THE USE OF A LINEAR SPLINE MODEL IN THE ANALYSIS
OF A REPEATED MEASURES EXPERIMENT THROUGH SAS

Rocco L. Brunelle and David W. Johnson
Lilly Research Laboratories

Many techniques can be used in analysing data from experiments that have repeated measures over time. One such technique is to reduce the vector of observations into a meaningful subset of responses that can be analyzed in a straightforward manner. Sanders (1) and Johnson and Brunelle (2) reduce the vector of observations through modeling techniques, and then use the estimates of the model parameters to examine treatment effects. Johnson and Brunelle further describe contrasts of clinical interest between elements in the vector of observations.

This paper presents another method for reducing the vector of observations into a subset of responses. Here, the responses over time are fit with a linear spline model with one knot, or change point. This technique can be used to model the response, and the resulting model parameters can be employed in further analyses of treatment effects.

Suppose that the vector of observed values follows a decreasing (or increasing) pattern over time and then a change in the pattern occurs. An example of this is shown in Figure 1. A linear spline model with one change point provides an excellent fit for this type of data.

Assume the $X_i$'s are ordered

$X_1 < X_2 < \ldots < X_n$,

then the linear spline model is

\begin{align*}
  y_i &= a_1 + b_1 X_i + e_i & \text{for } X_i < T, \ i = 1, \ldots, m \\
  y_i &= a_2 + b_2 (X_i - T) + e_i & \text{for } X_i > T, \ i = m+1, \ldots, n
\end{align*}

under the constraint that

$a_1 + b_1 T = a_2 + b_2 T$

or that $a_2 = a_1 + T(b_2 - b_1)$

where $y_i = \text{response at visit } X_i$

$a_1, a_2 = \text{intercepts of the two lines}$

$b_1, b_2 = \text{slopes of the two lines}$

$X_i = \text{visit number for the } i\text{th visit}$

$e_i = \text{random error}$

$T = \text{the knot or break point which is fixed}.$

Model (1) can be rewritten as

\begin{align*}
  y_i &= a_1 + b_1 X_i + e_i & \text{for } X_i < T, \ i = 1, \ldots, m \\
  y_i &= a_2 + b_2 X_i + e_i & \text{for } X_i > T, \ i = m+1, \ldots, n
\end{align*}

Thus, there are two intersecting straight lines which meet at $T$.

EXAMPLE DATA

\begin{figure}[h]
  \centering
  \includegraphics[width=\textwidth]{example_data.png}
  \caption{Example Data}
\end{figure}

The above model is similar to that of Ertel and Fowlkes (3) with one change point. The form of the equations for Ertel and Fowlkes' model is somewhat different than equations (2). Ertel and Fowlkes are primarily interested in estimating predicted values at the knots; whereas, the objective of this analysis is to estimate the slopes of the line segments.

In choosing $T$, Ertel and Fowlkes method of selecting $T$'s halfway between adjacent values of $X$ is used. That is,

$$T = \frac{X_m + X_{m+1}}{2}$$

for $m = 3, 4, \ldots, n-3$.

An additional restriction is made requiring at least three observations in each line segment giving $n-5$ possible values for $T$. This restriction is used in order to reduce the possibility of a line segment being forced to fit one spurious observation.

For each $T$, model (2) is fit to the vector of observed values and the SSE is calculated, where $\text{SSE} = \sum (y_i - \hat{y}_i)^2$. The $T$ with the smallest SSE is then chosen as the "best" fit. Figure 2 presents the best linear spline fit with $T = 9.5$ from the example data in Figure 1, and Figure 3 presents the plot of the SSE by $T$ for each of the $n-5$ fits.

To evaluate this technique, the performance of two potentially useful models, a piecewise linear regression model and a quadratic model, is investigated. The piecewise linear model is similar to model (1) without the restriction that the two line segments meet at $T$. Thus, this model has 5 parameters (2 intercepts, 2 slopes, and $T$); whereas, the linear spline model has 4 parameters (1 intercept, 2 slopes, and $T$). The parameters in the piecewise linear
The quadratic model used has the following form:

\[ y_i = b_0 + b_1(x_i - \bar{x}) + b_2(x_i - \bar{x})^2 + e_i \]

for \( i = 1, \ldots, n \)

where:
- \( y_i \) = response at visit \( x_i \)
- \( b_0 \) = intercept
- \( b_1 \) = linear component
- \( b_2 \) = quadratic component
- \( x_i \) = visit number for the \( i \)th visit
- \( e_i \) = random error.

The above three parameters are estimated in the usual fashion.

Figure 4 presents the fits for the piecewise linear and quadratic models for the example data. Both of these models seem to adequately fit the observed responses. However, the piecewise linear model has a disadvantage of not being continuous at \( T \). In this example, the discontinuity is not a serious problem, but one can foresee the possibility of the two line segments having a large disparity close to \( T \). For the quadratic model, the estimated parameters may be difficult to interpret. Centering the data at \( \bar{x} \) causes the linear component to estimate the line tangent to the model at \( \bar{x} \). Thus, a large initial decrease may not be reflected in the estimate of the linear component. Also the quadratic component may be difficult to interpret. A large positive value could represent an initial decrease and then a subsequent return to baseline. Further, care must be exercised in extrapolation beyond the actual data when the quadratic component is large.
Assuming that the underlying trends in the responses are continuous, the linear spline model seems to give a straightforward interpretation of the trends and does not share the problems of the piecewise regression model or the quadratic model.

