21 Lecture 21: March 21

Last time

- Added-variable plots
- Should unusual data be discarded

Today

- More response from midterm evaluations
- Diagnosing non-normality, non-constant error variance, and nonlinearity (JF chapter 12)

Studentized residual and its hypothesis test

Recall that the externally studentized residual

$$\hat{\epsilon}_i^* \equiv \frac{\hat{\epsilon}}{\hat{\sigma}_{(-i)}\sqrt{1-h_i}}$$

has an independent numerator and denominator and follows a t-distribution with $n-p-2$ degrees of freedom.

It is of our interest to pick the studentized residual $\hat{\epsilon}^*_{max}$ with the largest absolute value among $\hat{\epsilon}^*_1, \hat{\epsilon}^*_2, \ldots, \hat{\epsilon}^*_n$ to test for outlier. However, by doing so, we are effectively picking the biggest of n test statistics such that it is not legitimate simply to use t_{n-p-2} to find a p-value. We need a correction on the p-value because of multiple-comparisons.

Suppose that we have $p' = \Pr(t_{n-p-2} > |\hat{\epsilon}^*_{max}|)$, the p-value before correction. Then the Bonferroni adjusted p-value is $p = np'$.

What is the null hypothesis? Can you construct the exact p-value?

Non-normally distributed errors

The assumption of normally distributed errors is almost always arbitrary. Nevertheless, the central limit theorem ensures that, under very broad conditions, inference based on the least-squares estimator is approximately valid in all but small samples. Why concern about non-normal errors?

- For some types of error distributions, particularly those with heavy tails, the efficiency of least-squares estimation decreases markedly.
- Highly skewed error distributions, aside from their propensity to generate outliers in the direction of the skew, compromise the interpretation of the least-squares fit. This fit is a conditional mean (of Y given the Xs), and the mean is not a good measure of the center of a highly skewed distribution.

- A multimodal error distribution suggests that omission of one or more discrete explanatory variables that divide the data naturally into groups. An examination of the distribution of the residuals may motivate respecification of the model.

Note: The skewness α_3 is defined as $\alpha_3 \equiv \frac{\mu_3}{(\mu_2)^{3/2}}$ where μ_n denotes the nth central moment of a random variable X. The skewness measures the lack of symmetry in the pdf.

Quantile-comparison plot, JF 3.1.3

Quantile-comparison plots are useful for comparing an empirical sample distribution with a theoretical distribution, such as the normal distribution.

Let $P(x)$ represent the theoretical cumulative distribution function (cdf) with which we want to compare the data, that is $P(x) = \Pr(X \leqslant x)$. The quantile-comparison plot is constructed by:

1. Order the data values from smallest to largest, $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. The $X_{(i)}$ are called the order statistics of the sample.
2. By convention, the cumulative proportion of the data "below" $X_{(i)}$ is given by
$$P_i = \frac{i - \frac{1}{2}}{n}$$
3. Use the inverse of the cdf to find the value z_i corresponding to the cumulative probability P_i, that is
$$z_i = P^{-1}(\frac{i - \frac{1}{2}}{n})$$
4. Plot the z_i as horizontal coordinates against the $X_{(i)}$ as vertical coordinates. If X is sampled from the distribution P, then $X_{(i)} \approx z_i$.
   - if the distributions are identical except for location, then the plot is approximately linear with nonzero intercept, $X_{(i)} \approx \mu + z_i$
   - if the distributions are identical except for scale, then the plot is approximately linear with a slope different from 1, $X_{(i)} \approx \sigma z_i$
   - if the distributions differ both in location and scale but have the same shape, then $X_{(i)} \approx \mu + \sigma z_i$
5. It is often helpful to place a comparison line on the plot to facilitate the perception of departures from linearity. For a normal quantile-comparison plot (comparing the distribution of the data with the standard normal distribution), we can alternatively use the median as a robust estimator of μ and the interquartile range/1.39 as a robust estimator of σ.
6. We expect some departure from linearity because of sampling variation. It therefore assists interpretation to display the expected degree of sampling error in the plot. The

