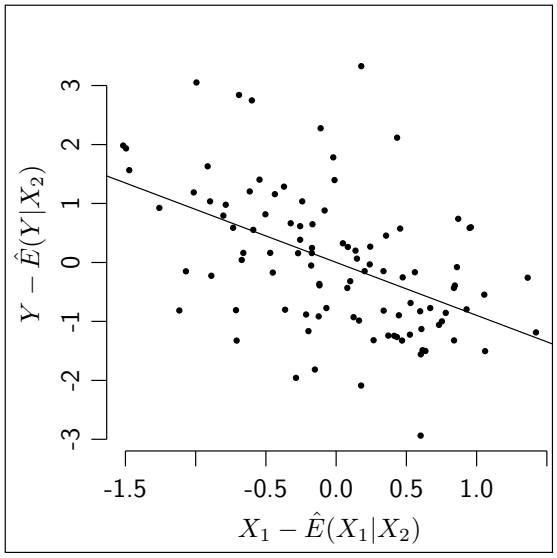
$$Y^* = \sum_i w_i Y_i^* \quad \text{and} \tag{3.11}$$

$$X^* = \sum_i \left(w_i \hat{\beta}_{ki}/\hat{\beta}_k \right) X_i^* \equiv \sum_i w_i^* X_i^*. \tag{3.12}$$

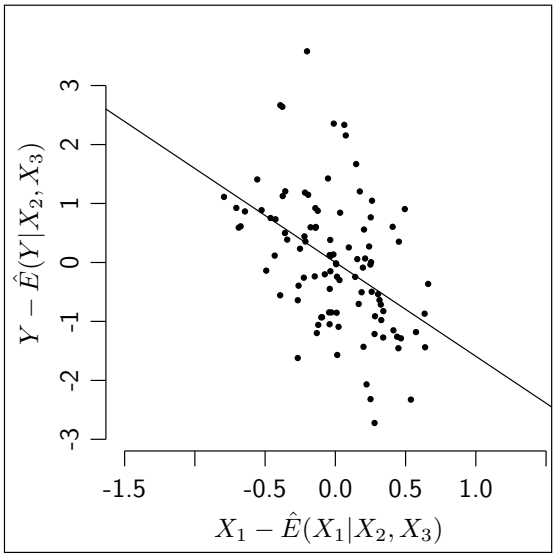
This plot is also consistent for the idealized plot, for as $n \to \infty$, $w_m \to 1$ and $\hat{\beta}_{km} \to \hat{\beta}_k$, where m is the index of the correct model.

Example 3.4

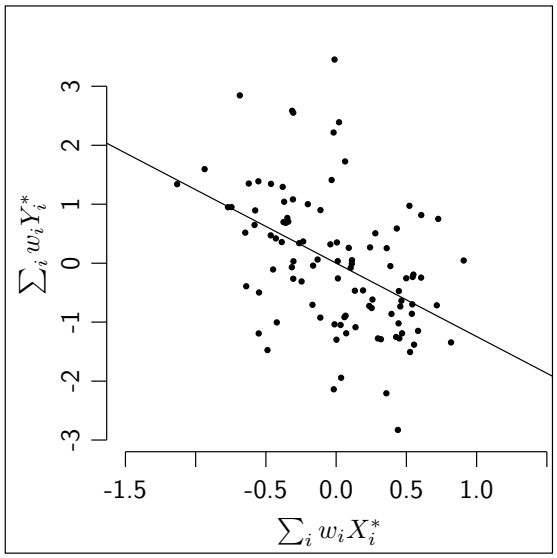
Figure 3.9 shows four estimated added-variable plots for the data from Example 3.2, using a weighted model with equal weights on two models: (1) the model including all three variables, and (2) the model only including X_1 and X_2. Plots are shown that fully ignore X_3, fully account for X_3, and somewhat account for X_3, based on these weights. Figure 3.9c uses the weighted average on both axes, while Figure 3.9d uses the weighted average only on the y-axis, and on the x-axis uses the weights in (3.12) instead. Here, these weights are are 0.64 and 0.36, as the estimates of β_1 for these two models are -1.59 and -0.9. The averaged versions are both compromises between the other two plots. □



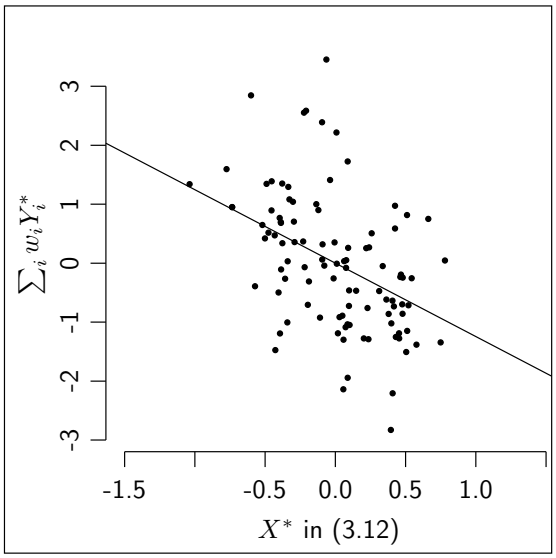
(a) Fully ignoring X_3.



(b) Fully accounting for X_3.



(c) Using full model weights on both axes to somewhat account for X_3.



(d) Using weights as in (3.12) to somewhat account for X_3.

Figure 3.9: Estimated added-variable plots for Example 3.4, accounting for X_3 in different ways.

3.4 Example: Beef consumption

Consider the following data set from the U. S. Department of Agriculture on beef consumption from 1925–1941, as well as possible predictors beef price, pork price and consumption, disposable income (both actual and adjusted for inflation), a food price index (both actual and adjusted for inflation) and a food consumption index. Plots of these variables by year are shown in Figure 3.10.

3.4.1 Model selection and inference

Suppose it is desired to display the effect of disposable income on beef consumption. This may be difficult because actual and adjusted disposable income are highly correlated, so it may be difficult to tell what part of the effect comes from the actual disposable income, and what part comes from the adjusted disposable income.

Consider first only the adjusted disposable income. Component-plus-residual plots and added-variable plots using three different models are shown in Figure 3.11 and Figure 3.12. The added-variable plots have lines added to them showing the estimated slope and the associated 95% confidence interval.

The first model considered is the full model, with all predictors. This model has an AIC value of 38.9 and an R^2 value of 0.991, so it fits very well, and is probably even overfit. According to the plots in Figure 3.11a and Figure 3.12a, an increase in adjusted income is associated with an increase in beef consumption, when all other variables, including actual income, are held constant. While the increase looks large in the component-plus-residual plot, in the added-variable plot it is not statistically significant, as a slope of zero falls within the confidence bounds.

The second model is the model without actual income, but with all the other predictors. This model has an AIC value of 39.04 and an R^2 value of 0.990, so by these measures it is very comparable to the full model. The plots for this model are

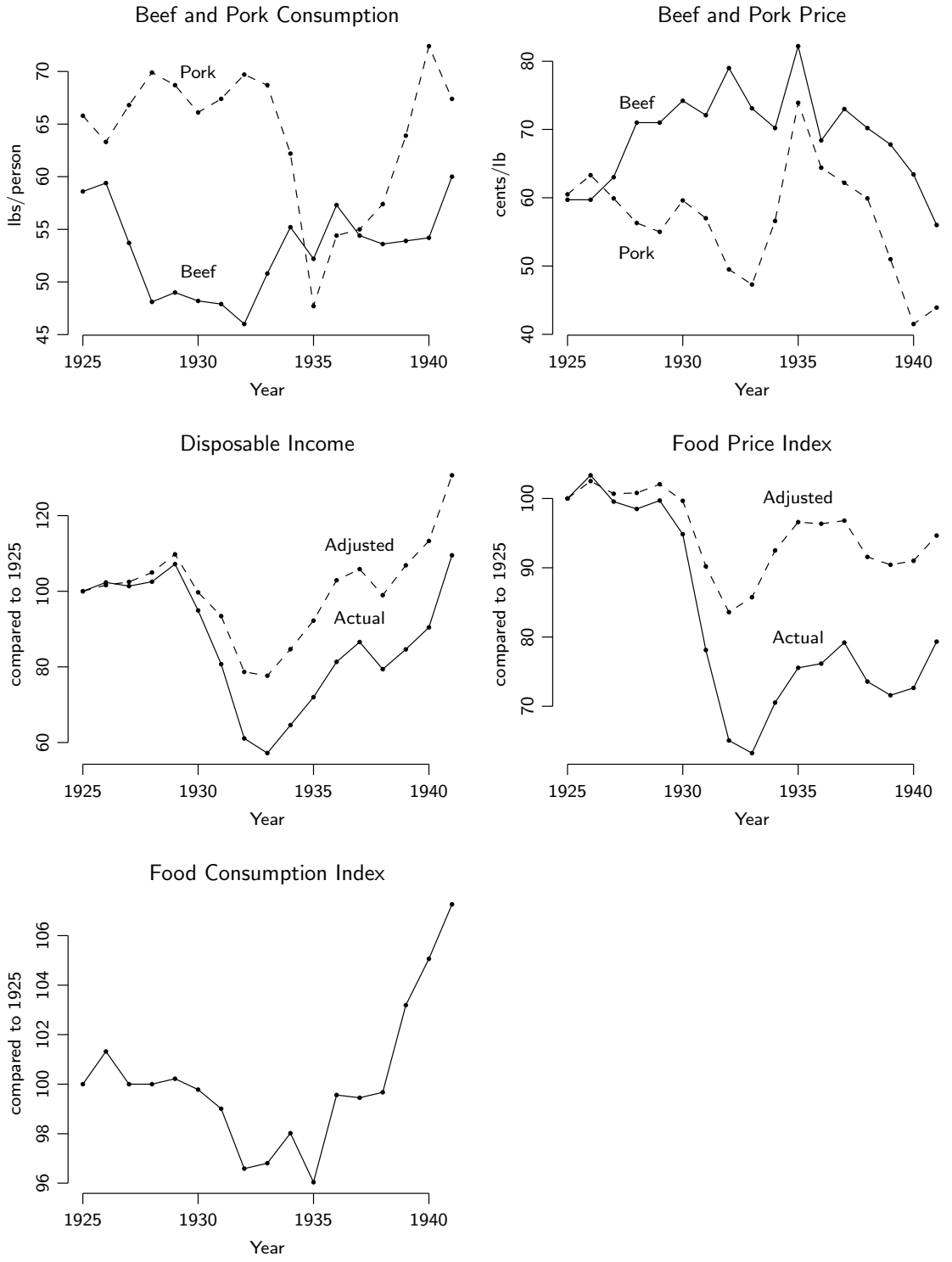
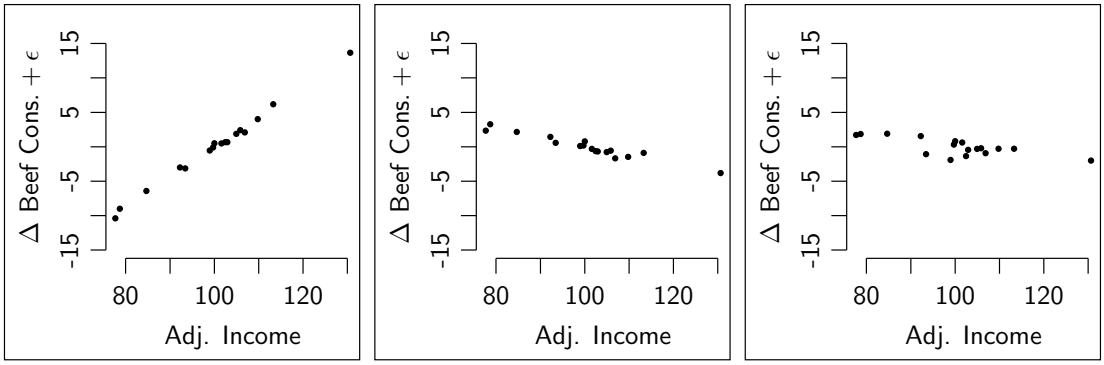


Figure 3.10: Seven variables about beef and pork, 1925–1941.



(a) Full model.

(b) Model without actual income.

(c) Model with only four predictors.

Figure 3.11: Component-plus-residual plots for adjusted income, constructed from three different models.

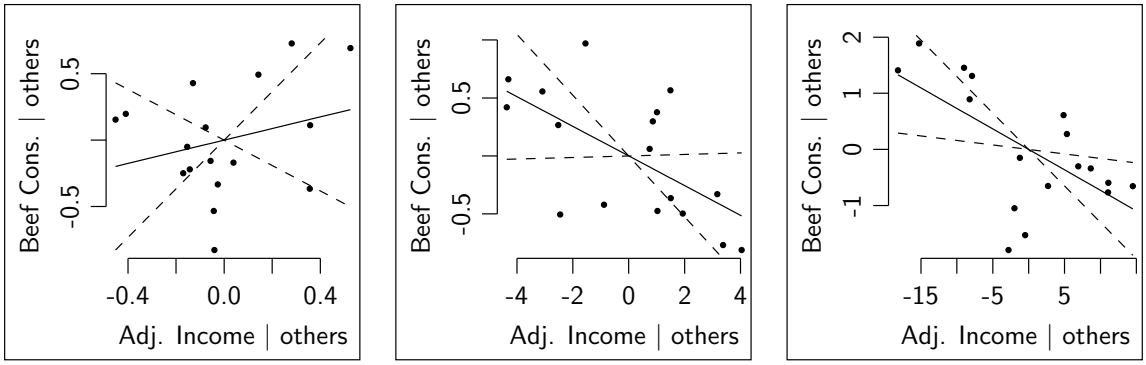
shown in Figure 3.11b and Figure 3.12b, and show that an increase in adjusted income is associated with a decrease in beef consumption, again when all other variables are held constant. The added-variable plot shows that this decrease is not statistically significant either.

The third model is the model with only beef price, pork price and consumption, and adjusted income. This model has an AIC value of 53.9 and an R^2 value of 0.96, so using the AIC measure, it should not be considered as good a fit, but since the R^2 value is still very high, the fit is still very good. Plots for this model, shown in Figure 3.11c, again show a slight decrease, but this time it is significant.

These three different models offer three different answers to the question of how adjusted disposable income affects beef consumption.

The component-plus-residual plots for both actual and adjusted income, in Figure 3.13, gives one explanation for these three different answers. Since adjusted and actual income are correlated, these plots together can show income is generally related to beef consumption.

For the full model, an increase in adjusted income is associated with a large increase in beef consumption, but an increase in actual income is associated with



(a) Full model.

(b) Model without actual income.

(c) Model with only four predictors.

Figure 3.12: Added-variable plots for adjusted income, constructed from three different models.

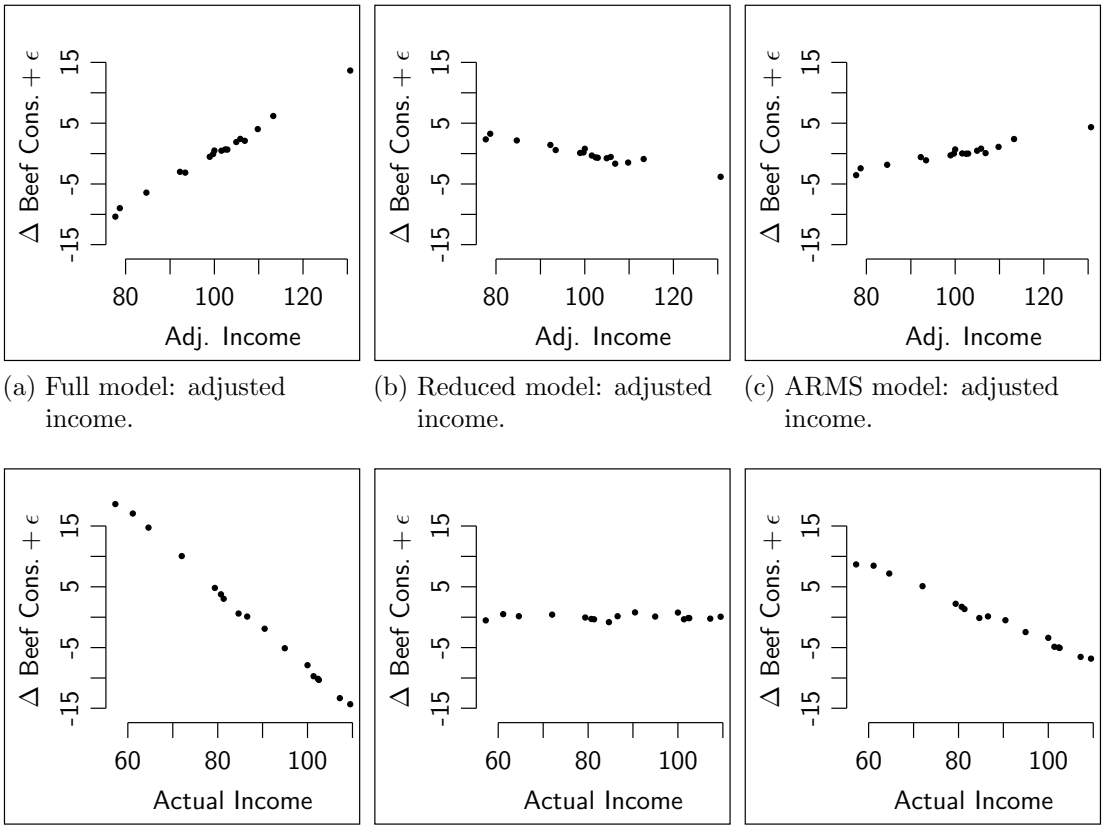
an even larger decrease in beef consumption. So the overall effect of an increase in disposable income is a slight decrease in beef consumption.

