Transformations to Additivity in Measurement Error Models

R. Stephen Eckert,1 Raymond J. Carroll,2 and Naisyin Wang2

1Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, Indiana 46285, U.S.A.
2Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.

Summary

In many problems, one wants to model the relationship between a response $Y$ and a covariate $X$. Sometimes it is difficult, expensive, or even impossible to observe $X$ directly, but one can instead observe a substitute variable $W$ that is easier to obtain. By far, the most common model for the relationship between the actual covariate of interest $X$ and the substitute $W$ is $W = X + U$, where the variable $U$ represents measurement error. This assumption of additive measurement error may be unreasonable for certain data sets. We propose a new model, namely $h(W) = h(X) + U$, where $h(\cdot)$ is a monotone transformation function selected from some family $\mathcal{H}$ of monotone functions. The idea of the new model is that, in the correct scale, measurement error is additive. We propose two possible transformation families $\mathcal{H}$. One is based on selecting a transformation that makes the within-sample mean and standard deviation of replicated $W$'s uncorrelated. The second is based on selecting the transformation so that the errors $(U)$'s fit a prespecified distribution. Transformation families used are the parametric power transformations and a cubic spline family. Several data examples are presented to illustrate the methods.

1. Introduction

Measurement error models concern the situation where one or more variables in a study cannot be measured exactly. We restrict our attention to the case where a single variable is measured with error. It is usually assumed that the relationship between the variable that is actually observed, $W$, and the true covariate of interest, $X$, is $W = X + U$, where $U$ represents measurement error. Fuller (1987) applied this additive model for measurement error to many classical linear models.

There are also other ways to model the relationship between $W$ and $X$, such as the multiplicative error model $W = X e^U$, which gives additivity in the logarithmic scale, i.e., $\log(W) = \log(X) + U$.

The idea behind both the additive and multiplicative error structure models is that, in the correct scale, measurement error is additive. The additive and the multiplicative error models are specific cases of a more general model $W = \mathcal{G}(X, U)$ for some function $\mathcal{G}$. In this article we consider the set of functions $\mathcal{G}$ such that $\mathcal{G}(X, U) = \mathcal{H}^{-1}(\mathcal{H}(X) + U)$, where $\mathcal{H}$ is a monotone function with inverse $\mathcal{H}^{-1}$.

Additivity underlies almost all the measurement error models and modeling techniques in the common case that $X$ is unobservable. The classical functional methods for ordinary regression (Fuller, 1987) and for general nonlinear models (Carroll, Ruppert, and Stefanski, 1995) essentially without exception assume additivity. Likelihood (structural) methods that naturally allow for the commonly occurring within-person replication of the $W$'s typically assume additivity in some scale with a known distribution for $U$.

For all of these reasons, finding a scale for additive measurement error is important. In this paper we investigate methods for determining an appropriate scale. Section 2 discusses two different

Key words: Errors-in-variables; Nonlinear models; Power transformations; Regression calibration; SIMEX; Spline transformations; Transform-both-sides.
methods for determining the correct scale for additivity of measurement error, the correlation method and the error distribution method. In Section 3 we describe the transformations used, and in Section 4 we describe their implementation. In Section 5 we present data examples to illustrate the methods.

2. Functional Transformations

In measurement error models, the literature makes a distinction between classical functional models, in which the values of unobserved true values of $X_i$, $i = 1, \ldots, n$, are considered to be a sequence of unknown fixed constants, and classical structural models, in which the values of $X$ are considered to be random variables. We believe that a more fruitful classification scheme is that of functional modeling, where no assumptions are made about the distribution of the $X_i$’s, and structural modeling, in which parametric assumptions are made about the distribution of the unknown $X$’s. For a full description of functional versus structural modeling, see Carroll et al. (1995, pp. 144–145).

Additive error models assume that there is a monotone function $h(\cdot)$ such that

$$h(W) = h(X) + U,$$  \hspace{1cm} (1)

where the random variable $U$ is independent of $X$. There is an essential difference between our work and that typical in transformations, namely that in our case $X$ cannot be observed, so that without any additional information, $h(\cdot)$ cannot be identified. In practice, this extra information comes from replicating the $W$’s, so that $(W_{ij})$ is observed for $i = 1, \ldots, n$ units and $j = 1, \ldots, J$ replicates per unit. The resulting errors $(U_{ij})$ are assumed to be independent of $X_i$, although they may be correlated either given $i$ or marginally.