standard error of the order statistic $X_{(i)}$ is

$$\text{SE}(X_{(i)}) = \frac{\hat{\sigma}}{p(z_i)}\sqrt{\frac{P_i(1-P_i)}{n}}$$

where $p(z_i)$ is the probability density function, pdf, corresponding to the CDF $P(z)$. The values along the fitted line are given by $\hat{X}_{(i)} = \hat{\mu} + \hat{\sigma} z_i$. An approximate 95% confidence "envelope" around the fitted line is, therefore,

$$\hat{X}_{(i)} \pm 2 \times \text{SE}(X_{(i)})$$

- Figure 21.1 plots a sample of $n = 100$ observations from a normal distribution with mean $\mu = 50$ and standard deviation $\sigma = 10$.

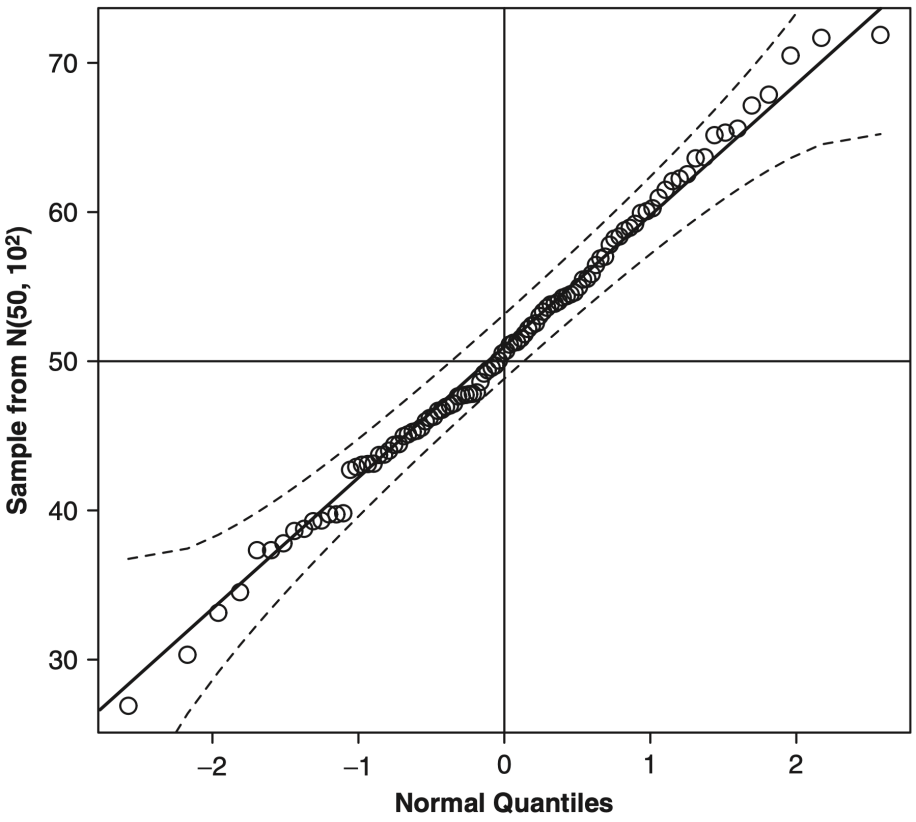

Figure 21.1: Normal quantile-comparison plot for a sample of 100 observations drawn from a normal distribution with mean 50 and standard deviation 10. The fitted line is through the quartiles of the distribution, the broken lines give a pointwise 95% confidence interval around the fit. JF Figure 3.8.

The plotted points are reasonably linear and stay within the rough 95% confidence envelope.

- Figure 21.2 plots a sample of $n = 100$ observations from the positively skewed chi-square distribution with 2 degrees of freedom. The positive skew of the data is reflected in points that lie *above* the comparison line in both tails of the distribution. (In contrast, the tails of negatively skewed data would lie *below* the comparison line.)

