19 Lecture 19: March 16

Last time

- Unusual and influential data (JF chapter 11)

Today

- Added-variable plots
- Should unusual data be discarded
- Diagnosing non-normality, non-constant error variance, and nonlinearity (JF chapter 12)

Added-variable plots

Unlike the case of SLR, the scatterplot with the response variable and one predictor gives only the marginal effect in MLR. Instead, the added-variable plot (also called a partial-regression plot or a partial-regression leverage plot) gives a graphical inspection over each dimension.

Let $\tilde{Y}_i^{(1)}$ represent the residuals from the least-squares regression of Y on all the Xs except X_1, in other words, the residuals from the following fitted regression equation:

$$Y_i = \tilde{\beta}_0^{(1)} + \tilde{\beta}_2^{(1)} X_{i2} + \cdots + \tilde{\beta}_p^{(1)} X_{ip} + \tilde{Y}_i^{(1)}$$

where the parenthetical superscript (1) indicates the omission of X_1 from the right-hand side of the regression equation. Likewise, $\check{X}_i^{(1)}$ is the residual from the least-squares regression of X_1 on all the other Xs:

$$X_{i1} = \check{\beta}_0^{(1)} + \check{\beta}_2^{(1)} X_{i2} + \cdots + \check{\beta}_p^{(1)} X_{ip} + \check{X}_i^{(1)}$$

Then, the residuals $\tilde{Y}_i^{(1)}$ and $\check{X}_i^{(1)}$ have the following interesting properties:

1. The slope from the least-squares regression of $\tilde{Y}_i^{(1)}$ on $\check{X}_i^{(1)}$ is simply the least-squares slope $\hat{\beta}_1$ from the *full* multiple regression.
2. The residuals from the simple regression of $\tilde{Y}_i^{(1)}$ on $\check{X}_i^{(1)}$ are the same as those from the full regression, that is
$$\tilde{Y}_i^{(1)} = \hat{\beta}_1 \check{X}_i^{(1)} + \hat{\epsilon}_i$$
3. The variation of $\check{X}_i^{(1)}$ is the *conditional variation* of X_1 holding the other Xs constant.

Figure 19.1 shows that the conditional variation is smaller than its marginal variation – much smaller when X_1 is strongly collinear with other Xs,