The reduced model plots tell a slightly different story; that the small decrease in beef consumption is due totally to the increase in adjusted disposable income, and that the corresponding increase in actual income has no effect on beef consumption.

These two models tell the same story about disposable income in general, but disagree about how that increase is associated with the two correlated predictors, actual and adjusted disposable income. While the models themselves can include some of this uncertainty, as in the added-variable plots, it is also important to be aware of the uncertainty about which model to choose, as the plots about one particular predictor variable and the resulting inferences can be very different between two models.

3.4.2 Plots using ARMS

Model combining methods can be useful in situations where multiple models seem appropriate, as they do not choose only one model, but combine multiple appropriate



(a) Full model: adjusted income.

(b) Reduced model: adjusted income.

(c) ARMS model: adjusted income.

(d) Full model: Actual Income

(e) Reduced model: Actual Income

(f) ARMS model: Actual Income.

Figure 3.13: Component-plus-residual plots for actual and adjusted disposable income, estimated using the full model, a reduced model, and the ARMS model.

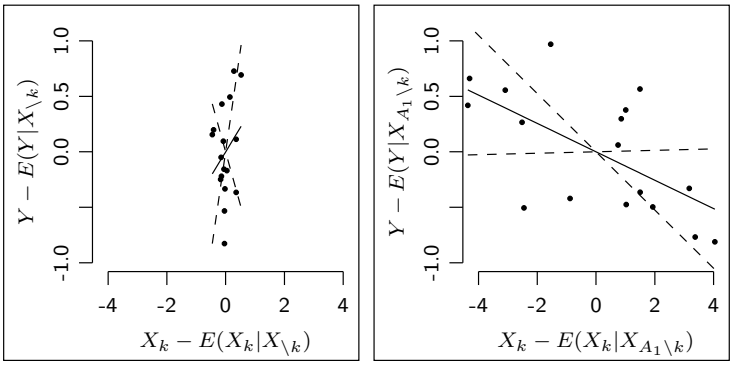
models using appropriate weights. For this data, a combined model was fit using ARMS, and various plots constructed.

Figure 3.13 shows component-plus-residual plots for actual and adjusted income estimated using the ARMS model, in addition to the plots estimated using the full model and the model without actual income. The ARMS plots average out the differences between models, as they are compromises between the two other plots, as well as plots from other submodels.

Figure 3.14 shows two added-variable plots for adjusted income constructed from two submodels and three plots constructed using the ARMS fit. The axes on all of these plots are common to assist in comparison. First, the importance of adjusted income is much smaller in the full model (Figure 3.14a) than in the model without actual income (Figure 3.14b); the range of the x-axis, which shows how much additional information is in the variable, is much smaller, for when actual income is omitted, adjusted income is the only source of information about income.

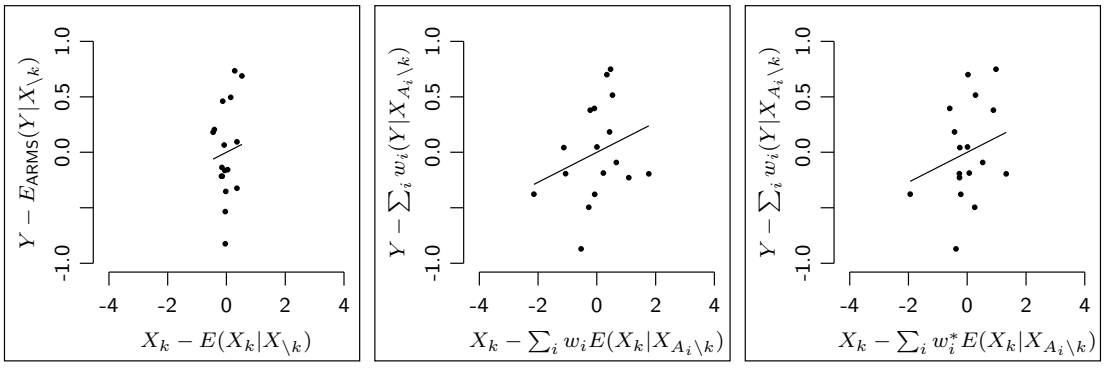
The added-variable plot using ARMS, when fully accounting for the other variables (Figure 3.14c), is very similar to the plot using the full model; indeed, the values on the x-axis are identical, as ARMS uses ordinary least squares to fit each submodel. The difference is on the y-axis; it shows a smaller slope than that in the full model because it compromises between the various submodels considered. This shows the same information as in the standard added-variable plot; the information unexplained by the other variables against the new information in adjusted income.

However, it can be misleading because perhaps actual income shouldn't be included in the model at all, and therefore should be considered one of the "other variables." The other two added-variable plots weight the x-axis according to the weights given by ARMS to remove some, but not all of the dependence on actual income, as well as some of the dependence on the other variables. It is a nice compromise between the plot for the full model and the plot for the model without actual



(a) Full model.

(b) Model using the set of variables without income (called A_1).



(c) ARMS model, estimated by accounting for all other variables.

(d) ARMS model, estimated by weighting other variables.

(e) ARMS model, estimated by the alternate weighting method; w_i^* are the weights as defined in (3.12).

Figure 3.14: Added-variable plots for adjusted income, using the full model, the model without actual income, and several methods for the ARMS model. The response Y is Beef Consumption and the variable of interest X_k is adjusted income.

income; this suggests that adjusted income is not as unimportant as it looks in the full model, nor as important as it looks in the submodel, but instead is somewhere in between.

3.5 Example: Island deforestation

In Chapter 1, the plots showing the importance and effect of each of the eight variables on island deforestation were made using straightforward linear methods. But several of the variables were correlated, and some seemed to have no effect, so it is unlikely that the best model includes all eight variables, but it may be difficult to determine which variables to include. So the data was analyzed using the ARMS method of model combining, and new plots produced.

The component-plus-residual plots are shown in Figure 3.15, and the added-variable plots, both when fully accounting for the other variables, and when weighting them appropriately, are shown in Figure 3.16 and Figure 3.17, respectively. These plots are very similar to the plots in Section 1.9; see Figure 1.19 and Figure 1.20.

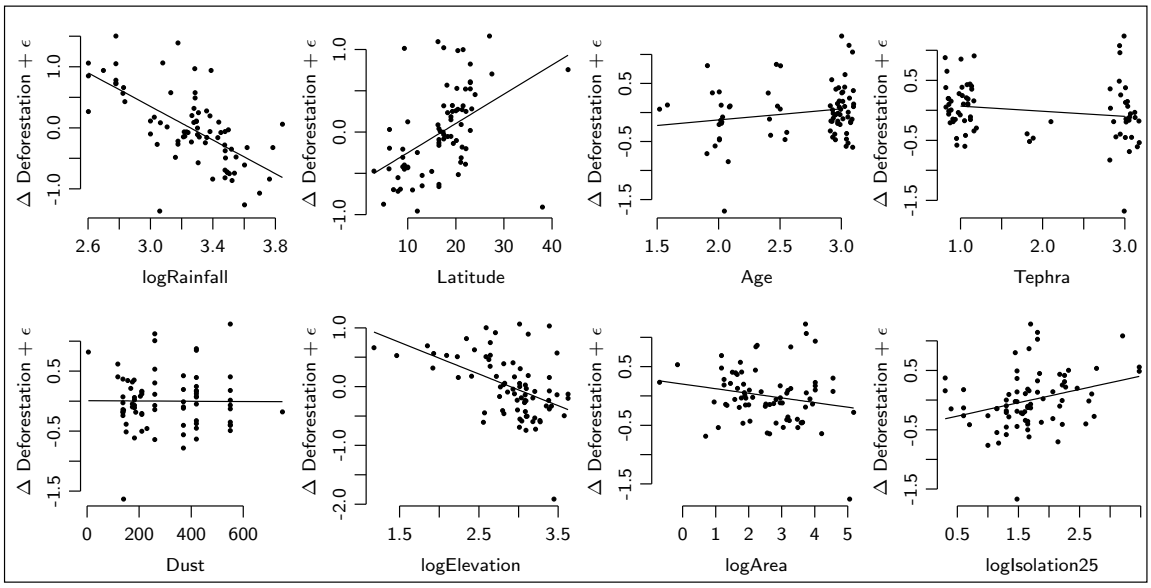


Figure 3.15: Component-plus-residual plots for the island deforestation data, using the ARMS model.

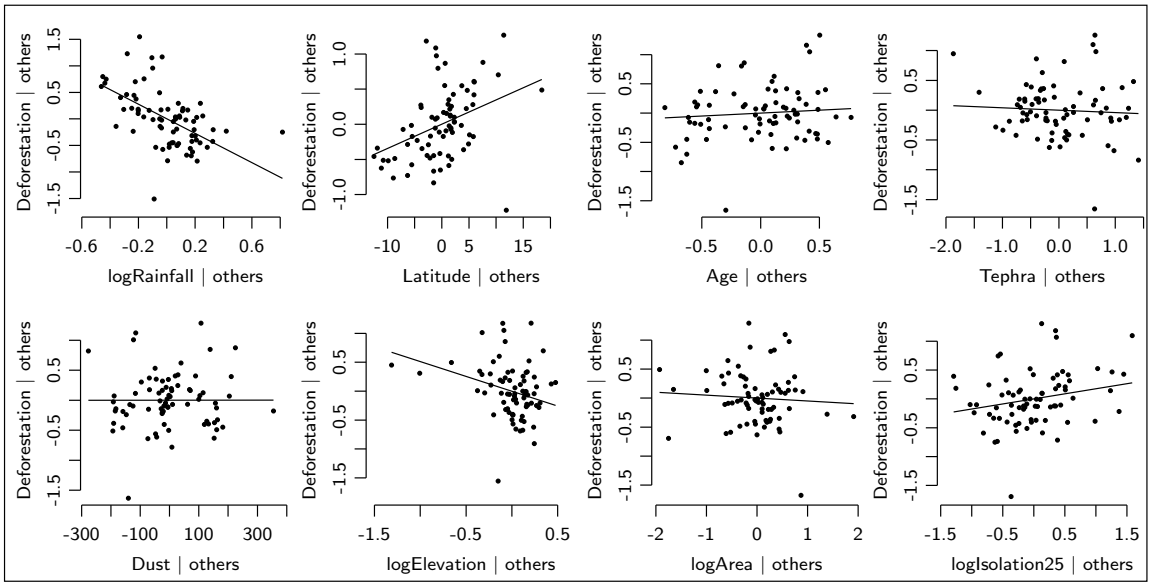


Figure 3.16: Added-variable plots for the island deforestation data, using the ARMS model and fully accounting for all predictors.

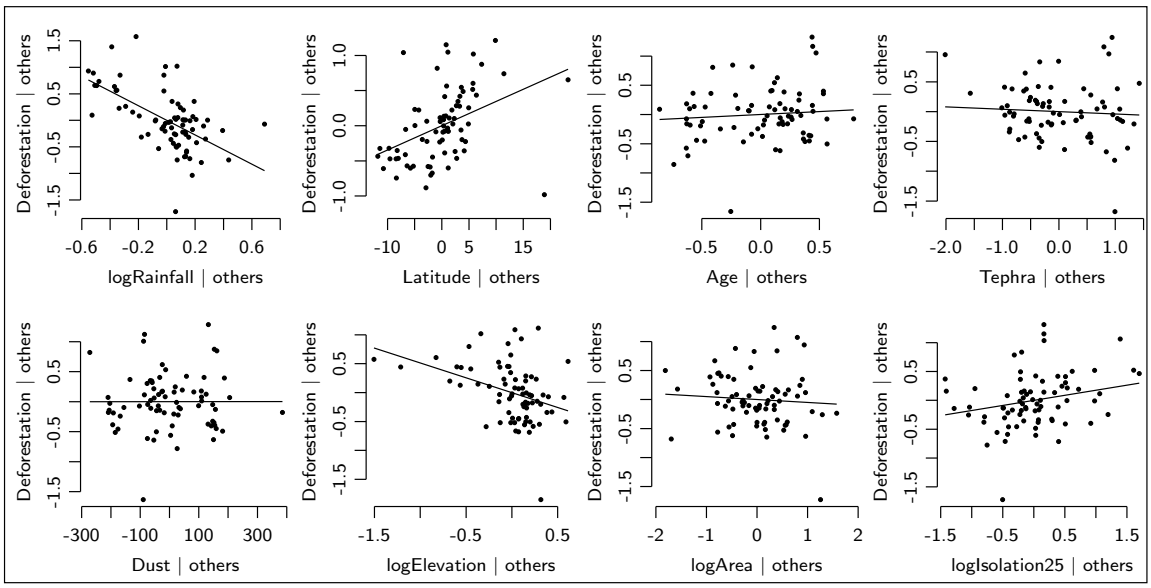


Figure 3.17: Added-variable plots for the island deforestation data, using the ARMS model and weighting the other predictors using the ARMS weights.

Chapter 4

Generalized linear models

While a ordinary linear model assumes $Y|X$ is normally distributed, with the mean equal to a linear combination of the predictors, a generalized linear model (McCullagh and Nelder, 1983) allows $Y|X$ to follow any exponential distribution, with the mean $\mu = E(Y|X)$ equal to some function of a linear combination of the predictors, so

$$g(\mu) = \beta' X.$$

These two generalizations allow for a much greater range of data to be properly analyzed, including binary data and count data.

In this chapter, the proposals for extending the added-variable plot to this generalized setting by Wang (1985) and O'Hara Hines and Carter (1993) will be reviewed and a new version proposed. Additionally, the extension of partial residual plots and CERES plots to this setting by Landwehr et al. (1984) and Cook and Croos-Dabrera (1998), respectively, will be briefly reviewed.

4.1 Added-variable plot

To explore what plots may be useful to show the effect of X_2 on the response Y, consider a Poisson regression example with the canonical log link. In this case, $Y \sim \text{Poi}(\mu)$ and $g(\cdot) = \log(\cdot)$, so $\log(\mu) = \beta' X$. To make the situation clear, let there be

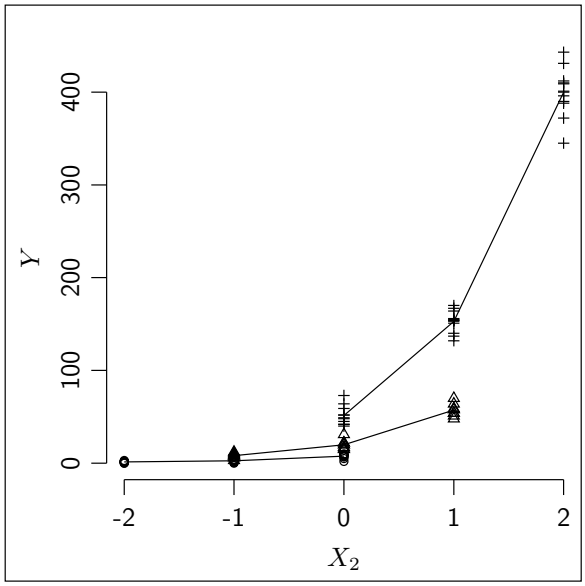


Figure 4.1: Net-effect plot for X_2, from a sample from a Poisson regression model.

two discrete predictors, where X_1 is equally likely to be -1, 0, or 1, and X_2 is equally likely to be $X_1 - 1$, X_1, or $X_1 + 1$. Let $\beta' X = 3 + X_1 + X_2$. Figure 4.1 shows a net-effect plot for a random sample of size 100. For all values of X_1, increasing X_2 is associated with an increase in Y, but the amount of that increase depends on X_1. For example, for points where $X_1 = -1$, shown with circles, X_2 causes only a small increase. But for points where $X_1 = 1$, shown with plus signs, X_2 causes a large increase.

A goal in constructing useful plots can be to combine these net-effect plots so the distribution of the variable on the y-axis, Y^*, given the variable on the x-axis, X^*, is independent of the other variables. In general, this goal is impossible to reach for generalized linear models because the response Y may be discrete. A binary response is the most extreme example.

So instead, the mean and the variance will be studied. Because the variance is not constant but instead may vary with the mean, controlling the mean and the variance simultaneously can result in plots that are difficult to interpret. So the primary focus

of this investigation will be to find plots where

$$E(Y^*|X^*, X_1) = E(Y^*|X^*), \tag{4.1}$$

and methods of either standardizing the variance or displaying it appropriately will be applied after these plots are found. Methods will include using the Pearson residuals on the y-axis, standardizing the variances around the mean function, and multiplying both axes by appropriate weights. This third method is appropriate only when the axes are related linearly, as otherwise it will distort the shape of the data points. Other methods suggested by Landwehr and Pregibon (1993) are to vary the size of the points with the size of the variances or to use a secondary plot with variance information.