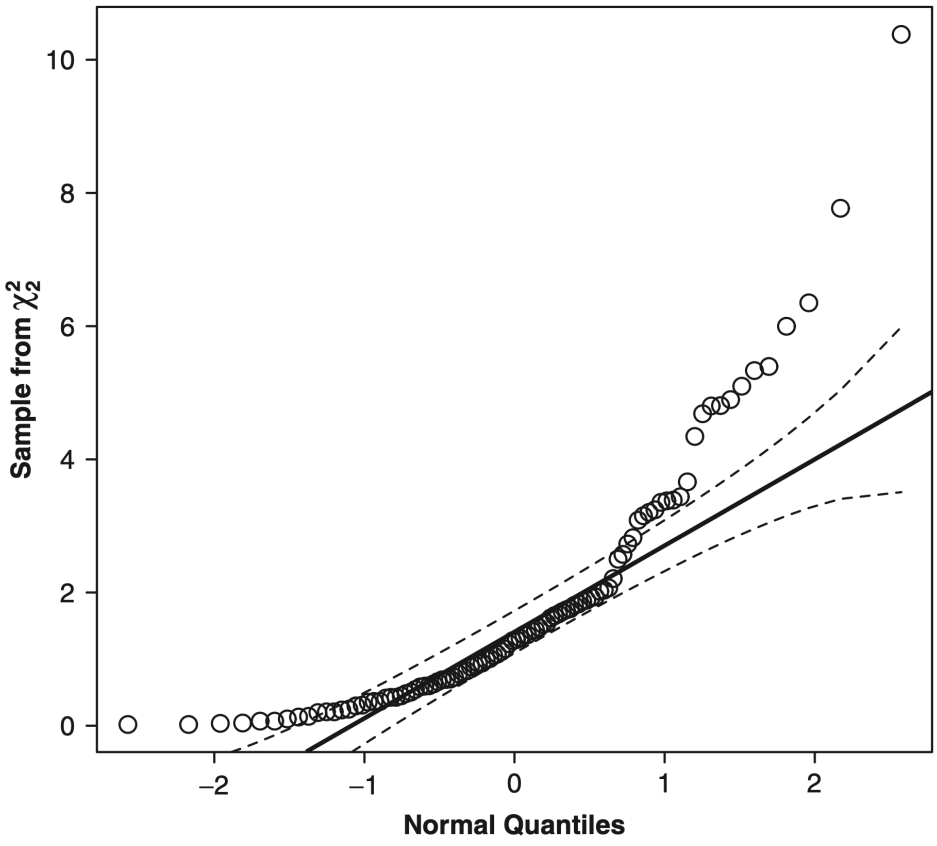

Figure 21.2: Normal quantile-comparison plot for a sample of 100 observations drawn from the positively skewed chi-square distribution with 2 degrees of freedom. JF Figure 3.9.

- Figure 21.3 plots a sample of $n = 100$ observations from the heavy-tailed t distribution with 2 degrees of freedom. In this case, values in the upper tail lie above the corresponding normal quantiles, the values in the lower tail below the corresponding normal quantiles.

