Published in final edited form as:

Stat Med. 2013 October 30; 32(24): 4306–4318. doi:10.1002/sim.5838.

A Flexible Model for the Mean and Variance Functions, with Application to Medical Cost Data

Jinsong Chen^a, Lei Liu^{a,b,*}, Daowen Zhang^c, and Ya-Chen T. Shih^d

^aDepartment of Preventive Medicine, Northwestern University, Chicago, U.S.A

^bRobert H. Lurie Comprehensive Cancer Center of Northwestern University, Chicago, U.S.A

^cDepartment of Statistics, North Carolina State University, Raleigh, NC, U.S.A

^dDepartment of Medicine, University of Chicago, Chicago, IL, U.S.A

Abstract

Medical cost data are often skewed to the right and heteroscedastic, having a nonlinear relation with covariates. To tackle these issues, we consider an extension to generalized linear models by assuming nonlinear associations of covariates in the mean function and allowing the variance to be an unknown but smooth function of the mean. We make no further assumption on the distributional form. The unknown functions are described by penalized splines, and the estimation is carried out using nonparametric quasi-likelihood. Simulation studies show the flexibility and advantages of our approach. We apply the model to the annual medical costs of heart failure patients in the clinical data repository (CDR) at the University of Virginia Hospital System.

Keywords

Generalized linear model; semiparametric regression; health econometrics; smoothing parameter; generalized cross validation

1. INTRODUCTION

Medical costs in the U.S. have been rising at a faster pace than general inflation for decades. A recent report from the National Institute for Health Care Management [1] shows that health care costs per person reached 8,100 U.S. dollars, for a total of 2.5 trillion U.S. dollars, which accounted for 17.6% of the gross domestic product in 2009. The rising health care costs are the main obstacle to securing medical coverage for all citizens in the United States [2]. Containing skyrocketing health care costs is one of the central targets of President Obama's health care reform plan.

The heightened interest in health care cost containment strategies increases the importance of statistical analysis of medical cost data. Medical cost data are routinely collected in billing records of hospitals and claims of health insurance plans (e.g., Medicare, Medicaid,

---

Correspondence to: Lei Liu, Department of Preventive Medicine, Northwestern University. 680 North Lakeshore Drive, Ste 1400. Chicago, IL 60611, U.S.A. lei.liu@northwestern.edu.

or commercial insurance). The wide availability of such data has motivated the development and application of state-of-the-art statistical and econometric methods in the past.

Although other summary measures of medical costs have been taken, e.g., median cost [3], mean cost is used overwhelmingly in health economics studies to measure the financial burden of a disease, as far as the total cost is concerned. We are thus interested in the mean medical cost in this paper.

Many previous studies of health care costs indicate that the estimate of mean response may be quite sensitive to heteroscedasticity and severe skewness. These data patterns are clearly demonstrated in Figure 1, which shows the histogram of annual medical costs for heart failure patients in the clinical data repository (CDR) at University of Virginia (UVA) Health System database. To overcome these problems, researchers historically relied on logarithmic or other transformation of response Y , followed by regressing the transformed Y on covariates X . The main drawback of transforming Y (e.g., $\log Y$) is that the regression no longer results in a model for the mean response in the original scale, which is the scale of interest for most applications. A retransformation is needed in order to draw inference about the mean response in the natural scale of Y [4]. However, the retransformation is complicated in the presence of heteroscedasticity [5]. In practical applications, any retransformation can potentially yield biased estimators unless considerable efforts are devoted to studying the appropriate form of heteroscedasticity. Another drawback of transformation of response is that it might mislead the hypothesis testing. For example, the t -test of equality of the mean of log-transformed costs tests the wrong hypothesis unless the variances in the log-scale are equal (Zhou et al., [6]).

To avoid the aforementioned issues associated with retransformation, researchers have focused on the use of generalized linear models (GLMs), e.g., [5, 7, 8, 9, 10]. Instead of estimating $E(\log Y)$, the GLMs model $\log E(Y)$ so we can obtain the mean medical cost easily. These models simultaneously describe the link and variance structure by pre-specified functions, e.g., log link with gamma error distribution. The link function provides the relationship between the linear predictor and the mean of the distribution function. However, the assumption of linear covariate associations in the link function is sometimes violated. This has led to the developments and applications of various generalized nonparametric models [11].

In the GLMs framework, consistency of the estimated regression coefficients depends on a correctly specified link function regardless of variance function. However, a mis-specified variance function could lead to a substantial loss of efficiency for parameter estimates. Chiou and Muller [12] considered the variance function as an unspecified function of the mean by using local polynomial or kernel smoothing in their estimation. Basu and Rathouz [13] proposed an extension to the estimating equations in the GLMs, with variance structure as a flexible parametric function of the mean.

In this paper, we propose a model with nonparametric components to relax the assumption of linear covariate associations in the mean function, and characterize the variance as an unspecified function of the mean. The unknown functions are assumed to be smooth and

estimated using penalized splines (P-splines). The estimates are obtained by maximizing the penalized nonparametric quasi-likelihood (PNQL). We call it PNQL instead of penalized quasi-likelihood (PQL) because the true variance function in the usual definition of quasi-likelihood [7] is replaced by a smooth, unspecified variance function.

The rest of the paper is organized as follows. The model is introduced in Section 2. Section 3 presents the estimation method and inference. In Section 4, we show the results of simulation studies. In Section 5 we apply our method to analyze the annual medical cost data of heart failure patients in the CDR database. Summary of our conclusions and implications of this work are presented in the final section.