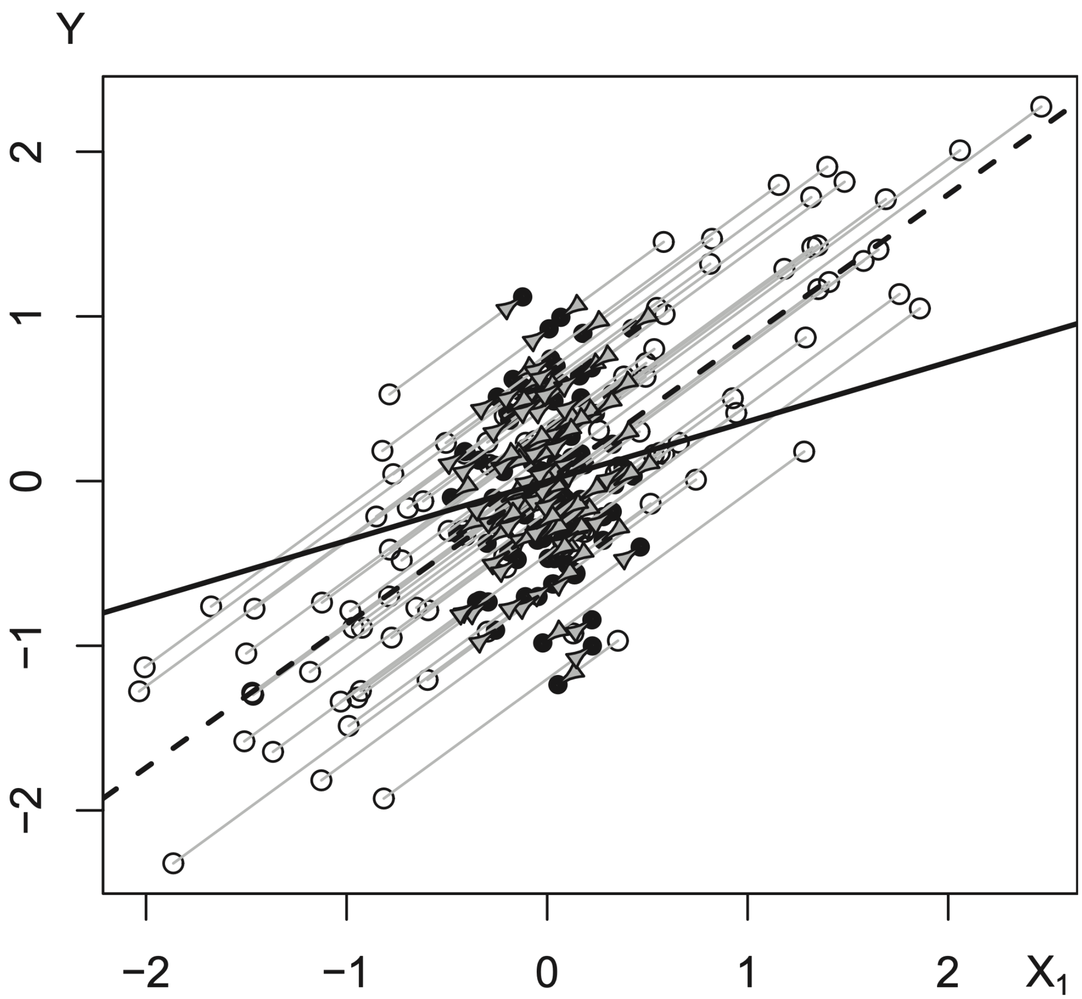

Figure 19.1: The marginal scatterplot (open circles) for Y and X_1 superimposed on the added-variable plot (filled circles) for X_1 in the regression of Y on X_1 and X_2. The variables Y and X_1 are centered at their means to facilitate the comparison of the two sets of points. The arrows show how the points in the marginal scatterplot map into those in the AV plot. In this contrived data set, X_1 and X_2 are highly correlated ($r_{12} = 0.98$), and so the conditional variation in X_1 (represented by the horizontal spread of the filled points) is much less than its marginal variation (represented by the horizontal spread of the open points). The broken line gives the slope of the marginal regression of Y on X_1 alone, while the solid line gives the slope $\hat{\beta}_1$ of X_1 in the MLR of Y on both Xs. JF Figure 11.9.

Figure 19.2 illustrates the added-variable plots using the Duncan's data.

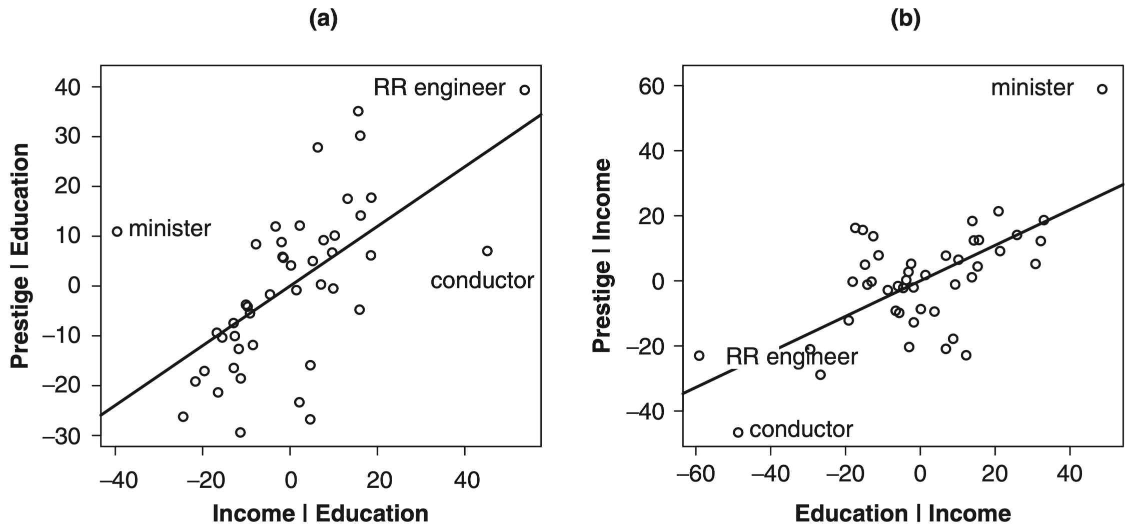

Figure 19.2: Added-variable plots for Duncan's regression of occupational prestige on the (a) income and (b) education levels of 45 US occupations in 1950. Three unusual observations, *miniters*, *conductors*, and *railroadengineers*, are identified on the plots. The added-variable plot for the intercept $\hat{\beta}_0$ is not shown. JF Figure 11.10.