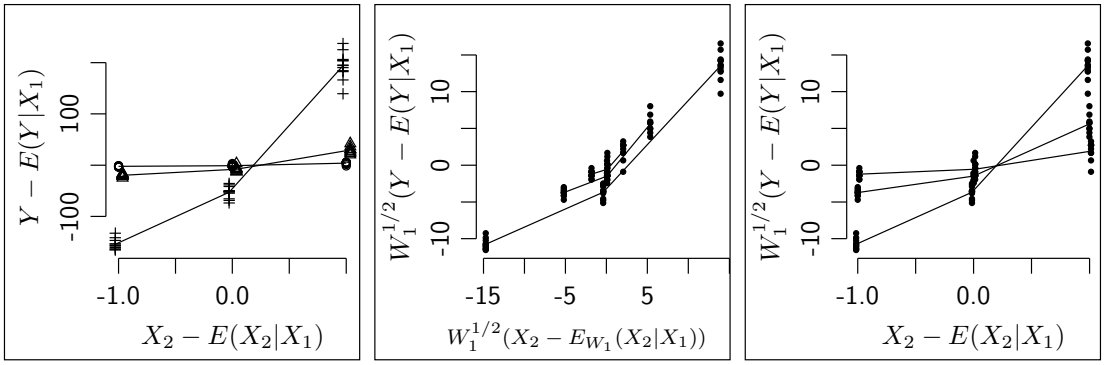
For the contexts thus far, useful plots have been found by adding or subtracting functions of the predictors from the axes. The varied effects of X_2 as shown in Figure 4.1 suggest that adding or subtracting may no longer be sufficient, and that applying some other function to Y may be required.

To begin, consider the plot shown in Figure 4.2a,

$$\{X_2 - E(X_2|X_1), Y - E(Y|X_1)\}, \tag{4.2}$$

where each slice is centered around its mean both vertically and horizontally by subtracting the conditional expected value. This plot is quite understandable, as it plots the amount unexplained by X_1 against the new information in X_2. However, both the mean function and the variance function are still dependent on X_1, as not only are the three lines for the three values of X_1 distinctly different, but so is the spread of the points around the three lines.

In general, the dependence of the mean on X_1 cannot be fixed. Consider the



(a) Centering around the means. (b) Wang version. (c) `Arc` version.

Figure 4.2: Three possible estimated added-variable plots, for a sample from a Poisson model.

position where $X^* = X_2 - E(X_2|X_1) = 0$. Then

$$\begin{aligned} E(Y^*|X^* = 0, X_1) &= E(Y - E(Y|X_1)|X_2 = E(X_2|X_1), X_1) \\ &= E(Y|X_1, X_2 = E(X_2|X_1)) - E(Y|X_1). \end{aligned} \tag{4.3}$$

When g is linear, $E(Y|X_1, X_2) = \beta_1 X_1 + \beta_2 X_2$, so

$$\begin{aligned} E(Y|X_1) &= E(\beta_1 X_1 + \beta_2 X_2|X_1) \\ &= \beta_1 X_1 + \beta_2 E(X_2|X_1) \\ &= E(Y|X_1, X_2 = E(X_2|X_1)), \end{aligned} \tag{4.4}$$

so $E(Y^*|X^* = 0, X_1) = 0$ for all X_1. But when g is not linear,

$$E(Y|X_1) \neq E(Y|X_1, X_2 = E(X_2|X_1)), \tag{4.5}$$

so $E(Y^*|X^* = 0, X_1)$ will not always be 0, and it instead depends on the concavity of g at $\hat{\mu}$, which in turn depends on X_1. Multiplying by a constant cannot remove

this dependence, as it can even be positive for some values and negative for others if g changes concavity, as it does when using a logit link.

Nonetheless, there are two existing methods that use this framework; the first is the added-variable plot proposed by Wang (1985), who multiplies both axes by appropriate weights. This method calculates $E(Y|X_1)$ by fitting a generalized linear model, leaving out X_2, and uses weights W_1 equal to the inverse of the fitted variances. The expected value of $X_2|X_1$ is also estimated using these weights. Because the weights are a function of only X_1, this expected value of this estimate is on average equal to the expected value without the weights, although it is not as efficient. The proposed plot is

$$\left\{W_1^{1/2}\left(X_2 - E_{W_1}(X_2|X_1)\right), W_1^{1/2}(Y - E(Y|X_1)\right\}, \tag{4.6}$$

and is shown in Figure 4.2b. In this example, both the mean function and the variance function still depend on X_1.

One of the reasons that this plot does not properly standardize the conditional variance of $Y^*|X^*$ is because the method of calculating the conditional mean and variance is appropriate only if X_2 has no effect. If X_2 does have an effect, the variance will be underestimated because the variation due to X_2 is not included in the model. The values on the y-axis of Figure 4.2b show that this has happened in this case. Multiplying by the square root of the weights results in Pearson residuals, which should mostly be within 2 standard deviations of the mean. The values on this axis range from -10 to 10, showing that the extra variation due to X_2 has not been accounted for. Also, this method of standardizing the variance is only appropriate when the fit is linear, as otherwise it can disguise the pattern of the points.

Additionally, this plot has little explanatory value; having $W_1^{1/2}(X_2 - E_{W_1}(X_2|X_1))$ on the x-axis does not help in understanding the role of X_2. Instead, the derivation of

this plot by Wang (1985) shows that it can be interpreted as the graphical version of the hypothesis test that $\beta_2 = 0$, and that a nonzero slope shows evidence for $\beta_2 \neq 0$. Because this hypothesis test assumes that β_2 really is 0 and X_2 has no effect, it is not surprising that it lacks explanatory value when β_2 is not 0 and X_2 does have an effect.

The second method also uses the Pearson residuals, as in the Wang (1985) plot, but only on the y-axis. This plot is

$$\left\{X_2 - E(X_2|X_1), W_1^{1/2}\big(Y - E(Y|X_1)\big)\right\}, \tag{4.7}$$

and is used in `Arc` (Cook and Weisberg, 1999a). It plots the Pearson residual of the fit only involving X_1 against the new information in X_2, and is shown in Figure 4.2c. It also has dependence on X_1 in both the mean and the variance.

Instead of using the usual scale for the response, consider instead the linear scale as used in the fitting of the generalized linear model. This iterative process can be understood as follows. The first step is converting the data to the linear scale of $g(\mu) = X\beta$. This cannot be done by using $g(Y)$, because it may not exist or be finite for all data points. For example, in logistic regression, $g(Y) = \log(Y)$, so for $Y = 0$, $g(Y)$ is not finite. Instead, a first-order Taylor approximation is used,

$$g(Y) \approx g(\hat{\mu}) + g'(\hat{\mu})(Y - \hat{\mu}),$$

where $\hat{\mu}$ is the current best estimate of μ. This value is called the adjusted dependent variable and will be denoted by Z. Now that the observations are on a linear scale, weighted least squares is used to make a new estimate of $\hat{\mu}$, with weights equal to the inverse of the fitted variances for each point. This process is repeated until the fitted

values do not change. At the end of this procedure,

$$\begin{aligned} Z &= \beta' X + g'(\hat{\mu})(Y - \hat{\mu}) \\ &= g(\hat{\mu}) + e \end{aligned} \tag{4.8}$$

where e are called the working residuals and are independent of Z and X.

Now instead of the plot of (4.2), which was on the response scale, consider the plot

$$\{X_2 - E(X_2|X_1), Z - E(Z|X_1)\} \tag{4.9}$$

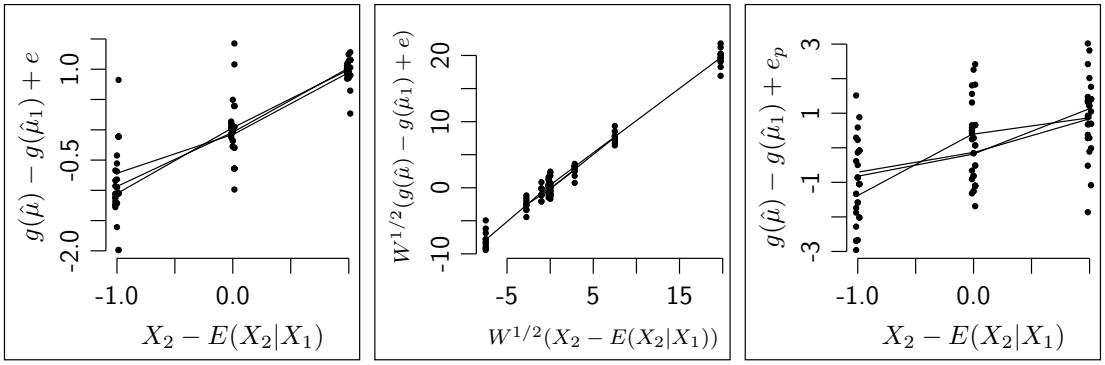
as shown in Figure 4.3a, which is on the linear scale. In this plot, the mean function does not depend on X_1, although the variance still does. This independence of the mean function can be derived; $Z = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + e$, so

$$\begin{aligned} E(Z|X_1) &= E(E(Z|X_1, X_2)|X_1) \\ &= E(\hat{\beta}_1 X_1 + \hat{\beta}_2 X_2|X_1) \\ &= \hat{\beta}_1 X_1 + \hat{\beta}_2 E(X_2|X_1) \end{aligned} \tag{4.10}$$

and

$$\begin{aligned} E(Y^*|X^*, X_1) &= E(Z - E(Z|X_1)|X^* = X_2 - E(X_2|X_1), X_1) \\ &= E(Z|X_1, X_2 = E(X_2|X_1) + X^*) - E(Z|X_1) \\ &= (\hat{\beta}_1 X_1 + \hat{\beta}_2(E(X_2|X_1) + X^*) - (\hat{\beta}_1 X_1 + \hat{\beta}_2 E(X_2|X_1)) \\ &= \hat{\beta}_2 X^*. \end{aligned} \tag{4.11}$$

.



(a) With working residuals. (b) Standardizing the variance by multiplying by weights. (c) Standardizing the variance by using Pearson residuals.

Figure 4.3: Three new estimated added-variable plots, for a sample from a Poisson model.

This plot can also be interpreted by writing it terms of $g(\mu)$; the y-axis is

$$Z - E(Z|X_1) = g(\hat{\mu}) - g(\hat{\mu}_1) + e \tag{4.12}$$

where $\hat{\mu}_1 = \hat{\beta}_1 X_1 + \hat{\beta}_2 E(X_2|X_1)$. So the y-axis is the amount that the fitted means on the linear scale change between using all the information about X_2 and using only $E(X_2|X_1)$, plus an error term on the linear scale. The x-axis is the new information in X_2, beyond that in X_1. So this plot shows how much information can be explained by X_1 and X_2 together beyond that explained by X_1 and $E(X_2|X_1)$. A nice feature of this plot is that the fitted slope is equal to $\hat{\beta}_2$, showing that for every unit increase in X_2 over $E(X_2|X_1)$, the response on the linear scale has an estimated increase of $\hat{\beta}_2$.

The variance in this plot still does depend on X_1. Depending on the purpose, this may be a good thing; it forces one to realize that the amount the response may change, as measured on the linear scale, is not constant for all values of the predictors. Still, two ways of standardizing the variance are shown in Figure 4.3b and Figure 4.3c. The first method is to multiply both axes by the appropriate weights, as in the Wang

(1985) plot. It is more successful here because the fit really is linear, so it maintains the pattern of the points. However, it loses interpretability; the values on the x-axis now are difficult to relate to the actual value of X_2.

The second method of standardizing the variance is to use the Pearson residuals,

$$e_p = (y - \hat{\mu})/\sqrt{\text{Var}(y)} \tag{4.13}$$

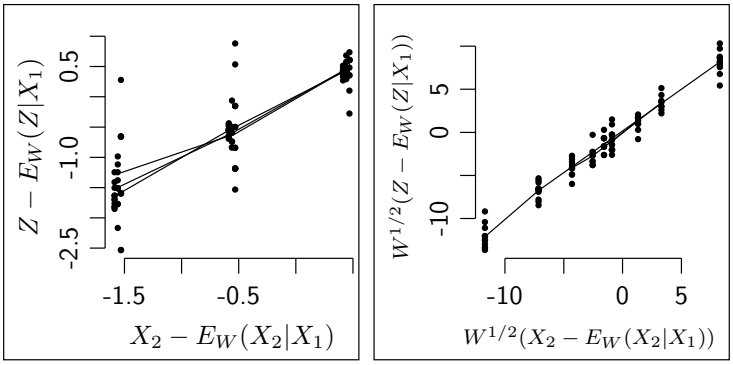
instead of the working residuals e in (4.12), shown in Figure 4.3c. Neither the mean or the variance of this plot depend on X_1, though it is unclear if it aids in understanding the situation. Indeed, it may be more confusing. The mean function remains on the linear scale, so the plot still shows that for every unit increase in X_2, the response changes by β_2 units on the linear scale. But the variance function is on the standardized scale, so looking at the points at X_2 near -1 in Figure 4.3c does not mean that the response is changing between -3 and 1 units on the linear scale, but that it is changing on average about -1 units on the linear scale, with variation of up to 2 standard deviations in each direction.

As in Section 2.4, a more general version of the added-variable plot of (4.9) is

$$\{X_2 - f(X_1), Z - E(Z|X_1, X_2 = f(X_1))\}. \tag{4.14}$$

As in (4.11), for this plot,

$$\begin{aligned} E(Y^*|X^*, X_1) &= E(Z - E(Z|X_1, X_2 = f(X_1))|X^* = X_2 - f(X_1), X_1) \\ &= E(Z|X_1, X_2 = f(X_1) + X^*) - E(Z|X_1, X_2 = f(X_1)) \\ &= \left(\hat{\beta}_1 X_1 + \hat{\beta}_2(f(X_1) + X^*\right) - \left(\hat{\beta}_1 X_1 + \hat{\beta}_2 f(X_1)\right) \\ &= \hat{\beta}_2 X^*. \end{aligned} \tag{4.15}$$



(a) Without multiplying by weights.

(b) With multiplying by weights.

Figure 4.4: Estimated added-variable plot suggested by O'Hara Hines and Carter (1993).

O'Hara Hines and Carter (1993) also noted problems with the added-variable plot of Wang (1985). Their focus was on its ability to show influential points, and they suggested using

$$\{X_2 - E_W(X_2|X_1), Z - E_W(Z|X_1)\} \tag{4.16}$$

with the variance corrected by multiplying both axes by the appropriate weights. The notation E_W refers to the expected value using the weights W equal to the inverse of the fitted variances. All of their expectations, as well as the weights, are estimated using the full model. This is an instance of the general version of (4.14) with $f(X_1) = E_W(X_2|X_1)$. It is not very interpretable for explanatory purposes, as it is difficult to find any meaning in fitting X_2 to X_1 given weights derived from the response. However, it does accurately represent the actual influence of each of the data points in the fit, so is quite appropriate for diagnostics. It also has a slope equal to $\hat{\beta}_2$. Figure 4.4a and Figure 4.4b show (4.16) without and with the weights.

4.2 Component-plus-residual plot

Partial residual plots for generalized linear models were first proposed by Landwehr et al. (1984). For the variable X_2, the component from X_2 and the residuals from a model with linear mean function

$$g(u) = \beta_1 X_1 + \beta_2 X_2$$

are plotted,

$$\{X_2, \hat{\beta}_2 X_2 + e\},$$

where $e = g'(\hat{\mu})(Y - \hat{\mu})$ are the working residuals. O'Hara Hines and Carter (1993) show that the slope of this plot does not match $\hat{\beta}_2$, and suggest that a weighted version

$$\{W^{1/2} X_2, \hat{\beta}_2 W^{1/2} X_2 + e_p\},$$

where $e_p = (y - \hat{\mu})/\sqrt{\text{Var}(y)}$ are the Pearson residuals, better shows influence. These plots for X_2 from the Poisson model are shown in Figure 4.5.

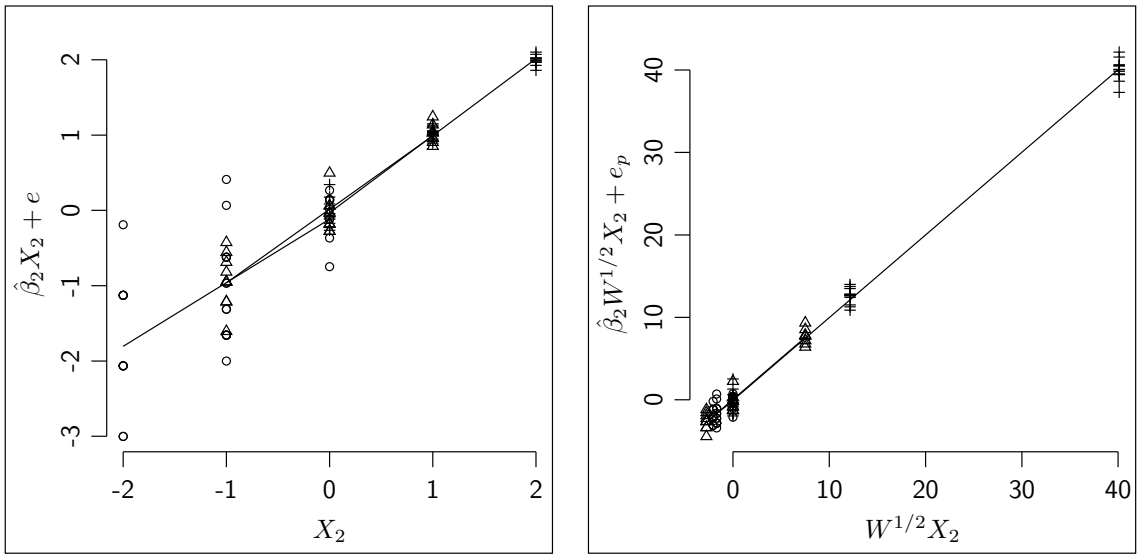
As in the additive model case, both of these versions require linearity of $E(X_1|X_2)$. Cook and Croos-Dabrera (1998) address this requirement by extending CERES plots to the generalized linear model, but find an another requirement is necessary: that the inverse link function g^{-1} stays away from its extremes so g is essentially linear. This plot uses the coefficients from the model with mean function

$$g(\mu) = \beta_1 X_1 + \beta_2 l(X_2),$$

where $l(X_2)$ is a function, perhaps multidimensional, that captures the behavior of $E(X_1|X_2)$. The CERES plot for X_2 is then

$$\{X_2, \hat{\beta}_2 l(X_2) + e\}.$$

Cook and Croos-Dabrera show this is unbiased for the true component of X_2 only when g is essentially linear and provide guidelines for commonly used link functions.