2. Model

Define by Y_i the medical cost for subject i , and $E(Y_i) = \mu_i$. We propose a generalized semiparametric model with an unknown variance function (GSUV) as follows:

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} + f_1(z_{1i}) + \dots + f_m(z_{mi}), \quad (1)$$

$$\text{Var}(Y_i) = \mathcal{V}(\mu_i), \quad (2)$$

where $\mathbf{x}_i$ denotes covariates of length p with linear regression coefficients $\boldsymbol{\beta}$, and $\mathbf{z}_i = (z_{1i}, \dots, z_{mi})^T$ is a $m \times 1$ covariate vector with unspecified functional relations $f_j(\cdot)$, $j = 1, \dots, m$, with the mean. In model (1), the link $g(\cdot)$ is a known monotone and differentiable function, e.g., log link for medical costs. This is in contrast to the “unknown link” models, where the link function $g(\cdot)$ either is unspecified (nonparametric), e.g., Chiou and Muller [12] and Zhou et al. [14], or involves some unknown parameters to be estimated, e.g., Basu and Rathouz [13] of a Box-Cox transformation form. In model (2), the variance $\mathcal{V}(\cdot)$ is modeled as an unknown but smooth function of μ . For medical cost data and many other situations, the functional form of the error variance is of secondary importance - we are mostly interested in the mean structure. Note in this model we make no further assumption on the specific distribution (e.g., Gamma) of Y .

Our model is especially appealing when there exist different functional forms for different covariates on the impact of the mean medical cost, whereas in the “unknown link” models, all covariates are restricted to having a linear relation with $g(\mu)$. Therefore, the GSUV provides flexibility in modeling the nonlinear association between response and continuous covariates (e.g., age) by introducing the nonparametric components $f_j(\cdot)$. Also, as we use a known log link for the mean medical cost, the interpretation of covariate associations in model (1) is straightforward. For example, we can interpret the covariate associations of $\mathbf{x}_i$ (e.g., treatment) as the percentage change (e.g., $(e^{\boldsymbol{\beta}} - 1) \times 100\%$) in the mean cost for the linear parameter $\boldsymbol{\beta}$. In contrast, the interpretation of covariate associations in the “unknown link” models could be complicated since it depends on the unknown link function, which has to be estimated from data. For example, Basu and Rathouz [13] relied on marginal and incremental effects to study the impact of covariate on the mean cost, which involves additional complicated derivation.

Second, since we make no assumptions on the specific form of the variance structure, the GSUV is more robust in fitting data with heteroscedasticity. This adaptability is especially useful in health cost studies. An alternative approach is to model the dispersion as a semiparametric function of multiple covariates, e.g.,

$\text{dispersion} = \mathbf{x}_i^T \boldsymbol{\alpha} + h_1(z_{1i}) + \dots + h_m(z_{mi})$ [15, 16, 17, 18, 19]. These models are more appealing when one is interested in the associations between covariates and the variance structure. However, while there could exist multiple smooth functions for nonlinear covariate associations in the dispersion function in their approach, our model is simpler in that there is only a single smooth function (of μ) in model (2). Thus, our model can avoid variable selection and model averaging for the variance function in their framework (e.g., Yau and Kohn, [15]).

Remark

In our model, the focus is the mean structure (1). The variance function is assumed to be a function of the mean. In health economics literature, Park's test [20] is commonly used for medical cost studies, i.e., regressing the log of the estimated residual squared (on the scale of the analysis) on the raw (untransformed) scale for y , to estimate and test a very specific form of heteroscedasticity. Our formulation is thus an extension to Manning and Mullahy [9] and Basu and Rathouz [13].

3. Methods

3.1. Penalized Nonparametric Quasi-likelihood

Suppose that the components of the response vector $\mathbf{y}$ are independent with mean vector $\boldsymbol{\mu}$ and variance vector $\sigma^2 V(\boldsymbol{\mu})$, where σ^2 may be unknown and $V(\cdot)$ is a known function. The quasi-likelihood [7, 21] is

$$Q(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^n \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma^2 V(t)} dt.$$

For our GSUV, we assume an unknown variance function $\mathcal{V}(\mu)$ instead of $\sigma^2 V(\boldsymbol{\mu})$ in $Q(\boldsymbol{\mu}; \mathbf{y})$. Since the variance function is unknown, we replace it with a nonparametrically estimated variance function $\hat{\mathcal{V}}(\cdot)$. Following Chiou and Muller [12], we propose the nonparametric quasi-likelihood

$$\tilde{Q}(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^n \int_{y_i}^{\mu_i} \frac{y_i - t}{\hat{\mathcal{V}}(t)} dt.$$

The unknown univariate function $f_j(\cdot)$ in our GSUV can be estimated by a P-spline [22]. Assume that

$$f_i(t) = \tau_{j1}t + \dots + \tau_{jq}t^q + \sum_{k=1}^{K_j} \tau_{j(q+k)}(t - h_{jk})_+^q,$$

where $\{h_{jk}\}_{k=1}^{K_j}$ are spline knots. The basis

$\mathbf{B}_j(t) = \left(t, \dots, t^q, (t - h_{j1})_+^q, \dots, (t - h_{jK_j})_+^q \right)^T$ is known as the truncated power basis of degree q . We choose $q = 2$ in this paper. Define the spline coefficient vector

$\boldsymbol{\tau}_j = \left(\tau_{j1}, \dots, \tau_{j(q+K_j)} \right)^T$. Then the mean function of our GSUV is

$$g(\mu) = \mathbf{x}^T \boldsymbol{\beta} + \mathbf{B}_1 \boldsymbol{\tau}_1 + \dots + \mathbf{B}_m \boldsymbol{\tau}_m = \mathbf{x}_\theta,$$

where $\mathbf{X} = (\mathbf{x}, \mathbf{B}_1, \dots, \mathbf{B}_m)$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}_1^T, \dots, \boldsymbol{\tau}_m^T)^T$. We do not include intercept in the P-spline representation of the nonparametric function to avoid the problem of un-identifiability.

Given $\hat{\mathcal{V}}(\cdot)$, the estimates of can be obtained by maximizing the following penalized nonparametric quasi-likelihood

$$\tilde{Q}(\boldsymbol{\mu}; \mathbf{y}) - \frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\Sigma} \boldsymbol{\theta},$$

where $\boldsymbol{\Sigma} = \text{diag}(0_{p \times p}, \lambda_1 \mathbf{K}_1, \dots, \lambda_m \mathbf{K}_m)$. In other words, we have

$\frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\Sigma} \boldsymbol{\theta} = \sum_{j=1}^m \frac{\lambda_j}{2} \boldsymbol{\tau}_j^T \mathbf{K}_j \boldsymbol{\tau}_j$, where $\mathbf{K}_j$ is the j th (positive semidefinite) penalty matrix with unknown smoothing parameter λ_j . In this article, we use $\mathbf{K}_j = \text{diag}(\mathbf{0}_{q \times q}, \mathbf{I}_{K_j \times K_j})$.

