Regression Example for a Curvilinear Relationship in Bivariate Data

**Example.** A random sample of 7 students provided the following bivariate data, where $X$ denotes the number of study hours for a difficult exam and $Y$ denotes the corresponding exam score.

<table>
<thead>
<tr>
<th>$X$ (hours)</th>
<th>3</th>
<th>2</th>
<th>7</th>
<th>4</th>
<th>4</th>
<th>6</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$ (score)</td>
<td>55</td>
<td>40</td>
<td>80</td>
<td>60</td>
<td>66</td>
<td>77</td>
<td>70</td>
</tr>
</tbody>
</table>

A scatterplot is given below.

![Scatterplot of score vs hours](image)

**Model I.**

Using the model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ with $\epsilon_i$ iid $N(0, \sigma^2)$, we obtain the estimated regression function

$$\hat{Y} = 29.750 + 7.7339X$$

with $R^2 = .931$ and a statistically significant linear regression relationship between $Y$ and $X$. 
However, the scatterplot indicates some curvature, and the standardized residual versus fitted plot below substantiates the evidence of curvature.

The original scatterplot and the residual plot suggest a regression function of the type

$$E(Y) = \beta_0 + \beta_1 \sqrt{X}$$

or

$$E(Y) = \beta_0 + \beta_1 \log X.$$

**Model II.**

Using the later type, our data set becomes

<table>
<thead>
<tr>
<th>$X'$ (ln hours)</th>
<th>1.09861</th>
<th>0.69315</th>
<th>1.94591</th>
<th>1.38629</th>
<th>1.38629</th>
<th>1.79176</th>
<th>1.60944</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$ (score)</td>
<td>55</td>
<td>40</td>
<td>80</td>
<td>60</td>
<td>66</td>
<td>77</td>
<td>70</td>
</tr>
</tbody>
</table>

Regressing $Y$ on $X' = \ln(X)$, we obtain

$$\hat{Y} = 18.577 + 32.080 X'$$

with $R^2 = .981$ and a statistically significant linear regression relationship between $Y$ and $X'$. A plot of standardized residuals versus fitted values is given below.
Note that the previous curvilear pattern in the residual plot is less apparent.

Suppose we wanted to estimate the expected score for all students who study $X = 4$ hours.

For model I, the 95% confidence interval for $E(Y|X = 4)$ is given by $(56.70, 64.67)$.

For model II, the 95% confidence interval for $E(Y|X' = \ln(4))$ is given by $(61.006, 65.093)$.

Note the much narrower confidence interval for model II. A similar reduction in interval width occurs for prediction intervals. For simple linear regression this fact is indicated by the increase in $R^2$ for the second model since a higher $R^2$ implies a lower error mean square (MSE). Recall that $\sqrt{\text{MSE}}$ is a factor in the margin of error for interval estimates.
Assessing the normality assumption for the random errors, we obtain the normal scores for the residuals and plot them against the residuals.

The correlation coefficient for ordered residuals and their normal scores is .986. When testing

\[ H_0: \text{the random errors are normally distributed} \]

vs.

\[ H_a: \text{the random errors are not normally distributed} \]

using \( \alpha = .05 \), we reject \( H_o \) if the correlation coefficient is smaller than .898 (see Table in appendix of Kutner et al.). Therefore, we do not have sufficient evidence to reject the assumption of normality for the random errors.