(a) Ordinary partial residual plot.

(b) Weighted partial residual plot.

Figure 4.5: Partial residual plots for X_2, sliced over X_1, for a sample from a Poisson model.

Chapter 5

Interactions

It has been shown how various plots are useful for determining and displaying the effect of individual variables when the underlying model is linear or additive, even when complex weighted methods are used to fit the model. But in many cases, additivity does not hold and interactions exist between two or more of the predictors.

In some of these cases, the marginal relationship of Y and X_1 may be of interest, even when there is an interaction present and

$$F(Y|X_1, X_2 = x_{20})$$

is completely different than

$$F(Y|X_1, X_2 = x_{21}).$$

This interaction is irrelevant when investigating this marginal relationship, and the marginal response plot is sufficient. If, however, the conditional relationships are of interest, the plots developed so far may not be sufficient or appropriate. Plots useful for these cases will now be investigated.

5.1 Net-effect plots

The net-effect plots of Section 1.3 are again useful here. To show the net-effect plot for X_1, the marginal plot of $\{X_1, Y\}$ is constructed, using symbols to show which data

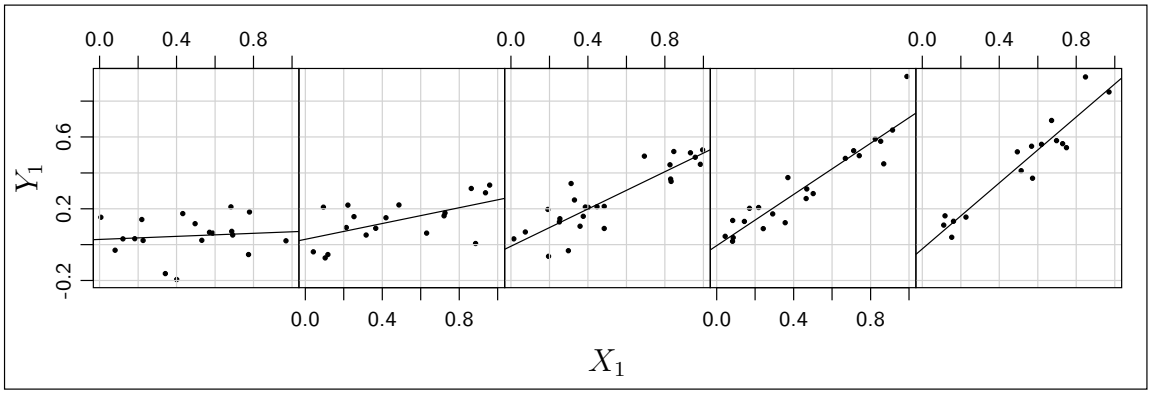


Figure 5.1: Conditioning plots for X_1, split over five ranges of X_2 values.

points have similar values of X_2. Overall, this plot shows the marginal effect of X_1, and for each symbol type, it shows the conditional effect for that range of X_2 values. This is especially useful when interactions are present, because the conditional effect of X_1 will change as X_2 changes.

Example 5.1

Let X_1 and X_2 be independent random variables, uniformly distributed between 0 and 1. Suppose that the response is $Y_1 = X_1X_2 + 0.1\epsilon$, where $\epsilon \sim N(0, 1)$. The net-effect plot for X_1 given X_2 is shown in Figure 5.1. Here X_2 is sliced into five equal parts, and for clarity, the five corresponding plots are shown separately, using a conditioning plot (Cleveland, 1993). Fitted lines have also been added to emphasize the effects.

The effect of X_1 depends on X_2, because the relationship between X_1 and Y_1 is different in the five subplots. In the first plot, X_2 varies between 0 and 0.2, and because $E(Y_1|X) = X_1X_2$, the slope of the fitted line is small, between 0 and 0.2. And in the last plot, X_2 varies between 0.8 and 1, so the slope is larger, between 0.8 and 1. □

This type of plot is similar to the interaction plot used when the predictors are factors, as shown by Cook and Weisberg (1999a, pp. 302–303). However, in that case, there is no need to slice the other variable, as then the subplots can be made for each value of the factor.

But just as in the linear and additive cases, these plots can be quite complex and difficult to interpret. When the underlying model is additive and the conditional effects are the same, it was possible to combine the plots together in various ways to get a simpler plot that was conditionally independent of the other variables. But because the conditional effects are different when interactions are present, these methods must be reconsidered to determine if and how they might still be useful.

5.2 Component-plus-residual plot

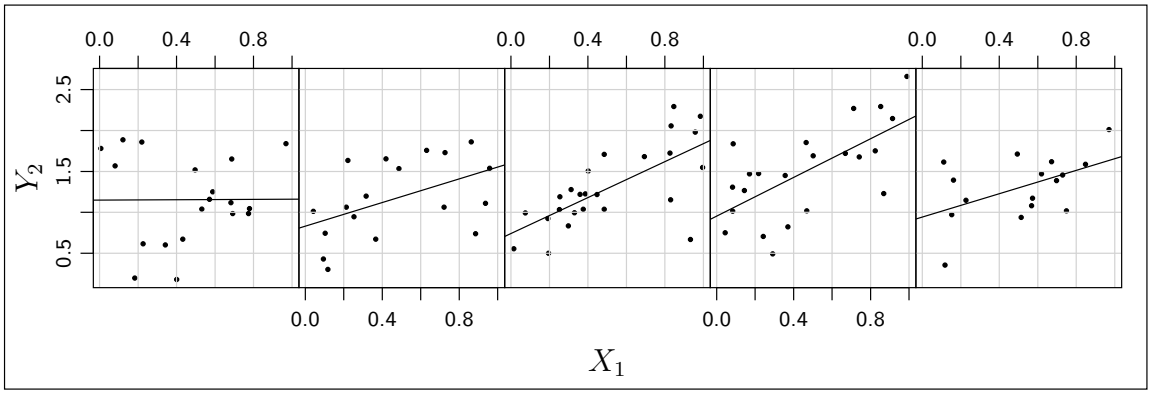
The component-plus-residual plot can be especially useful when two of the variables interact only with each other, so the mean function is

$$E(Y|X) = f_1(X_1, X_2) + f_2(X_3, \ldots, X_p).$$

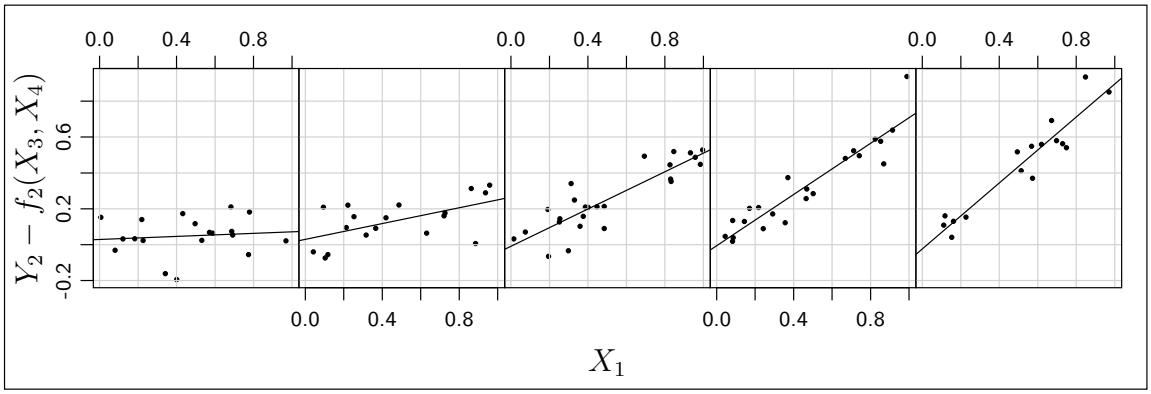
Here the variability due to variables $X_3, \ldots, X_p$ may overwhelm the variability due to X_1 and X_2, so the effects of X_1 and X_2 may not show up in a net-effect plot. An alternative is to create a three-dimensional component-plus-residual plot for X_1 and X_2 to only show the effect of X_1 and X_2; this plot has $Y - f_2(X_3, \ldots, X_p)$ on the vertical axis and X_1 and X_2 on the horizontal axes. As three dimensional plots can be difficult to visualize without the ability to view the plot from multiple angles, as on the printed page, an alternate is to use conditioning plots to show what various slices of the three-dimensional plot look like.

Example 5.2

Suppose $Y_2 = X_1X_2 + X_3 + X_4 + 0.1\epsilon$, where each of the X_i's are independent Unif$(0, 1)$



(a) Net-effect plot for X_1 given X_2, using a conditioning plot.



(b) Component-plus-residual plot for X_1 and X_2, using a conditioning plot.

Figure 5.2: Net-effect and component-plus-residual plots for X_1 and X_2 from Example 5.2; the component-plus-residual plot makes the effect of the interaction between X_1 and X_2 clearer.

random variables, and ϵ is a independent standard normal. A net-effect plot for X_1 given X_2, where X_2 is sliced into five equal regions, is shown in Figure 5.2a. Although the relationship between X_1, X_2, and Y_2 is the same as the relationship in Example 5.1, it is not clear visually because of the extra variability added by X_3 and X_4. Fitting the linear model with interactions only between X_1 and X_2, the coefficients for X_3 and X_4 are 1.02 and 1.02, respectively. Figure 5.2b shows the component-plus-residual plot for X_1 and X_2, by conditioning on the same five regions of X_2 as in previous plots. The behavior of the interaction is now clearer; for small values of X_2, the effect of X_1 is small, but it increases as X_2 increases. □

When a variable X_k is involved in an interaction, a one variable component-plus-residual plot usually is not the right thing to do, as it shows the effect of X_k after accounting for the other variables. When interactions are present, this effect is different depending on what the other variables are, so it is impossible to remove the effects of the other variables without also removing some of the effect of X_k.

Nevertheless, there are two ways this might be done; the first is to make the plot as if the underlying model was additive. Consider splitting the interaction up into three parts, so

$$f(X_1, X_2) = f_1(X_1) + f_2(X_2) + f_{12}(X_1, X_2),$$

where f_{12} contains no additive functions of X_1 or X_2. Then if the underlying model is additive, $f_{12} = 0$, and the component-plus-residual plot for X_1 plots $Y - f_2(X_2)$, or $f_1(X_1) + f_{12}(X_1, X_2) + \epsilon$, against X_1. But the conditional expectation of the y-axis given the x-axis is then

$$E(f_1(X_1) + f_{12}(X_1, X_2) + \epsilon | X_1) = f_1(X_1) + E(f_{12}(X_1, X_2)|X_1)$$

which still depends on X_2, so the dependence on X_2 has not been removed.

Alternatively, the full effect of the other variable could be subtracted from Y, including how it may change with the variable of interest; this plot would be $Y - f_2(X_2) - f_{12}(X_1, X_2) = f_1(X_1) + \epsilon$. This plot no longer conditionally depends on X_2, as

$$E(f_1(X_1) + \epsilon | X_1) = f_1(X_1).$$

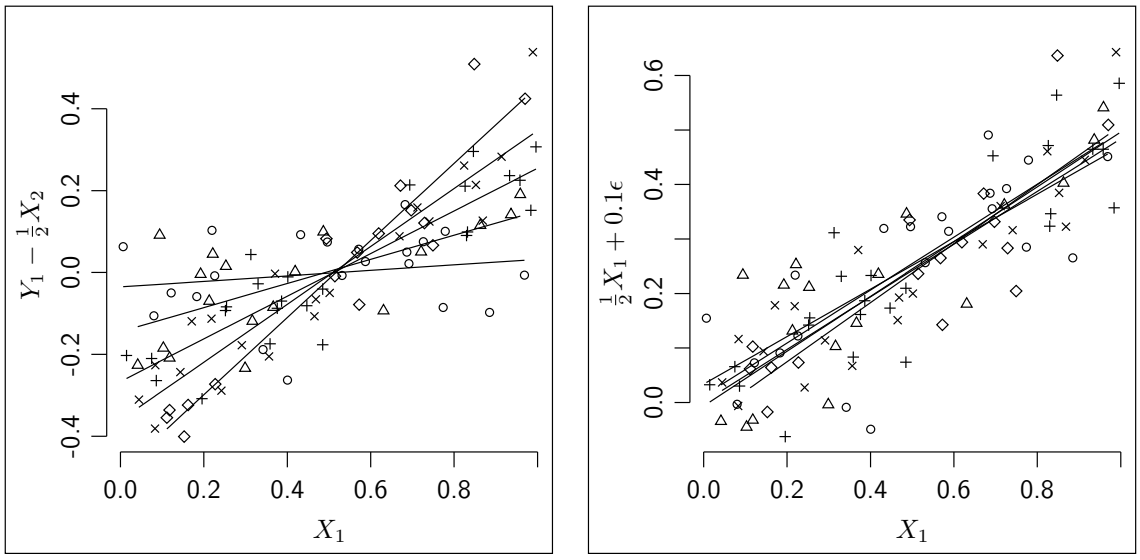
However, it does not show the full effect of the variable of interest, only the additive part. When only the additive part is of interest, this may be useful; but it must be remembered that the variable has a more complex relationship with the response than is pictured.

Example 5.3

Consider the data set in Example 5.1 where $Y_1 = X_1 X_2 + 0.1\epsilon$ and X_1 and X_2 are Unif$(0, 1)$. In this case, the interaction can be split up as follows:

$$Y_1 = \tfrac{1}{4} + \tfrac{1}{2}X_1 + \tfrac{1}{2}X_2 + \left(X_1 - \tfrac{1}{2}\right)\left(X_2 - \tfrac{1}{2}\right) + 0.1\epsilon.$$

The first option of removing only the additive part of X_2 by plotting $Y - \frac{1}{2}X_2$ against X_1 is shown in Figure 5.3a, with slices over X_2. The varying slopes of the slices show that the X_2 dependence has not been removed. The second option is shown in Figure 5.3b. As the slopes of the slices are now the same, the dependence on X_2 has been removed; but this plot now only shows the additive part of X_1, with slope of $\frac{1}{2}$. When fully accounting for X_2, as in the net-effect plot of Figure 5.1, the slope instead varies from 0 to 1. □



(a) Component-plus-residual plot for X_1, removing only additive part of X_2. Slices over X_2 show that the X_2 dependence has not been removed.

(b) Component-plus-residual plot for X_1, leaving only additive part of X_1.

Figure 5.3: Component-plus-residual plots for X_1 from Example 5.3, using two different ways of accounting for X_2 and the interaction between X_1 and X_2.

5.3 Added-variable plot

When the variable X_k is involved in an interaction with another variable, the general added-variable plot of

$$\{E(Y|X) - E(Y|X_{\setminus k}), Y - E(Y|X_{\setminus k})\}$$

or the detrended version with a slope of zero, is also appropriate. It plots the residuals from the model without X_k against the change in fitted values between the models with X_k and without X_k.

Additionally, when the fitting method used can either include interactions or not, an added-variable plot can be made for the interaction itself. Letting M_{int} be the model where X_k is allowed to interact with other variables, and M_{add} the model where it is not, then the added-variable plot is

$$\{E_{\text{int}}(Y|X) - E_{\text{add}}(Y|X), Y - E_{\text{add}}(Y|X_{\setminus k})\},$$

where E_{int} and E_{add} are the expected values with and without the interaction, respectively.

5.3.1 Interactions compared with correlated predictors

It can be instructive to compare the effect of an interaction with that of correlated predictors when displaying the effect of individual predictors using general added-variable plots. These plots have the change in fitted values with and without the predictor on the x-axis; the y-axis is either the residuals from the fit without it, which is the change in fitted values plus the residuals from the full model, or to detrend it to have a zero slope, the residuals from the full model.

When comparing several predictors, the range of the x-axis, which is the change in fitted values, is one of the easily interpreted parts of the plot. When the range of the x-axis is small, the effect of the variable can be considered small as it doesn't

change the fit very much; and when it is larger, the effect of the variable can be considered larger. However, both interactions and correlated predictors can complicate this interpretation, in opposite ways.

