

Transformations of Covariates for Longitudinal Data

WESLEY K. THOMPSON AND MINGE XIE

Department of Statistics, Rutgers University, Piscataway, NJ 08855

HELENE R. WHITE

Center of Alcohol Studies, Rutgers University, Piscataway, NJ 08855

SUMMARY

This paper develops a general approach for dealing with parametric transformations of covariates for longitudinal data, where the responses are modeled marginally and generalized estimating equations (GEE) are used for estimation of regression parameters. We propose an iterative algorithm for obtaining regression and transformation parameters from estimating equations, utilizing existing software for GEE problems. The algorithmic technique is closely related to that used in the Box-Tidwell transformation in classical linear regression, but we develop it under the GEE setting and for more general transformation functions. We provide supporting theorems for consistency and asymptotic Normality of the estimates. Inference between two nested models is also considered. This methodology is applied to two data sets. One consists of pill dissolution data, the other is taken from the Pittsburgh Youth Study (PYS). The PYS is a prospective longitudinal study of the development of delinquency, substance use, and mental health in male youth. We use the model-based parametric approach to examine the association between alcohol use at an early stage of adolescent development and delinquency over the course of adolescence.

Some key words: Longitudinal Data, Fractional Polynomials, Box-Tidwell Transformation, Generalized Estimating Equations.

1. INTRODUCTION

1.1 Covariate Transformations in Longitudinal Data: Background and Motivations

Transformations of variables, including both dependent and independent variables, are commonly applied when formulating regression models. Among the best known transformations of continuous dependent variables in classical linear regression models are the Box-Cox

and variance stabilization transformations. The transformation of independent variables (covariates) is as important in practice. One such class of covariate transformations, termed “fractional polynomials” by Royston and Altman (1994), is applicable to generalized linear models. Fractional polynomials are an extension of the Box-Tidwell power transformation for covariates. Box and Tidwell (1962) developed an iterative parametric technique for estimating the parameters of this covariate transformation in the classical Gaussian linear regression setting. We propose to use parametric approaches in this paper to study the transformations of covariates for longitudinal data.

Very closely related to covariate transformation is nonparametric regression, since both techniques treat the response as a nonlinear function of the covariates. In nonparametric regression, several smoothing techniques have been developed in the past twenty years and the research in the field is still very active. Recently, Lin and Carroll (2000, 2001) have extended nonparametric regression techniques to deal with longitudinal data, where the responses are modeled marginally and *generalized estimating equations* (GEE) (Zeger and Liang, 1986) are used for estimation. We consider our parametric technique to be a complement to the nonparametric approaches developed by Lin and Carroll (2000, 2001).

While the strengths of nonparametric regression are well documented, there are advantages to using a parametric approach. For example, the parametric modeling approach is simple, and is more easily explained to non-statistician researchers in fields such as the medical or social sciences. Our primary motivation for developing the parametric transformation technique outlined in this paper arose from fitting regression models to data from the social sciences (the Pittsburgh Youth Study, described below). We wish to use a regression technique for these data which gives a more flexible class of models than does simple polynomial regression, while at the same time preserving economy of form.

There are also advantages to parametric approaches from a more purely statistical point of view. First, if the true response function is closely approximated by a function of known form, one can recover that form using a parametric model; this is not possible when using nonparametric regression. We give an example of this situation in Section 5. Second, parametric approaches are applicable in cases where nonparametric regression techniques are not. For example, nonparametric regression techniques often require that the values of the covariates are from a distribution with density function bounded away from zero (see, e.g., Stone, 1982). Sometimes, however, the covariates take only a few values with gaps

between them; for instance, covariates in well-designed experiments may have only a few dose or temperature levels, or there may be a small number of pre-determined age groups in social science studies. In these cases, the parametric approaches can still be used, but not the smoothing techniques. Third, the convergence rate of estimates in nonparametric regression is \sqrt{nh} (where $h = h_n \rightarrow 0$ is the bandwidth), which is slower than \sqrt{n} . As a result, nonparametric regression normally requires more data to obtain a good fit than does parametric regression. Fourth, some issues with nonparametric regression for longitudinal data still require further investigation. Lin and Carroll (2000), for example, pointed out some results that are “surprising”; e.g., “...the asymptotically most efficient estimator of a nonparametric function is obtained by entirely ignoring the correlation within each cluster.” Finally, the parametric approach we develop in this paper is easily executed using any off-the-shelf program which implements the GEE method for longitudinal data regression.

Our methodology can also be applied to cross-sectional data, or some other simpler settings. We develop our methodology under the more complicated setting of longitudinal data, because our motivating application takes the form of repeated categorical responses.

1.2 Motivating Application

The motivating application for this paper is the Pittsburgh Youth Study (PYS), a prospective longitudinal study of the development of delinquency, substance use, and mental health problems in Pittsburgh inner-city male youth (Loeber *et al.*, 1998). The PYS researchers selected three cohort groups from first, fourth, and seventh grade students in the city of Pittsburgh public schools during 1987 and 1988. Within each cohort approximately 500 boys were randomly selected, 250 from the students considered most anti-social, and another 250 from the general student population. The assessments of student behaviors, mental health, and family information were obtained through self-reported questionnaires, court records, and subject and parent interviews. For this study, we use only the oldest cohort, who were in the seventh grade at screening. These subjects were followed up at six-month intervals for four additional assessments and then at yearly intervals for another three assessments. We use a model-based approach in this paper to study the connections between delinquent behavior (yes or no) during the course of adolescence and alcohol use at an early stage of adolescent development. Alcohol use and delinquent behavior may be related because they co-occur during the same development period. Research findings have confirmed

this hypothesis (see, e.g., White, 1990, 1997). The nature of this relationship, however, is not very well understood. It is hoped that the covariate transformation procedure set forth in this paper will help shed some light on this problem.

The format of this paper is outlined as follows. In Section 2, we present the longitudinal regression model, in which the marginal means depend upon transformations of the independent variables. In Section 3, we describe an iterative technique for estimating the transformation parameters along with the regression parameters. In Section 4, we propose a score test for testing transformation parameter values. In Section 5, we illustrate our methodology by applying the technique to a subset of data taken from the PYS, and to an example due to Crowder (1996). In Section 6, we include some concluding remarks. Due to space limitation, the Appendices are posted on the journal's web-site <http://biostatistics.oupjournals.org/> or at <http://stat.rutgers.edu/~mxie/biostat/>.