The issue we address in this paper is estimating the transformation function $h(\cdot)$. We propose two different methods, both of which are truly functional modeling methods in that they make no assumptions about the distribution of $X$, so that the methods are robust to the distribution of the predictor.

There are two general methods we propose, correlation methods and error distribution methods. These two methods are derived from the properties of the transformation model (1), as follows.

**Property 1.** Define the within-person mean $\bar{W}_i(h)$ and the within-person standard deviation $s_i(h)$ as

$$\bar{W}_i(h) = J^{-1} \sum_{j=1}^{J} h(W_{ij}); \quad s_i(h) = \left[ (J-1)^{-1} \sum_{j=1}^{J} \left( h(W_{ij}) - \bar{W}_i(h) \right)^2 \right]^{1/2},$$

respectively. Under model (1), if the errors are symmetrically distributed, then $\bar{W}_i(h)$ and $s_i(h)$ are uncorrelated. Thus, the correlation method selects the transformation $h(\cdot)$ so that the sample correlation for $\bar{W}_i(h)$ and $s_i(h)$ equals zero. Ruppert and Aldershof (1989), Box, Hunter, and Hunter (1978), and Solomon and Cox (1992) each mention correlation-type methods in different contexts.

**Property 2.** For the correct transformation, the distribution of

$$h(W_1) - h(W_2) = U_1 - U_2$$  \hspace{1cm} (2)

does not depend on $X$. In particular, if the $U$’s are multivariate normal, so too is (2). If $(U_1, \ldots, U_J)$ have a multivariate $t$-distribution with $k$ degrees of freedom, then (2) is a multiple of a $t$-distribution with $k$ degrees of freedom. If the $U$’s are independent with a mixture of two mean zero normal distributions, (2) has a symmetric mixture of three normal distributions. These ideas suggest that the second class of methods, the error distribution methods, transform so that the terms in (2) follow one of the distributions mentioned. We address distributional shape via the Anderson–Darling (Anderson and Darling, 1954) and Filliben correlation (Filliben, 1975) statistics. There are situations when a distributional model for the measurement error is desirable or even essential. Carroll et al. (1995) describe several techniques that require that the measurement error distribution be specified, including simulation–extrapolation, corrected scores, conditional scores, and likelihood techniques.

3. Two Families of Transformations

3.1 Definition of the Power Transformation

The power transformation family was described in Box and Cox (1964). The transformations in this family are indexed by the scalar parameter $\theta$ and have the form...
\[ h(u \mid \Theta, \xi) = \begin{cases} (u^\theta - 1)/\theta, & \theta 
eq 0 \\ \log(u), & \theta = 0. \end{cases} \]

Power transformations are monotone for fixed \( \theta \). We have found that the restricted shape of power transformations limits their utility somewhat in our context, and for that reason we describe below an alternative family.

3.2 Definition of the Spline Transform

The transformation family \( \mathcal{H} \) that we consider is the set of all zero-intercept, cubic piecewise-polynomial spline functions with knots at \( \xi = (\xi_1, \ldots, \xi_p) \). Transformations from this family have the form

\[ h(u \mid \Theta, \xi) = \theta_1 u + \theta_2 u^2 + \theta_3 u^3 + \sum_{k=1}^p \theta_{k+3} (u - \xi_k)^3_+ , \tag{3} \]

where \( (a)^3_+ = a^3 \) if \( a > 0 \) and is 0 otherwise. Note that the transformation given in (3) does not have the usual constant term, which is not identifiable. The parameter vector \( \Theta = (\theta_1, \ldots, \theta_{p+3}) \).

In general, the problem of picking knot points \( \xi \) is a difficult one (Eubank, 1988). One approach is to fix the knots as uniformly distributed among the percentiles of the data. We modified this approach slightly to generate extra knots in the tails, as follows.

1. Let the combined data vector be \( \tilde{W}^* \), including all replicates across all units.
2. Let \( \tilde{\xi}_1 = \tilde{W}^*(0.005) \), where \( W^*(p) \) is the pth sample percentile of \( \tilde{W}^* \).
3. Let \( \tilde{\xi}_2 = W^*(0.01), \tilde{\xi}_{p-1} = W^*(0.99) \), and \( \tilde{\xi}_p = W^*(0.995) \).
4. Let \( \tilde{\xi}_i = W^*(0.01 + (p - 3)^{-1}(i - 1)(0.99 - 0.01)), \) \( i = 3, \ldots, p - 2 \).

