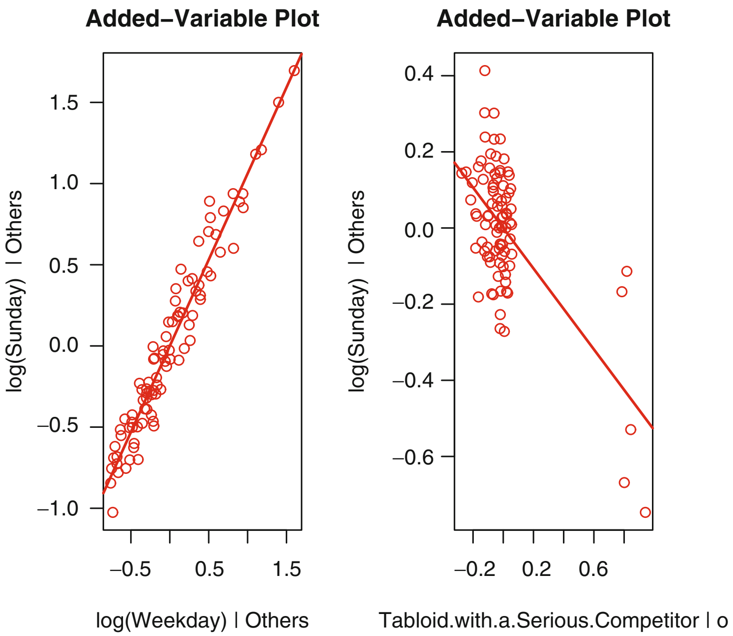

Figure 6.32 Added-variable plots for model (6.18)

6.3 Graphical Assessment of the Mean Function Using Marginal Model Plots

We begin by briefly considering simple linear regression. In this case, we wish to visually assess whether

$$Y = \beta_0 + \beta_1 x + e \tag{6.19}$$

models E($Y|x$) adequately. One way to assess this is to compare the fit from (6.19) with a fit from a general or nonparametric regression model (6.20) where

$$Y = f(x) + e \tag{6.20}$$

There are many ways to estimate f nonparametrically. We shall use a popular estimator called loess, which is based on local linear or locally quadratic regression fits. Further details on nonparametric regression in general and loess in particular can be found in Appendix A.2.

Under model (6.19), $\mathrm{E}_{\mathrm{M}_1}(Y \mid x) = \beta_0 + \beta_1 x,$ while under model (6.20), $\mathrm{E}_{\mathrm{F}_1}(Y \mid x) = f(x).$ Thus, we shall decide that model (6.19) is an adequate model if $\hat{\beta}_0 + \hat{\beta}_1 x$ and $\hat{f}(x)$ agree well.

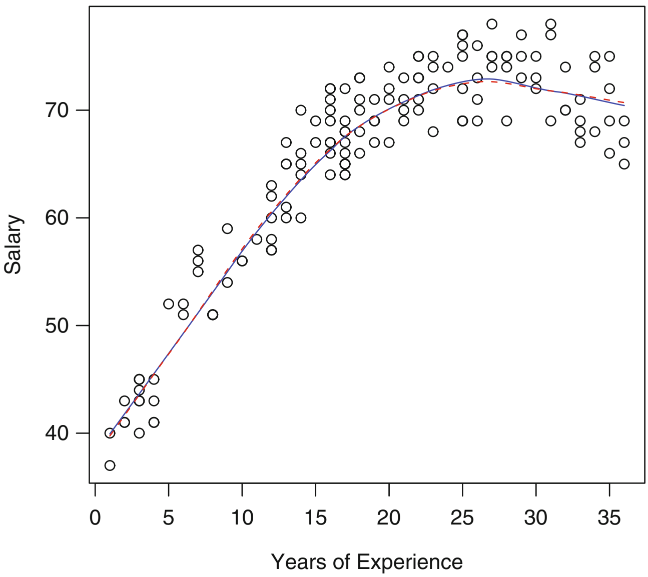

Figure 6.34 A plot of the professional salary data with quadratic and loess fits

The challenge for the approach we have just taken is how to extend it to regression models based on more than one predictor. In what follows we shall describe the approach proposed and developed by Cook and Weisberg (1997).

Marginal Model Plots

Consider the situation when there are just two predictors x_1 and x_2. We wish to visually assess whether

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \tag{M1}$$

models E($Y|x$) adequately. Again we wish to compare the fit from (M1) with a fit from a nonparametric regression model (F1) where

$$Y = f(x_1, x_2) + e \tag{F1}$$

Under model (F1), we can estimate $\mathrm{E}_{\mathrm{F}_1}(Y \mid x_1)$ by adding a nonparametric fit to the plot of Y against x_1. We want to check that the estimate of $\mathrm{E}_{\mathrm{F}_1}(Y \mid x_1)$ is close to the estimate of $\mathrm{E}_{\mathrm{M}_1}(Y \mid x_1)$.