2. LONGITUDINAL MODEL WITH TRANSFORMED COVARIATES

Suppose we have n individuals, each of which is observed t_i times, $i = 1, \dots, n$, resulting in response vector $\mathbf{y}_i = (y_{i1}, \dots, y_{it_i})^T$. Along with each response y_{ij} is an associated $p+1$ vector of covariates $\mathbf{x}_{ij} = (1, x_{ij1}, \dots, x_{ijp})^T$, so that $X_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{it_i})^T$ is the $t_i \times (p+1)$ matrix of covariates for subject i . In the following, we assume that the responses are marginally distributed according to an exponential family, so that the density of Y_{ij} is given by

$$f(y_{ij}) = \exp\{\{y_{ij}\theta_{ij} - a(\theta_{ij}) + b(y_{ij})\}/\phi\},$$

where $\theta_{ij} = u(\eta_{ij})$ (with u a known injective function) and η_{ij} is the linear predictor. Under this model, the marginal mean is $E(Y_{ij}) = \mu_{ij} = a'\{u(\eta_{ij})\}$. In the usual notation for generalized linear models, $h^{-1} = a'\{u(\cdot)\}$ is the inverse link function. The marginal variance is $\text{Var}(Y_{ij}) = \sigma_{ij}^2 = a''\{u(\eta_{ij})\}\phi$, where the scale parameter ϕ may or may not be fixed. We are interested in the case where the linear predictor η_{ij} is related to the covariates through transformations with known functional form but indexed by a vector of unknown parameters.

Suppose the transformation of the k th covariate x_{ijk} is an R^1 to R^{s_k} mapping: $x_{ijk} \rightarrow \Psi_k(x_{ijk}, \gamma_k)$, where the form of the $s_k \times 1$ vector function $\Psi_k(x_{ijk}, \gamma_k)$ is known or predetermined and $q_k \times 1$ vector γ_k is the unknown transformation parameter. Write $\Psi_{ij} =$

$(1, \Psi_1(x_{ij1}, \gamma_1)^T, \dots, \Psi_p(x_{ijp}, \gamma_p)^T)^T$. We assume that the linear predictor takes the form

$$\eta_{ij} = \Psi_{ij}^T \beta,$$

where $\beta = (\beta_0, \beta_1^T, \dots, \beta_p^T)^T$ is an $s + 1$ vector of regression parameters, $s = \sum_{k=1}^p s_k$, and β_k is the s_k -dimensional regression parameter for transformed covariate vector $\Psi_k(x_{ijk}, \gamma_k)$. In the model, the transformations depend on $q = \sum_{k=1}^p q_k$ unknown parameters $\gamma = (\gamma_1^T, \dots, \gamma_p^T)^T$. One mild condition we place on the transformation function $\Psi_k(x_{ijk}, \gamma_k)$ is

(C1) For any fixed x_{ijk} in the domain set of the k th covariate, the first derivative of $\Psi_k(x_{ijk}, \gamma_k)$ with respect to γ_k , denoted as $\Psi_k^{[1]}(x_{ijk}, \gamma_k)$, exists for γ_k in the admissible parameter set Γ , and each element of $\Psi_k^{[1]}(x_{ijk}, \gamma_k)$ is bounded above and below from infinity. More generally, we allow that there may exist a finite (discrete) set \mathcal{F} in Γ for which $\Psi_k(x_{ijk}, \gamma_k)$ is not continuous and $\Psi_k^{[1]}(x_{ijk}, \gamma_k)$ does not exist.

Note that both $\Psi_k(x_{ijk}, \gamma_k)$ and γ_k are vectors. Thus, $\Psi_k^{[1]}(x_{ijk}, \gamma_k)$ is a $s_k \times q_k$ matrix whose (u, v) th element is the derivative of the u th element of $\Psi_k(x_{ijk}, \gamma_k)$, with respect to the v th element of γ_k . The results that we develop in the next section work for transformations which satisfy condition **(C1)**.

In the simulation studies and examples, we will concentrate on power transformations of covariates. In this case, the γ parameters will be the unknown power or powers applied to the covariates before performing a linear regression. In particular, we consider a special class of power transformations known as fractional polynomials that is described by Royston and Altman (1994). Fractional polynomials are low-degree polynomials in the covariates with the power of each term equal to some possibly non-integral value. They can provide a large variety of possible shapes for modeling the response function. The fitting procedure of Royston and Altman (1994) limits the possible powers for fractional polynomials to a small discrete set of integers and fractions, whereas the method outlined in this paper puts no restraint on the powers.

Fractional polynomials are defined as follows. For the k th covariate x_{ijk} , $k = 1, \dots, p$, let $s_k \geq 1$ be the number of terms (degree) of the polynomial, and let $\gamma_k = (\gamma_{k1}, \dots, \gamma_{ks_k})^T$ be the vector of powers for each term. A fractional polynomial of degree s_k for continuous positive covariate x_{ijk} is

$$\phi_{s_k}(x_{ijk}, \beta_k, \gamma_k) = \{\Psi_k(x_{ijk}, \gamma_k)\}^T \beta_k$$

where β_k is a $s_k \times 1$ vector of regression parameters and $\Psi_k(x_{ijk}, \gamma_k) = (\psi_{k1}(x_{ijk}, \gamma_{k1}), \dots, \psi_{ks_k}(x_{ijk}, \gamma_{ks_k}))^T$, with $\gamma_{k1} \leq \gamma_{k2} \leq \dots \leq \gamma_{ks_k}$. Here, from Royston and Altman (1994)

$$\psi_{k1}(x_{ijk}, \gamma_{k1}) = x_{ijk}^{(\gamma_{k1})}$$

and its u th element, $\psi_{ku}(x_{ijk}, \gamma_{ku})$, for $u = 2, \dots, s_k$, is recursively defined as

$$\psi_{ku}(x_{ijk}, \gamma_{ku}) = \begin{cases} x_{ijk}^{(\gamma_{ku})} & , \gamma_{ku} > \gamma_{k,(u-1)} \\ \psi_{k,(u-1)}(x_{ijk}, \gamma_{ku}) \log x_{ijk} & , \gamma_{ku} = \gamma_{k,(u-1)}. \end{cases}$$