The issue of obtaining a monotone transformation function is difficult. For a set of data and knot points \( \xi \), describing the set \( \{ \Theta : h(u + \delta \mid \Theta, \xi) - h(u \mid \Theta, \xi) \geq 0, \forall \delta > 0 \} \) is a nontrivial analytical problem. Certainly, requiring that \( \Theta \geq 0 \), \( i = 1, \ldots, p + 3 \), is sufficient to obtain a monotone transformation, but this is clearly overly restrictive.

Our solution to the problem of obtaining a monotone transform is to create a set of \( M + 1 \) grid points \( \{ \tau_0 < \tau_1 < \ldots < \tau_M \} \) on which we require that

\[ h(\tau_i \mid \Theta, \xi) - h(\tau_{i-1} \mid \Theta, \xi) \geq 0, \quad i = 1, \ldots, M. \tag{4} \]

This allows us to use existing programs for constrained optimization (see Section 4.1 for details). We usually set \( M = 22 \); let \( \tau_0 \) through \( \tau_9 \) be the 1st, 2nd, 3rd, ..., and 9th percentiles of the combined data vector \( \tilde{W}^* \); let \( \tau_{10} \) through \( \tau_{32} \) be the 91st, 92nd, 93rd, ..., and 99th percentiles of \( \tilde{W}^* \); and let \( \tau_{33} \) through \( \tau_{72} \) be set at evenly spaced percentiles of \( \tilde{W}^* \) between the 9th and the 91st, similar to step 4 in the procedure describing the selection of knot points. We have found that, through careful selection of these grid points, the final transformation will be monotone through the range of the data.

4. Correlations and Distributions

We now address implementation of the correlation and distribution methods. To shorten the discussion we focus on the spline transformation family. All the techniques are easily applied to the power transformation family (we restrict the power parameter to \( -3 < \theta < 3 \)).

Using Property 1 of Section 2, the correlation methods find the transformation that makes the within-person mean and standard deviation have zero sample correlation. We have found unique zeros in our applications, although in principle this need not be the case.

4.1 Assessing the Need for a Transformation Using Correlation and Powers

An important question to answer is whether the data indicate that a transformation would be appropriate. One way to answer this question is to use the power family and create a confidence interval for \( \theta_0 \), the value at which the population correlation between the within-person mean and standard deviation equals zero. One rejects the "no transform" null hypothesis if the interval does not contain 1. We constructed an asymptotic confidence interval for \( \theta_0 \) using the delta method and sandwich covariance estimate via standard techniques, but while the confidence level for this interval is correct asymptotically, the convergence to the nominal level is quite slow. We also considered confidence intervals using resampling techniques described in Efron and Tibshirani (1993). Figure 1 shows the results of a simulation study comparing the asymptotic confidence interval to a confidence interval created using the bootstrap estimate of standard error. We studied
other resampling-based confidence intervals with the same result: the delta-method confidence interval with sandwich covariance matrix estimate converges to its nominal level much more slowly than any of the bootstrap methods.

4.2 Assessing Distributional Shape

We consider the spline transformation to normality when there are exactly two replicates. The overall goal is to find a vector \( \hat{\theta} \) that, for a given data set, makes the differences \( E_i = h(W_{i2} \mid \hat{\theta}, \xi_j) - h(W_{i2} \mid \hat{\theta}, \xi_1) \) look as "normal" as possible, while satisfying the constraints given in (4). Actually, one need not specify a normal distribution for the measurement error. We investigate both bivariate t-distributions and normal mixture distributions later in this article. There are several ways to check for normality of the differences \( E_i \) for a given value of \( \theta \). We have chosen to use the probability plot correlation coefficient (PPC) described in Filliben (1975), which is closely related to the Shapiro-Wilk W statistic described in Shapiro and Wilk (1965). The basic idea is to calculate the correlation coefficient for a QQ-plot of the \( E_i \). The closer the empirical distribution of the \( E_i \) is to a normal distribution, the closer the PPC for the \( E_i \) should be to 1. Hence, our method of estimating \( \theta \) is to find the value \( \hat{\theta} \) that, subject to the constraints in (4), maximizes \( \gamma(E) \), where \( E = (E_1, \ldots, E_n)^T \) and \( \gamma(v) \) is the PPC for the vector \( v = (v_1, \ldots, v_n)^T \).