The score function for the penalized nonparametric quasi-likelihood is

$$\sum_{i=1}^n \mathbf{D}_i^T \hat{\mathcal{V}}_i^{-1}(y_i - \mu_i) - \boldsymbol{\Sigma} \boldsymbol{\theta}, \quad (3)$$

where $\mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{B}_1 \boldsymbol{\tau}_1 + \dots + \mathbf{B}_m \boldsymbol{\tau}_m)$, $\hat{\mathcal{V}}_i = \hat{\mathcal{V}}(\mu_i)$, and $\mathbf{D}_i = \boldsymbol{\mu}_i' / \boldsymbol{\theta}$.

In our GSUV, we do not assume any specific form of the variance function $\mathcal{V}(\cdot)$, i.e., $\mathcal{V}(\cdot)$ is estimated from data. Define $\varepsilon_i^2 = (Y_i - \mu_i)^2$, we have $E[\varepsilon_i^2] = \mathcal{V}(\mu_i)$ based on model (2). Consequently, we can form a variance model as:

$$\varepsilon_i^2 = \mathcal{V}(\mu_i) + \varepsilon_i, \quad (4)$$

where ε_i is an error term with $E[\varepsilon_i] = 0$. Note that we should plug in the estimated values $\hat{\varepsilon}_i^2$ and $\hat{\mu}_i$, respectively, in (4) to estimate the unknown functional form in $\mathcal{V}(\cdot)$.

3.2. Estimation Procedure

To obtain the estimates of $\beta, f_j(\cdot)$, and $\mathcal{V}(\cdot)$, we develop the iterative estimation procedure as follows:

- Step 0** Initialize $\hat{\beta}$ and $\hat{\tau}_j$ which are estimated by quasi-likelihood using a known variance, e.g., $\mathcal{V}(\cdot)=1$.
- Step 1** Estimate $\mathcal{V}(\cdot)$ by minimizing the penalized least square

$$\sum_i \left\{ \hat{\varepsilon}_i^2 - \mathcal{V}(\hat{\mu}_i) \right\}^2 - J_\lambda(\mathcal{V}),$$

where $\hat{\varepsilon}_i^2 = (Y_i - \hat{\mu}_i)^2$, and $J_\lambda(\cdot)$ is a penalty function with smoothing parameter λ . Note here we have only 1 smooth function, which can be estimated by conventional nonparametric methods, e.g., P-spline with quadratic penalty $\mathbf{a}^T \mathbf{K} \mathbf{a}$ as in Wood [23], where $\mathbf{a}$ is a vector of spline coefficients, and $\mathbf{K}$ is similar the $\mathbf{K}_j$, one of the penalty matrices described above.

- Step 2** Given $\mathcal{V}(\cdot)$, estimate $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}_1^T, \dots, \boldsymbol{\tau}_m^T)^T$ by solving the penalized nonparametric quasi-score function:

$$\sum_{i=1}^n \mathbf{D}_i^T \hat{\mathcal{V}}_i^{-1} \left\{ y_i - g^{-1}(\mathbf{X}\boldsymbol{\theta}) \right\} - \boldsymbol{\Sigma}\boldsymbol{\theta} = 0.$$

Iterating between Steps 1 and 2 until convergence will yield the estimates. A similar algorithm was adopted in Chiou and Muller [12] for a quasi-likelihood regression with link and variance functions both being unknown. We note Step 0 provides consistent estimates of initial values for $\hat{\beta}$ and $\hat{\tau}_j$ when the link function is correct. As a result, we have not encountered the convergence issue in our simulation studies and application.

Estimates in Step 1 can be obtained using existing software, such as SAS Mixed Procedure [22]. We use gam() function with P-spline in the mgcv package of R [23]. To obtain the estimates in Step 2, we adopt the Newton-Raphson method with Fisher scoring. Given the estimate in the l th iterate, the $(l+1)$ th estimate is updated by:

$$\hat{\theta}_{l+1} = \hat{\theta}_l + \left\{ \sum_{i=1}^n \mathbf{D}_i^T(\hat{\theta}_l) \hat{\mathcal{V}}_i^{-1}(\hat{\theta}_l) \mathbf{D}_i(\hat{\theta}_l) + \boldsymbol{\Sigma} \right\}^{-1} \left\{ \sum_{i=1}^n \mathbf{D}_i^T(\hat{\theta}_l) \hat{\mathcal{V}}_i^{-1}(\hat{\theta}_l) (y_i - \hat{\mu}_i) - \boldsymbol{\Sigma} \hat{\theta}_l \right\}.$$

3.3. Smoothing Parameter and Knots Selections

Selecting appropriate smoothing parameters plays an important role in nonparametric smoothing. When using penalized splines, smoothing parameters control the trade-off between goodness of fit to the data and smoothness of the fitted splines. To implement the estimation procedure introduced in Section 3.2, we need to select two sets of smoothing parameters: the smoothing parameter for variance function $\mathcal{V}(\cdot)$; and the smoothing parameter for regression functions $f_j(t)$. We use gam() function with P-spline in mgcv package of R to estimate the variance function. Generalized cross validation (GCV) is used

for smoothing parameter selection in `gam()`, providing efficient automatic smoothing parameter selection [24]. However, GCV can not be directly applied for PNQL. Consequently, we concentrate on smoothing parameter selection for regression function $f_j(t)$ in model (1).