The notation $x^{(\gamma)}$ denotes the Box-Tidwell power transformation

$$x^{(\gamma)} = \begin{cases} x^\gamma & , \gamma \neq 0 \\ \log x & , \gamma = 0 \end{cases}.$$

For example, $\phi_1(x, \beta, .5) = \beta_1 x^{.5}$, whereas $\phi_3(x, \beta, 0, 1, 1) = \beta_1 \log x + \beta_2 x + \beta_3 x \log x$. We note that $\phi_1(x, \beta, \gamma) = \beta_1 x^{(\gamma)}$ is the Box-Tidwell power transformation and $\phi_m(x, \beta, 1, 2, \dots, m) = \beta_1 x + \beta_2 x^2 + \dots + \beta_m x^m$ is a polynomial transformation of degree m . The class of fractional polynomials is much larger than the Box-Tidwell power transformation and conventional polynomials.

In Section 5.2, we analyze a subset of data taken from the PYS. More specifically, we study the connection between the binary response vector of property crimes prevalence and alcohol drinking frequency in the year prior to the first assessment. Property crimes prevalence is measured at six yearly assessments and thus is cluster correlated by subject. As will be seen in Section 5.2, using conventional polynomial regression leads to fits which are not completely satisfactory. To preserve a relatively simple model structure but increase the class of possible model shapes, we instead fit a fractional polynomial to the data. In this instance, the resulting fit reflects the behavior of the data much better than the usual polynomial models do.

3. PARAMETER ESTIMATION

Due to practical limitations, the joint likelihood for the observations is sometimes not specified when using marginal models. If the joint likelihood is not specified, we cannot use maximum likelihood to estimate parameter values. We can, however, estimate the transformation parameters using estimating equations. Let our model be as in Section 2, with

$\eta_{ij} = \{\Psi_{ij}\}^T \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the $s + 1$ vector of regression coefficients and Ψ_{ij} depends on the $q \times 1$ vector of transformation parameters $\boldsymbol{\gamma}$. Let $\boldsymbol{\xi} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$. We can estimate the regression coefficients and transformation parameters simultaneously by solving the following $s + 1 + q$ estimating equations

$$U_n(\boldsymbol{\xi}) = \sum_{i=1}^n D_i(\boldsymbol{\xi})^T V_i^{-1}(\boldsymbol{\xi}) \{\mathbf{y}_i - \boldsymbol{\mu}(\boldsymbol{\xi})\} = \mathbf{0}. \quad (3.1)$$

Here $V_i = A_i^{1/2} R_i(\boldsymbol{\alpha}) A_i^{1/2}$ is a $t_i \times t_i$ working covariance matrix, $A_i = \text{diag}\{\text{Var}(Y_{i1}), \dots, \text{Var}(Y_{it_i})\}$, and $R_i(\boldsymbol{\alpha})$ is a working correlation matrix, chosen by the researcher, and parameterized by vector $\boldsymbol{\alpha}$. Further, $D_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\xi} = A_i \Delta_i Z_i$ is a $t_i \times (s + 1 + q)$ matrix, with $\Delta_i = \text{diag}(d\theta_{i1}/d\eta_{i1}, \dots, d\theta_{it_i}/d\eta_{it_i})$, and $Z_i = \partial \boldsymbol{\eta}_i / \partial \boldsymbol{\xi}$. Note that the $t_i \times (s + q + 1)$ matrix Z_i contains the unknown transformation parameter $\boldsymbol{\gamma}$, which is different from the usual GEEs. The first $s + 1$ equations are for estimating regression parameters $\boldsymbol{\beta}$. When the transformation parameters are known, they are exactly the usual GEEs for regression parameter $\boldsymbol{\beta}$. The last q equations are for estimating the transformation parameter $\boldsymbol{\gamma}$.

Equations (3.1) may be solved directly by the Newton-Raphson algorithm or by some other technique for solving nonlinear equations to obtain estimate $\hat{\boldsymbol{\xi}}$, but it would involve the derivative of $U_n(\boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}$ and intensive programming effort. Alternatively, we propose a simple iterative algorithm, utilizing the existing software for solving GEEs, to obtain estimates from equations (3.1). The iterative algorithm, similar to that of Box and Tidwell (1962), depends upon estimating parameters in the first order Taylor's expansion of η_{ij} .

More specifically, for $\boldsymbol{\gamma}$ around a given value, say $\boldsymbol{\gamma}^{(0)}$, we have approximately

$$\eta_{ij} = \{\Psi_{ij} |_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(0)}}\}^T \boldsymbol{\beta} + \sum_{k=1}^p \left[\{\Psi_k^{[1]}(\mathbf{x}_{ij}, \boldsymbol{\gamma}^{(0)})\}^T \boldsymbol{\beta}_k \right]^T (\gamma_k - \gamma_k^{(0)}),$$

where the $s_k \times 1$ vector $\boldsymbol{\beta}_k$ is the sub-vector of $\boldsymbol{\beta}$ which corresponds to Ψ_k , the vector of transformations of the k th covariate x_{ijk} . At each step, the iterative algorithm uses two GEEs to update the parameter estimates $\boldsymbol{\xi}^{(new)} = (\{\boldsymbol{\beta}^{(new)}\}^T, \{\boldsymbol{\gamma}^{(new)}\}^T)^T$. Denote the current estimates as $\boldsymbol{\beta}^{(curr)}$ and $\boldsymbol{\gamma}^{(curr)}$. First, we obtain the updated values of $\boldsymbol{\beta}^{(new)}$ by fitting the standard GEEs with linear predictor η_{ij} as

$$\eta_{ij} = \{\Psi_{ij} |_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(curr)}}\}^T \boldsymbol{\beta}. \quad (3.2)$$

