Chapter 12: Diagnostics

12.1 Non-normally Distributed Errors

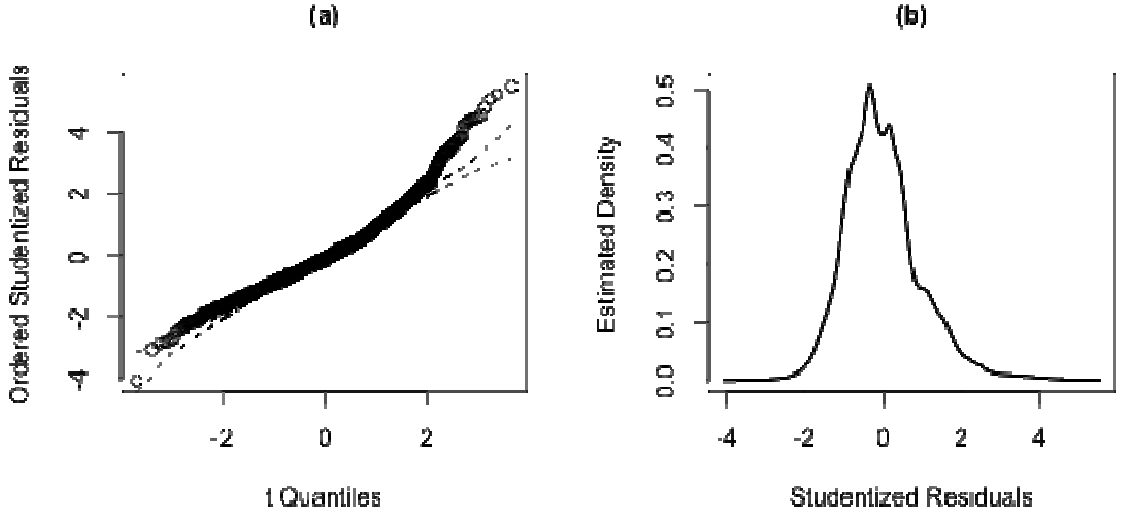
Notes:

- First, remember that it is the error term, the conditional distribution of Y given X's, that is of concern (not the marginal distribution).
- Least squares estimation is generally *robust* to departures from normality in the sense that hypothesis tests and confidence intervals are approximatley correct when normality is violated, but the *efficiency* (minimum variance property) is not robust (in comparison with possible nonlinear estimators) under non-normality, including heavy tails (outliers).
- *Conditional mean*, which is what we address with a regression model and normal (symmetrically distributed) errors is often not a good measure of the center of a highly *skewed* distribution.
- A *multimodal* distribution may suggest omission of one or more categorical explanatory variables.

Formal tests of normality:

- Shapiro-Wilk test (`shapiro.test`) --- related to a QQ-plot in that it is derived from observed order statistics and expected values of order statistics for a normal distribution.
- Kolmogorov-Smirnov test (`ks.test`) --- $D_n = \sup_x \left| F_n(x) - F(x) \right|$, where $F_n(x)$ is the empirical distribution of a sample and $F(x)$ is the cumulative distribution function of the reference distribution of interest. The preferred variant, accounting for effects of estimation of mean and variance, is the Lilliefors test (`lillie.test` in package nortest).