In this maximization problem, both the constraints and the objective function have simple matrix expressions. Given the data \( \{W_{ij}\} \), \( i = 1, \ldots, n \), \( j = 1, 2 \), and a set of knot points \( \xi_i \), define the matrices \( D \) and \( C \) as

\[
D_{ik} = \begin{cases} 
W_{i1}^k - W_{i2}^k & i = 1, \ldots, n; k = 1, \ldots, 3 \\
(W_{i1} - \xi_{k-3})^3_{+} - (W_{i2} - \xi_{k-3})^3_{+} & i = 1, \ldots, n; k = 4, \ldots, p + 3 
\end{cases}
\]

\[
C_{ik} = \begin{cases} 
\tau_i^k - \tau_i^{k-1} & i = 1, \ldots, M; k = 1, \ldots, 3 \\
(\tau_i - \xi_{k-3})^3_{+} - (\tau_{i-1} - \xi_{k-3})^3_{+} & i = 1, \ldots, M; k = 4, \ldots, p + 3.
\end{cases}
\]

Thus, the maximization problem is to find \( \max_{\theta} \gamma(D\theta) \) subject to \( C\Theta \geq 0 \), where by \( C\Theta \geq 0 \) we mean that each element of \( C\Theta \) is nonnegative. The constrained maximization is accomplished using the FORTRAN program NPSOL (Gill et al., 1986).

![Coverage Probabilities for Theta = 0.5](image)

**Figure 1.** Coverage probabilities for both the asymptotic confidence interval and the confidence interval that uses the bootstrap standard error.
In modeling data such as the examples we discuss in Section 5, it is possible that the error distribution may be something other than normal. We consider alternate distributions for the measurement error, specifically the bivariate $t_k$-distributions (Johnson and Kotz, 1972) for $k = 20, 10, 8, 6, 4, 3$, and find separate transformations for each possible error distribution. Note that the bivariate $t_k$-distributions are such that if $(U_1, U_2) \sim \text{bivariate } t_k$, then $(U_1 - U_2)/\sqrt{2} \sim \text{univariate } t_k$. The modification to the PPC statistic is simple—one calculates the correlation coefficient for the QQ-plot of the specified distribution instead of the normal distribution. As an additional check, for each transformation we calculate the Anderson–Darling A statistic (AD) for the vector of differences $E$ (Anderson and Darling, 1954).

We found with most of our examples that the spine transformation based on the error method transforms the data such that the error distribution is either normal or “nearly normal,” i.e., a bivariate $t$-distribution with either 20 or 10 degrees of freedom, with the nonnormality being attributable to a small number of points in the difference vector $E$. Another reasonable way to model the data is to assume that the measurement error is distributed as a two-component normal mixture distribution, with the measurement error for a (relatively small) number of data pairs being generated by a normal distribution with slightly heavier tails. We selected four normal mixture distributions, each chosen to have the same first four moments as a univariate $t_k$-distribution, for $k = 20, 10, 8, 6$, and 4, respectively. We use the shorthand $\text{NM}(k)$ to refer to such a normal mixture distribution. More information about the $\text{NM}(k)$-distributions is given in the Appendix.

4.3 The Spine Transform with More Than Two Measurements Per Individual

Unlike with correlations, the error distribution methods that model the distribution of the differences given in (2) do not have an easy direct definition for the case of $J > 2$. There are a variety of possibilities, including transformations so that the within-person sample standard deviation has the distribution of a sample standard deviation of a candidate error model, in which case the results of the previous subsection apply. Alternatively, one may wish to analyze the data pairwise, as this can often point out unusual replicates. Here we describe such a pairwise implementation.

In order to select the optimal $\Theta$ value, we must first determine the appropriate distribution for $U$, and then optimize with respect to that measurement error distribution. We select the distribution for $U$ by some preliminary analyses on two columns of data. If the data are measurements on the same individual taken over time, then it makes some sense to use the two columns of data for which the measurements are earliest apart chronologically.