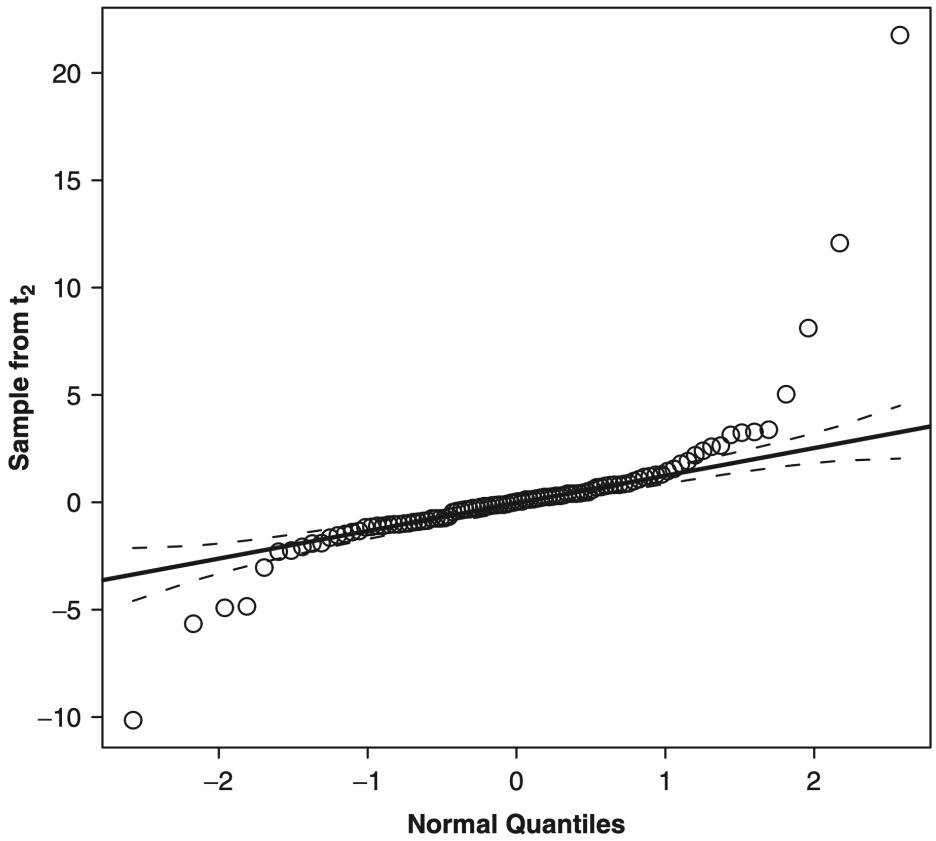

Figure 21.3: Normal quantile-comparison plot for a sample of 100 observations drawn from heavy-tailed t-distribution with 2 degrees of freedom. JF Figure 3.10.

- Figure 21.4 shows the normal quantile-comparison plot for the distribution of infant mortality. The positive skew of the distribution is readily apparent.

