Diagnostics and Remedial Measures: Continue.

Recall:

In the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1,2,...,n$$

$$E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2 \quad and \quad Cov(\varepsilon_i, \varepsilon_j) = 0 \ for\ all\ i \neq j.$$

Then

$$E(Y_i) = \beta_0 + \beta_1 X_i \qquad and \qquad Var(Y_i) = \sigma^2.$$

The point estimates of β_0, β_1 are

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = b_0 = \bar{Y} - b_1 \bar{X}$$

$$\hat{\beta}_1 = b_1 = \sum_{i=1}^{n} K_i Y_i, \quad K_i = \frac{\left(X_i - \bar{X}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}, \quad \hat{\beta}_0 = \sum_{i=1}^{n} L_i Y_i, \quad L_i = \frac{1}{n} - \bar{X} K_i$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}, \quad Var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2}\right],$$

The unbiased estimate of σ^2 is

$$s^2 = MSE = \widehat{\sigma^2} = \frac{SSE}{n-2}, \quad SSE = \sum_{i=1}^{n} e_i^2, \quad e_i = Y_i - \hat{Y}_i, \quad i = 1,2,...$$

When a regression model, such as the simple linear regression model is considered for an application, we can usually not be certain in advance that the model is appropriate for that application. Anyone, or several, of the features of the model (conditions), such as linearity of the regression function or normality of the error terms, may not be appropriate for the particular data at hand. Hence, it is important to examine the aptness of the model for the data before inferences based on that model are undertaken. In this part, we discuss some simple graphic methods for studying the appropriateness of a model.

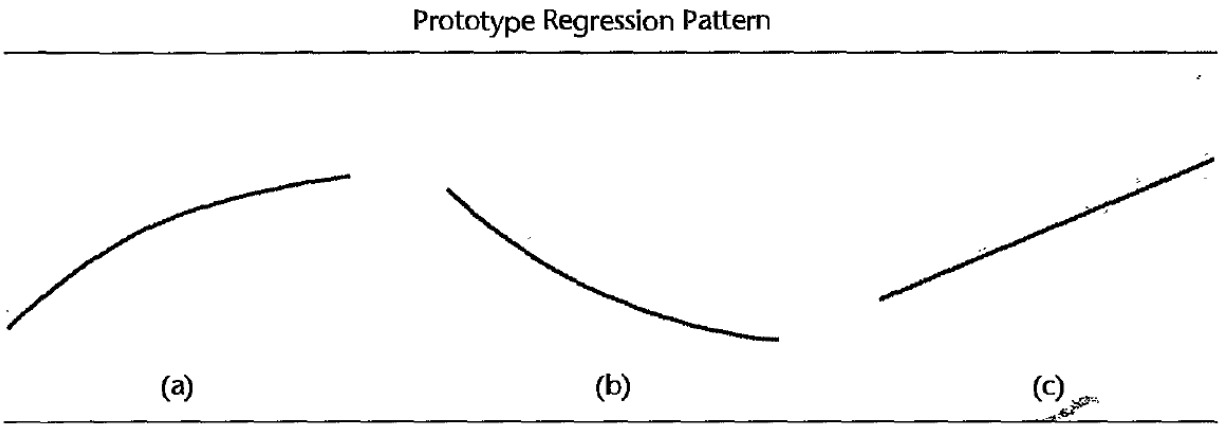
We also consider some remedial techniques that can be helpful when the data are not in accordance with the conditions of regression model. One of these techniques is the transformations.

Transformations for Nonnormality and Unequal Error Variances

Unequal error variances and nonmorality of the error terms frequently appear together. To remedy these departures from the simple linear regression model, we need a transformation on Y, since the shapes and spreads of the distributions of Y need to be changed. Such a transformation on Y may also at the same time help to linearize a curvilinear regression relation. At other times, a simultaneous transformation on X may be needed to obtain or maintain a linear regression relation.

Frequently, the nonmorality and unequal variances departures from regression model take the form of increasing skewness and increasing variability of the distributions of the error terms as the mean response E {Y} increases. For example, in a regression of yearly household expenditures for vacations (Y) on household income (X), there will tend to be more variation and greater positive skewness (i.e., some very high yearly vacation expenditures) for high-income households than for low-income households, who tend to consistently spend much less for vacations.

The figure below also presents some simple transformations on Y that may be helpful for these cases. Several alternative transformations on Y may be tried, as well as some simultaneous transformations on X. Scatter plots and residual plots should be prepared to determine the most effective transformation(s).



Example (page 132)

Data on age (X) and plasma level of a polyamine (Y) for a portion of the 25 healthy children in a study are presented in columns 1 and 2 of Table 3.8. These data are plotted in Figure 3.16a as a scatter plot. Note the distinct curvilinear regression relationship, as well

as the greater variability for younger children than for older ones.

Child i	(1) Age X_i	(2) Plasma Level Y_i	(3) $Y_i' = \log_{10} Y_i$
1	0 (newborn)	13.44	1.1284
2	0 (newborn)	12.84	1.1086
3	0 (newborn)	11.91	1.0759
4	0 (newborn)	20.09	1.3030
5	0 (newborn)	15.60	1.1931
6	1.0	10.11	1.0048
7	1.0	11.38	1.0561
...	...	...	...
19	3.0	6.90	.8388
20	3.0	6.77	.8306
21	4.0	4.86	.6866
22	4.0	5.10	.7076
23	4.0	5.67	.7536
24	4.0	5.75	.7597
25	4.0	6.23	.7945

On the basis of the prototype regression pattern in Figure 3.15b, we shall first try the logarithmic transformation Y' = log10(Y). The transformed Y values are shown in column 3 of Table 3.8. Figure 3.16b contains the scatter plot with this transformation. Note that the transformation not only has led to a reasonably linear regression relation, but the variability at the different levels of X also has become reasonably constant. To further examine the reasonableness of the transformation Y' = log10(Y), we fitted the

simple linear regression model (2.1) to the transformed Y data and obtained:

$$Y' = 1.135 - .1023X$$

A plot of the residuals against X is shown in Figure 3.16c, and a normal probability plot of the residuals is shown in Figure 3.16d. The coefficient of correlation between the ordered residuals and their expected values under normality is .981. For a = .05, Table B.6 indicates that the critical value is .959 so that the observed coefficient supports the assumption of normality of the error terms. All of this evidence supports the appropriateness of regression model (2.1) for the transformed Y data.