Then, treating $\boldsymbol{\beta}^{(new)}$ as given, we obtain the updated $\boldsymbol{\gamma} = (\{\gamma_1\}^T, \dots, \{\gamma_p\}^T)^T$ values, say $\boldsymbol{\gamma}^{(new)} = (\{\gamma_1^{(new)}\}^T, \dots, \{\gamma_p^{(new)}\}^T)^T$, by solving the standard GEEs but with linear predictor

replaced by

$$\eta_{ij} = \omega_{ij}^{(curr)} + \sum_{k=1}^p \left[\{\Psi_k^{[1]}(\mathbf{x}_{ij}, \boldsymbol{\gamma}^{(curr)})\}^T \boldsymbol{\beta}_k^{(new)} \right]^T \boldsymbol{\gamma}_k, \quad (3.3)$$

where $\omega_{ij}^{(curr)} = \{\Psi_{ij} |_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(curr)}}\}^T \boldsymbol{\beta}^{(new)} - \sum_{k=1}^p \left[\{\Psi_k^{[1]}(\mathbf{x}_{ij}, \boldsymbol{\gamma}^{(curr)})\}^T \boldsymbol{\beta}_k^{(new)} \right]^T \boldsymbol{\gamma}_k^{(curr)}$ is the offset term. The offset term $\omega_{ij}^{(curr)}$ depends only on $\boldsymbol{\beta}^{(new)}$ and $\boldsymbol{\gamma}^{(curr)}$, both of which are available at this stage. In the above discussion, we implicitly assume $\Psi_k^{[1]}$ exists at $\boldsymbol{\gamma}^{(curr)} \notin \mathcal{F}$. When $\boldsymbol{\gamma}^{(curr)}$ belongs to the discrete set \mathcal{F} , it is necessary to handle the estimation of $\boldsymbol{\gamma}$ somewhat differently. See Appendix A for how to deal with this situation in the case of the Box-Tidwell power transformation and fractional polynomials. However, in practice, we have found that this is not an issue when the starting value is not from \mathcal{F} . Note that we can usually choose the starting value outside of the set \mathcal{F} . We iterate this updating scheme until the updated estimates $\boldsymbol{\xi}^{(new)}$ in two consecutive steps are close.

Equations (3.1) are Fisher consistent. Clearly, the parameter estimates obtained from (3.1) are consistent. We summarize the related asymptotic results in the following theorem. The proof is provided in Appendix B, along with a set of easily verifiable, information matrix-based conditions. The main conditions mirror those of Xie and Yang (2002).

Theorem 1 *Under some mild conditions (for example, those listed in Appendix B.1), there exists a sequence $\hat{\boldsymbol{\xi}}_n$ of r.v.'s such that $P\{U_n(\hat{\boldsymbol{\xi}}_n) = \mathbf{0}\} \rightarrow 1$ and $\hat{\boldsymbol{\xi}}_n \xrightarrow{P} \boldsymbol{\xi}_0$, as $n \rightarrow \infty$. Also,*

$$\{V_\xi |_{\xi=\xi_0}\}^{-1/2} (\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0) \xrightarrow{L} N(0, I), \quad \text{as } n \rightarrow \infty,$$

where $\boldsymbol{\xi}_0$ is the true parameter value and the covariance matrix V_ξ is given by

$$V_\xi = n \left(\sum_{i=1}^n D_i^T V_i^{-1} D_i \right)^{-1} \left\{ \sum_{i=1}^n D_i^T V_i^{-1} \text{cov}(\mathbf{Y}_i) V_i^{-1} D_i \right\} \left(\sum_{i=1}^n D_i^T V_i^{-1} D_i \right)^{-1}. \quad (3.4)$$

When the iterative algorithm is convergent, we obtain estimates $\hat{\boldsymbol{\xi}}$ of both regression and transformation parameters. The estimated transformed covariates for the ij th response are then given by $\hat{\boldsymbol{\Psi}}_{ij} = (1, \{\Psi_1(\mathbf{x}_{ij1}, \hat{\boldsymbol{\gamma}}_1)\}^T, \dots, \{\Psi_p(\mathbf{x}_{ijp}, \hat{\boldsymbol{\gamma}}_p)\}^T)^T$ and the estimated $t_i \times (s + 1 + q)$ design matrix of Z_i is

$$\hat{Z}_i = \begin{pmatrix} \{\hat{\boldsymbol{\Psi}}_{i1}\}^T & \hat{\boldsymbol{\beta}}_1^T \boldsymbol{\Psi}_1^{[1]}(x_{i11}, \hat{\boldsymbol{\gamma}}_1) & \dots & \hat{\boldsymbol{\beta}}_p^T \boldsymbol{\Psi}_p^{[1]}(x_{i1p}, \hat{\boldsymbol{\gamma}}_p) \\ \vdots & \vdots & \dots & \vdots \\ \{\hat{\boldsymbol{\Psi}}_{it_i}\}^T & \hat{\boldsymbol{\beta}}_1^T \boldsymbol{\Psi}_1^{[1]}(x_{it_i1}, \hat{\boldsymbol{\gamma}}_1) & \dots & \hat{\boldsymbol{\beta}}_p^T \boldsymbol{\Psi}_p^{[1]}(x_{it_ip}, \hat{\boldsymbol{\gamma}}_p) \end{pmatrix}, \quad i = 1, \dots, n.$$

The matrix \hat{Z}_i can be obtained from the outputs of the two GEEs-fittings in the last iteration of the algorithm proposed above; the first $s + 1$ columns from fitting model (3.2) and the last q columns from fitting model (3.3). The variance estimates of $\hat{\xi}$ can be obtained through formula (3.4), where we substitute $\hat{\xi}$ for ξ and \hat{Z}_i for Z_i in D_i and V_i and substitute $\text{cov}(\mathbf{Y}_i) = (\mathbf{Y}_i - \boldsymbol{\mu}_i(\hat{\xi}))(\mathbf{Y}_i - \boldsymbol{\mu}_i(\hat{\xi}))^T$ for $\text{cov}(\mathbf{Y}_i)$.