The added-variable plot for income in Figure 19.2(a) reveals three observations that exert substantial leverage on the income coefficient:

- *minister*, whose income is unusually low given the educational level of the occupation
- *conductor*, whose income is unusally high given education
- *railroad engineer*, whose income is relatively high given education.

Remember that the horizontal variable in this added-variable plot is the residual from the regression of income on education, and thus values far from 0 in this direction are for occupations with incomes that are unusually high or low given their levels of education.

Should unusual data be discarded?

In practice, although problematic data should not be ignored, they also should not be deleted automatically and without reflection:

- It is important to investigate *why* an observation is unusual. Truly "bad" data (e.g., an error in data entry) can often be corrected or, if correction is not possible, thrown away. When a discrepant data point is correct, we may be able to understand why the observation is unusual. For Duncan's data, for example, it makes sense that ministers enjoy prestige not accounted for by the income and educational levels of the occupation and for a reason not shared by other occupations. In a case like this, where an outlying observation has characteristics that render it unique, we may choose to set it aside from the rest of the data.

- Alternatively, outliers, high-leverage points, or influential data may motivate model respecification, and the pattern of unusual data may suggest the introduction of additional explanatory variables. We noticed, for example, that both conductors and railroad engineers had high leverage in Duncan's regression because these occupations combined relatively high income with relatively low education. Perhaps this combination of characteristics is due to a high level of unionization of these occupations in 1950, when the data were collected. If so, and if we can ascertain the levels of unionization of all of the occupations, we could enter this as an explanatory variable, perhaps shedding further light on the process determining occupational prestige.

- Except in clear-cut cases, we are justifiably reluctant to delete observations or to respecify the model to accommodate unusual data. Some researchers reasonably adopt alternative estimation strategies, such as robust regression, which continuously downweights outlying data rather than simply discarding them. Because these methods assign zero or very small weight to highly discrepant data, however, the result is generally not very different from careful application of least squares, and , indeed, robust-regression weights can be used to identify outliers.

- Finally, in large samples, unusual data substantially alter the results only in extreme instances. Identifying unusual observations in a large sample, therefore, should be regarded more as an opportunity to learn something about the data not captured by the model that we have fit, rather than as an occasion to reestimate the model with the unusual observations removed.

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 <u>skewness α_3</u> 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, $X_{(1)}, X_{(2)}, \dots, X_{(n)}$. 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
$$\mathrm{SE}(X_{(i)}) = \frac{\hat{\sigma}}{p(z_i)}\sqrt{\frac{P_i(1 - P_i)}{n}}$$
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,
$$\hat{X}_{(i)} \pm 2 \times \mathrm{SE}(X_{(i)})$$

- Figure 19.3 plots a sample of $n = 100$ observations from a normal distribution with mean $\mu = 50$ and standard deviation $\sigma = 10$.