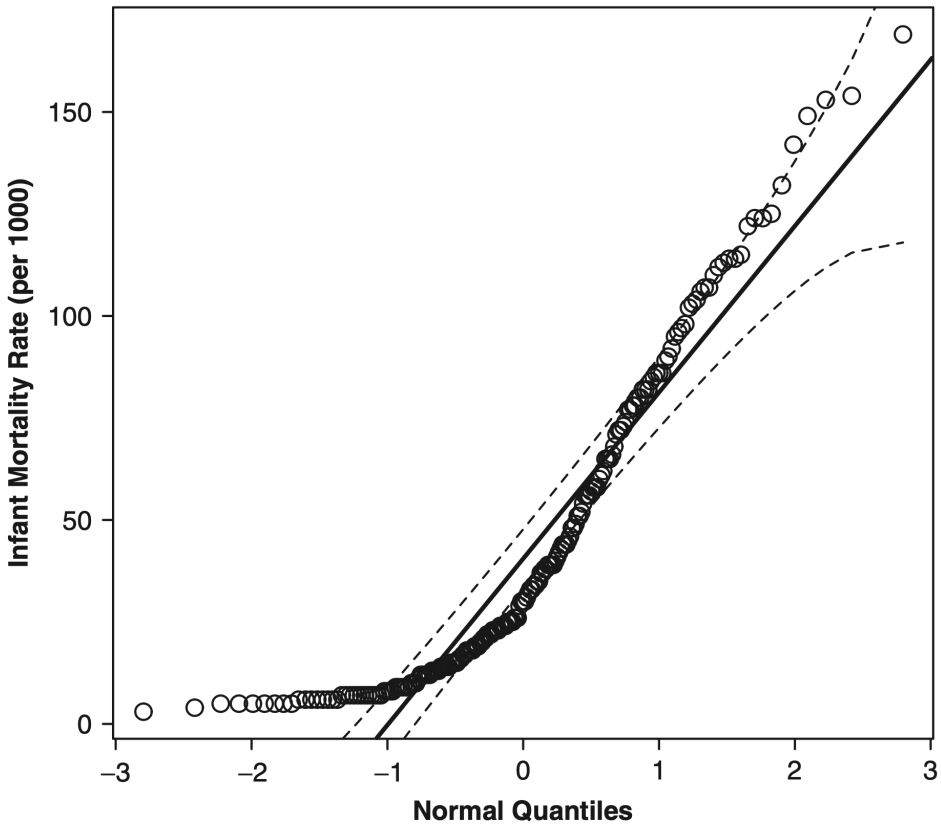

Figure 21.4: Normal quantile-comparison plot for the distribution of infant mortality. Note the positive skew. JF Figure 3.11.

Nonconstant error variance

One of the assumptions of the regression model is that the variation of the response variable around the regression surface (the error variance) is everywhere the same:

$$\mathrm{Var}(\epsilon) = \mathrm{Var}(\mathrm{Y}|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p}) = \sigma^2_\epsilon$$

Constant error variance is often termed homoscedasticity, and similarly, nonconstant error variance is termed heteroscedasticity. We detect nonconstant error variances through graphical methods.

Residual plots

Because the least square residuals have unequal variance even when the constant variance assumption is correct:

$$\mathrm{Var}(\hat{\epsilon}_\mathrm{i}) = \sigma^2(1 - \mathrm{h}_\mathrm{i}).$$

It is preferable to plot studentized residuals against fitted values. A pattern of changing spread is often more easily discerned in a plot of absolute studentized residuals, $|\hat{\epsilon}^*_i|$, or

squared studentized residuals, $\hat{\epsilon}_i^{*2}$, against $\hat{Y}$. If the values of $\hat{Y}$ are all positive, then we can plot $\log|\hat{\epsilon}_i^*|$ against $\log\hat{Y}$. Figure 21.5 shows a plot of studentized residuals against fitted values and spread-level plot of studentized residuals, several points with negative fitted values were omitted.

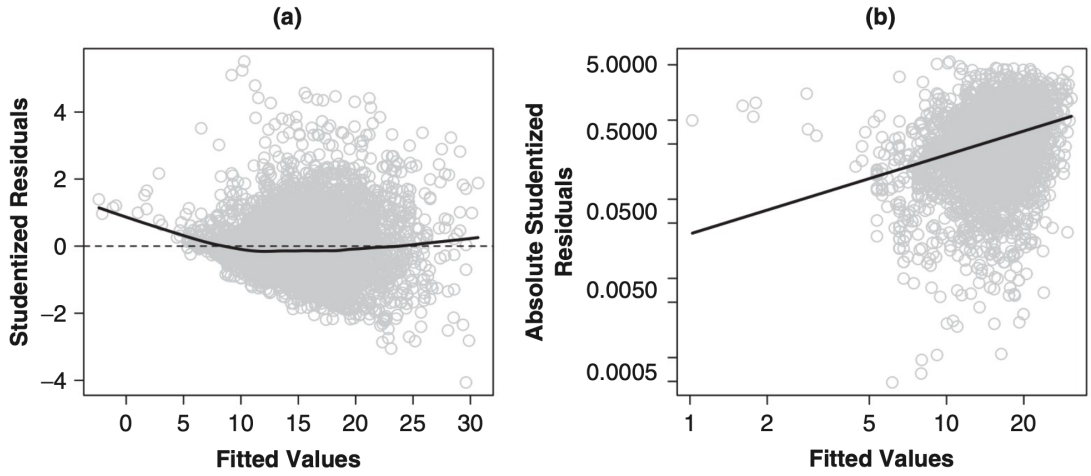

Figure 21.5: (a) Plot of studentized residuals versus fitted values and (b) spread-level plot for studentized residuals. JF Figure 12.3.