At the first step of this iterative algorithm, it is necessary to choose an initial guess for the transformation parameter γ . Denote this initial guess by $\boldsymbol{\gamma}^{(0)} = (\{\gamma_1^{(0)}\}^T, \dots, \{\gamma_p^{(0)}\}^T)^T$. Box and Tidwell (1962) suggested that $\boldsymbol{\gamma}^{(0)}$ can typically be chosen so that $\Psi_k(x_{ijk}, \gamma_k^{(0)}) = x_{ijk}$, $k = 1, \dots, p$. For more general transformations this is not always possible, and starting values may need to be chosen using some other criterion. If the dimension q of γ is not too large, it is possible to obtain an initial estimate $\boldsymbol{\gamma}^{(0)}$ through a grid search if we limit our possible starting values for γ to some reasonably small discrete set \mathcal{Q} . More specifically, for each $\gamma_u \in \mathcal{Q}$, we perform one iteration of the estimation procedure in (3.2) and (3.3), with $\boldsymbol{\gamma}^{(curr)} = \gamma_u$. From this, we can obtain an estimate $\hat{\zeta}(\gamma_u)$ for $\gamma - \gamma_u$. The $q \times 1$ vector $\hat{\zeta}(\gamma_u)$ has estimated covariance matrix $\hat{V}_{\hat{\zeta}(\gamma_u)}$, which is the $q \times q$ lower-right submatrix of formula (3.4), estimated as above, assuming that the true transformation parameter is γ_u . Then, the initial guess for γ is $\boldsymbol{\gamma}^{(0)} = \text{argmin}_{\gamma_u \in \mathcal{Q}} \{\hat{\zeta}(\gamma_u)^T \hat{V}_{\hat{\zeta}(\gamma_u)}^{-1} \hat{\zeta}(\gamma_u)\}$. In practice, a quick estimate can be obtained by choosing $\boldsymbol{\gamma}^{(0)} \in \mathcal{Q}$ such that $\|\hat{\zeta}(\boldsymbol{\gamma}^{(0)})\| \leq \|\hat{\zeta}(\gamma_u)\|$ for all $\gamma_u \in \mathcal{Q}$.

We carry out two simulation studies to evaluate the performance of the estimates $\hat{\gamma}$ and the coverage percentages of the (asymptotic) 95% confidence intervals, calculated by the iterative procedure outlined in this section. The data sets in the first study follow the Gaussian marginal model $Y_{ij} = \beta_0 + \beta_1 x_{ij}^{(\gamma)} + \epsilon_{ij}$. The ϵ_{ij} are generated with two types of within-subject correlation structure: exchangeable and first-order autoregressive (AR-1). For each data set, we calculate the (asymptotic) 95% confidence interval for $\hat{\gamma}$. The resulting proportion of confidence intervals containing the true value of γ is typically close to 95%, except that when the within-subject correlation is large, the working independence assumption leads to serious over-coverage. The second simulation study consists of binary responses with marginal mean modeled as $\mu_{ij} = P(Y_{ij} = 1 | x_{ij}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \exp(\eta_{ij}) / \{1 + \exp(\eta_{ij})\}$, where $\eta_{ij} = \beta_0 + \phi_2(x_{ij}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. Here, $\phi_2(x_{ij}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ denotes a fractional polynomial of degree 2. The median estimates of $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are close to the true values. The proportion of times that the true values of γ_1 and γ_2 fall within the 95% confidence intervals is again close to 95%. Details of the two simulation studies can be found in Appendix C.

4. HYPOTHESIS TESTING

We consider the problem of choosing between two nested models. Our main concern is to select between two nested sets of parametric transformations of covariates. The proposed tests, however, also apply to other types of nested models.

As above, let $\boldsymbol{\xi} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ be the combined vector of regression coefficients and transformation parameters. Rearranging the vector if necessary, write $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^T, \boldsymbol{\xi}_2^T)^T$, where $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are $r \times 1$ and $(s+1+q-r) \times 1$ vectors of transformation parameters and/or regression coefficients, respectively. We consider testing the null hypothesis H: $\boldsymbol{\xi}_1 = \boldsymbol{\xi}_0$. We propose to use the “working” score test statistic

$$T_S = n^{-1}[\mathbf{U}_{\boldsymbol{\xi}_1}\{\boldsymbol{\xi}_0, \hat{\boldsymbol{\xi}}_2(\boldsymbol{\xi}_0)\}]^T \hat{W}_{\boldsymbol{\xi}_1} \mathbf{U}_{\boldsymbol{\xi}_1}\{\boldsymbol{\xi}_0, \hat{\boldsymbol{\xi}}_2(\boldsymbol{\xi}_0)\}, \quad (4.1)$$

where $\mathbf{U}_{\boldsymbol{\xi}_1}(\boldsymbol{\xi}) = \sum_{i=1}^n \{\partial \boldsymbol{\mu}_i(\boldsymbol{\xi}) / \partial \boldsymbol{\xi}_1\} \mathbf{V}_i^{-1}(\boldsymbol{\xi}) \{\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\xi})\}$. Here, $\hat{W}_{\boldsymbol{\xi}_1}$ and $\hat{\boldsymbol{\xi}}_2(\boldsymbol{\xi}_0)$ are the estimates of $W_{\boldsymbol{\xi}_1}$ and $\boldsymbol{\xi}_2$; both estimated under the null hypothesis, and the matrix $W_{\boldsymbol{\xi}_1}$ is the principal $r \times r$ submatrix of $W_{\boldsymbol{\xi}} = n \left(\sum_{i=1}^n D_i^T \mathbf{V}_i^{-1} D_i \right)^{-1}$. This score statistic allows for the modeling of within-cluster correlation.

One might form a score test statistic differently, say T_S^* , by using the covariance estimate $\hat{V}_{\boldsymbol{\xi}_1}$ of $\mathbf{U}_{\boldsymbol{\xi}_1}\{\boldsymbol{\xi}_0, \hat{\boldsymbol{\xi}}_2(\boldsymbol{\xi}_0)\}$ in place of $\hat{W}_{\boldsymbol{\xi}_1}$ in equation (4.1). However, T_S^* often suffers from computational instability, because of the instability of the estimate $\hat{V}_{\boldsymbol{\xi}_1}$, especially when the number of independent clusters is not very large; also see, Rotnitzky and Jewell (1990). From our experience, we prefer using T_S to T_S^* .

As in Rotnitzky and Jewell (1990), we also have in our setting the following theorem.

Theorem 2 *Under some mild conditions (for example, those in Appendix B), $T_S \xrightarrow{\mathcal{L}} \sum_{k=1}^{s_1} a_k \chi_{1,k}^2$, where the $\chi_{1,k}^2$ are independent χ^2 variates and the weights $a_1 \leq a_2 \leq \dots \leq a_{s_1}$ are the eigenvalues of $V_{\boldsymbol{\xi}_1} W_{\boldsymbol{\xi}_1}^{-1}$.*

The proof of this theorem is sketched out in Appendix B. If the within-cluster covariance has been correctly modeled, the weights a_k should be close to one.