In Step 2, we use the approximate generalized cross-validation (AGCV) to select the optimal smoothing parameter for $f_j(t)$. We call it AGCV rather than GCV because it is derived based on the nonparametric quasi-likelihood. We define the AGCV score as:

$$AGCV = \frac{n \|\mathbf{W}^{1/2} \{\mathbf{y} - \mathbf{A}\mathbf{y}\}\|^2}{[n - \text{tr}\{\mathbf{A}\}]^2}$$

where $\mathbf{W} = \text{diag}[1/\{g'(\mu_i)^2 \mathcal{V}_i\}]$, and $\mathbf{A} = \mathbf{X} \left(\mathbf{X}^T \mathbf{W} \mathbf{X} + \sum_{j=1}^m \lambda_j \mathbf{K}_j \right)^{-1} \mathbf{X}^T \mathbf{W}$. See Appendix for the justification of AGCV in detail. The optimal smoothing parameter is found by minimizing the AGCV score over a grid of values of λ_j . Yu and Ruppert [25] used a 30-point grid where the values of $\log_{10}(\lambda_j)$ are equally spaced between -6 to 7 . Based our simulated and application data, we find that the 20-point grid from -6 to 3 is a reasonable choice.

Of note, the smoothing parameter selection is done separately in Steps 1 (for the variance function) and 2 (for the mean function). This is also the common practice for such type of models, e.g., Chiou and Muller [12], Rigby and Stasinopoulos [16], Gijbels et al. [19]. A simultaneous smoothing parameter selection procedure for both steps merits further consideration.

It is recommended [25] that the knots be placed at equally-spaced sample quantiles of the predictor variable. For example, if there are 9 knots, they could be placed at the 10th percentile, 20th percentile, ... of the values of predictor variable. Ruppert [26] has a detailed study on the choice of total number of knots. For smooth and either monotonic or unimodal regression functions, 10 to 20 knots are often adequate. If the regression function is less smooth somewhere, then it is important to place a knot near it.

3.4. Inference

Following McCullagh and Nelder [7] and incorporating the penalty matrix [27], we estimate the covariance matrix of $\hat{\boldsymbol{\theta}}$ as follows:

$$V(\hat{\boldsymbol{\theta}}) = \left(\sum_{i=1}^n \hat{\mathbf{D}}_i^T \hat{\mathcal{V}}_i^{-1} \hat{\mathbf{D}}_i + \boldsymbol{\Sigma} \right)^{-1} \left(\sum_{i=1}^n \hat{\mathbf{D}}_i^T \hat{\mathcal{V}}_i^{-1} (y_i - \hat{\mu}_i)^2 \hat{\mathcal{V}}_i^{-1} \hat{\mathbf{D}}_i \right) \left(\sum_{i=1}^n \hat{\mathbf{D}}_i^T \hat{\mathcal{V}}_i^{-1} \hat{\mathbf{D}}_i + \boldsymbol{\Sigma} \right)^{-1}. \quad (5)$$

This sandwich estimate can be used for joint confidence regions and for hypothesis testing, such as the Wald test. For example, if a test of the null hypothesis of $H_0: \mathbf{R}\boldsymbol{\theta} - \mathbf{q}_0 = 0$ is desired, where $\mathbf{R}$ is a $d_1 \times \dim(\boldsymbol{\theta})$ matrix of full rank $d_1 = \dim(\boldsymbol{\theta})$, then the test can be based on the Wald statistic,

$$(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{q}_0)^T (\mathbf{R}\mathbf{V}(\hat{\boldsymbol{\theta}})\mathbf{R}^T)^{-1} (\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{q}_0),$$

which has a chi-squared limiting distribution with d_1 degrees of freedom.

The estimated smoothing function is $\hat{f}_j(z) = \mathbf{B}_j \hat{\boldsymbol{\tau}}_j$. The variance of $\hat{f}_j(\cdot)$ can be estimated via $\hat{\text{Var}}\{\hat{f}_j(\cdot)\} = \mathbf{B}_j^T V(\hat{\boldsymbol{\theta}})_{\tau_j \tau_j} \mathbf{B}_j$, where $V(\hat{\boldsymbol{\theta}})_{\tau_j \tau_j}$ is the $(q + K_j) \times (q + K_j)$ block matrix of $V(\hat{\boldsymbol{\theta}})$.

4. Simulation

Simulation studies are conducted to assess the performance of our estimation method. Two settings are generated from gamma and overdispersed Poisson distributions, respectively. In both settings, we generate 1000 datasets, each with sample size $n = 200$. For illustrative purpose, we only consider one nonlinear function in simulation.

Example 1

The simulated data are analogous to medical cost data with a continuous response variable. The response Y is generated from a gamma distribution with the density function

$$\frac{1}{\gamma^\alpha \Gamma(\alpha)} y^{\alpha-1} \exp(-y/\gamma),$$

where $\alpha = \mu^3$, $\gamma = 1/\mu^2$, $E(Y) = \mu$, and $V(Y) = 1/\mu$. Note this variance is of the power variance form $(V(Y_i) = \theta_1 \mu_i^{\theta_2})$ in Basu and Rathouz [13]. We assume a log link with $\log(\mu) = \beta x + f(z)$, where $\beta = 1$. The continuous covariate Z is uniformly distributed over $[0, 1]$, and $f(z) = -0.5 + \sin(2\pi z)$. The linear covariate X follows distribution $N(0, 1)$.

Example 2

We simulate data with an overdispersed Poisson response variable to show the wide application of our method. The overdispersed Poisson data are generated via a Gamma-Poisson mixture. The log link is used: $\log(\mu) = \beta x + f(z)$, where $f(z) = 5 + \{\sin(2\pi z)\}/2$ and $\beta = 1$. Covariate Z is uniformly distributed over $[0, 1]$, and X follows distribution $N(0, 1)$. The underlying variance function is $V(Y) = \mu(1 + 6\mu)$. Note this variance is of the quadratic variance form $(V(Y_i) = \theta_1 \mu_i + \theta_2 \mu_i^2)$ in Basu and Rathouz [13].

We fit each dataset using three methods: (1) our PNQL method assuming an unknown variance function; (2) the PQL method assuming the correct parametric variance structure (termed ‘‘PQL-cv’’); and (3) the PQL method with a misspecified parametric variance structure (termed ‘‘PQL-icv’’): we assume $V(Y) = \mu$ for both gamma and overdispersed Poisson data. The PQL method with correct variance function is the best method (the gold standard) and serves as a criterion for assessing the performance of other methods.