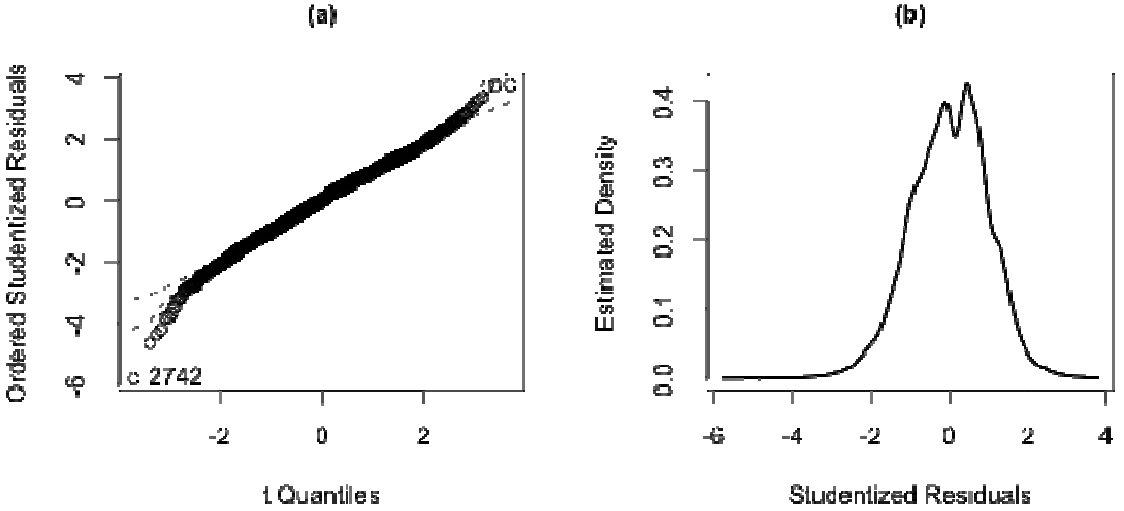
Figure 12.1 QQ-plot of *studentized residuals* of regression of wages on sex, age and education (SLID data) vs *t-distribution* on n-k-2 df, with pointwise 95% simulated confidence envelope (sect 12.1.1). (`qq.plot` of the car package)



Review question: What is the difference between standardized and studentized residuals?

Figure 12.2 QQ-plot of *studentized residuals* of regression of *log wages* on sex, age and education (SLID data)

Fox notes that cube root does a better job of reducing long left tail, but he prefers the more easily interpreted log transformation. So do I.



12.2 Nonconstant variance

12.2.1 Residual plots

Figure 12.3 (a) Studentized residuals vs fitted values and (b) “spread-level plot” (with “robust linear regression fit”)

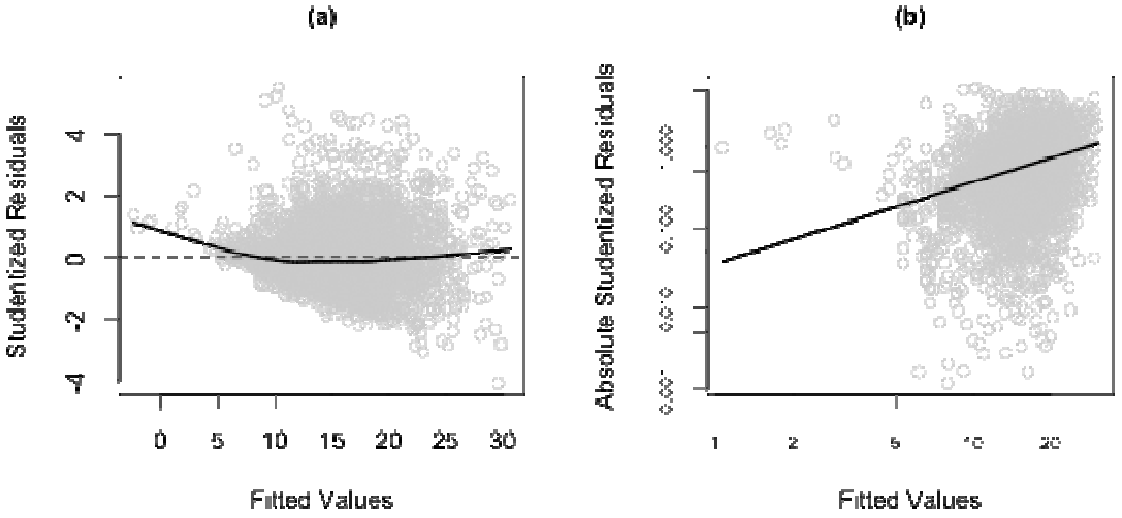
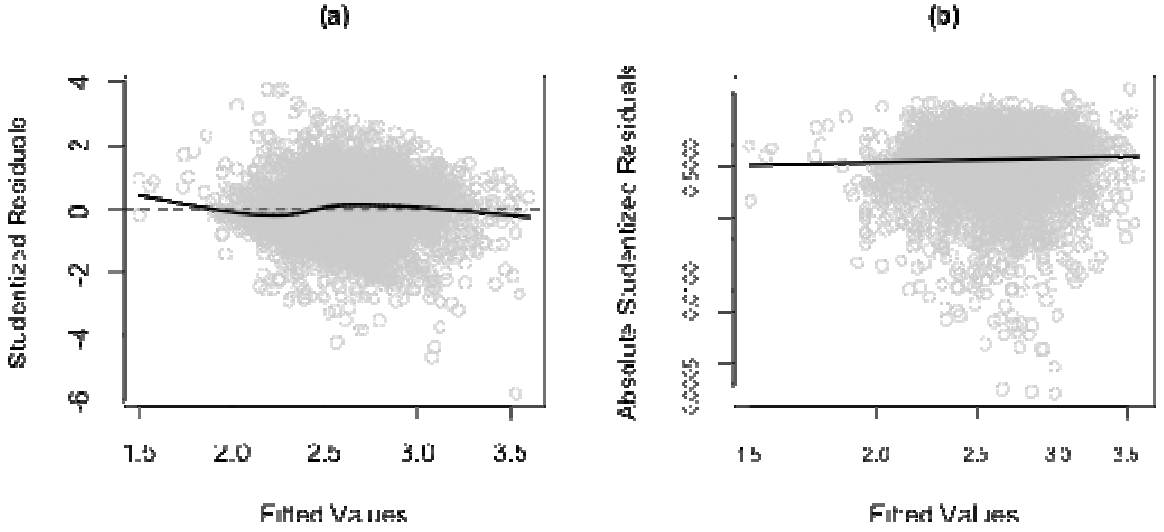