In practice, the eigenvalues of $V_{\boldsymbol{\xi}_1} W_{\boldsymbol{\xi}_1}^{-1}$ can be approximated by estimating the matrices $V_{\boldsymbol{\xi}_1}$ and $W_{\boldsymbol{\xi}_1}$ as above, with parameters estimated assuming the null hypothesis is true. If $r > 1$, some authors (e.g., Rotnitzky and Jewell, 1990) have suggested using T_S/\bar{a} , where \bar{a} is

the mean of the eigenvalues of $\hat{V}_{\xi_1} \hat{W}_{\xi_1}^{-1}$. Under the null hypothesis, T_S/\bar{a} is approximately χ^2 distributed with r degrees of freedom. The exact quantiles of a weighted χ^2 distribution can also be computed. Using Ruben's series expansion (1965), for instance, Xie (1998) developed an algorithm to calculate the quantiles of a weighted χ^2 distribution up to 10^{-4} precision. Finally, bootstrap simulations can also be used to obtain the quantiles of the distribution of T_S , where they are estimated by the quantiles of the bootstrapped sample distributions of T_S . The bootstrap approach does not involve estimating the eigenvalues of $V_{\xi_1} W_{\xi_1}^{-1}$.

In the first simulation study detailed in Appendix C, we also evaluate the performance of the proposed working score test. The working score test rejects the null hypothesis $H: \gamma = 1$ each time when the true value used in the simulations are $\gamma = .5, 0$, or -1 ; this also matches confidence interval coverages. When the null hypothesis $H: \gamma = 1$ is true, the rejection rates of the working score test is 0.062, slightly higher than the expected rate of .05.

5. EXAMPLES

5.1 Pill Dissolution Data

We first present a brief example from Crowder (1996). The data consist of measurements of times to various fractions (0.90, 0.70, 0.50, 0.30, 0.25 and 0.10) of the pills to dissolve on $N = 17$ pills. Based on scientific considerations for the diffusion of a small tablet in a large reservoir, Crowder derived the model $\mu_{ij} = \beta_0 + \beta_1(1 - f_j^{2/3})$, for $i = 1, \dots, 17, j = 1, \dots, 6$. Here, μ_{ij} is the mean time for fraction f_j of pill i to dissolve. We carry out the estimation using the identity link $\eta_{ij} = \theta_{ij}$. We assume that the mean time has structure $\mu_{ij} = \beta_0 + \beta_1 f_j^\gamma$. Using the iterative estimation procedure with exchangeable working correlation, an estimate of $\hat{\gamma} = 0.69$ is produced in three iterations, which is extremely close to the value $2/3$ predicted from theoretical considerations. In contrast to nonparametric procedures, we can recover the true form of the transformation using the parametric approach. This example is also a case where nonparametric procedures may not be appropriate, because of the discrete nature of the covariate under consideration and the small number of independent subjects.

5.2 The Pittsburgh Youth Study

In this section, we present an analysis of a data set taken from the oldest age group of the PYS. The subset of data we look at follows 506 boys over six yearly assessments from ages 13 through 14 (early adolescence) to ages 18 through 19 (late adolescence). One of the questions of interest in the PYS is the determination of factors present at the initial assessment which are predictive of delinquent behavior over the course of adolescence. Research findings, based on Pearson correlations, suggest that drug and alcohol use and delinquency are highly correlated (White, 1997, White, *et al.*, 1999). Here, we wish to use the model-based parametric approach to examine the association between a subject's alcohol use at an early stage of development and his propensity for committing delinquent acts throughout adolescence.

The response variable for this example is prevalence of property crimes in the full year prior to the current assessment. Prevalence is a categorical variable, with value equal to 1 if the subject committed any one or more property crimes in the past year, and 0 otherwise. The binary response vector for subject i is $\mathbf{y}_i = (y_{it_1}, \dots, y_{it_6})^T$, $i = 1, \dots, 506$, where response y_{it_j} is the prevalence (0 or 1) of property crimes (PROP) committed by the i th subject in the year prior to the j th assessment.

The covariate variable under consideration is alcohol frequency (AlcF) at the first assessment (i.e., at ages 13 through 14). The measurement is the number of times (self-reported) that subject i consumed beer, wine or hard liquor in the year prior to the first assessment. In the following analysis, we use the standardized covariate $x_i = (\text{AlcF}_i + 1) / \sqrt{\text{Var}(\text{AlcF})}$. We also include a time component in the model to account for changes in PROP over time within each subject. The actual time covariate we use is the ordinal value of the measurement standardized to have unit variance, i.e., $z_{ij} = z_j = j / \sqrt{3.5}$, $j = 1, \dots, 6$.

Figure 1(a) plots the mean property crime prevalence versus the x_i . For each level of x_i , the proportion of positive responses for PROP is displayed. The time aspect of the data is suppressed in this plot and each subject's data are summarized in one point on the graph. PROP clearly shows an increasing tendency as x_i increases. The median value of x_i corresponds to $\text{AlcF}_i = 1$, or one drink prior to the first year on study; roughly 40% of the subjects consumed no alcoholic drinks prior to the first year on study.

Since the responses are binary, we fit a logistic regression model to the data, $P(Y_{ij} = 1) = \mu_{ij} = \exp(\eta_{ij}) / \{1 + \exp(\eta_{ij})\}$, where η_{ij} is the linear predictor. Concentrating on the alcohol frequency covariate x_i , we fit models of the type $\eta_{ij} = \beta_0 + \beta_1 z_j + \{\Psi_{ij}(x_i, \gamma)\}^T \beta$.