The results of estimates of β s are shown in Table 1. For both gamma and overdispersed Poisson data, the gold standard - the PQL-cv performs the best. Our PNQL performs very similarly in terms of biases and standard deviation to the gold standard. However, the PQL-icv yields slightly larger biases and much higher variation. We also assess the performance of the asymptotic covariance formula (5). In Table 1, “SD” is the sampling standard deviation of 1000 coefficient estimates, “SE” is the average of the estimated standard errors, calculated by formula (5). We notice that the standard error formula works well for both gamma and overdispersed Poisson data.

Figure 2 shows point-wise biases, mean squared errors (MSE), and empirical coverage probabilities of estimated curves $\hat{f}(z)$ for the gamma data. It can be seen that the estimated curve using the PQL-icv shows larger bias compared with the other two, although all biases are relatively small in magnitude. The average coverage probabilities are 95.3%, 95.2%, and 97.7% for PQL-cv, PNQL, and PQL-icv, respectively. The same plots are shown in Figure 3 for the overdispersed Poisson data. We observe the similar pattern to Figure 2. The estimated curve using the PQL-icv shows larger bias compared with the other two, although still small in magnitude. The average coverage probabilities are 95.7%, 95.7%, and 97.2% for PQL-cv, PNQL, and PQL-icv, respectively. For both gamma and overdispersed Poisson data, the point-wise MSE and coverage probabilities obtained by our method are very close or even overlaps with the gold standard. However, the MSE of the PQL-icv is much larger than that of our method and the gold standard.

Based on the results, we notice that the PQL method with the misspecified variance function does not have substantial impact on the consistency of the estimates, but results in loss of efficiency in estimating both linear coefficients and smoothing functions. The performance of our PNQL is reasonably well compared with the PQL with the correct variance function. For a single dataset in our simulation studies, the implementation procedure runs around 50 seconds for a 20-point grid search of smoothing parameter (3.30GHz CPU, 16G RAM). The running time doubles for a 40-point grid search. We acknowledge a drawback of our method is the increased computational burden for the smoothing parameter selection for multiple nonparametric functions due to the grid search method introduced in Section 3.3.

We conduct more simulation studies for other types of outcome, e.g., binary data with a logistic link. The findings from these simulations are similar and thus ignored (available upon request).

Per reviewers’ suggestions, we compare our method to a standard method and a naive method. We use the `gam()` function in the `mgcv` package in R as the standard method (Wood [23]), adopting penalized splines which assume a parametric variance structure and a parametric distribution. The “family” option in the `gam()` function allows us to specify the parametric gamma and Poisson distributions (no overdispersed Poisson distribution available). The variance functions are restricted to be μ^2 and μ for gamma and poisson distributions respectively in the `gam()` function. We find out the `mgcv` estimates have relatively larger bias and standard deviation compared with the results of PQL-cv and PNQL. For example, the MSEs of PQL-cv, PNQL, PQL-icv, and `mgcv` estimates for β are 0.030, 0.034, 0.083, and 0.138 for the gamma data, and 0.902, 0.966, 1.609, and 1.498 for

the overdispersed poisson data, respectively. The estimated nonlinear function from mgcv has relatively larger point-wise biases and mean squared errors, and poorer coverage probabilities (results available upon request). The results are consistent with our expectations. First, mgcv performs much worse than the PQL-cv and PNQL methods when it misspecifies the variance function. Second, between PQL-icv and mgcv, we find that for the overdispersed Poisson data when PQL-icv and mgcv both misspecify the variance function as $V(Y) = \mu$, the result for mgcv is slightly better than that of PQL-icv since mgcv assumes a parametric distribution (Poisson), while PQL-icv makes no parametric assumption on the distributional form. While for gamma data, mgcv assumes $V(Y) = \mu^2$, which is further away from the truth ($V(Y) = 1/\mu$) than PQL-icv ($V(Y) = \mu$). As a result, the PQL-icv method yields much better result than mgcv.

For the naive method, we compare our method to GLM of Manning and Mullahy [9]. For Setting 1, we fit the data by the quasilielihood method [9] which uses Z , Z^2 and Z^3 (i.e., $\hat{\beta}_0 + \hat{\alpha}_0 z + \hat{\alpha}_1 z^2 + \hat{\alpha}_2 z^3$) to approximate $f(z)$. Such approximation, however, shows a poor performance (results available upon request), indicating that limitation of this method, especially when there exist complicated nonlinear functions.

5. Application

Heart failure is a chronic illness where the heart fails to deliver oxygen-rich blood to the body's other cells. It is the only cardiac disease with rising prevalence, affecting a total of over 5 million patients in America. Heart failure is also one of the most expensive health care problems in the U.S.. The economic burden for heart failure in 2010 is 39.2 billion U.S. dollars [28].

We apply our method to the analysis of annual medical costs for heart failure patients. The data are extracted from the CDR database from UVa Health System. Our dataset includes 1370 patients aged from 60 to 90 years whose first treatment of heart failure at UVa Health System was in 2004 (ICD9 diagnosis code 428.xx). The data consist of medical care costs incurred within one year follow-up. The summary of this cohort is given in Table 2. Medical costs are highly skewed to the right and heteroscedastic (median 9,298 U.S. dollars; mean 22,287 U.S. dollars).

GSUV with a log link function is used to fit the cost data. The response is annual costs in U.S. dollars. The covariate vector $\mathbf{x}_i$ in the linear term includes four predictors: gender (1 for male, 0 for female), white (1 for white, 0 for non-white), whether the patient received inpatient service within the 12-month study period (1 for yes, 0 for no), and death during the study period (1 for yes, 0 for no). The continuous predictor in the nonlinear term z_i is age.

The estimated coefficients for linear predictors are shown in Table 3. The standard errors for testing the significance of the coefficients are computed using formula (5). The association between gender and medical costs is not significant ($p=0.858$). The associations for “white”, “inpatient”, and “death” are significant. The mean annual medical cost for white patients is 20% less than other races ($p = 0.005$), indicating the possibility of racial disparity in heart failure. Such racial disparity is in agreement with a study by Jha et al. [29], showing that

black women less often received appropriate preventive therapy and adequate risk factor control despite a greater coronary heart disease event risk. Consequently, they were at a more severe disease stage when treated in a hospital, resulting in higher costs. Being hospitalized increases the annual medical cost by 3.9 times ($p < 0.0001$), while patients dead within the one year follow-up period spent 45% more in medical costs than those alive ($p = 0.045$).