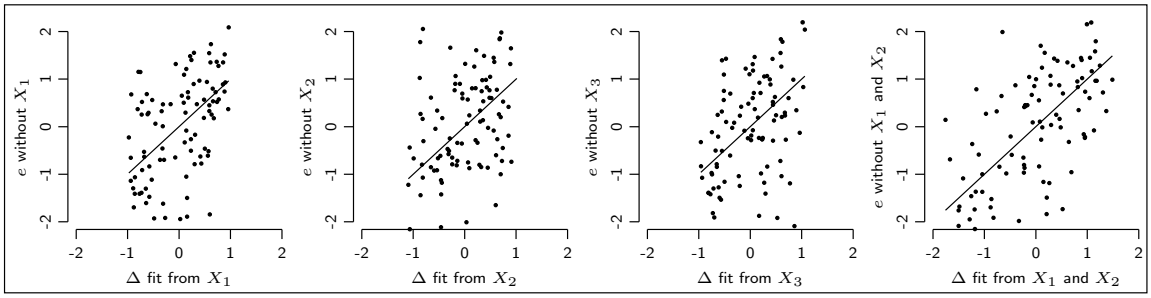
When two predictors are correlated, the change in fitted values from each predictor may be small, because the information contained in one is mostly contained in the other. Conversely, when two predictors interact, the change in fitted values from each predictor may be large, because removing one predictor may also remove most of the influence of the other as well.

Example 5.4

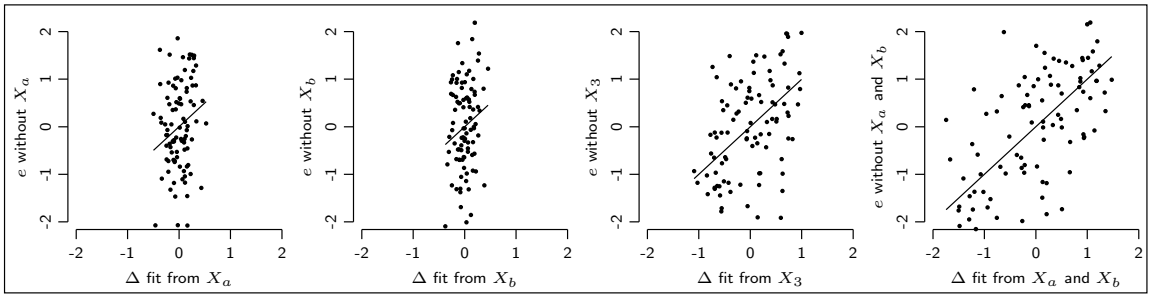
Consider the following three situations, each with three predictors. First, let X_1, X_2, X_3, be independent Unif$(-1, 1)$ random variables and $Y = X_1 + X_2 + X_3 + \epsilon$, where $\epsilon \sim N(0, 1)$. Then each predictor is related to the response in exactly the same way. General added-variable plots for each variable are shown in Figure 5.4a, along with the added-variable plot for the combined effect of X_1 and X_2. The range of the x-axis is about the same for the plots for each of the three variables.

Now, let $X_a = X_1 + X_2 + 0.2\epsilon_2$ and $X_b = X_1 + X_2 - 0.2\epsilon_2$, where $\epsilon_2 \sim N(0, 1)$. Here X_a and X_b are correlated, with $\rho = 0.9$. Consider the model with the three predictors X_a, X_b, and X_3; overall it is just as good as the previous model because $X_a + X_b = X_1 + X_2$, but because of the correlation the apparent effect of X_a and X_b will be smaller. General added-variable plots for each variable are shown in Figure 5.4b, along with the added-variable plot for the combined effect of X_a and X_b. The range of the x-axis of the X_a and X_b plots are now smaller than the range for the X_3 plot. However, the combined effect of X_a and X_b remains the same as the effect of X_1 and X_2 in the uncorrelated case.

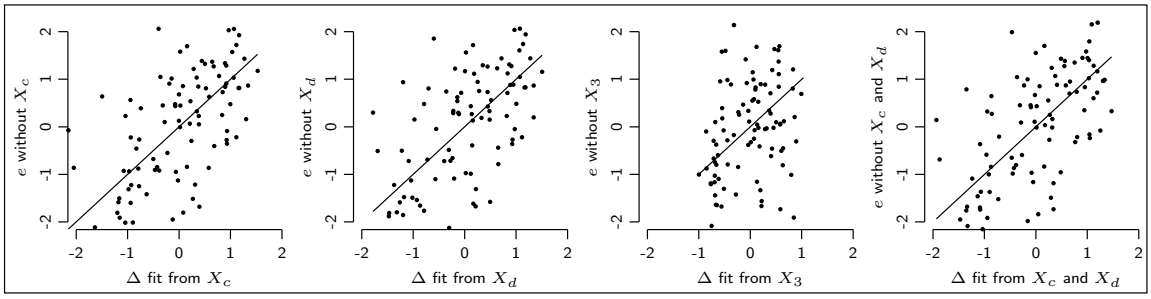
Finally, let X_c be a random sample from $\{-1, 1\}$, and let $X_d = X_c(X_1 + X_2)$, and consider the model with the variables X_c, X_d, and X_3. The true model is now



(a) For uncorrelated predictors, with no interaction.



(b) For correlated predictors, with no interaction.



(c) For uncorrelated predictors, with an interaction.

Figure 5.4: General added-variable plots for the situations in Example 5.4, showing the effect of interactions and correlation.

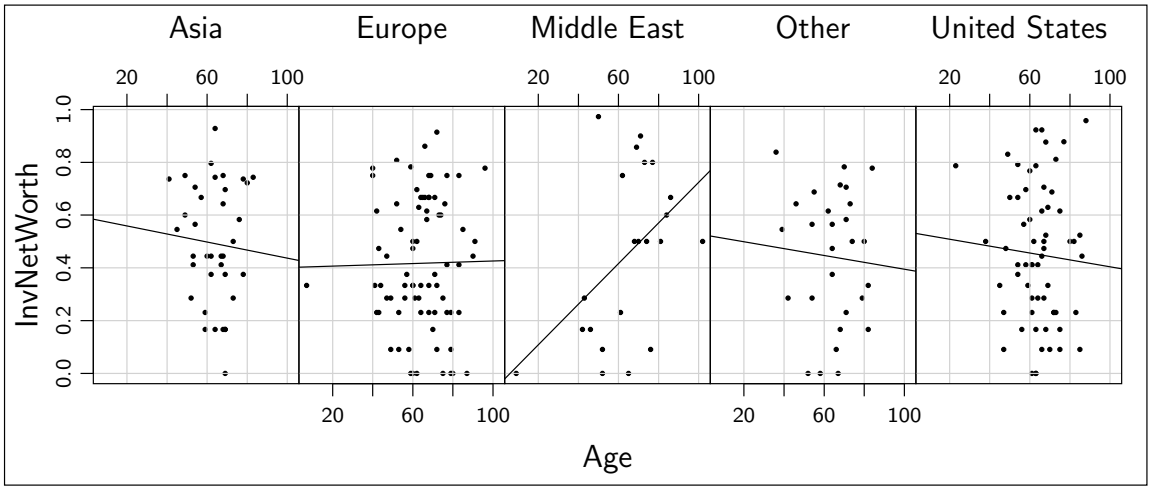


Figure 5.5: Net-effect plot for Age, from the billionaires data set.

$Y = X_c \times X_d + X_3$, so there is now an interaction in the model. General added-variable plots for each variable are shown in Figure 5.4c, along with the added-variable plot for the combined effect of X_c and X_d. Now, the range of x-axis for the X_c and X_d plots are larger than that for the X_3 plot, while the combined effect again remains the same as in the previous cases. □

5.4 Example: Billionaires

Each year, Fortune magazine publishes a list of the world's billionaires. In 1992 the list included 233 individuals, along with their net worth, age, and region (Asia, Europe, Middle East, United States, and Other). The net worth variable is very skewed, so the inverse transformation $1 - 1/\text{net worth in billions}$, called InvNetWorth, was used to make it more symmetric.

Figure 5.5 shows a net-effect plot for age conditioned over the five regions. The effect of age is minimal for each region except the Middle East, where older individuals tend to have more wealth, indicating an interaction between age and region. Figure 5.6

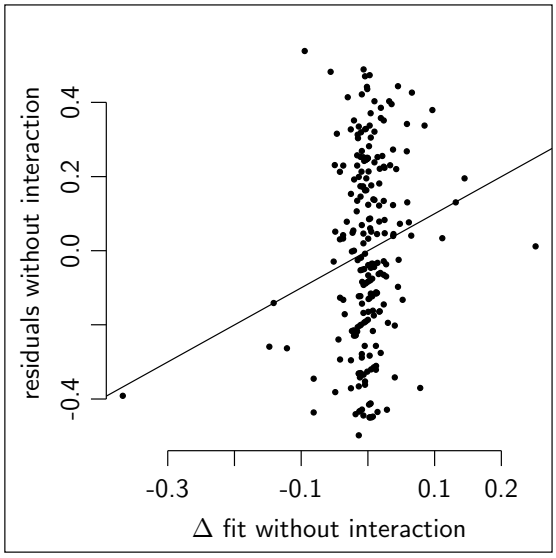


Figure 5.6: Added-variable plot for the interaction between Age and Region, from the billionaires data set.

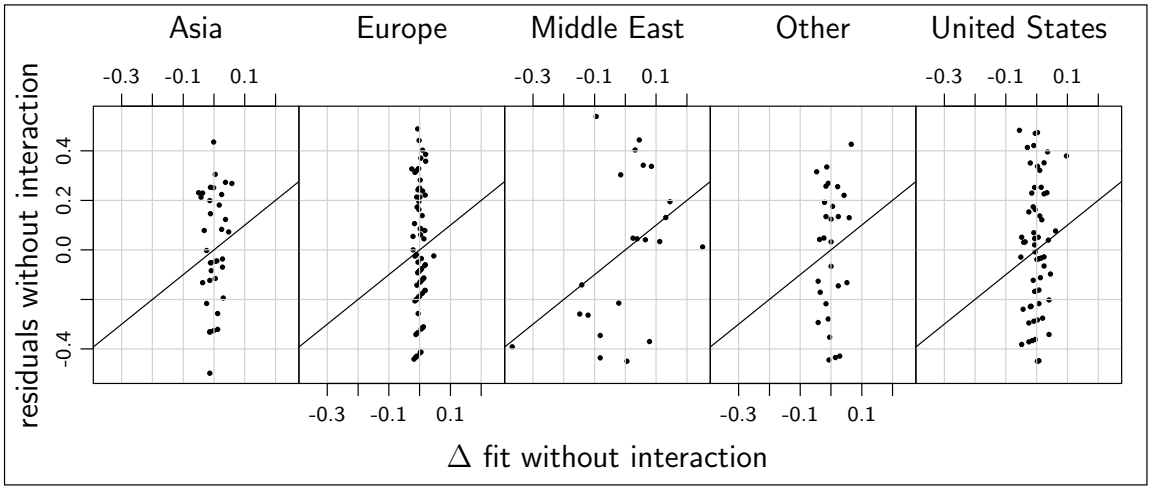


Figure 5.7: Added-variable plot for the interaction between Age and Region, conditioned on Region, from the billionaires data set.

shows the general added-variable plot for this interaction with a line of slope one added. Most of the points are centered in the middle of the plot and so the interaction has a small effect on them, but a few fall farther out. Figure 5.7 shows this same added-variable plot, but conditioned over the five regions; the points that change most when the interaction is added are shown to be those from the Middle East region.

Chapter 6

Black box methods

Many methods and algorithms exist for modeling the relationship between a set of predictors and a given response that do not rely on assumptions such as linear or additivity, as most of the methods used so far do. Many of these are quite complex and difficult to understand. To the end user, in fact, they often seem like black boxes, which take a set of predictors as input, and give a predicted value for the response as output.

Tree based methods are good examples of these kinds of algorithms. A tree based method is an algorithm that creates a decision tree for predicting the value of either a categorical or continuous response from a given set of predictor values. Unlike conventional statistical methods, tree based methods are nonparametric and nonlinear, so are not concerned with estimating a few population parameters using statistical inference, but instead with making the best prediction about the response. This flexibility allows these types of models to fit data that may be difficult or impossible to fit using conventional methods. Two of the most popular algorithms are CART (Breiman et al., 1984), and C4.5 (Quinlan, 1993).

In one sense, the results from these methods are easy to understand; to get a prediction from a set of values, just follow the if-then decisions given by the tree. It is clear what the method is doing and how the prediction is found. However, because

the tree may split on an individual predictor multiple times and in multiple branches, the particular effect of that predictor may be difficult to understand.

The basic idea of tree based methods is to divide the data into several parts, or branches, by splitting on one of the predictors, where the split is chosen to minimize prediction error. The process continues by splitting the branches until the tree is considered optimal. The method of calculating prediction error depends on the type of response; for categorical responses, there are several measures of node impurity, such as misclassification error, and for continuous responses, a common measure is mean squared error. The prediction error can also be weighted to give more or less importance to individual data points, or to a particular type of error.

In principle, splitting could continue until the tree perfectly explains the data set, with the exception of data points that have identical predictors but different responses. However, a tree like this will not be useful because it will be just as complicated as the original data set, and it will not be accurate for predicting new observations because the tree will have fit all characteristics of the particular data set, even those which are solely due to random variation. So the size of the finished tree must be restricted, by stopping construction early, by pruning the final branches back, or by some combination of these procedures.

To stop the construction, splits are not made to branches that are too small, measured as a fraction of either the entire data set, or for classification trees, of the subset of a particular categorical response. To decide which branches to prune back, several cross-validation variations have been proposed.

Breiman (2001) extended this tree methodology by combining multiple trees together, calling the new result a random forest. An advantage of this method is that it does not usually have the sharp breaks that a single tree would have. However, the resulting forest is more difficult to interpret and understand than a single tree because there is no longer a straightforward decision tree describing the relationship

between the predictors and the response.

In another variation, Friedman and Popescu (2005) explore rule-based methods, which do not use the entire tree that has been built, but instead use the individual if-then statements, or rules. The prediction model is the optimal weighted combination of the rules formed by creating many possible trees.

These methods all produce algorithms for predicting the response Y for a particular set of predictor variables X, but this prediction does not necessarily correspond to the conditional expected value $E(Y|X)$; for example, it might instead correspond to the conditional median or mode. To account for this possibility, in this section, E will more generally denote the prediction resulting from the black box.

6.1 Added-variable plot

In general, these black box methods allow for interactions; if they are present, the general added-variable plot

$$\{E(Y|X) - E(Y|X_{\setminus k}), Y - E(Y|X_{\setminus k})\}$$

will be the only appropriate two-dimensional plot for showing the effect of X_k, as discussed in Chapter 5. This plot can also be made in its alternate form of

$$\{E(Y|X) - E(Y|X_{\setminus k}, X_k = E(X_k|X_{\setminus k})), Y - E(Y|X_{\setminus k}, X_k = E(X_k|X_{\setminus k}))\},$$

as described in Section 2.4.

If a term X_k enters the model linearly, where

$$E(Y|X) = \beta_k X_k + f_{\setminus k}(X_{\setminus k}),$$

an ordinary added-variable plot can be constructed from the black box, again using either the change in predicted values when ignoring X_k,

$$\{X_k - E(X_k|X_{\setminus k}), Y - E(Y|X_{\setminus k})\}$$

or the change in predicted values when setting $X_k = E(X_k|X_{\setminus k})$,

$$\{X_k - E(X_k|X_{\setminus k}), Y - E(Y|X_{\setminus k}, X_k = E(X_k|X_{\setminus k}))\}.$$

6.2 Component-plus-residual plot

If a term enters additively, so

$$E(Y|X) = f_k(X_k) + f_{\setminus k}(X_{\setminus k}), \tag{6.1}$$

then a component-plus-residual plot is also appropriate for X_k. But since a black box method does not necessarily fit f_k and $f_{\setminus k}$ separately, the usual method of plotting $f_k(X_k) + \epsilon$ is impossible. In this case the alternate version of the plot described in Section 2.4,

$$\{X_k, Y - E(Y|X_k, X_{\setminus k} = C)\}, \tag{6.2}$$

must be used. Here C is a fixed constant of dimension equal to $X_{\setminus k}$.

In a perfectly additive model, the choice of C is unimportant, as it will affect the result only by a constant. However, even when the underlying model is strictly additive, it is likely that there will be slight interactions in the fitted black box model. But the effects of these interactions should be small, so averaging over various values for C would achieve a good estimate of f_k.

One way to do this is to use values of C corresponding to the actual measured values for $X_{\setminus k}$. This is what Friedman (2001) recommends, calling it a *partial dependence plot*. In general, the partial dependence of a function $g(x_1, x_2)$ on x_1 is defined as the marginal expectation over x_1, resulting in a function of only x_2,

$$F(x_2) = E_{x_1}(g(x_1, x_2)) \tag{6.3}$$

which then in some way summarizes how $g(x_1, x_2)$ depends on x_2, and can be visualized with the plot of $\{x_2, F(x_2)\}$. This plot was used by Friedman (2001) and Friedman and Popescu (2005). In the additive context of (6.1), the partial dependence of X_k is

$$\begin{aligned} F(X_k) &= E_{X_{\setminus k}}(f_k(X_k) + f_{\setminus k}(X_{\setminus k})) \\ &= f_k(X_k) + E_{X_{\setminus k}}(f_{\setminus k}(X_{\setminus k})). \end{aligned}$$

Since the expected value is a constant, this results in the component f_k as used in the component-plus-residual plot. Interestingly, this method also works when the true model is multiplicative and

$$E(Y|X) = f_k(X_k) \times f_{\setminus k}(X_{\setminus k});$$

the partial dependence is then

$$\begin{aligned} F(X_k) &= E_{X_{\setminus k}}(f_k(X_k) \times f_{\setminus k}(X_{\setminus k})) \\ &= f_k(X_k) \times E_{X_{\setminus k}}(f_{\setminus k}(X_{\setminus k})). \end{aligned}$$

In practice, the empirical distribution of $X_{\setminus k}$ must be used to take this expectation, using the observed data values $\{X_{\setminus k}\}_i$. The result is

$$\begin{aligned} F(X_k) &= E_{X_{\setminus k}}(E(Y|X_k, X_{\setminus k})) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(Y|X_k, X_{\setminus k} = \{X_{\setminus k}\}_i). \end{aligned} \tag{6.4}$$

This is identical to averaging the plots created by (6.2) when using each of the values of $X_{\setminus k}$ for C.