We implement the preliminary analyses in two stages. In the first stage we select two columns of data and find separate estimates $\hat{\Theta}_k$, $k = \infty, 3, 4, 6, 8, 10, 20$, where $\hat{\Theta}_\infty$ is the value of $\Theta$ that maximizes $\gamma(E)$ for normally distributed measurement error and $\hat{\Theta}_k$, $k > 0$, is the value of $\Theta$ that maximizes $\gamma(E)$ for measurement error with a bivariate $t_k$-distribution. For each value of $\hat{\Theta}_k$, we examine the PPC and AD statistics for the difference vector $E$, and for $\hat{\Theta}_\infty$ we also examine the PPC and AD statistics for the $\text{NM}(k)$-distribution for $k = 6, 8, 10, 20$. We also calculate the intra-individual mean/standard deviation correlation for the two selected data columns for each $\hat{\Theta}_k$.

Using the calculations in the first stage as a guide, we then select an appropriate $\hat{\Theta}_k$, say $\hat{\Theta}_*$, for additional analysis in the second stage. In this stage, we apply the transformation $h(v | \hat{\Theta}_*, \xi)$ to every data value $W_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, J$, and then do an analysis of the differences of the transformation for each possible pair of columns. In this difference analysis, we calculate PPC and AD statistics and their $p$-values for the normal distribution, bivariate $t_k$-distributions with $k = 3, 4, 6, 8, 10, 20$, and $\text{NM}(k)$ for $k = 6, 8, 10, 20$. We also calculate the intra-individual mean/standard deviation correlation for each pair of columns of transformed data.

By combining the two stages of analysis, we can select an appropriate distribution for the measurement error $U$. We can then define quantities as follows to find an optimal $\hat{\Theta}$ value. Specifically, if $\gamma((u_1, \ldots, u_n)^T)$ is the PPC function for the specified error distribution, let

$$A_{km}(\Theta) = h(W_{jk} | \Theta, \xi) - h(W_{jm} | \Theta, \xi)$$

$$Q_{km}(\Theta) = \gamma\left(\{A_{1km}(\Theta), \ldots, A_{nkm}(\Theta)\}^T\right)$$

$$\overline{Q}(\Theta) = \frac{2}{J(J - 1)} \sum_{k=1}^{J-1} \sum_{m=k+1}^{J} Q_{km}(\Theta)$$

$$\hat{Q}(\Theta) = \text{median}(Q_{km}(\Theta)), \quad 1 \leq k \leq J - 1, \quad k < m \leq J$$
We can then consider both $\hat{\Theta}$, the numerical maximizer of $Q(\Theta)$, and $\tilde{\Theta}$, the numerical maximizer of $\tilde{Q}(\Theta)$.

5. Examples

5.1 Urinary Sodium Chloride Data

The urinary sodium chloride data (USC) are discussed in Liu and Liang (1992). In a study attempting to relate the incidence of hypertension with urinary sodium, overnight urine samples were taken from 397 men on 7 consecutive nights. The data from days 1–6 were available to us. Because the data have a very high autocorrelation, we examined the data from days 1 and 6, which have the least correlation in the errors and hence presumably the most stable statistical properties.

The estimated power transform from the correlation method was $\hat{\theta}_0 = 2.304$, with bootstrap confidence interval $[1.520, 2.688]$, thus indicating the need for a transformation. We tested the differences of the power transformed data for normality, and found a PPC p-value of 0.967 and an AD statistic p-value of 0.801. In both cases the null hypothesis is that the difference vector $E$ has a normal distribution, with low p-values indicating nonnormality. Hereafter, we shall say that a data vector "passes" a given test (either PPC or AD) for a certain distribution if the p-value for the calculated statistic is greater than 0.10. Thus, the difference vector $E$ from the power transformation "passes" both the PPC and the AD tests for normality.

Table 1 shows the results of the error distribution method using cubic splines for estimating the transform. Each row in the table gives the transformation, the criterion for optimization, the within-person mean and standard deviation sample correlation, the distribution under which the PPC and AD statistics are computed, and their corresponding p-values. One can see that the differences from either spline transformation clearly pass the PPC and AD tests for normality, with acceptably low within-person sample mean versus standard deviation correlation.

Figure 2 compares the correlation method power transformation and the error distribution method spline transformation. The circles in the graph represent percentiles of the data, from the 1st to the 99th. Each transformation has been standardized to the same scale. For this data set, the power transformation and the spline transformation were almost identical.

We repeated the analysis using all pairs of days and all 6 days together, and with one exception the answers were similar. The exception occurs for the pair of days (5, 6), which seem to behave together quite differently from all the others. We have no explanation for this behavior.