It is apparent from both graphs that the residual spread tends to increase with the level of the response, suggesting a violation of constant error variance assumption.

Weighted-least-squares estimation

Weighted-least-squares (WLS) regression provides an alternative approach to estimation in the presence of nonconstant error variance. Suppose that the errors from the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ are independent and normally distributed, with zero means but *different* variances: $\epsilon_i \sim N(0, \sigma_i^2)$. Suppose further that the variances of the errors are known up to a constant of proportionality σ_ϵ^2, so that $\sigma_i^2 = \sigma_\epsilon^2/w_i^2$. Then the likelihood for the model is

$$L(\beta, \sigma_\epsilon^2) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)^T\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\beta)\right]$$

where $\mathbf{\Sigma}$ is the covariance matrix of the errors,

$$\mathbf{\Sigma} = \sigma_\epsilon^2 \times \text{diag}\{1/w_1^2, \ldots, 1/w_n^2\} \equiv \sigma_\epsilon^2\mathbf{W}^{-1}$$

The maximum-likelihood estimators of β and σ_ϵ^2 are then

$$\hat{\beta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$$
$$\hat{\sigma}_\epsilon^2 = \frac{\sum(w_i\hat{\epsilon}_i)^2}{n}$$

Correcting OLS standard errors for nonconstant variance

The covariance matrix of the ordinary-least-squares (OLS) estimator is

$$\begin{aligned}\mathbf{Var}\left(\hat{\beta}\right) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Var}\left(\mathbf{Y}\right)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma_\epsilon^2(\mathbf{X}^T\mathbf{X})^{-1}\end{aligned}$$

under the standard assumptions, including the assumption of constant error variance, $\mathbf{Var}\left(\mathbf{Y}\right) = \sigma_\epsilon^2\mathbf{I}_n$. If, however, the errors are heteroscedastic but independent then $\mathbf{\Sigma} \equiv \mathbf{Var}\left(\mathbf{Y}\right) = \text{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$, and

$$\mathbf{Var}\left(\hat{\beta}\right) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

White (1980) shows that the following is a consistent estimator of $\mathbf{Var}\left(\hat{\beta}\right)$

$$\tilde{\text{Var}}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{\Sigma}}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

with $\hat{\mathbf{\Sigma}} = \text{diag}\{\hat{\sigma}_1^2, \ldots, \hat{\sigma}_n^2\}$, where $\hat{\sigma}_i^2$ is the OLS residual for observation i.

Subsequent work suggested small modifications to White's coefficient-variance estimator, and in particular simulation studies by Long and Ervin (2000) support the use of

$$\tilde{\text{Var}}^*(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{\Sigma}}^*\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

where $\hat{\mathbf{\Sigma}}^* = \text{diag}\{\hat{\sigma}_i^2/(1-h_i)^2\}$ and h_i is the hat-value associated with observation i. In large samples, where h_i is small, the distinction between $\tilde{\text{Var}}(\hat{\beta})$ and $\tilde{\text{Var}}^*(\hat{\beta})$ essentially disappears.

A rough *rule* is that nonconstant error variance seriouly degrades the least-squares estimator only when the ratio of the largest to smallest variance is about 10 or more (or, more conservatively, about 4 or more).

Nonlinearity

If $\mathbf{E}\left(\mathbf{Y}|\mathbf{X}\right)$ is not linear in $\mathbf{X}$ (in other words, $\mathbf{E}\left(\epsilon|\mathbf{X}\right) \neq 0$ for some x), $\hat{\beta}$ may be biased and inconsistent. Usually we employ "linearity by default" but we should try to make sure this is appropriate: **detect** non-linearities and **model** them accurately.

