Diagnostics and Remedial Measures

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 Nonlinear Relation Only

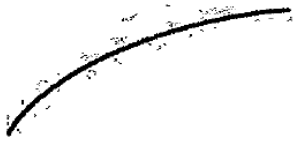
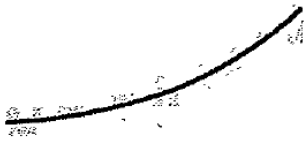
We first consider transformations for linearizing a nonlinear regression relation when the distribution of the error terms is reasonably close to a normal distribution and the error terms have approximately constant variance. In this situation, transformations on X should be attempted. The reason why transformations on Y may not be desirable here is that a transformation on Y, such as $Y' = \sqrt{y}$, may materially change the shape of the distribution

of the -error terms from the normal distribution and may also lead to substantially differing

error term variances.

The following Figure contains some prototype nonlinear regression relations with constant error variance and also presents some simple transformations on X that may be helpful to linearize the regression relationship without affecting the distributions of Y. Several alternative transformations may be tried. Scatter plots and residual plots based on each transformation should then be prepared and analyzed, to decide which transformation is most effective

Prototype Nonlinear Regression Patterns with Constant Error Variance and Simple Trans-formations of *X*.

	Prototype Regression Pattern	Transformations of X	
(a)		$X' = \log_{10} X$	$X' = \sqrt{X}$
(b)		$X' = X^2$	$X' = \exp(X)$
(c)		$X' = 1/X$	$X' = \exp(-X)$

Example (page 129)

Data from an experiment on the effect of number of days of training received (X) on performance (Y) in a battery of simulated sales situations are presented in the following table, columns 1 and 2, for the 10 participants in the study. A scatter plot of these data is shown. Clearly the regression relation appears to be curvilinear, so the simple linear regression model does not be appropriate. Since the variability at the different X levels appears to be fairly constant, we shall consider a transformation on X. Based on the prototype plot in (a), we shall consider initially the· square root transformation

$X' = \sqrt{x}$ The transformed values are shown below:

Sales Trainee i	(1) Days of Training X_i	(2) Performance Score Y_i	(3) $X_i' = \sqrt{X_i}$
1	.5	42.5	.70711
2	.5	50.6	.70711
3	1.0	68.5	1.00000
4	1.0	80.7	1.00000
5	1.5	89.0	1.22474
6	1.5	99.6	1.22474
7	2.0	105.3	1.41421
8	2.0	111.8	1.41421
9	2.5	112.3	1.58114
10	2.5	125.7	1.58114

To examine further whether the simple linear regression model (2.1) is appropriate now, we fit it to the transformed X data. The regression calculations with the transformed X data are carried out in the usual fashion, except that the predictor variable now is X'. We obtain the following fitted regression function:

Y = -10.33 + 83.45X'

Form this model, we see that

1- No strong indications of substantial departures from normality are indicated by this plot.

2- The coefficient of correlation between the ordered residuals and their expected values under normality, .979. is substantially larger and supports the reasonableness of normal error terms. Thus, the new simple linear regression model appears to be appropriate here for the transformed data.

The fitted regression function in the original units of X can easily be obtained, if desired:

$$Y = -10.33 + 83.4S\ \sqrt{x}$$

Scatter Plots and Residual Plots—Sales Training Example.