Figure 12.4 Same plots for studentized residuals from regression of log wages.



12.2.2 Weighted-least-squares estimation

If we know how the variance of the response changes, there is an alternative approach to estimation with nonconstant variance. The current approach concerns unweighted sum of squares: $\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2$

Suppose $\text{var}(y_i \mid \mathbf{x}_i) = \text{var}(\varepsilon_i) = \sigma_i^2$. With nonconstant variance we might consider a *weighted sum of squares*:

$$\sum_{i=1}^{n}\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma_i}\right)^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

where $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}$.

$$\frac{\partial}{\partial \beta}(SS) = 2\left(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} = \boldsymbol{0}$$

so $\hat{\boldsymbol{\beta}}_w = \left(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$ is the weighted least squares estimator.

The text writes $\boldsymbol{\Sigma} = \sigma_\varepsilon^2 \text{diag}\left(1/w_1^2, \ldots, 1/w_n^2\right) = \sigma_\varepsilon^2 \boldsymbol{W}^{-1}$, so that we can also write

$$\hat{\boldsymbol{\beta}}_w = \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}$$

and

$$\text{var}\left(\hat{\boldsymbol{\beta}}_w\right) = \hat{\sigma}_\varepsilon^2\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}$$

The text notes that the weighted least squares estimator is also the maximum-likelihood estimator because in this case the likelihood is

$$L(\boldsymbol{\beta}, \sigma_{\varepsilon}^{2}) = \frac{1}{(2\pi)^{n/2} \left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)^{T} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{y} - \boldsymbol{X\beta}\right)\right\}$$

and maximizing this is equivalent to minimizing the weighted least squares criterion.

The text suggests a procedure due to White for estimating $\boldsymbol{\Sigma}$ when it is not known (sect 12.2.3). I won't recommend this for you.

In fact, we will make the most use of the expressions given here in the case where $\boldsymbol{\Sigma}$ is not a diagonal matrix with unequal variances along the diagonal, but where $\boldsymbol{\Sigma}$ is a full matrix with constant variances along the diagonal but non-zero entries off the diagonal, corresponding to *covariances* among the residuals--- i.e., the case of non-independent data. See Chapter 16 on time series.