Lowess smoother, JF 2.3

We can employ local averaging plots to help with diagnostics. Lowess method is in many respects similar to local-averaging smoothers, except that instead of computing an average Y-value within the neighborhood of a focal x, the lowess smoother computes a *fitted* value based on a locally weighted least-squares line, giving more weight to observations in the neighborhood that are close to the focal x than to those relatively far away. The name "lowess" is an acronym for *lo*cally *we*ighted *s*catterplot *s*moother and is sometimes rendered as *loess*, for *lo*cal regr*ess*ion.

Component-plus-residual plots

Added-variable plots, introduced for detecting influential data, can reveal nonlinearity. However, the added-variable plots are not always useful for locating a transformation:

- The added-variable plot adjusts X_j for the other Xs.
- The *unadjusted* X_j is transformed in respecifying the model.

Moreover, Cook (1998, Section 14.5) shows that added-variable plots are biased toward linearity when the correlations among the explanatory variables are large.

Component-plus-residual plots (also called *partial-residual plots*) are often an effective alternative. The component-plus-residual plots are not as suitable as added-variable plots for revealing leverage and influence, though. The component-plus-residual plots are constructed by

1. Compute residuals from full regression:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

2. Compute "linear component" of the partial relationship:

$$C_i = \hat{\beta}_j X_{ij}$$

3. Add linear component to residual to get partial residual for the jth explanatory variable

$$\hat{\epsilon}_i^{(j)} = \hat{\epsilon}_i + C_i = \hat{\epsilon}_i + \hat{\beta}_j X_{ij}$$

4. Plot $\hat{\epsilon}_{\cdot}^{(j)}$ against $X_{\cdot j}$

Figure 21.6 shows the component-plus-residual plots for the regression of log wages on variables (age, education and sex) of the 1994 wave of Statistics Canada's Survey of Labour and Income Dynamics (SLID) data. The SLID data set includes 3997 employed individuals who were between 16 and 65 years of age and who resided in Ontario.

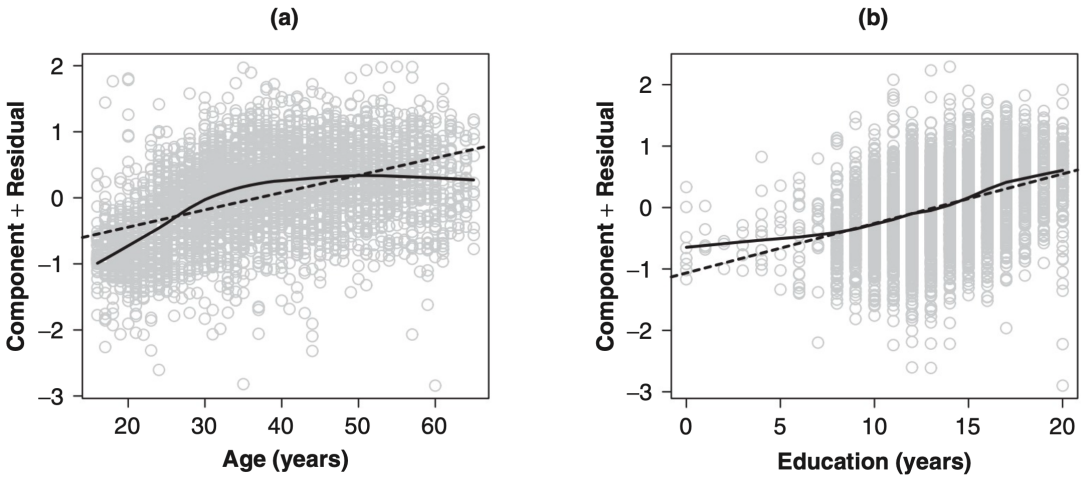

Figure 21.6: Component-plus-residual plots for age and education in SLID regression of log wages on these variables and sex. The solid lines are for lowess smooths with spans of 0.4, and the broken lines are for linear least-squares fits. JF Figure 12.6.