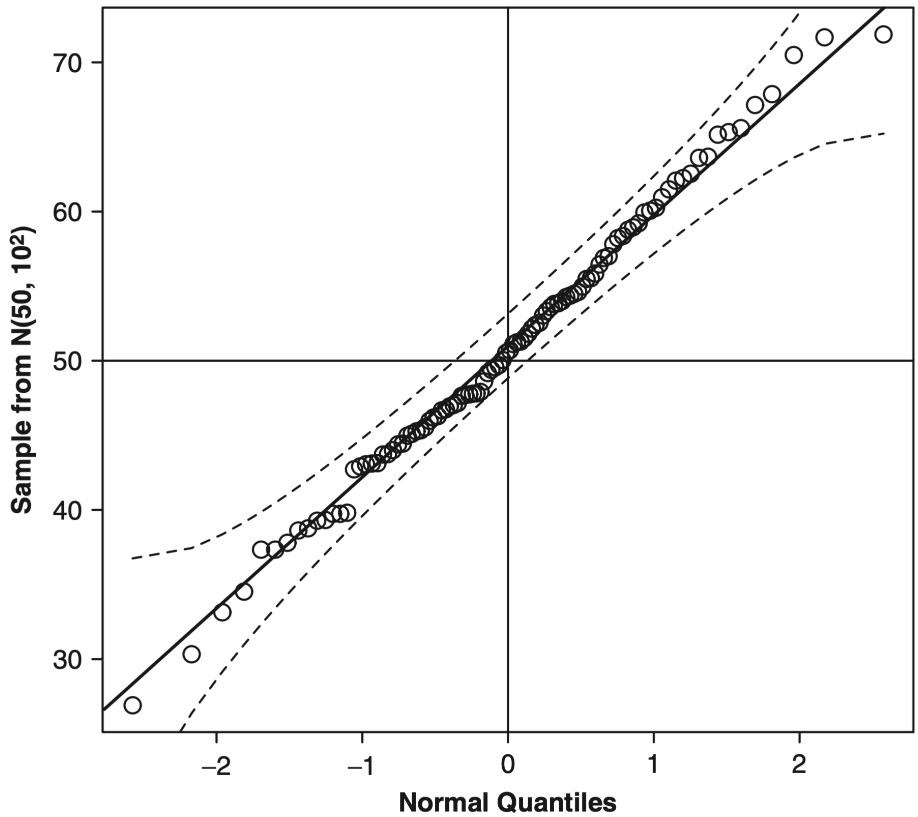

Figure 19.3: Normal quantile-comparison plot for a sample of 100 observations drawn from a normal distribution with mean 50 and standard deviation 10. The fitted line is through the quartiles of the distribution, the broken lines give a pointwise 95% confidence interval around the fit. JF Figure 3.8.

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

- Figure 19.4 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.)