5.2 Framingham Heart Study

The Framingham heart study measured various factors such as age, smoking habits, and blood pressure for 1,615 men aged 31–65, attempting to link these factors to the presence of coronary heart disease. The data we analyze here are two systolic blood pressure (SBP) measurements, the first of which is the average of two SBP measurements taken during a physical exam, and the second of which is the average of two SBP measurements taken at another physical exam 2 years later. We actually pretransform the data by analyzing $\log(\text{SBP} - 50)$, which is a modification of the transformation originally suggested by Cornfield (1962) and which we will designate as LSBP.

For the pretransformed LSBP variable, using the correlation method with power transformation, we found $\hat{\theta}_0 = 1.726$ with 90% bootstrap confidence interval $[1.113, 2.339]$ and 95% confidence interval $[0.996, 2.459]$.

For the error distribution method using the spline transformation, Table 2 shows the usual statistics for the transformations of the LSBP data to additivity with various error distributions. There are a number of points to note. The power transformation using the correlation method results in differences that are nonnormal and do not pass tests for the t-distribution with 10 degrees of freedom. The spline transformation using the correlation method does pass the $t_{10}$ and

<table>
<thead>
<tr>
<th>Transform</th>
<th>Optimization criterion</th>
<th>Mean/ std. correlation</th>
<th>Error dist. comparison</th>
<th>PPC p-value</th>
<th>AD p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power ($\hat{\theta} = 2.304$)</td>
<td>Correlation</td>
<td>0.00</td>
<td>Normal</td>
<td>0.967</td>
<td>0.801</td>
</tr>
<tr>
<td>Spline</td>
<td>Correlation</td>
<td>0.00</td>
<td>Normal</td>
<td>0.963</td>
<td>0.788</td>
</tr>
<tr>
<td></td>
<td>PPC(normal)</td>
<td>0.01</td>
<td>Normal</td>
<td>0.968</td>
<td>0.808</td>
</tr>
</tbody>
</table>

Table 1 Comparing transformations to different error distributions for the USC data. The spline transformations used eight knot points.
Figure 2. Graph of transformations for urinary sodium chloride (USC) data. The dashed vertical lines show the locations of the knot points.

NM(10) distribution tests. The spline transformation that attempts to fit a normal distribution to the differences is unsuccessful in doing so, at least with this number of knots. All of these calculations suggest that the errors are heavier-tailed than the normal distribution. The spline transformation under the error distribution method for the NM(10) distribution is shown in Figure 3.

5.3 CSFII Data

Our third example involves the Continuing Survey of Food Intakes for Individuals (CSFII) study data set (Thompson et al., 1992). This data set contains information on nutrient intakes for 2,134 women. The data contain multiple measurements for each woman for a variety of daily dietary components such as vitamin A, vitamin C, amount of saturated fat, total calories, etc. Four measurements for each component were gathered for each woman. The first measurement was based on an extensive interview, and the subsequent three measurements were based on follow-up telephone interviews.

We analyze one dietary component from the CSFII data, percent calories from fat, by considering the second and fourth measurements for each woman in the study. We choose not to use the

<table>
<thead>
<tr>
<th>Transform</th>
<th>Optimization criterion</th>
<th>Mean/std. correlation</th>
<th>Error dist. comparison</th>
<th>PPC p-value</th>
<th>AD p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power (( \theta = 1.726 ))</td>
<td>Correlation</td>
<td>0.00</td>
<td>Normal</td>
<td>&lt;0.005</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( t_{10} )</td>
<td>0.071</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NM(10)</td>
<td>0.088</td>
<td>0.384</td>
</tr>
<tr>
<td>Spline</td>
<td>Correlation</td>
<td>0.00</td>
<td>Normal</td>
<td>&lt;0.005</td>
<td>0.149</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NM(10)</td>
<td>0.165</td>
<td>0.211</td>
</tr>
<tr>
<td></td>
<td>PPC(normal)</td>
<td>-0.085</td>
<td>Normal</td>
<td>&lt;0.005</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td></td>
<td>PPC(( t_{10} ))</td>
<td>-0.112</td>
<td>( t_{20} )</td>
<td>0.979</td>
<td>0.461</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NM(10)</td>
<td>0.952</td>
<td>0.354</td>
</tr>
</tbody>
</table>
first measurement because it was gathered in a different manner than the last three. The power transformation using the correlation method yields an estimate $\delta_0 = 1.056$, with bootstrap confidence interval $[0.948, 1.164]$. However, as shown in Table 3, the differences of the no-transform model fail both tests for normality. The no-transform differences do pass both PPC and AD tests for $t_{10}$ and $\text{NM}(10)$ distributions.