Under model (M1)

$$\mathrm{E}_{\mathrm{M}_1}(Y \mid x_1) = \mathrm{E}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \mid x_1) = \beta_0 + \beta_1 x_1 + \beta_2 \mathrm{E}(x_2 \mid x_1)$$

Notice that this last equation includes the unknown $\mathrm{E}_{\mathrm{M}_1}(x_2 \mid x_1)$ and that in general there would be $(p-1)$ unknowns, where p is the number of predictor variables in model (M1). Cook and Weisberg (1997) overcome this problem by utilizing the following result:

$$\mathrm{E}_{\mathrm{M}_1}(Y \mid x_1) = \mathrm{E}\left[\mathrm{E}_{\mathrm{M}_1}(Y \mid x) \mid x_1\right] \tag{6.24}$$

The result follows from the well-known general result re conditional expectations. However, it is easy and informative to demonstrate the result in this special case. First, note that

$$\mathrm{E}_{\mathrm{M}_1}(Y \mid x) = \mathrm{E}_{\mathrm{M}_1}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \mid x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

so that

$$\mathrm{E}\left[\mathrm{E}_{\mathrm{M}_1}(Y \mid x) \mid x_1\right] = \mathrm{E}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \mid x_1) = \beta_0 + \beta_1 x_1 + \beta_2 \mathrm{E}(x_2 \mid x_1)$$

matching what we found on the previous page for $\mathrm{E}_{\mathrm{M}_1}(Y \mid x)$.

Under model (M1), we can estimate $\mathrm{E}_{\mathrm{M}_1}(Y \mid x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ by the fitted values $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. Utilizing (6.24) we can therefore estimate $\mathrm{E}_{\mathrm{M}_1}(Y \mid x_1) = \mathrm{E}\left[\mathrm{E}_{\mathrm{M}_1}(Y \mid x) \mid x_1\right]$ by estimating $\mathrm{E}\left[\mathrm{E}_{\mathrm{M}_1}(Y \mid x) \mid x_1\right]$ with an estimate of $\mathrm{E}\left[\hat{Y} \mid x_1\right]$.

In summary, we wish to compare estimates under models (F1) and (M1) by comparing nonparametric estimates of $\mathrm{E}(Y \mid x_1)$ and $\mathrm{E}\left[\hat{Y} \mid x_1\right]$. If the two nonparametric estimates agree then we conclude that x_1 is modelled correctly by model (M1). If **not** then we conclude that x_1 is **not** modelled correctly by model (M1).

Example: Modelling defective rates (cont.)

Recall from earlier in Chapter 6 that interest centres on developing a model for Y, Defective, based on the predictors x_1, Temperature; x_2, Density and x_3, Rate. The data can be found on the book web site in the file defects.txt.

The first model we considered was the following:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e \tag{6.25}$$

The left-hand plot in Figure 6.35 is a plot of Y against x_1, Temperature with the loess estimate of $\mathrm{E}(Y \mid x_1)$ included. The right-hand plot in Figure 6.35 is a plot of $\hat{Y}$ against x_1, Temperature with the loess estimate of $\mathrm{E}\left[\hat{Y} \mid x_1\right]$ included.

The two curves in Figure 6.35 do not agree with the fit in the left-hand plot showing distinct curvature, while the fit in the right-hand plot is close to a straight line. Thus, we decide that x_1 is **not** modelled correctly by model (6.25).

In general, it is difficult to compare curves in different plots. Thus, following Cook and Weisberg (1997) we shall from this point on include both nonparametric curves on the plot of Y against x_1. The plot of Y against x_1 with the loess fit for Y against x_1 and the loess fit for $\hat{Y}$ against x_1 both marked on it is called a **marginal model plot** for Y and x_1.