Specifically, we compare (1) a straight-line model, $\eta_{ij} = \beta_0 + \beta_1 z_j + \beta_2 x_i$, (2) the best-fitting polynomial model, $\eta_{ij} = \beta_0 + \beta_1 z_j + \beta_{2,1} x_i + \beta_{2,2} x_i^2 + \beta_{2,3} x_i^3$, and (3) a first degree fractional polynomial model, $\eta_{ij} = \beta_0 + \beta_1 z_j + \beta_2 x_i^{(\gamma)}$. The parameter estimates and the values of the Z and robust Z statistics obtained from fitting these models are given in Table 1. The working covariance matrix V_i in the GEE fitting is chosen to be exchangeable, i.e., $\text{corr}(Y_{it_j}, Y_{it_k}) = \rho$, for $t_j \neq t_k$. The fitted lines for all three models are displaced in Figure 1(a), where we have set z_j to be its mean value. The fit obtained from Model (1) indicates a modest positive correlation between increased alcohol consumption and property crime prevalence. The best-fitting polynomial model, Model (2), indicates that there is some curvature in the response function which is not being captured in the straight-line model, since all the terms in the cubic polynomial are strongly significant. Finally, in fitting Model (3), we estimate the transformation parameter γ as in Section 3, obtaining $\hat{\gamma} = 0.04$. The (asymptotic) 95% confidence interval for γ is $[-0.018, 0.062]$. For simplicity, we choose the log transformation, resulting in the model $\eta_{ij} = \beta_0 + \beta_1 z_j + \beta_2 \log x_i$. We also test fractional polynomial models of degree 1 versus degree 2 using the score test statistic T_S , which indicates no improvement in fit for the more complex model.

The three different models lead to differing interpretations of the influence of the AlcF covariate on PROP, as can be seen in Figure 1(a). In Model (1), μ_{ij} starts off relatively high at AlcF = 0, with $\mu_{ij} = .28$, and increases monotonically but slowly towards $\mu_{ij} = 1$ for very heavy drinkers. Model (2) also starts relatively high with $\mu_{ij} = .24$ when $x_i = 0$; however, the fitted responses are not monotonically increasing as x_i increases, which seems to be an artifact produced by fitting a cubic polynomial. Model (3) starts off with μ_{ij} essentially 0, rises quite rapidly, and quickly levels off to about .75 for the heaviest drinkers in the sample. The monotonicity in x_i of Model (3) lends itself to a simpler and more intuitive interpretation than does the result obtained from Model (2), i.e., higher alcohol use leads to higher rates of delinquency. Model (3) also suggests that there is a something of a dichotomy between the subjects with AlcF = 0 and those subjects with AlcF > 0, which agrees with some practice in the field. Models (1) and (2), while giving a picture of positive correlation between the two variables, do not suggest such a stark difference between users and non-users of alcohol before the first assessment.

McCullagh and Nelder (1989) suggested a diagnostic plot (attributed to Pregibon) of *partial residuals* versus covariate x as a diagnostic tool for checking the correct scale for

x . Define the partial residual as $u = v - \hat{\eta} + \hat{\beta}x$, where $\hat{\eta}$ is the fitted linear predictor for the model including covariate x , and v is the adjusted dependent variable (also known as “working response”), given by $v = \hat{\eta} + (y - \hat{\mu})(d\eta/d\mu)$. If the scale of the covariate x is correct, the points in the partial residual plot should be approximately linear. In our study, we need to check the adequacy of transformed cubic and fractional polynomial models. We modify the partial residual plot by instead plotting $u^* = v - \hat{\eta} + \hat{T}(x)$ versus $\hat{T}(x)$, where $\hat{T}(x)$ is the fitted sum of all transformation terms of the covariate x . In particular, in Model (1) through (3), $\hat{T}(x)$ is $\hat{\beta}_1x$, $\hat{\beta}_1x + \hat{\beta}_2x^2 + \hat{\beta}_3x^3$, and $\hat{\beta}_1 \log x$, respectively. For the same reason, if the terms of the covariate x are adequate, the points in the modified partial residual plot should approximately fall around a line through the origin with unit slope.

In Figures 1(b) through 1(d), we display the diagnostic plots for Model (1) (linear in x_i), Model (2) (cubic in x_i) and Model (3) (linear in $\log x_i$). In these figures, the points plotted are the means of the partial residuals, where the mean is taken for each unique point of $\hat{T}(x)$ (the graphs of the non-aggregated partial residuals, not shown here, are a little messier but lead to essentially the same conclusions). Model (1) has a mean partial residual of -29 which is included (labeled as “A”) but not shown to scale. The partial residual plot for Model (3) seems preferable to the first two plots. This provides further confirmation that the log transformed covariate is a substantial improvement over a straight-line model and is preferable to the cubic polynomial model.

6. DISCUSSION

We have proposed a procedure for estimating the parameters of transformations of covariates when the form of the transformation is known. Transforming covariates in such a fashion provides a complement to non-parametric regression, with several advantages in certain situations. One of the primary advantages is the ease with which the resulting model can be explained to non-statistician researchers. The specific class of power transformations we have focused on, fractional polynomials, has advantages over a strictly parametric approach as well. In the analysis of the PYS data, for example, the log transformation indicated by the procedure fits much better than the straight-line model, while avoiding the more complicated, non-monotonic behavior of the best-fitting polynomial model (cubic). Also

note that while we have focused on power transformations in this paper, the estimation procedure is applicable to a broad class of possible non-linear transformations.

As with all statistical modeling, care should be taken when selecting the type of transformation applied to the covariates. The pill dissolution data model taken from Crowder (1996) is an example where the form of the transformation was determined from scientific considerations, independently of the data. In this case our transformation procedure worked well in recovering the true response function. Of course this is not always possible; for example, data from the social sciences, like the PYS, are usually not as amenable to this type of analysis. We believe, however, that the transformation technique outlined in this paper remains valuable in such situations.

Acknowledgements

The authors wish to thank Drs. Loeber and Stouthamer-Loeber for kindly allowing us to use the Pittsburgh Youth Study (PYS) data. The authors also wish to thank the editor and the reviewers for their constructive suggestions. The PYS was funded by grants from the National Institute on Drug Abuse (DA 41101), the National Institute on Mental Health (NIMH 50778) and the Office of Juvenile Justice and Delinquency Prevention (OJJDP 96-MU-FX-0012). Xie's research was partly supported by NSF grant DMS 9803273. Correspondence concerning this paper should be addressed to mxie@stat.rutgers.edu.

References

- Box, G.E.P and Tidwell, P.W. (1962). Transformation of the independent variables. *Technometrics* **4** 531–550.
- Crowder, M.J. (1996). Keep timing the tablets: Statistical analysis of pill dissolution rate data. *Applied Statistics* **45** 323–334.
- Lin, X. and Carroll, R.J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Amer. Statist. Assoc.* **95** 520–534.
- Lin, X. and Carroll, R.J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *J. Amer. Statist. Assoc.* **96** 1045–56.