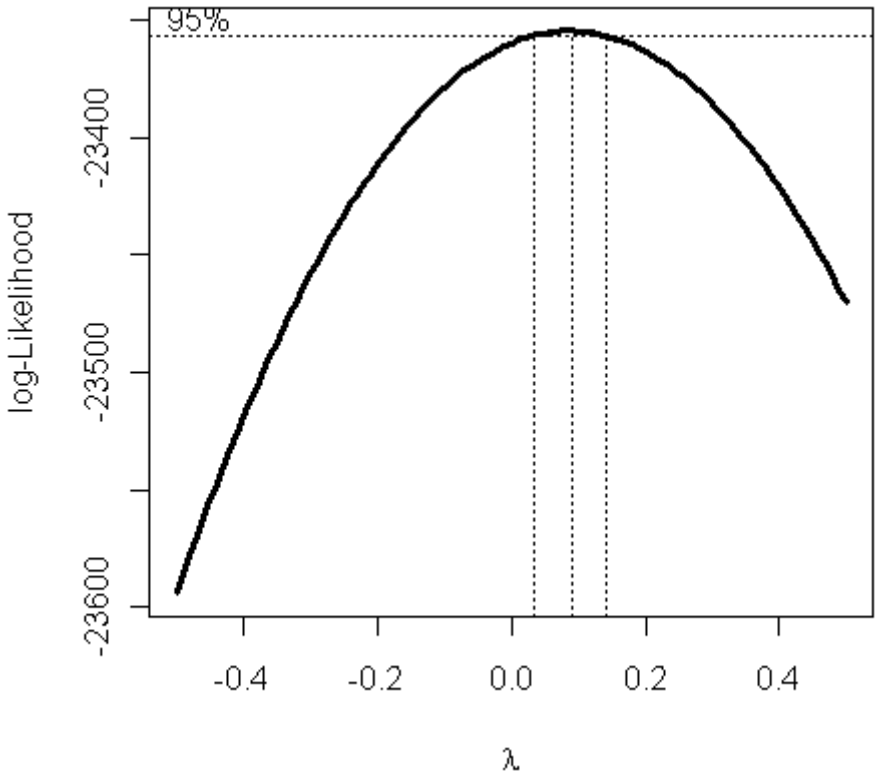
The text makes one final useful comment in sect 12.2.4. It notes that the effects of nonconstant variance on the efficiency of (ordinary, or unweighted) least squares are not too great unless the ratio of the largest error variance to the smallest error variance is quite large, perhaps a factor of 10. The recommended conservative rule of thumb is not to worry too much unless the *ratio of variances exceeds 4*. (This was the rule-of-thumb in the Stat 421/502 lecture notes.) If we have grouped data we can actually compute these variances; otherwise we have to judge this from scatterplots or make groups in order to compute sample variances.

Note however, that the most common scenario of nonconstant variance has $\sigma_i \propto x_i$, or more generally, $\sigma_i \propto E(y_i \mid x_{1i}, x_{2i}, \ldots)$, and in this case we look for *variance-stabilizing transformations* as discussed in Chapter 4.

Section 12.5.1 covers the estimation of power transformations, the *Box-Cox family*, by maximum likelihood.

Fig 12.14

```
> mod0 <- lm(wages ~ sex + age + education)
> windows(width=5, height=5)
> bc <- boxcox(mod0, lambda=seq(-0.5, 0.5,
by=.01))
> lines(bc, lwd=3)
> bc$x[which.max(bc$y)]
[1] 0.09
> range(bc$x[max(bc$y) - bc$y < 1.92])
[1] 0.04 0.14
> cor(box.cox(wages, 0), box.cox(wages,
0.09))
[1] 0.9996324
>
```



12.3 Nonlinearity

12.3.1 Component-Plus-Residual plots

Motivation:

2D scatterplots aren't guaranteed to detect nonlinear structure in multidimensional regression problems (sect 12.3.3).

Suppose $Y_i = \alpha + f(x_{i1}) + \beta_2 x_{i2} + \cdots + \varepsilon_i$

but we fit $Y_i = \alpha' + \beta_1' x_{i1} + \beta_2' x_{i2} + \cdots + \varepsilon_i'$

The *partial residuals* for the *j*th explanatory variable are

$$E_i^{(j)} = E_i + \hat{\beta}_j x_{ij}$$

These add back the linear component of the (estimated) partial relationship between Y and X_j to the least squares residuals (which may include an unmodeled nonlinear component).

The *component-plus-residual plot* (also a partial-residual plot) is a plot of x_{ij} vs $E_i^{(j)}$.

By construction, the slope of the simple linear regression of $E_i^{(j)}$ on x_{ij} is the multiple regression coefficient estimate $\hat{\beta}_j$ --- but nonlinearity may be apparent as well.