The estimated curve for age is presented in the left panel of Figure 4. There is a nonlinear age effect: annual medical costs increase from age 60, reach a maximum value around age 70, then decrease and level off after age 85. This finding agrees well with Liu [30] who found the quadratic association between age and monthly medical cost of heart failure patients. An intuitive explanation is: (i) starting from age 60, medical costs increased since the physical conditions of the patients deteriorated as age increased; (ii) after age 70, older patients with heart diseases were often treated less aggressively, resulting in lower medical costs. For example, Gatsonis et al. [31] showed less frequent utilization of coronary angiography for elderly patients with an acute myocardial infarction. Stukel et al. [32, 33] showed that younger patients with heart diseases were more likely to receive invasive treatment and medical therapy. The plot on the right panel of Figure 4 is the estimated variance curve. It can be seen that the variance increases with the mean, indicating clear evidence of heteroscedasticity. The variance has a nonlinear relation with the mean which might not be described easily by a parametric (e.g., polynomial) model. Our method is flexible to capture the nonlinear association between response and predictors, and the dependence of the variance on the mean.

After seeing the left panel of Figure 4, we also fit the medical cost data using GSUV assuming a parametric quadratic effect of age in the mean function, with an unknown variance function. The results are shown in the right panel of Table 3. Clearly there is a strong indication for a quadratic association for age. It can be seen that the estimates of the other linear covariate associations are similar to those using GSUV with an unspecified function of age. Our model can thus be used as an exploratory tool to fit medical cost data more flexibly.

Per referee's suggestions, we fit the data assuming the variance is a linear function of the mean, $\text{Var}(Y_i) = 2\mu_i$, as is the case roughly in the right panel of Figure 4. We find the estimates of linear coefficients have larger standard errors compared to Table 3, which may be an indication of the inappropriate variance structure. The association for race becomes non significant (results available upon request). We also fit the data with additional interaction terms involving death, none of which is significant (results available upon request).

6. Discussion

Motivated by the study of medical costs for heart failure patients, we have proposed a generalized semiparametric model with an unknown variance function. This model allows some of the predictors to be modeled linearly, with others being modeled nonlinearly. We use the penalized nonparametric quasi-likelihood for estimation, and develop an AGCV

score for selecting optimal smoothing parameters. Assuming an unknown but smooth variance function provides flexibility and robustness in fitting various data as shown in the simulation studies. This is especially useful when dealing with the problems of heteroscedasticity and severe skewness that are common in medical cost data.

In this paper, we use the truncated power basis due to its ease to implement. Although there is some concern on the possible correlation in the truncated power basis, the penalty used in our research is equivalent to assuming those coefficients have independent normal distribution with mean zero and constant variance λ_j^{-1} for the j th nonparametric function. This penalty will stabilize the computation regardless of the correlation in the truncated power basis functions. In our future research, we are interested in working on other basis functions, e.g., B-spline and cubic smoothing spline.

There are two issues of model selection. In our simulation studies, we are interested in the selection of appropriate variance functional form. For instance, Manning and Mullahy [9] has discussion about specification and diagnosis for variance functions. Our method is an extension of PQL with correct variance function (PQL-cv), relying on the data to estimate the variance function. In other words, our method is more general while retaining the similar performance of PQL-cv. The other issue is the selection of covariates in the mean function (1). We have not addressed the problem of variable selection in this paper, it will be worth considering variable selection in future research.

As raised by a reviewer, in the Application study, “inpatient” can be endogenous. Unfortunately, the information available in the CDR database would not allow us to apply econometric methods, e.g., instrumental variables [34], to address the endogeneity issue. We caution that the regressions should be treated as estimating a conditional mean with no clear behavioral interpretations, but nevertheless, providing a useful illustration of the method in a real-world setting.

In this paper we are interested in modeling cross-sectional medical costs, e.g., annual medical costs. We have not considered the analysis of incomplete or censored medical costs, where there could exist an induced dependent censoring on medical costs. Many authors, e.g., [35, 36, 37, 38, 39, 40, 41], tackled this issue by the inverse probability of censoring weighting method proposed by Robins and Rotnitzky [42]. In our future research, we can adopt this approach with inverse weight of censoring probability to our model framework. There is also a growing interest in the temporal trend of longitudinal medical cost, e.g., monthly medical costs, see Liu and colleagues [30, 43, 44, 45] and Yabroff et al. [46]. As containing growing health care costs is the single most important fiscal issue in the U.S. [47], we expect considerable theoretical and numerical innovations in modeling medical costs in the coming years. These new developments will also help applied researchers for the practical problems encountered currently, such as the availability of easy-to-use software and public sharing data resources. The CDR database from UVa is available and subject to approval at <http://www.medicine.virginia.edu/clinical/departments/phs/informatics/cdr-page>, and the R code of our method is given as a supplemental material in the journal’s website.

A. Justification of AGCV Score

Estimates are obtained by maximizing the penalized nonparametric quasi-likelihood. Given the estimate $\hat{\theta}$ in the l th iterate, the $(l + 1)$ th estimate is updated by:

$$\hat{\theta}_{l+1} = \hat{\theta}_l + \left\{ \sum_{i=1}^n \mathbf{D}_i^T(\hat{\theta}_l) \hat{\mathcal{V}}_i^{-1}(\hat{\theta}_l) \mathbf{D}_i(\hat{\theta}_l) + \Sigma \right\}^{-1} \left\{ \sum_{i=1}^n \mathbf{D}_i^T(\hat{\theta}_l) \hat{\mathcal{V}}_i^{-1}(\hat{\theta}_l) (y_i - \hat{\mu}_i) - \Sigma \hat{\theta}_l \right\}. \quad (6)$$