In an additive model

$$f_k(X_k) + e = Y - f_{\setminus k}(X_{\setminus k}),$$

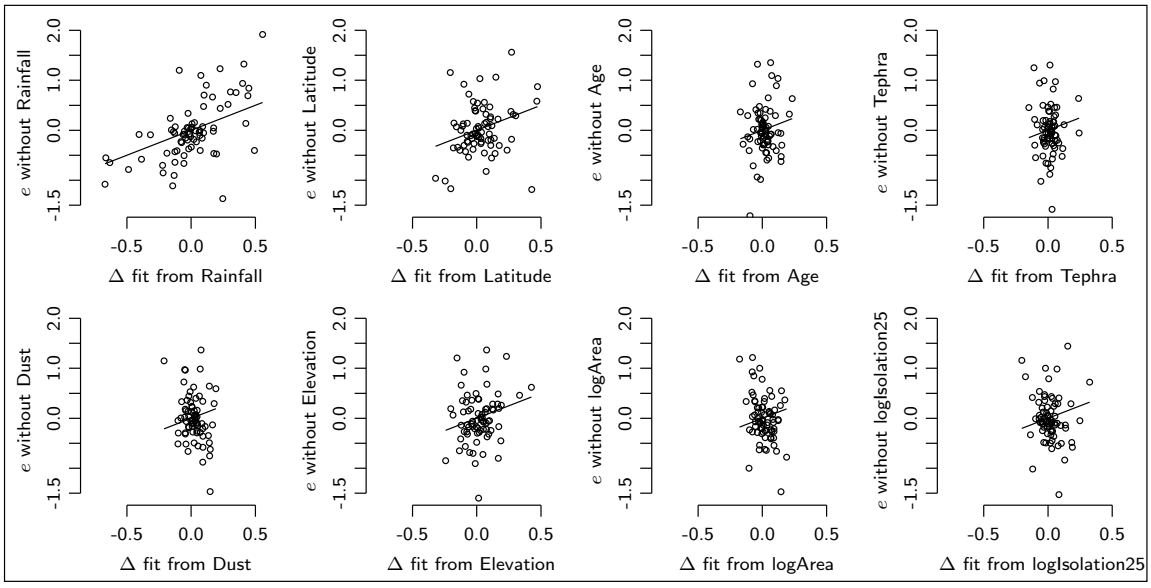
so a component-plus-residual plot can also be made by finding the partial dependence on $X_{\setminus k}$ and subtracting it from Y. If the model is truly additive, these two plots would be identical; any interactions between X_k and $X_{\setminus k}$ would cause them to differ.

These methods hold for X_k of dimension $q > 1$, although the plots are $(q+1)$-dimensional, so are usually only useful when $q \leq 2$. These multidimensional plot can be useful to look for interactions between two predictors.

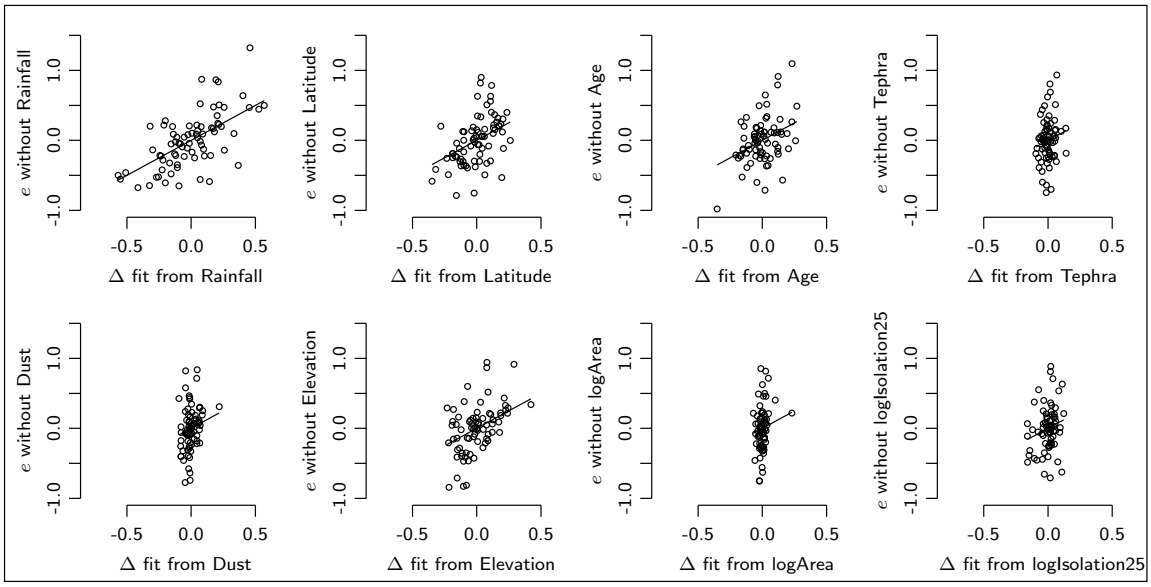
6.3 Example: Island deforestation

The island deforestation data from previous examples was refit using the random forest methodology of Breiman (2001). General added-variable plots for all eight variables are shown in Figure 6.1a, with the plots from the `gam` method also shown for comparison in Figure 6.1b. Both are constructed by refitting the model without the term in question and calculating the change in fitted values. For the most part, they are similar; both show Rainfall to have a large effect, Latitude, Age, and Elevation to have moderate effects. The effect of the rest are smaller, though generally larger in the random forest model than in the `gam` model; Area in particular. The random forest plots are also more spread out, as the residuals from this fitting method are larger. This does not necessarily mean that the fit is worse; remember that the `gam` method overfit to a few points in the previous analysis. In particular, there was one outlier in the random forest plots that is not in the `gam` plots; it is at the bottom of the Dust and Elevation plots. This is the site on the North Island of New Zealand; it has a smaller deforestation value (2) than would otherwise be expected using this model. It does not appear as an outlier in the `gam` model because it was one of the sites with large area identified in Section 2.5 that the model seemed to overfit to.

Component-plus-residual plots formed using the partial dependence method are shown in Figure 6.2a, with the plots from the `gam` method again shown for comparison

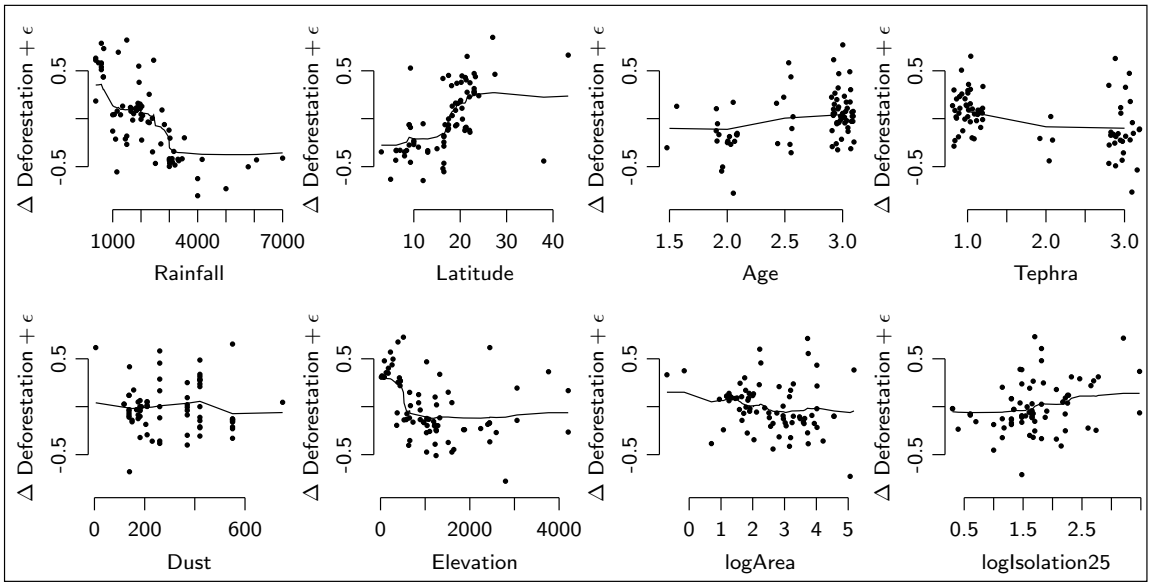


(a) Using the random forest fit.

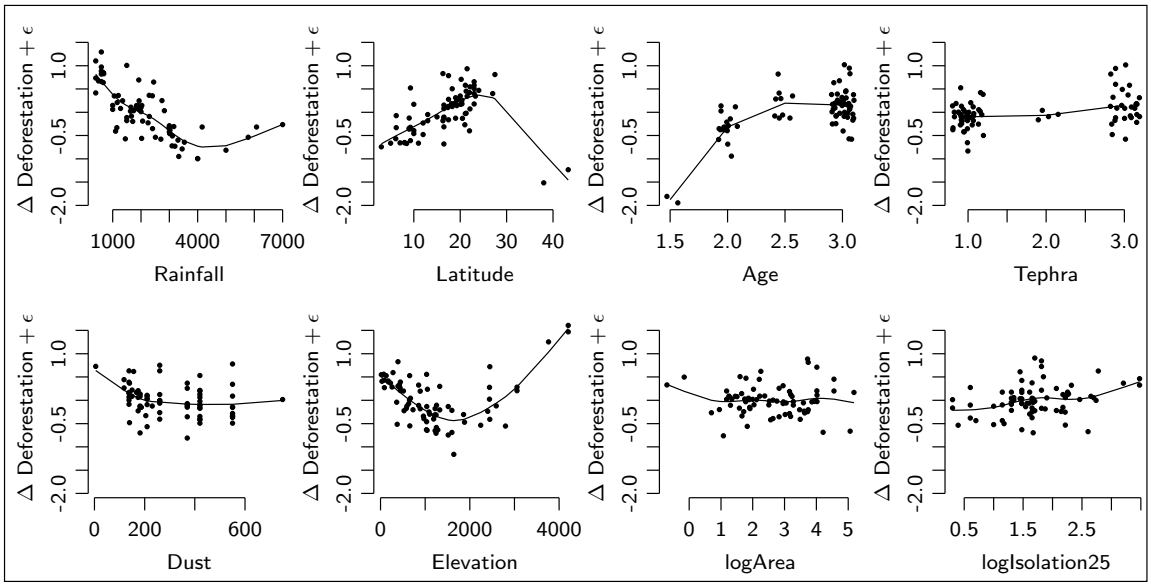


(b) Using the `gam` fit.

Figure 6.1: General added-variable plots for the island deforestation data, using the random forest and `gam` fits.

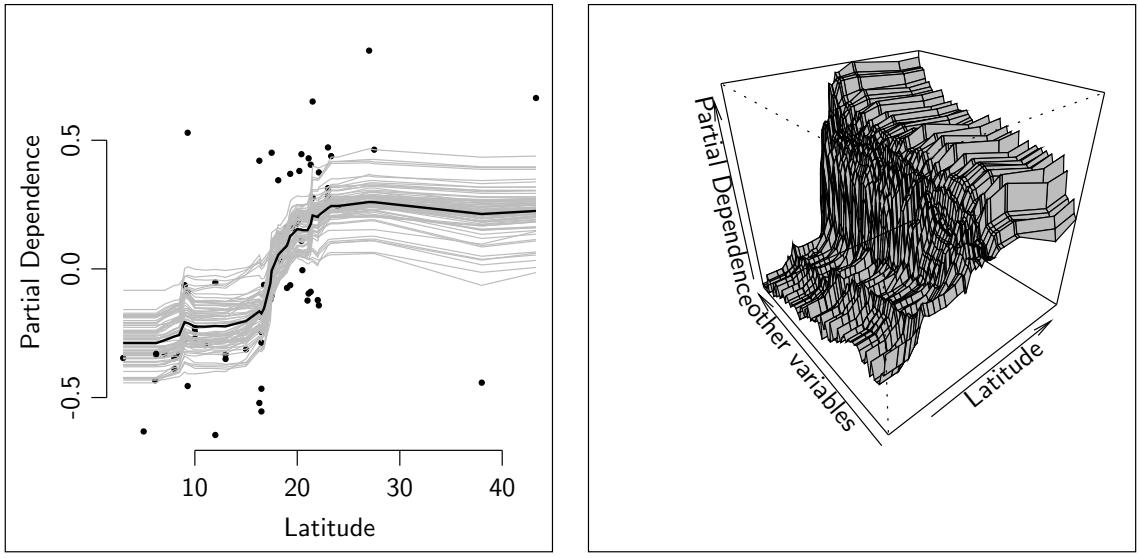


(a) Using the random forest fit.



(b) Using the `gam` fit.

Figure 6.2: Component-plus-residual plots for the island deforestation data, using the random forest and `gam` fits.



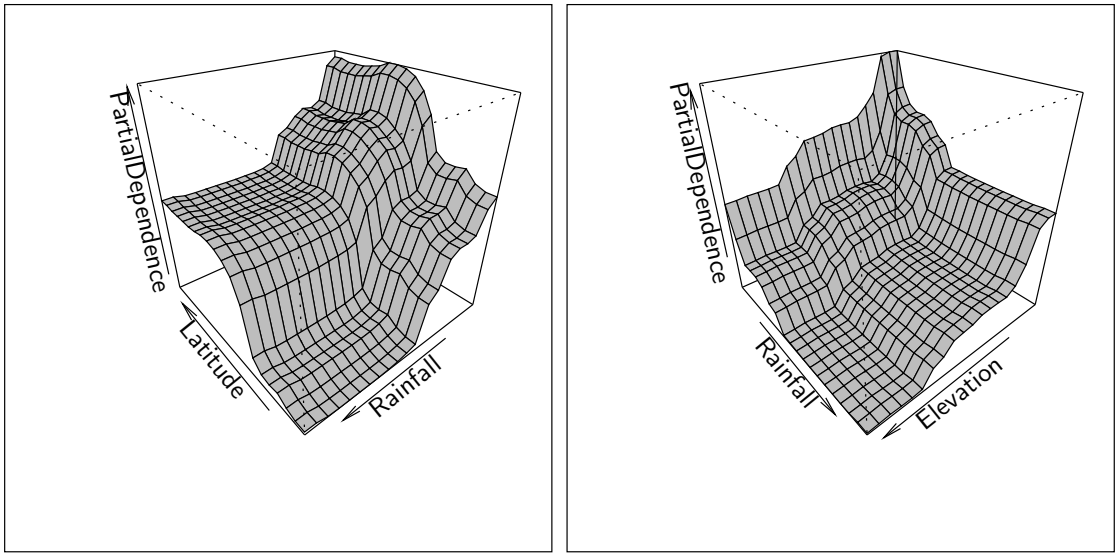
(a) Shown in grey, with the average in black, and the residuals added.

(b) Shown in a three-dimensional plot.

Figure 6.3: The partial dependences for Latitude calculated for each data point individually; these are averaged to get the overall partial dependence.

in Figure 6.2b. The fits from the random forest are for the most part again similar to the fits from the `gam` method, though there are some differences. In particular, where the `gam` method seemed to overfit to a few points, such as the sites with high Latitude, high Elevation, or low Age, the random forest method does not. Again, Rainfall, Latitude, and Elevation have the strongest relationship with Deforestation.

To explore further how these partial dependence plots were formed, Figure 6.3a shows the component-plus-residual plot for Latitude using a black line and black points, with the partial dependence calculated individually for each individual data point shown with gray lines. Each of these lines is centered with its mean at zero, so some of the lines show less of an effect than the average, and some more. Figure 6.3b shows these lines in three dimensions, ordered by increasing range. This might be evidence of an interaction between Latitude and other variables, if there is any systematic relationship between these lines and another variable.



(a) For Latitude and Rainfall. (b) For Elevation and Rainfall.

Figure 6.4: Two-dimensional partial dependence plots for two predictor combinations of the island deforestation data.

To investigate interactions, two-dimensional partial dependence plots can be created. Plots for Latitude and Rainfall and for Elevation and Rainfall are shown in Figure 6.4. In these plots, the effect of Rainfall does not change as Latitude or Elevation changes, so there is no evidence for an interaction involving Rainfall.

Finally, Figure 6.5 shows marginal response plots, with lines showing marginal loess fits and alternate marginal mean functions from the random forest fit. The behavior is similar.