But the *added variable plot* (or partial regression plot) of chapter 11 also has slope $\hat{\beta}_j$. Why this new plot?

First, **Fig 12.5** showing the ordinary residual plots are not enough.

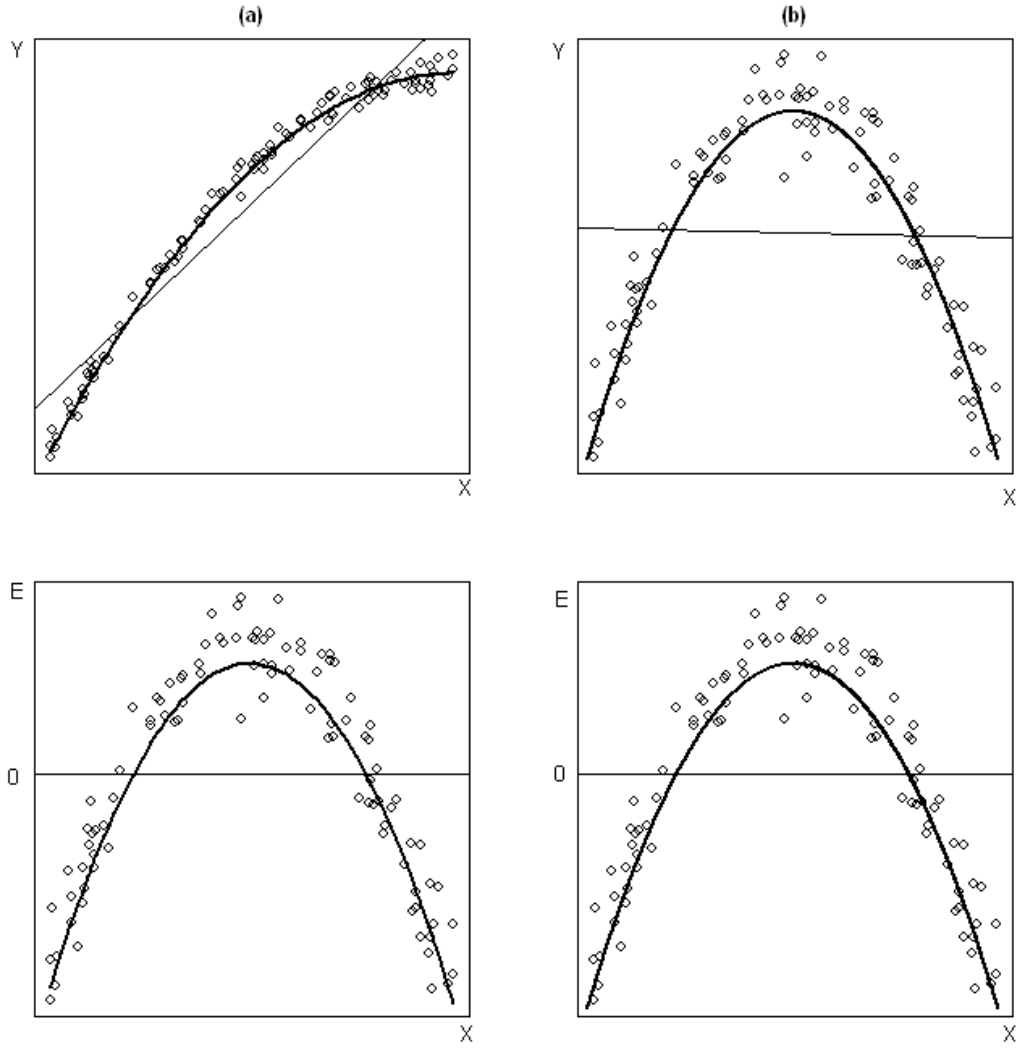
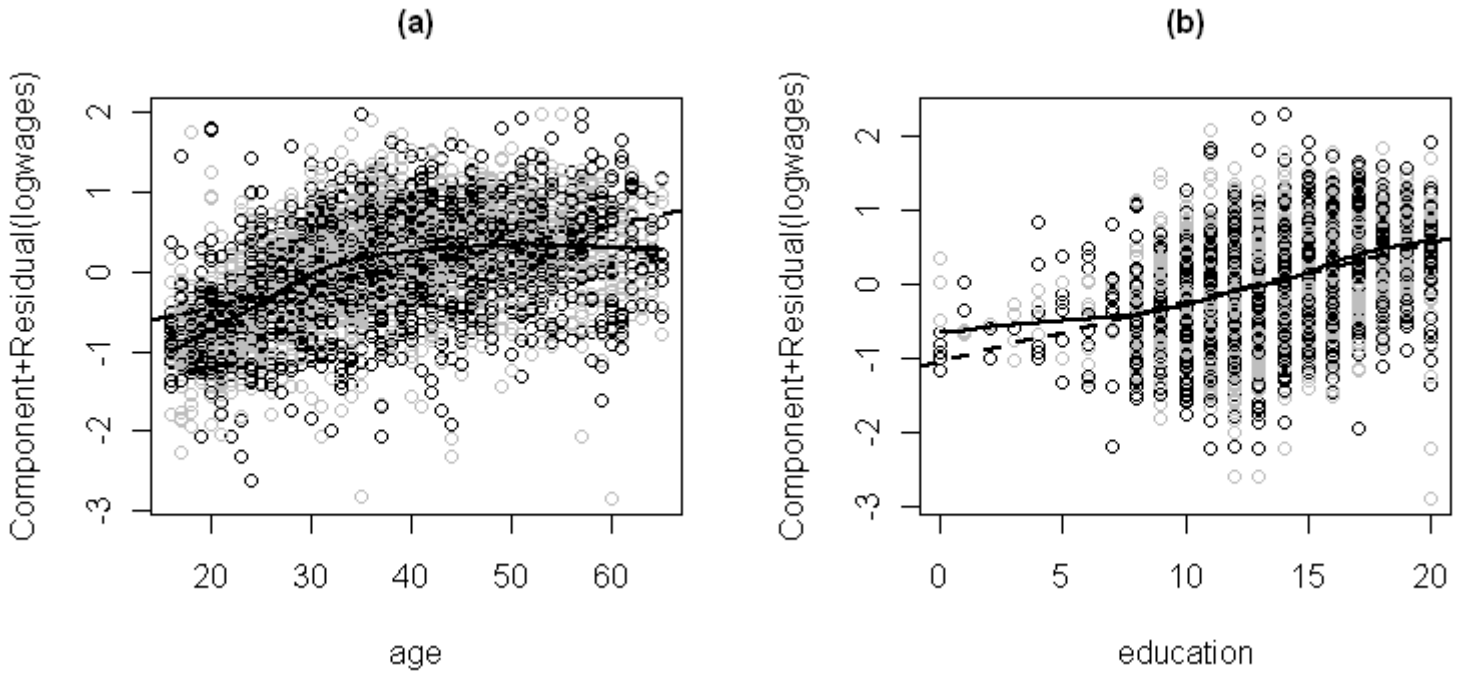