Denote $\mathbf{Z} = \mathbf{X}\hat{\theta}$, and $\mathbf{D}_i = \mathbf{X}_i \mathbf{X}_i^T$, where $\mathbf{X}_i = \mu_i / \eta_i = 1/g'(\mu_i)$. Equation (6) can be represented as

$$\begin{aligned} \hat{\theta}_{l+1} &= \hat{\theta}_l + \left\{ \sum_{i=1}^n \mathbf{X}_i^T \Delta_i^T(\hat{\theta}_l) \hat{\mathcal{V}}_i^{-1}(\hat{\theta}_l) \Delta_i(\hat{\theta}_l) \mathbf{X}_i + \Sigma \right\}^{-1} \left\{ \sum_{i=1}^n \mathbf{X}_i^T \Delta_i^T(\hat{\theta}_l) \hat{\mathcal{V}}_i^{-1}(\hat{\theta}_l) (y_i - \hat{\mu}_i) - \Sigma \hat{\theta}_l \right\} \\ &= \hat{\theta}_l + \left\{ \sum_{i=1}^n \mathbf{X}_i^T W_i(\hat{\theta}_l) \mathbf{X}_i + \Sigma \right\}^{-1} \left\{ \sum_{i=1}^n \mathbf{X}_i^T W_i(\hat{\theta}_l) \Gamma_i + (\hat{\theta}_l)(y_i - \hat{\mu}_i) - \Sigma \hat{\theta}_l \right\} \end{aligned}$$

where $W_i = 1/\{g'(\mu_i)^2 V_i\}$, and $\Gamma_i = g'(\mu_i)$.

This is equivalent to finding $\hat{\theta}_{l+1}$ by solving the weighted penalized least squares problem

$$\text{minimize} \|\mathbf{W}^{1/2}(\mathbf{z}_l - \mathbf{X}\hat{\theta})\|^2 + \frac{1}{2} \hat{\theta}^T \Sigma \hat{\theta},$$

where $\mathbf{W} = \text{diag}(W_i)$, $\Gamma = \text{diag}(\Gamma_i)$, and $\mathbf{z}_l = \mathbf{X}\hat{\theta}_l + \Gamma(\hat{\theta}_l) \{y - \mu(\hat{\theta}_l)\}$.

Consequently, we propose our AGCV as

$$AGCV = \frac{n \|\mathbf{W}^{1/2} \{y - \mathbf{A}y\}\|^2}{[n - \text{tr}\{\mathbf{A}\}]^2}.$$

$\mathbf{A}$ is the influence (hat) matrix for the model, i.e. $\tilde{\mathbf{W}}\tilde{\eta} = \mathbf{A}(\tilde{\mathbf{W}}\tilde{\mathbf{z}})$, where $\tilde{\mathbf{W}}$ and $\tilde{\mathbf{z}}$ are evaluations of the last iterate.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the editor, the associate editor, and the three referees for valuable suggestions. We are grateful to Dr. Jason Lyman, Mr. Mac Dent and Mr. Ken Scully at clinical data repository of the University of Virginia for preparing the medical cost data. This research is partly supported by the NIAAA grant RC1 AA 019274 and AHRQ grant R01 HS020263.

References

1. National Institute for Health Care Management. Understanding US health care spending: NICHM Foundataion data brief July 2011. NIHCM Foundation; 2011. Available at <http://nihcm.org/images/stories/NIHCM-CostBrief-Email.pdf>
2. Obama, B. Released from the Office of the Press Secretary. The White House; 2009. Remarks by the president on reforming the health care system to reduce costs. Available at http://www.whitehouse.gov/the_press_office/Remarks-by-the-President-on-Reforming-the-Health-Care-System-to-Reduce-Costs/
3. Bang H, Tsiatis A. Median regression with censored cost data. *Biometrics*. 2002; 58:643–649. [PubMed: 12229999]
4. Duan N. Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*. 1983; 78:605–610.
5. Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics*. 1998; 17:283–295. [PubMed: 10180919]
6. Zhou XH, Melfi CA, Hui SL. Method for comparison of cost data. *Annals of Internal Medicine*. 1997; 127:752–756. [PubMed: 9382393]
7. McCullagh, P.; Nelder, JA. *Generalized Linear Models*. Chapman Hall; New York: 1989.
8. Blough DK, Madden CW, Hornbrook MC. Modeling risk using generalized linear models. *Journal of Health Economics*. 1999; 18:153–171. [PubMed: 10346351]
9. Manning WG, Mullahy J. Estimating log models: to transform or not to transform? *Journal of Health Economics*. 2001; 20:461–494. [PubMed: 11469231]
10. Manning WG, Basu A, Mullahy J. Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics*. 2005; 20:465–488. [PubMed: 15811539]
11. Green, PJ.; Silverman, BW. *Nonparametric Regression and Generalized Linear Models*. Chapman Hall; New York: 1994.
12. Chiou JM, Muller HG. Nonparametric Quasi-likelihood. *The Annals of Statistics*. 1999; 27:36–64.
13. Basu A, Rathous PJ. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics*. 2005; 6:93–109. [PubMed: 15618530]
14. Zhou XH, Lin H, Johnson E. Non-parametric heteroscedastic transformation regression models for skewed data with an application to health care costs. *Journal of Royal Statistical Society, Series B*. 2009; 70:1029–1047.
15. Yau P, Kohn R. Estimation and variable selection in nonparametric heteroscedastic regression. *Statistics and Computing*. 2003; 13:191–208.
16. Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. *Applied Statistics*. 2005; 54:507–554.
17. Nott D. Semiparametric estimation of mean and variance functions for non-Gaussian data. *Computational Statistics*. 2006; 21:603–620.
18. Leng C, Zhang W, Pan J. Semiparametric mean-covariance regression analysis for longitudinal data. *Journal of the American Statistical Association*. 2010; 105:181–193.
19. Gijbels I, Prosdociimi I, Claeskens G. Nonparametric estimation of mean and dispersion functions in extended generalized linear models. *Test*. 2010; 19:580–608.
20. Park R. Estimation with heteroscedastic error terms. *Econometrica*. 1966; 34:888.
21. Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*. 1974; 61:439–447.
22. Ruppert, D.; Wand, MP.; Carroll, RJ. *Semiparametric Regression*. Cambridge; New York: 2003.
23. Wood SN. Modeling and smoothing parameter estimation with multiple quadratic penalties. *Journal of Royal Statistical Society, series B*. 2000; 62:413–428.
24. Wood, SN. *Generalized Linear Models: An Introduction with R*. Chapman & Hall; New York: 2006.
25. Yu Y, Ruppert D. Penalized spline estimation for partial linear single-index models. *Journal of the American Statistical Association*. 2002; 97:1042–1054.