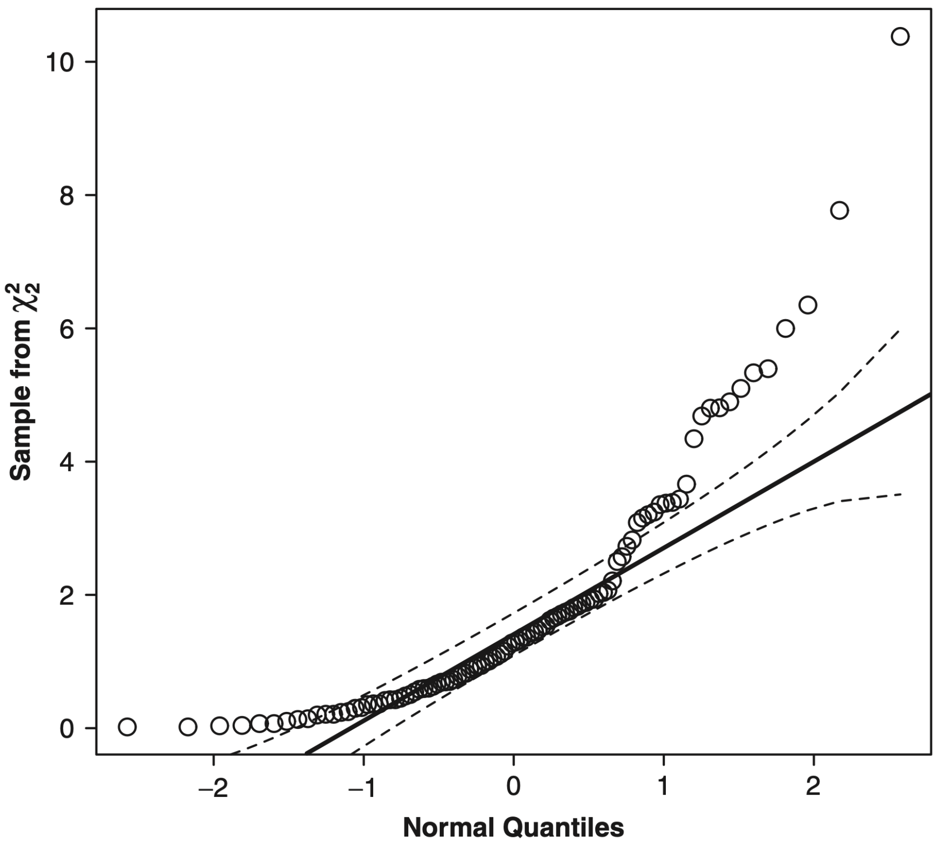

Figure 19.4: Normal quantile-comparison plot for a sample of 100 observations drawn from the positively skewed chi-square distribution with 2 degrees of freedom. JF Figure 3.9.

- Figure 19.5 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.

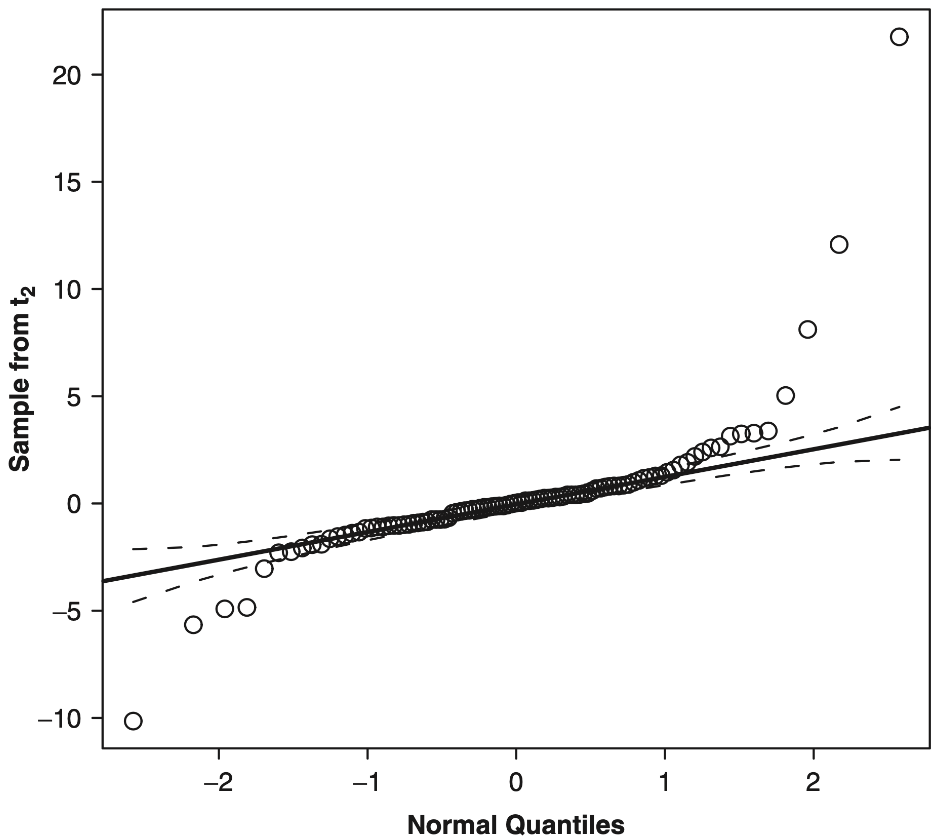

Figure 19.5: Normal quantile-comparison plot for a sample of 100 observations drawn from heavy-tailed t-distribution with 2 degrees of freedom. JF Figure 3.10.

- Figure 19.6 shows the normal quantile-comparison plot for the distribution of infant mortality. The positive skew of the distribution is readily apparent.