Figure 12.6 Component-plus-residual plots for age and education in the SLID regression of log wages on these variables and sex. Solid lies are lowess smooths (span = 0.4); dashed lines are linear least-squares fits. (Plots generated using car function cr.plot.)



But the *added variable plot* (or partial regression plot) of chapter 11 also has slope $\hat{\beta}_j$. Why this new plot?

- They can be extended to represent fitted quadratic effects (Fig 12.7) and fitted interaction effects (Figs 12.9-12.11).
- They are better for graphical diagnosis of nonlinearity.

Partial residuals $E_i^{(j)} = E_i + \hat{\beta}_j x_{ij}$ are estimates of

$$\varepsilon_i^{(1)} = \beta_j' x_{ij} + \varepsilon_i'$$

We would like these partial residuals to estimate (or reveal) any nonlinear structure. That is

$$f(x_{ij}) + \varepsilon_i$$

One can show that the partial residuals do estimate $f(x_{ij}) + \varepsilon_i$

if

(1) $f(x_{ij})$ is linear after all (see $\varepsilon_i^{(1)}$ above)

or if

(2) other explanatory variables are linearly related to x_{ij}

If there are nonlinear relationships between $\boldsymbol{x}_j$ and other $\boldsymbol{x}_k$, the component-plus-residual plot may not reveal the true nonlinear structure.

So, the first suggestion is to transform all the $\boldsymbol{x}_k$ s so that there are linear relationships among them.