- Loeber, R., Farrington, D.P., Stouthamer-Loeber, M. and Van Kammen, W.B. (1998). *Antisocial Behavior and Mental Health Problems: Explanatory Factors in Childhood and Adolescence*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Rotnitzky, A. and Jewell, N. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models with cluster correlated data. *Biometrika* **77** 485-497.
- Royston P. and Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modeling. *Appl. Statist.* **43** 429-467.
- Rubens, H. (1962). Probability content of regions under spherical normal distributions. IV: The distribution of homogeneous and non-homogeneous quadratic functions of normal variables. *The Annals of Mathematical Statistics*, **33** 542-570.
- Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, **10** 1040-1053.
- White, H.R. (1990). The drug use-delinquency connection in adolescence. In *Drugs, Crime, and Criminal Justice* ed. Weisheit 215-256. Cincinnati, OH: Anderson Publishing Co.
- White, H.R. (1997). Alcohol, illicit drugs, and violence. In *Handbook of Antisocial Behavior* eds. Stoff, Breiling and Maser 511-523. New York: John Wiley and Sons Press.
- White, H.R., Loeber, R., Stouthamer-Loeber, M. and Farrington, D. (1999). Developmental associations between substance use and violence. *Develop. and Psychopath.* **11** 785-803.
- Xie, M. (1998). Quantile evaluation of weighted sum of χ^2 distributions (with FORTRAN code). (unpublished research notes).
- Xie, M. and Yang, Y (2002). Asymptotics for generalized estimating equations with large cluster sizes. *Ann. Statist.* (in press).
- Zeger S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42** 121-130.

Table 1: Parameter Estimates and Z-Values for PYS Models

	Straight-line model			Cubic model			Log-transformed model		
	coef	Z	robust Z	coef	Z	robust Z	coef	Z	robust Z
Intcpt	-1.19	-11.24	-9.61	-1.41	-12.32	-11.87	-.27	-1.95	-1.88
z_j	.11	3.08	2.69	.11	3.12	2.69	.11	3.08	2.67
x_i	.42	4.33	2.24	1.46	5.77	6.01	-	-	-
x_i^2	-	-	-	-.32	-3.64	-4.35	-	-	-
x_i^3	-	-	-	.02	2.89	3.70	-	-	-
$\log(x_i)$	-	-	-	-	-	-	.43	7.83	7.67

The Z-value is calculated from the naive covariance estimate. The robust Z-value is calculated from the (robust) sandwich covariance estimate.

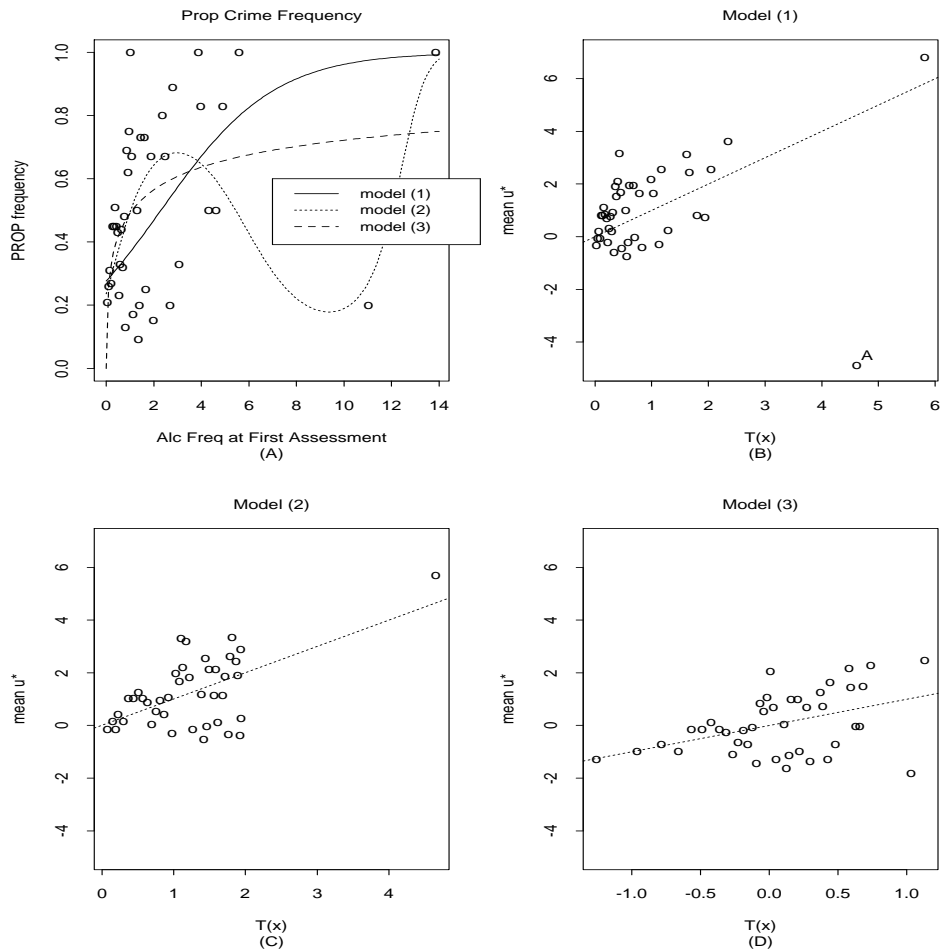


Figure 1(a): Mean property crime prevalence vs. modified alcohol frequency, with fitted lines for Models (1),(2), and (3). Figure 1(b): Mean partial residuals vs. $\hat{T}(x_i) = \hat{\beta}_1 x_i$. Mean partial residual with value of -29, labeled as “A”, is not shown to scale. Figure 1(c): Mean partial residuals vs. $\hat{T}(x_i) = \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \hat{\beta}_3 x_i^3$ for Model (2). Figure 1(d): Mean partial residuals vs. $\hat{T}(x_i) = \hat{\beta}_1 \log x_i$ for Model (3). In Figures 1(b) through 1(d), a line through the intercept with unit slope is included for comparison.