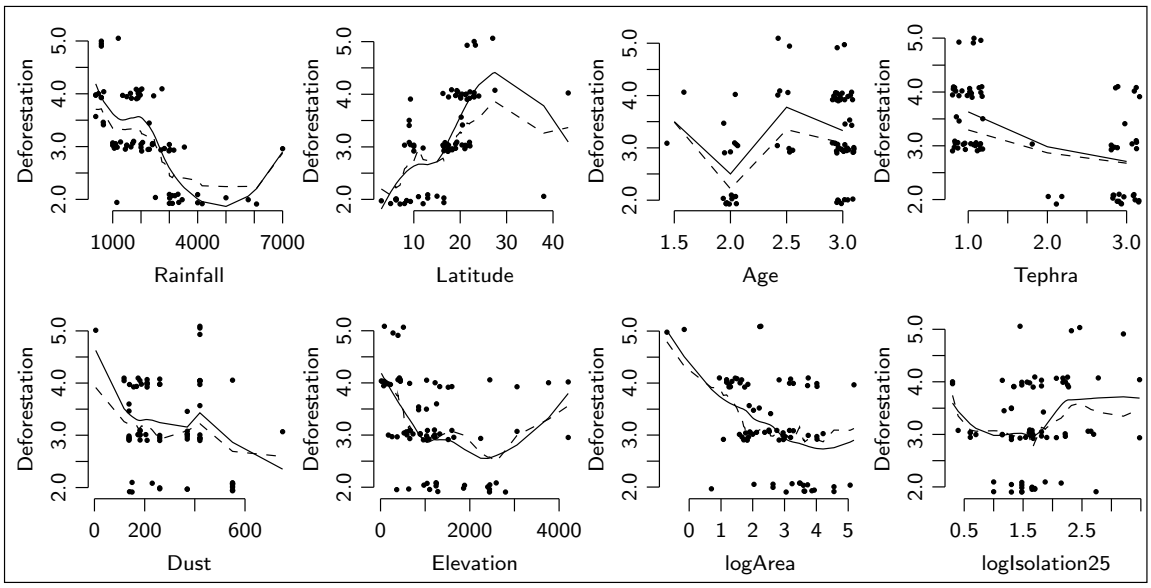


Figure 6.5: Marginal response plots for all variables in the island deforestation data set, with loess fits (solid lines) and alternate marginal mean functions for the random forest fit (dashed lines) added.

Chapter 7

Variable importance

Many methods of measuring variable importance are based on decomposing the R^2 value into contributions from each predictor. Several of these methods are discussed in this chapter, and a new method based on ARMS introduced. However, these methods rely on the predictors being a random sample from a population. For just as the visual sense given by the component-plus-residual plots changes when the range of a predictor changes, so does the change in R^2 due to that predictor. Nevertheless, these types of methods are presently used to measure variable importance, and so will be discussed here.

7.1 By decomposing R^2

Grömping (2007) provides a review of various variable importance measures for additive linear regression models that are based on decomposing R^2, and compares two measures that he deems the best. There are four criteria he considers useful, they are:

1. *Proper decomposition*: the contributions to R^2 by each variable should sum to the R^2 of the full model.

2. *Non-negativity*: each contribution should be non-negative.

3. *Exclusion*: if the coefficient of a given predictor is 0, the contribution should also be 0.

4. *Inclusion*: if the coefficient of a given predictor is nonzero, the contribution should also be nonzero.

When the correlation between the predictors are all zero, this is a simple problem, as the contributions do not "overlap" in any way. When the correlation is non-zero, but the predictors can be ordered in some way, the contributions can be divided by considering the amount explained by the first, the amount explained by the second given the first, the amount explained by the third given the first two, and so on.

In the general case, there is correlation and no obvious ordering. According to Grömping, some analysts use the ordering given by an automated forward or backward selection of variables, e.g. by stepwise methods or the lasso. But these approaches have been criticized because the ordering can be quite arbitrary.

Another option for what to do is to compute the contributions for each possible permutation of the variables and compute an average; see Lindeman et al. (1980) and Kruskal (1987). This is the first measure compared by Grömping. However, this technique does not satisfy the exclusion criteria, even in the population. Consider a spurious predictor (i.e., one with a nonzero coefficient) that is correlated with a nonspurious predictor. Then in the permutations where the spurious predictor is chosen before the nonspurious predictor, it will have a nonzero contribution.

Grömping claims that in some cases this might be the right behaviour, such as when a change in the spurious predictor causes a change in the nonspurious predictor. Here it might be desired to highlight the effect of the spurious predictor by allowing its importance to be nonzero.

The second measure compared by Grömping,proposed by Feldman (2005), does satisfy the exclusion criteria. It again uses the importances from each possible per-

mutation of the predictors, but computes the overall importance using a weighted average, where the weight for a particular permutation $(r_1, \ldots, r_p)$ is proportional to

$$\prod_{i=1}^{p-1} (\text{evar}(\text{all}) - \text{evar}(\{r_1, \ldots, r_i\}))^{-1} \tag{7.1}$$

where $\text{evar}(\{S\})$ is the variance explained by the variables in S. Thus if in a given permutation, the first regressor already captures a large part of the variance, that permutation will be given a large weight.

This procedure is a compromise between the procedure that averages over all permutations and the procedure based on a single permutation from an automated forward or backward model selection procedure.

7.2 Using model combining

An alternate way of compromising between equal weights for all permutations and full weight on one permutation would be to use a model averaging procedure such as ARMS by finding the importances for each submodel by using equal weights, and then computing a weighted average using the ARMS weights.

This asymptotically satisfies the exclusion criteria because if a coefficient is zero, the weight of models including that term will go to zero. Inclusion also holds asymptotically because ARMS always includes terms with non-zero coefficients. Proper decomposition would hold in the following sense; the sum of the contributions would not sum to the R^2 given by the model with all predictors, but instead to the weighted average of the R^2 values for each model, using the ARMS weights. This could be considered a better measure of the actual R^2 as the model with all the variables actually overfits to the actual data. In the cases investigated here, these two R^2 values are comparable; in cases where they are not, they could be compared by standardizing

so that the contributions always sum to one.

7.3 Comparing importance measures

Grömping compared the method with equal weights, called LMG, after Lindeman et al., and the method with weights given in (7.1), called PMVD, or proportional marginal variance decomposition, for several simulated settings. These settings, and a few more, have been recreated, and these methods compared with using ARMS to determine the weights.

The first set of simulations tested two parameter vectors and two correlation structures for various values of ρ; true R^2 was 0.9 for each. There were 100 repetitions, and the mean importance of each predictor is shown in Figure 7.1. The four simulations are shown in the four columns. They were:

1. $\beta = (1, 1, 1, 1)^T$ and $\text{corr}(X_j, X_k) = \rho$. Each predictor has equal importance, and all three methods agree.

2. $\beta = (1, 1, 1, 1)^T$ and $\text{corr}(X_j, X_k) = \rho^{|j-k|}$. The predictors have equal coefficients, but are given different importance because of the correlation structure. There are slight differences between LMG and PMVD; ARMS is quite close to LMG.

3. $\beta = (4, 1, 1, 0.3)^T$ and $\text{corr}(X_j, X_k) = \rho$. As the correlation increases, the importances from LMG become more equal, despite the differing coefficients. This is much less noticable in the PMVD method. ARMS behaves like LMG, as all the variables should be included in the model.

4. $\beta = (4, 1, 1, 0.3)^T$ and $\text{corr}(X_j, X_k) = \rho^{|j-k|}$. Same overall behavior as the previous simulation, but with different curvatures.

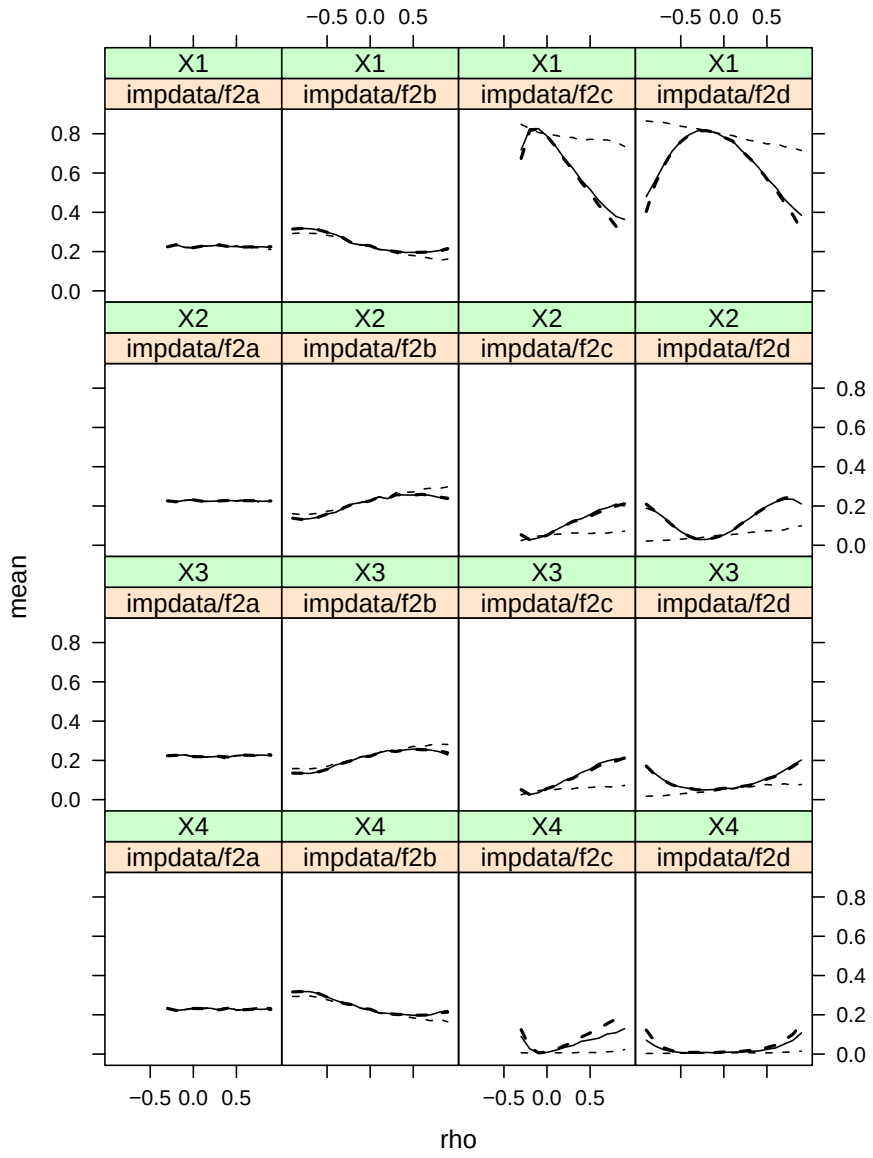


Figure 7.1: Proportions of R^2 allocated to each regressor for LMG (thick dotted), PMVD (thin dotted) and ARMS (thin line) for various ρ values. Each column (a through d) is a distinct simulation; each row is a different predictor (X_1 through X_4). The parameter vector was $(1, 1, 1, 1)^T$ for a and b and $(4, 1, 1, 0.3)^T$ for c and d. The correlation for a and c was ρ, for b and d it was $\text{corr}(X_j, X_k) = \rho^{|j-k|}$.

For these simulations, ARMS acted like LMG; this is not surprising as all the variables should have been included in the model.

The next set of simulations investigated the variation of the responses; Grömping claimed that this was the most interesting outcome of his simulations. The interquartile ranges for simulations with seven different parameter vectors are shown in Figure 7.2. Grömping noted that LMG usually had a smaller variation than PMVD. ARMS is usually between the two.

As ARMS should behave like PMVD when there are spurious variables, the parameter vector $(1, 0, 1, 0, 1, 0, 1, 0)$ was used, so there were four parameters with equal coefficients and four with zero coefficients. The two correlation structures were $\text{corr}(X_j, X_k) = \rho$ and $\text{corr}(X_j, X_k) = \rho^{|j-k|}$.

Figure 7.3 shows the average importances calculated over 100 simulations. ARMS does act more like PMVD in these examples, though it is still almost always between the two estimates. An interesting exception is X_3 and X_5 in the changing correlation simulation (Figure 7.4a) where the ARMS importance stays constant across ρ, unlike LMG, which increases, or PMVD, which decreases.

However, while the ARMS estimate usually acts more like PMVD in terms of the mean importance, Figure 7.4 shows that the ARMS variance does not increase like PMVD, but instead is usually as small as that of the LMG estimate.

For situations where there are many predictors, of which some are spurious but correlated with nonspurious predictors, using the ARMS weights to construct the importance measure can result in importances that meet the exclusion property by giving a small importance to the spurious predictors, but without the larger variance seen when using the PMVD procedure.

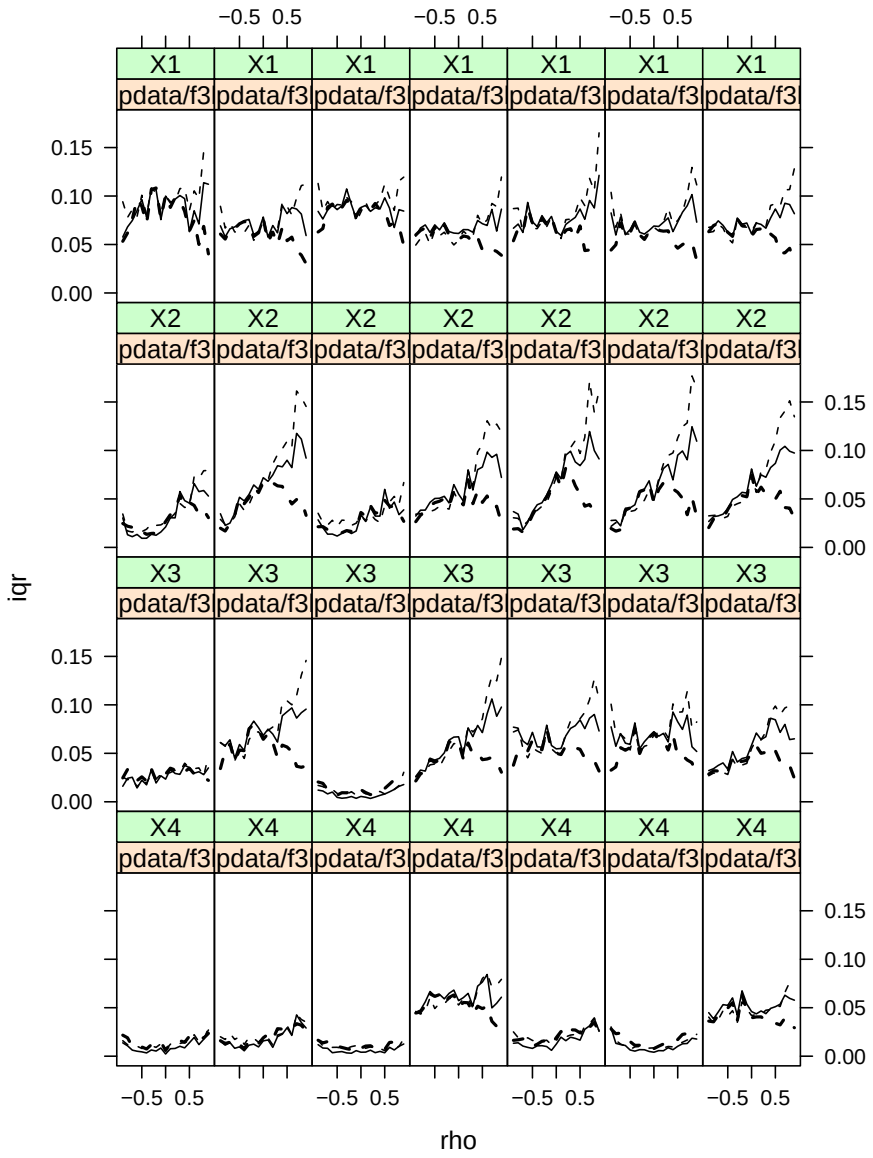
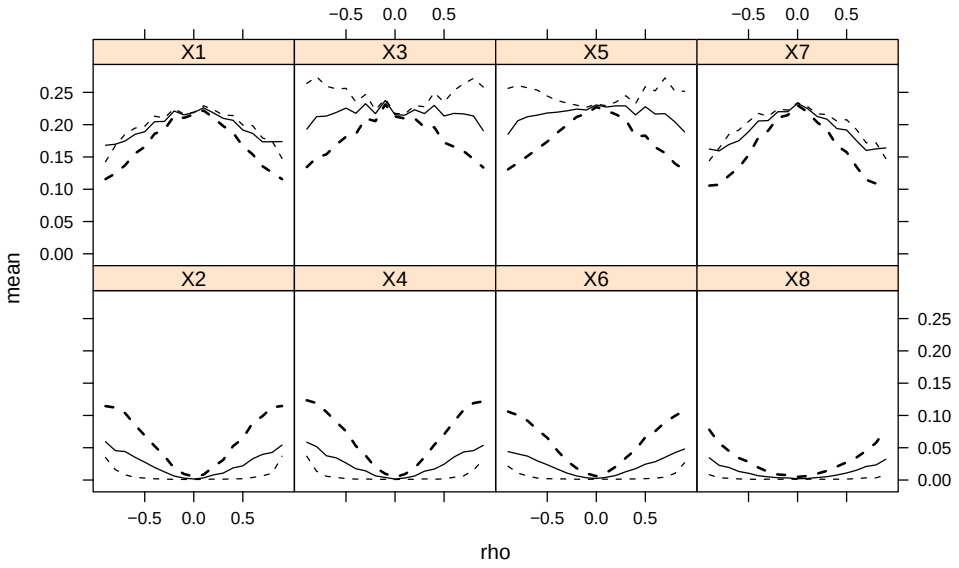
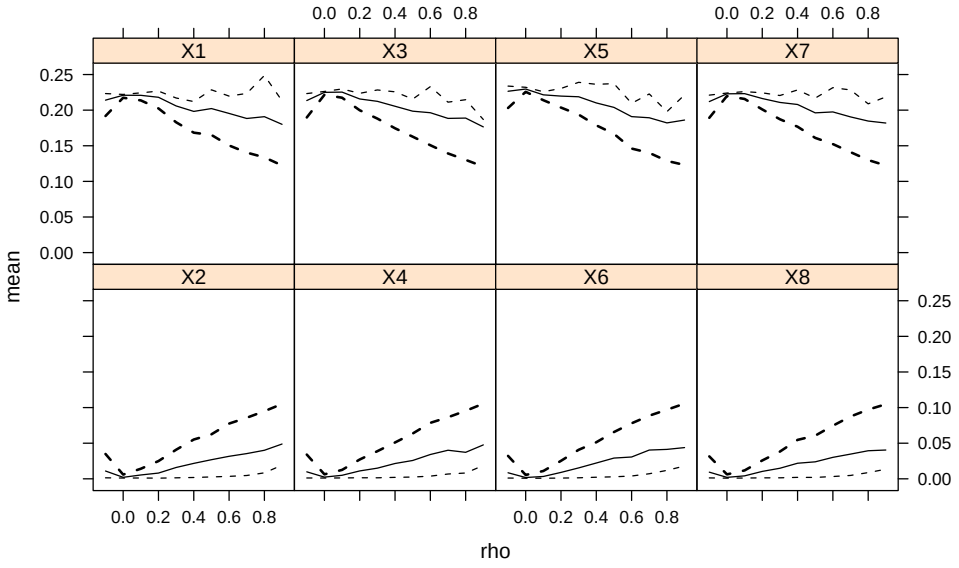


Figure 7.2: Interquartile ranges from 100 simulations for LMG (thick dotted), PMVD (thin dotted), and ARMS (thin line). The actual R^2 is 0.25 and the correlation structure is $\text{corr}(X_j, X_k) = \rho^{|j-k|}$.