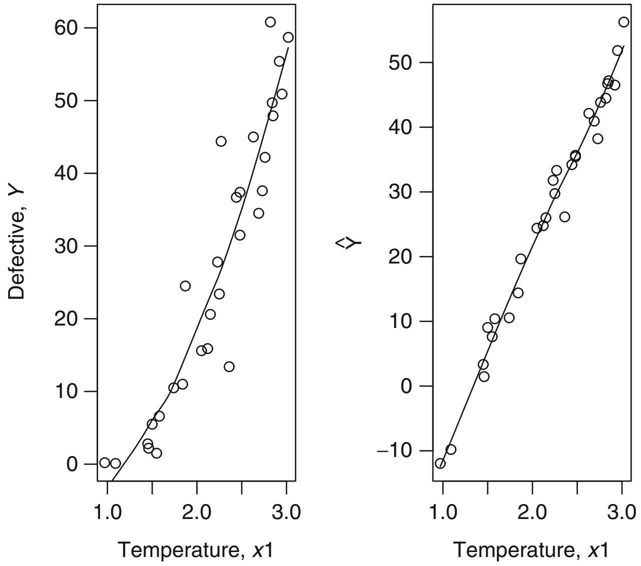

Figure 6.35 Plots of Y and $\hat{Y}$ against x_1, Temperature

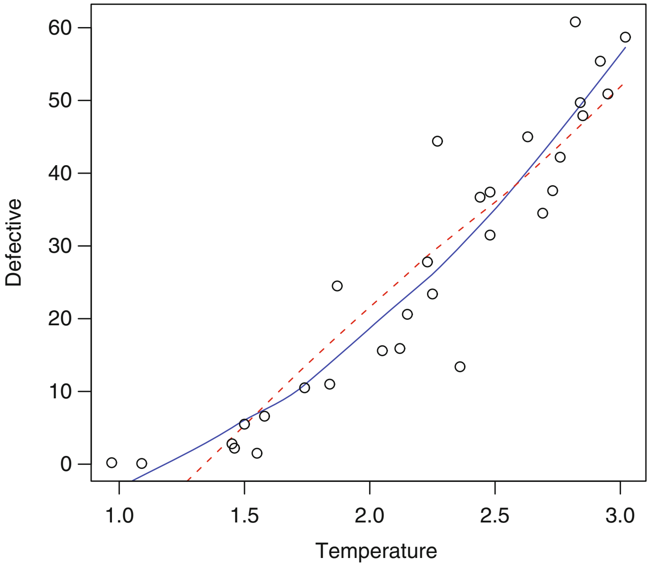

Figure 6.36 A marginal mean plot for Defective and Temperature

Figure 6.36 contains a **marginal model plot** for Y and x_1. The solid curve is the loess estimate of $E(Y \mid x_1)$ while the dashed curve is the loess estimate of $E\left[\hat{Y} \mid x_1\right]$. It is once again clear that these two curves do not agree well.

It is recommended in practice that marginal model plots be drawn for each predictor (except dummy variables) and for $\hat{Y}$. Figure 6.37 contains these recommended marginal model plots for model (6.25) in the current example. The two fits in each of the plots in Figure 6.37 differ markedly. In particular, each of the nonparametric estimates in Figure 6.37 (marked as solid curves) show distinct curvature which is not present in the smooths of the fitted values (marked as dashed curves). Thus, we again conclude that (6.25) is not a valid model for the data.

We found earlier that in this case, both the inverse response plot and the Box-Cox transformation method point to using a square root transformation of Y. Thus, we next consider the following multiple linear regression model

$$Y^{0.5} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e \tag{6.26}$$

Figure 6.38 contains the recommended marginal model plots for model (6.26) in the current example. These plots again point to the conclusion that (6.26) is a valid model for the data.

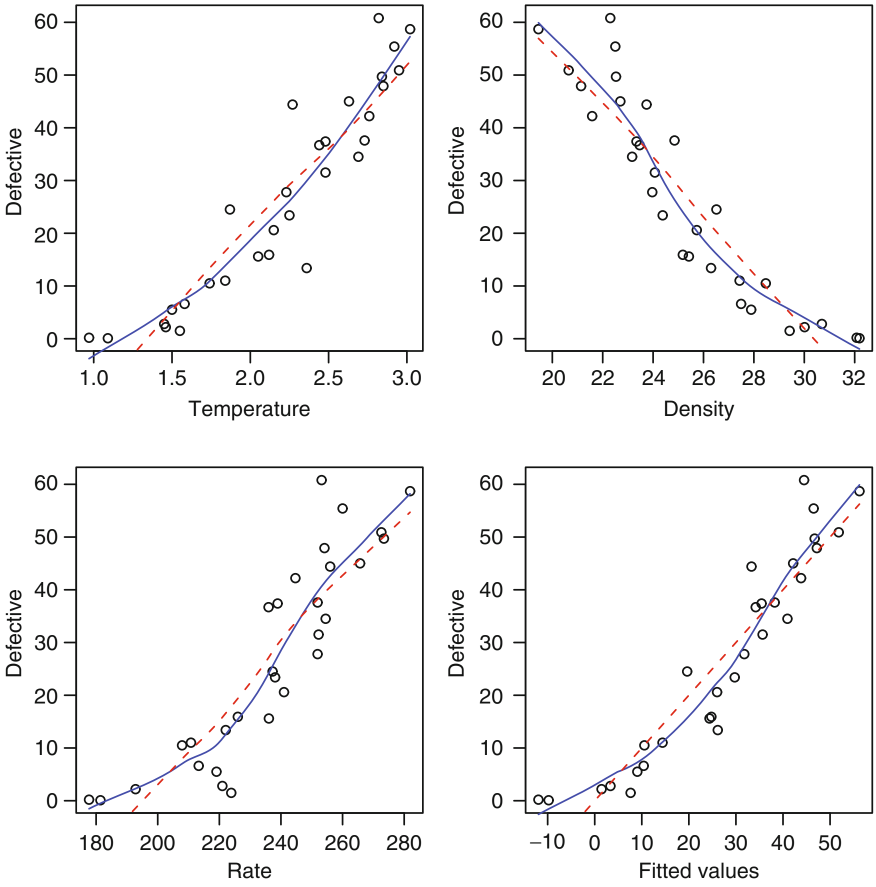

Figure 6.37 Marginal model plots for model (6.25)