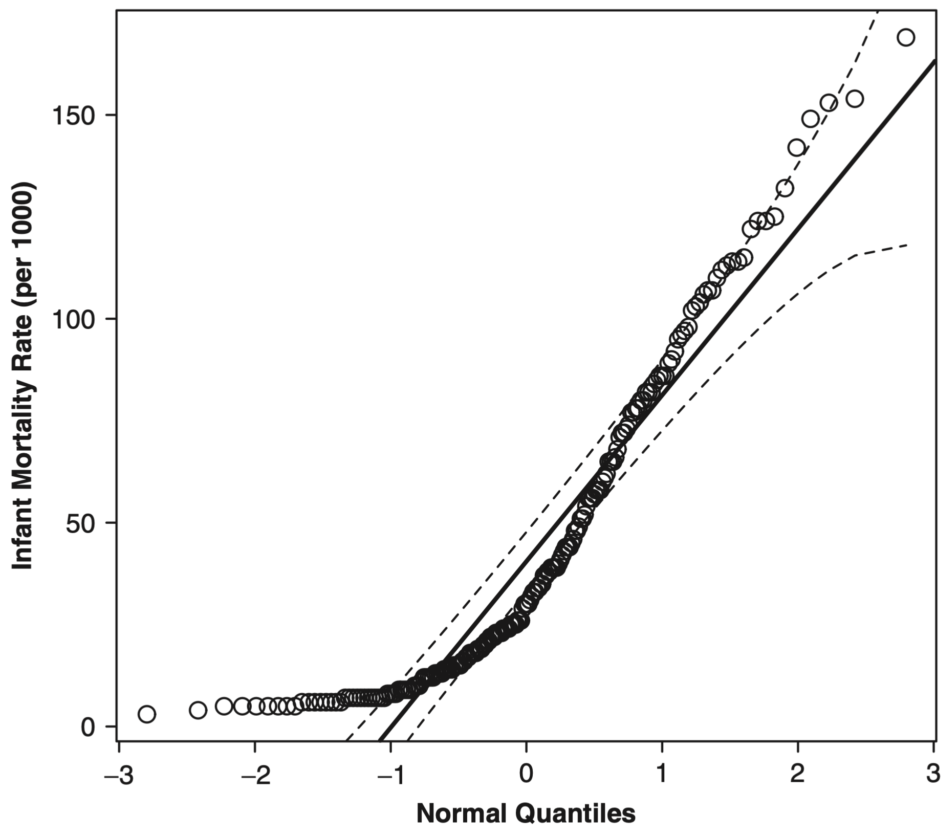

Figure 19.6: Normal quantile-comparison plot for the distribution of infant mortality. Note the positive skew. JF Figure 3.11.

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 19.7 shows a plot of studentized residuals against fitted values and spread-level plot of studentized residuals, several points with negative fitted values were omitted.

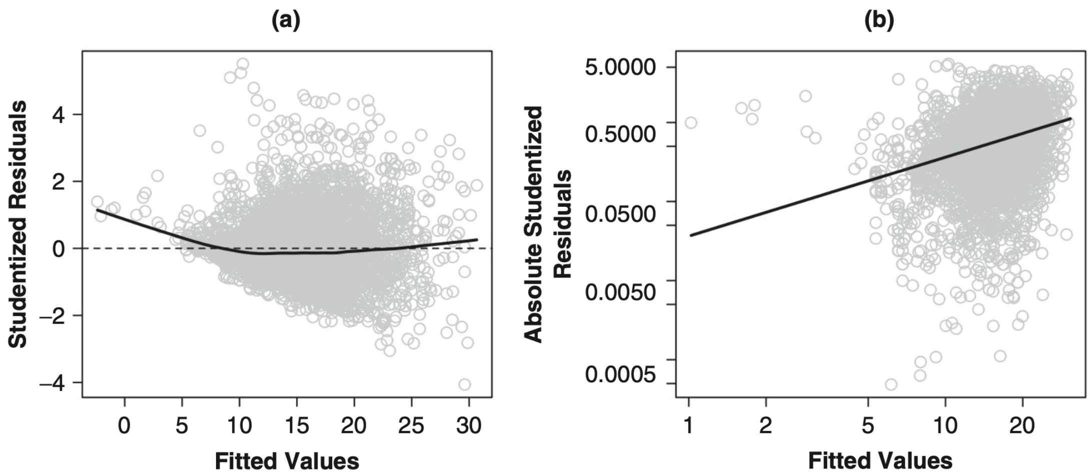

Figure 19.7: (a) Plot of studentized residuals versus fitted values and (b) spread-level plot for studentized residuals. JF Figure 12.3.

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,

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

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).