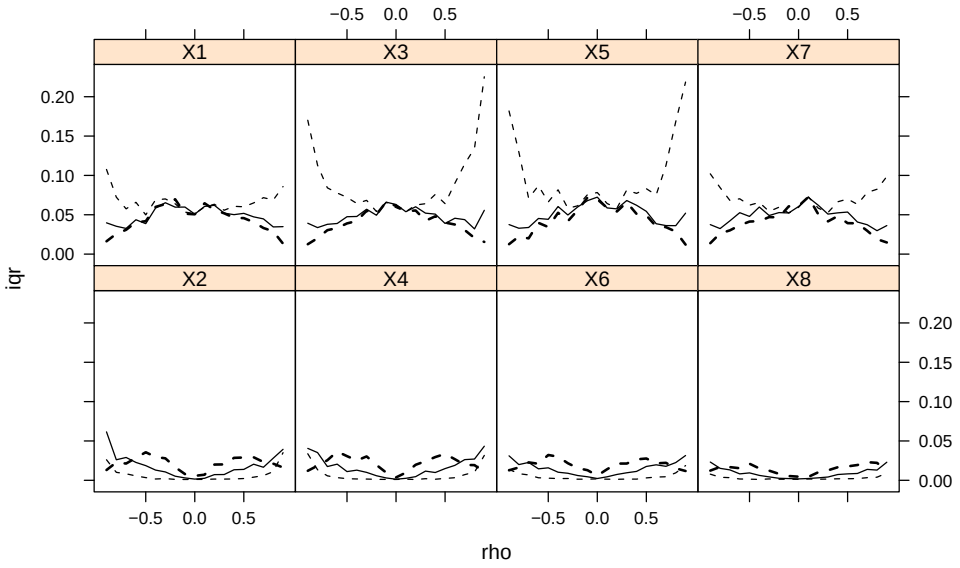


(a) Using correlation structure $\text{corr}(X_j, X_k) = \rho$.

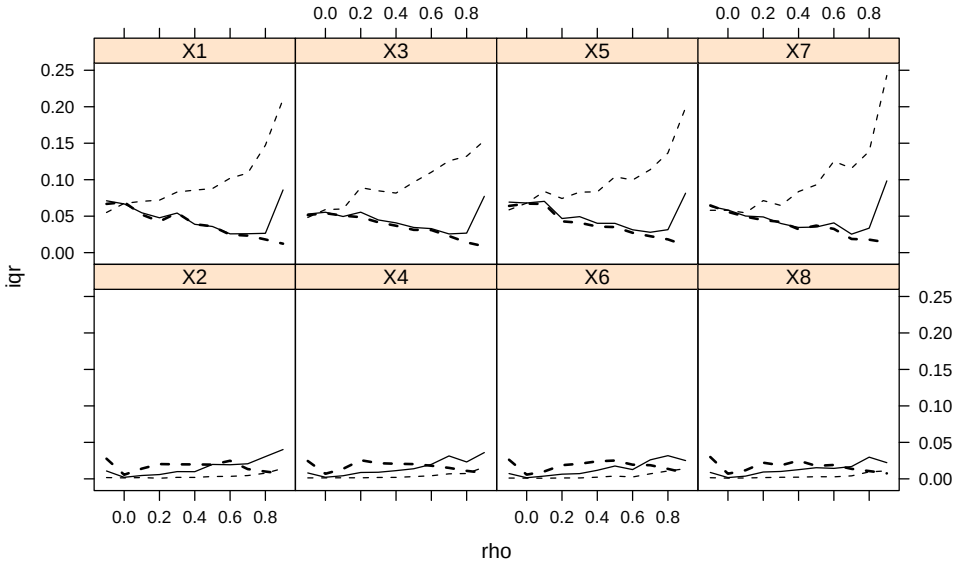


(b) Using correlation structure $\text{corr}(X_j, X_k) = \rho^{|j-k|}$.

Figure 7.3: Average importance values for the simulations using the parameter vector $(1, 0, 1, 0, 1, 0, 1, 0)$, for LMG (thick dotted), PMVD (thin dotted) and ARMS (thin line) methods.



(a) Using correlation structure $\text{corr}(X_j, X_k) = \rho$.



(b) Using correlation structure $\text{corr}(X_j, X_k) = \rho^{|j-k|}$.

Figure 7.4: Interquartile ranges for the simulations using the parameter vector $(1, 0, 1, 0, 1, 0, 1, 0)$, for LMG (thick dotted), PMVD (thin dotted) and ARMS (thin line) methods.

Chapter 8

Conclusions and future work

This research has investigated graphical methods of determining predictor importance and effect. First, a criteria for determining and evaluating useful plots for this purpose was developed; namely, that for a plot to show how an individual predictor is related to the response, the variable on the vertical axis should be conditionally independent of the other predictors, given the variable on the horizontal axis. Removing the dependence on the other predictors makes the relationship between the variable of interest and the response clearer.

Of the many plots that satisfy this criteria, plots with especially interpretable axes were investigated further, and used to show several different aspects of the relationship between the predictor of interest and the response; most importantly, the marginal relationship and the conditional relationship after conditioning on the other predictors. The effect of the variable of interest can look very different when the other predictors are conditioned on, and certain plots show one relationship, and certain plots the other.

To show the marginal relationship, the marginal response plot was first considered, but it does not satisfy the desired condition, as it includes variability from the other predictors. A new plot, called the marginal-plus-residual plot, was proposed to remove this variability and clarify the marginal relationship. Additionally, a new alternate

marginal plot was also discussed; for a given model function, it shows the relationship between the predictor and the response when the other predictors are set to their expected values given the predictor of interest. This plot is interesting because it shows some kind of compromise between the marginal and conditional relationships; it can show exactly the marginal or the conditional relationship, or something in between, depending on the relationships between the predictors.

Two well-known plots for showing the conditional relationship are the component-plus-residual plot and the added-variable plot. These were derived from the desired criteria, and their usefulness explored in several contexts, including linear models, additive models, generalized linear models, models with interactions, and models fit using "black box" methods. Of particular note is a new added-variable plot developed for use with generalized linear models that successfully removes the conditional dependence of the other variables. Additionally, an alternate added-variable plot was developed, and the relationship between the ARES plot and the added-variable plot was explored.

Of special interest was determining how to construct these plots when using a model combining method, such as ARMS. For the component-plus-residual plot, a straightforward weighting of the plots from each of the submodels met the appropriate criteria, and was shown to be an excellent picture of the effect of the predictor of interest, as it averages over the possible effects given by the various submodels. However, for the added-variable plot, several options are possible. One option is to condition on all the predictors, resulting in a plot that can be interpreted just like a regular added-variable plot. But if some of the predictors should not be conditioned on, this may understate the effect of the predictor of interest. So a second option is to construct the plot by only partially conditioning on the variables not included in one or more submodel, using the weights from the model combining method. This shows a compromise between the plot when fully conditioning, and the plot when fully not

conditioning, and is helpful because which variables should be removed is unknown.

The weights produced by a model combining method such as ARMS were also applied to variable importance measures, which aim to reduce the relative contribution of each predictor to a single number, rather than showing a more complex graphical representation. Although this is straightforward when the predictors are independent, when correlations and interactions between the predictors are present, tradeoffs must be made, and several competing methods exist. A modification of one of these methods was proposed, using the weights from a model combining method. It was shown in simulation that this modification had better properties than the initial method.

There are several ways this research could be extended. First, these plots could be explored in other settings, including mixed models, or cases with categorical predictors and responses. Another avenue for research is to continue to explore the usefulness of the apparently new plots developed in this thesis. In the examples shown here, they do seem helpful, but exactly which plots are helpful when, and how to determine those times, remains open.

Appendix A

Computing

All computing for this thesis was done in `R` version 2.5.1 (R Development Core Team, 2007), with two additional packages. These will be available on the author's website at `http://www.stat.umn.edu/~arendahl`.

First, the `ARMS` package, written by Sanford Weisberg with assistance from the author, was developed to run the ARMS procedure, as proposed by Yuan and Yang (2005). It can run ARMS on a full linear model, using linear models of subsets of the variables as submodels. The number of subsets to screen, the number of random splits, the size of the split, and when to screen can all be controlled. Output includes a list of the selected submodels, their weights, and their coefficients. Additionally, the coefficients, fitted values, and residuals of the full weighted model can all be easily accessed.

Secondly, the `akrplots` package was developed to facilitate building local net-effect plots and several related plots, including added-variable plots, component-plus-residual plots, marginal plots, and partial dependence plots. These specific plots can be built for linear models, additive models from the `gam` library, ARMS models, and random forest models. The added-variable plots for generalized linear models are also available. The code for added-variable plots and component-plus-residual plots is based on that in the `car` package (Fox, 2007).

This package also can construct local net-effect plots, two variables are specified for the axes of the scatterplot, with the option of slicing the plot on a third variable; this slicing can be done evenly, by quantiles, by unique values, or directly specified. The points in each slice are each given a different symbol. Additionally, linear, loess, and gam fits can be added, either for each slice or overall. This package also can save the information from a plot in a variable, to be replotted or modified later.

Additional packages used included `gam` (Hastie, 2006) and `mgcv` (Wood, 2006), for fitting additive models; `randomForest` (Liaw and Wiener, 2002), for the random forest models, `relaimpo` (Grömping, 2007), for the relative importance calculations, and `scatterplot3d` (Ligges and Mächler, 2003), for the three-dimensional plots.

Finally, most plots were created in the xfig format and separated into graphical and textual elements, which were then converted into pdf and LaTeX formats, respectively, and overlaid to get the final result. The remaining plots were created directly in the pdf format.

References

Atkinson, A. C. (1985). *Plots, Transformations, and Regression: an Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford University Press.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons.

Berk, K. N. and Booth, D. E. (1995). Seeing a curve in multiple regression. *Technometrics*, 37:385–398.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80:580–598.

Brieman, L. (1995). Stacked regressions. *Machine Learning*, 24:49–64.

Brieman, L. (1996). Bagging predictors. *Machine Learning*, 26:123–140.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 17:453–510.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 158:419–444.

Cleveland, W. S. (1993). *Visualizing Data*. Summit Press.

Cook, R. D. (1993). Exploring partial residual plots. *Technometrics*, 35:351–362.

Cook, R. D. (1995). Graphics for studying net effects of regression predictors. *Statistica Sinica*, 5:689–708.

Cook, R. D. (1996). Added-variable plots and curvature in linear regression. *Technometrics*, 38:275–278.

Cook, R. D. and Croos-Dabrera, R. (1998). Partial residual plots in generalized linear models. *Journal of the American Statistical Association*, 93:730–739.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall Ltd.

Cook, R. D. and Weisberg, S. (1989). Regression diagnostics with dynamic graphics. *Technometrics*, 31:277–291.

Cook, R. D. and Weisberg, S. (1991). Added variable plots in linear regression. In Stahel, W. and Weisberg, S., editors, *Directions in Robust Statistics and Diagnostics. Part I*, pages 47–60. Springer-Verlag Inc.

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*. John Wiley & Sons.

Cook, R. D. and Weisberg, S. (1997). Graphics for assessing the adequacy of regression models. *Journal of the American Statistical Association*, 92:490–499.

Cook, R. D. and Weisberg, S. (1999a). *Applied Regression Including Computing and Graphics*. John Wiley & Sons.

Cook, R. D. and Weisberg, S. (1999b). Graphs in statistical analysis: Is the medium the message? *The American Statistician*, 53:29–37.

Cox, D. (1958). *The Planning of Experiments*. John Wiley & Sons.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis (Second Edition)*. John Wiley & Sons.

Ezekiel, M. (1924). A method of handling curvilinear correlation for any number of variables. *Journal of the American Statistical Association*, 19:431–453.

Feldman, B. (2005). Relative importance and value. Unpublished manuscript. Online at `http://www.prismanalytics.com/docs/RelativeImportance050319.pdf`.

Fox, J. (2007). *car: Companion to Applied Regression*. R package version 1.2-2.

Freedman, D., Pisani, R., and Purves, R. (1978). *Statistics*. W. W. Norton & Co Inc.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Friedman, J. H. and Popescu, B. E. (2003). Importance sampling learning ensembles. Technical report, Stanford University, Department of Statistics.

Friedman, J. H. and Popescu, B. E. (2005). Predictive learning via rule ensembles. Technical report, Stanford University, Department of Statistics.

Grömping, U. (2007). Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. *The American Statistician*, 61(2):139–147.

Grömping, U. (2007). *relaimpo: Relative importance of regressors in linear models*. R package version 1.2.

Hastie, T. (2006). *gam: Generalized Additive Models*. R package version 0.98.

Hastie, T. and Tibshirani, R. (1999). *Generalized Additive Models*. Chapman & Hall Ltd.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–401.

Johnson, B. W. and McCulloch, R. E. (1987). Added-variable plots in linear regression. *Technometrics*, 29:427–433.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Kruskal, W. (1987). Relative importance by averaging over orderings. *The American Statistician*, 41:6–10.

Landwehr, J. M. (1986). Using residual plots to detect nonlinearity in multiple regression. Statistical Research Report 15, AT&T Bell Laboratories.

Landwehr, J. M. and Pregibon, D. (1993). Comments on "Improved added variable and partial residual plots for the detection of influential observations in generalized linear models". *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 42:16–20.

Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, 79:61–71.

Larsen, W. A. and McCleary, S. J. (1972). The use of partial residual plots in regression analysis. *Technometrics*, 14:781–790.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Ligges, U. and Mächler, M. (2003). Scatterplot3d - an r package for visualizing multivariate data. *Journal of Statistical Software*, 8(11):1–20.

Lindeman, R. H., Merenda, P. F., and Gold, R. Z. (1980). *Introduction to Bivariate and Multivariate Analysis*. Scott, Foresman & Co.

Mallows, C. L. (1986). Augmented partial residuals. *Technometrics*, 28:313–319.

Mansfield, E. R. and Conerly, M. D. (1987). Diagnostic value of residual and partial residual plots. *The American Statistician*, 41:107–116.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. Chapman & Hall Ltd.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression: A Secondary Course in Statistics*. Addison-Wesley Publishing Co Inc.

Oehlert, G. (1992). A note on the delta method. *The American Statistician*, 46(1).

O'Hara Hines, R. J. and Carter, E. M. (1993). Improved added variable and partial residual plots for the detection of influential observations in generalized linear models. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 42:3–16.

Pardoe, I. B. (2001). *A Bayesian Approach to Regression Diagnostics*. PhD thesis, University of Minnesota School of Statistics.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rolett, B. and Diamond, J. (2004). Environmental predictors of pre-european deforestation on pacific islands. *Nature*, 431:443–446.

Seo, H. S. (1999). A dynamic plot for the specification of curvature in linear regression. *Computational Statistics & Data Analysis*, 30:221–228.

Simpson, E. H. (1951). The interpretation of interaction in contigency tables. *Journal of the Royal Statistical Society, Series B: Methodological*, 13:238–241.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B: Methodological*, 36:111–147.

Wang, P. C. (1985). Adding a variable in generalized linear models. *Technometrics*, 27:273–276.

Weisberg, S. (1985). *Applied Linear Regression*. John Wiley & Sons.

Wolpert, D. (1992). Stacked generalization. *Neural Networks*, 5:241–259.

Wood, F. S. (1973). The use of individual effects and residuals in fitting equations to data. *Technometrics*, 15:677–695.

Wood, S. N. (2006). *Generalized Additive Models*. Chapman & Hall Ltd.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, 96(454):574–588.

Yuan, Z. and Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association*, 100(472):1202–1214.