The spline transformations with 5 knots pass the normality tests with an acceptably low within-person sample mean and standard deviation correlation of 0.055 for the transformed data values. The graph of the spline transformation using the error distribution method is given in Figure 4.

### 6. Discussion and Concluding Remarks

We have presented two methods for transforming the data to achieve additive measurement error. The correlation method transforms so that the sample correlation between the within-person mean and standard deviation equals zero, while the error distribution method transforms so that differences have a specified distribution. Within each method we used power transformations and transformations based on cubic splines. A question that may arise is, “why not just transform the data to normality?” Such a method has been suggested by Nusser et al. (1997), who also use power transformations and cubic splines. This method, which we call the marginal method, selects $h(\cdot)$

### Table 3

<table>
<thead>
<tr>
<th>Transform</th>
<th>Optimization criterion</th>
<th>Mean/ std. correlation</th>
<th>Error dist. comparison</th>
<th>PPC p-value</th>
<th>AD p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>N/A</td>
<td>-0.028</td>
<td>Normal</td>
<td>&lt;0.005</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\text{NM}(20)$</td>
<td>0.031</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\text{NM}(10)$</td>
<td>0.543</td>
<td>0.109</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$t_{10}$</td>
<td>0.765</td>
<td>0.401</td>
</tr>
<tr>
<td>Power ($\theta = 1.056$)</td>
<td>Correlation</td>
<td>0.00</td>
<td>Normal</td>
<td>&lt;0.005</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>Spline</td>
<td>Correlation</td>
<td>0.00</td>
<td>Normal</td>
<td>0.130</td>
<td>0.329</td>
</tr>
<tr>
<td></td>
<td>PPC(normal)</td>
<td>0.05</td>
<td>Normal</td>
<td>0.750</td>
<td>0.360</td>
</tr>
</tbody>
</table>
such that $h(W_{i1}), i = 1, \ldots, n$, is approximately normally distributed. Thus, it transforms the data to normality instead of transforming the errors to normality.

There is no intrinsic reason that the marginal method must find the "right" or "wrong" answer. Indeed, in many examples marginal methods will yield transformations that pass both our correlation and error distribution criteria. One drawback of marginal methods that is important in measurement error modeling can be seen by once again considering the concepts of functional and structural modeling. The methods of transformation we have suggested are functional, by which we mean that they make no explicit assumptions about the distribution of the unobservable $X$. This makes sense in the context of measurement error models because of the emphasis in that field of functional modeling on estimating regression parameters.

Unlike our methods, marginal approaches are explicitly structural and can depend in a strong way on the distribution of $X$. For example, consider the case that no transformation is necessary, so that $W = X + U$, $h(u) = u$, and $U$ is normally distributed. Marginal methods transform so that $W$ is normally distributed, and hence they will properly conclude that no transformation is necessary only if $X$ is also normally distributed. This does not mean that marginal methods have no value—far from it—but only that one needs some care in employing them. As a noteworthy example of such care, in their applications Nussel et al. (1997) also check what we call Properties 1 and 2 in Section 2.

One point to keep in mind is that if there are $J = 2$ replicates, then plots of the within-person standard deviation versus the mean will have an odd shape if a significant number of $W$'s approach a lower bound. For example, if the lower bound is zero and if $W_1 \approx 0$, then the standard deviation $\approx W_2/2^{1/2}$ while the mean is $\approx W_2/2$, so that the plot of the standard deviation against the mean will in effect be bounded by a line with intercept zero and slope $2^{1/2}$.

Finally, there is no guarantee that one can find a single transformation that will achieve additivity as measured by the correlation method with a normal or nearly normal error distribution as measured by the error distribution method. The Framingham data using power transformations provide a good example of this issue. Ruppert and Aldershof (1982) addressed this issue in a different context, and suggested estimating parameters either as a weighted average of the correlation and error distribution methods or by weighting their estimating equations. This is an interesting issue for further exploration.
ACKNOWLEDGEMENTS

This research was supported by the National Cancer Institute (CA-87030). We are extremely grateful to Professor H. J. Newton for help in numerical optimization.