To examine the performance of these models in practice, data obtained in a clinical trial comparing two antiarthritis treatments are used. This trial is a parallel study with 10 patients receiving drug A and 10 patients receiving drug B. One of the parameters measured in this study is morning stiffness severity – an index from 0 (None) to 20 (Terrible). Figures 5 and 6 present plots of this response, over the course of the study period, obtained from two patients in which the above three models give conflicting results.

The interpretations of the trends among the three models are quite different for both of these patients. Due to the discontinuity at T, the piecewise linear model may be misrepresenting the trends over time. This model tends to place too much emphasis on potential outliers. Overall, the quadratic model seems to fit well; however, the resulting parameters may be difficult to interpret and this model does not put much emphasis on the initial drop in the responses for both patients. Again the linear spline model fits well and the resulting parameters are easily understood.
Having fit the three models to all the individual patients, the resulting parameter estimates can be used to evaluate the treatments. Table 1 presents the descriptive statistics for selected parameters from each of the three models. Also, the mean plot of the actual severity scores by visit for both treatments is presented in Figure 7.

**MEAN PLOTS BY DRUG**

![Mean plots by drug](image)

**FIGURE 7**

Both the linear spline and piecewise linear models give similar mean estimates of the break point $T$ and of the two slopes for both drugs. An initial decrease is observed for both drugs which seems more pronounced for drug B than drug A (slope 1). However, a slight decreasing trend is still evident during the remainder of the study for drug A but is not evident for drug B (slope 2). The results from the quadratic model seem to indicate an overall decreasing trend with some curvature. Both drugs seem to be similar in the overall decreasing trend and drug B seems to have slightly more curvature than is evident for drug A.

In summary, for the example data above, the linear spline model seems to give an intuitively better fit to the data than either the piecewise linear model or the quadratic model. Also, the resulting parameters from the linear spline model are easily understood by the experimenters; whereas, interpretations of the resulting parameters may be difficult for the piecewise linear model and the quadratic model.

**COMPUTATIONS:** In the appendix is a listing of the SAS statements used to generate the linear spline parameters. This routine can handle missing values and outputs the parameter estimates into a data set for further analysis.

**TABLE 1**

<table>
<thead>
<tr>
<th>Model</th>
<th>Drug</th>
<th>Theta</th>
<th>Slope 1</th>
<th>Slope 2</th>
<th>SSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Spline</td>
<td>A</td>
<td>Mean</td>
<td>-1.53</td>
<td>-0.11</td>
<td>32.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S.D.</td>
<td>(1.36)</td>
<td>(1.06)</td>
<td>(30.93)</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>Mean</td>
<td>-2.57</td>
<td>0.37</td>
<td>41.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S.D.</td>
<td>(2.19)</td>
<td>(0.77)</td>
<td>(35.91)</td>
</tr>
<tr>
<td>Piecewise Linear</td>
<td>A</td>
<td>Mean</td>
<td>-1.85</td>
<td>-0.52</td>
<td>23.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S.D.</td>
<td>(1.55)</td>
<td>(1.67)</td>
<td>(24.51)</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>Mean</td>
<td>-2.34</td>
<td>-0.18</td>
<td>25.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S.D.</td>
<td>(2.64)</td>
<td>(2.12)</td>
<td>(24.98)</td>
</tr>
<tr>
<td>Quadratic*</td>
<td>A</td>
<td>Mean</td>
<td>--</td>
<td>0.05</td>
<td>39.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S.D.</td>
<td>--</td>
<td>(0.09)</td>
<td>(31.78)</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>Mean</td>
<td>--</td>
<td>0.14</td>
<td>54.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S.D.</td>
<td>--</td>
<td>(0.10)</td>
<td>(36.78)</td>
</tr>
</tbody>
</table>

*Slope 1 = linear component, Slope 2 = quadratic component*
PROC GLM can also be used to obtain the model parameters for all three models. For the linear spline and piecewise linear models, first transform the data for each $T$ by $Z = X - T$ and then create two groups: Group 1 for $Z < 0$ and Group 2 for $Z > 0$. For the two models the SAS statements are:

**Linear Spline Model** (forces the two lines to meet)

```sas
PROC GLM; CLASSES GROUP;
MODEL Y = X(GROUP)/SOLUTION;
```

**Piecewise Linear Model** (fits two separate lines)

```sas
PROC GLM; CLASSES GROUP;
MODEL Y = GROUP X(GROUP)/SOLUTION;
```

The above transformations and the models must then be rerun for each possible value of $T$.

If there are a large number of response vectors to analyze and/or a large number of possible change points $T$, then using PROC GLM may be inefficient. Also, it is difficult to output the resulting parameters into a data set for further analysis using this method.

PROC NLIN can also be used to fit the linear spline model using segmented regression. Here, the break point $T$ becomes one of the estimated parameters. If only a few models have to be fit, using PROC NLIN would be preferred. However, with many patients and multiple response variables to be modeled, this method may be inefficient. Also, all the potential problems of non-linear regression (non-convergence, poor starting values, excessive output, etc.) exist each time the procedure is run.

**REFERENCES**