RÉSUMÉ

Dans de nombreux problèmes, on souhaite modéliser la relation entre une réponse $Y$ et une covariable $X$. Il est parfois difficile, coûteux, voire impossible d’observer directement $X$, mais on peut observer à la place une variable de remplacement, plus facile à obtenir. Le modèle de loin le plus courant pour représenter la relation entre la covariable d’intérêt $X$ et sa remplaçante $W$ est $W = X + U$, où la variable $U$ représente l’erreur de mesure. Cette hypothèse d’additivité de l’erreur de mesure peut être déraisonnable pour certaines données. Nous proposons un nouveau modèle, à savoir $h(W) = h(X) + U$ où $h(\cdot)$ est une transformation monotone choisie dans une famille $\mathcal{H}$ de fonctions monotones. Le nouveau modèle repose sur l’idée d’une erreur additive sur la bonne échelle de mesure. Nous proposons deux familles $\mathcal{H}$ de transformations possibles, l’une fondée sur le choix d’une transformation qui annule la corrélation entre la moyenne et l’écart-type intra-échantillon d’observations répétées de $W$, l’autre fondée sur le choix de la transformation pour laquelle les erreurs $U$ suivent une distribution spécifiée à l’avance. Les transformations utilisées appartiennent à la famille paramétrique des fonctions puissances ou à une famille de fonctions splines cubiques. Plusieurs exemples illustrent les méthodes présentées.

REFERENCES


Received December 1995; accepted August 1996.
Mixture Normals

If $X \sim t_k$ and $k > 4$, the first and third moments equal zero and the second and fourth moments are $E[X^2] = k/(k - 2)$ and $E[X^4] = 3k^2/(k^3 - 6k + 8)$, respectively. A corresponding mixture normal density with the same first four moments is defined as follows. It has density

$$f(y) = \sum_{j=1}^{2} (\pi_j/\sigma_j) \phi\left(\frac{y - \mu_j}{\sigma_j}\right),$$

where $\phi(\cdot)$ is the standard normal density function, $\sigma_1^2 = 1$, $\sigma_2^2 = (2k)/(k - 4)$, $\pi_1 = k^2/(k^2 + 2k - 8)$, and $\pi_2 = 1 - \pi_1$.

Details of the Algorithm

The following are the steps for optimizing the PPC statistic with respect to the coefficient vector $\Theta$. Assume that we have data $Y_{ij}$, $i = 1, \ldots, n$, $j = 1, 2$, a vector of knot points $\xi = (\xi_1, \ldots, \xi_p)^T$, and a specified measurement error distribution $U$. We will use the notation $Y_i$, $i = 1, 2$, to denote the vector $(Y_{i1}, \ldots, Y_{in})^T$, and $Y^* = (Y_1^T, Y_2^T)^T$. Define the matrices $D$ and $C$ as in (5).

1. Let $\sigma_d$ be the theoretical standard deviation of $U_{i1} - U_{i2}$.
2. Let $s_Y$ be the sample standard deviation of $Y^*$.
3. Define $W_{ij} = Y_{ij}/s_Y$.
4. Generate random values $\hat{\Theta}_m$, $\hat{\Theta}_m = (\hat{\Theta}_{m,1}, \ldots, \hat{\Theta}_{m,p+3})^T$ for $m = 1, \ldots, 1000$:
   (a) Let $\hat{\Theta}_{m,i}, i = 1, \ldots, p + 3$, be independent uniform $[-1, 1]$;
   (b) Let $s_{dm}$ be the sample standard deviation of the elements of $D\hat{\Theta}_m$;
   (c) Multiply each element of $\hat{\Theta}_m$ by the factor $\sigma_d/s_{dm}$;
   (d) Test for $C\hat{\Theta}_m \geq 0$. If any element of $C\hat{\Theta}_m$ is negative, "throw out" $\hat{\Theta}_m$ and generate $\hat{\Theta}_{m+1}$ in step (e) above;
   (e) Calculate $\gamma_m = \gamma(D\hat{\Theta}_m)$, the PPC statistic.
5. Use the value of $\hat{\Theta}_m$ that gave the maximum $\gamma_m$ as the starting value in the numerical optimization program NPSOL to find $\max_{\Theta} \gamma(\Theta)$. 