26. Ruppert D. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*. 2002; 11:735–757.
27. Lin X, Zhang D. Inference in generalized additive mixed models by using smoothing splines. *Journal of Royal Statistical Society, Series B*. 1999; 61:381–400.
28. Lloyd-Jones D, Adams RJ, Brown TM, et al. Heart Disease and Stroke Statistics - 2010 Update. A Report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation*. 2010; 121:e1–e170. [PubMed: 20048228]
29. Jha AK, Varosy PD, Kanaya AM, Hunninghake DB, Hlatky MA, Waters DD, Furberg CD, Shlipak MG. Differences in medical care and disease outcomes among black and white women with heart disease. *Circulation*. 2003; 18:1089–1094. [PubMed: 12939228]
30. Liu L. Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Statistics in Medicine*. 2009; 28:972–986. [PubMed: 19040210]
31. Gatsonis C, Epstein AM, Newhouse JP, Normand SL, McNeil BJ. Variations in the utilization of coronary angiography for elderly patients with an acute myocardial infarction: an analysis using hierarchical logistic regression. *Medical Care*. 1995; 33:625–642. [PubMed: 7760578]
32. Stukel TA, Lucas FL, Wennberg DE. Long-term outcomes of regional variations in intensity of invasive vs medical management of Medicare Patients with acute myocardial infarction. *Journal of the American Medical Association*. 2005; 293:1329–1337. [PubMed: 15769966]
33. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *Journal of the American Medical Association*. 2007; 297:278–285. [PubMed: 17227979]
34. Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics*. 2008; 27:531–43. [PubMed: 18192044]
35. Bang H, Tsiatis A. Estimating medical costs with censored data. *Biometrika*. 2000; 87:329–343.
36. Lin DY. Proportional means regression for censored medical costs. *Biometrics*. 2000; 56:775–778. [PubMed: 10985215]
37. Lin DY, Feuer EJ, Etzioni R, Wax Y. Estimating medical costs from incomplete follow-up data. *Biometrics*. 1997; 53:419–434. [PubMed: 9192444]
38. Lin DY, Ying Z. Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *Journal of the American Statistical Association*. 2001; 96:103–126.
39. Lin DY. Regression analysis of incomplete medical cost data. *Statistics in Medicine*. 2003; 22:1181–1200. [PubMed: 12652561]
40. Gardiner JC, Luo ZH, Bradley CJ, Sirbu CA, Given CW. A dynamic model for estimating changes in health status and costs. *Statistics in Medicine*. 2006; 25:3648–3667. [PubMed: 16416401]
41. Zhao HW, Bang H, Wang H, Pfeifer PE. On the equivalence of some medical cost estimators with censored data. *Statistics in Medicine*. 2007; 26:4520–4530. [PubMed: 17380543]
42. Robins, J.; Rotnitzky, A. *AIDS Epidemiology-Methodological Issues*. Boston: Birkhauser; 1992. Recovery of information and adjustment for dependent censoring using surrogate markers; p. 297–331.
43. Liu L, Wolfe RA, Kalbfleisch JD. A shared random effects model for censored medical costs and mortality. *Statistics in Medicine*. 2007; 26:139–155. [PubMed: 16526006]
44. Liu L, Conaway MR, Knaus WA, Bergin J. A random effects four-part model, with application to correlated medical costs. *Computational Statistics and Data Analysis*. 2008; 52:4458–4473.
45. Liu L, Huang XL, O'Quigley J. Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics*. 2008; 64:950–958. [PubMed: 18162110]
46. Yabroff KR, Warren JL, Schrag D, Angela M, Marie T, Martin BL. Comparison of approaches for estimating incidence costs of care for colorectal cancer patients. *Medical Care*. 2009; 47(Suppl 1):S56–S63. [PubMed: 19536010]
47. Orszag, P. Medicare trustees to America: Bend the curve! Statements released from the Office of Management and Budget. The White House; 2009. Available at <http://www.whitehouse.gov/omb/blog/09/05/12/MedicareTrusteestoAmericaBendtheCurve/>

Figure 1. Histogram of Annual Medical Costs (in 10,000 U.S. dollars) for Heart Failure Patients in the UVa Health System. The solid line is the estimated density.

Figure 2.

Simulation Results of Estimated Functions for Gamma Data. Left: plot of point-wise bias; Middle: plot of point-wise mean squared error; Right: plot of point-wise empirical coverage probabilities. PNQL estimate: dash dotted line; PQL estimate with the correct variance function: dashed line; PQL estimate with a misspecified variance function: dotted line.

Figure 3.

Simulation Results of Estimated Functions for Overdispersed Poisson Data. Left: plot of point-wise bias; Middle: plot of point-wise mean squared error; Right: plot of point-wise empirical coverage probabilities. PNQL estimate: dash dotted line; PQL estimate with the correct variance function: dashed line; PQL estimate with a misspecified variance function: dotted line.

Figure 4.

Curve Estimation for UVa Heart Failure Data. The plot on the left is the estimated curve for age (in years) with 95% point-wise confidence interval. The plot on the right is the estimated variance function. “mu” represents the mean medical cost, in the unit of 10,000 U.S. dollars.

Table 1

Estimates of Linear Coefficients in the Simulation Studies. PQL-cv: PQL method assuming the correct parametric variance structure; PNQL: penalized nonparametric quasi-likelihood method assuming an unknown variance function; PQL-icv: PQL method with a misspecified parametric variance structure; SD: sampling standard deviation of estimates; SE: average of the estimated standard errors; CP: empirical coverage probability.

	Gamma Data				Overdispersed Poisson Data					
	$\mathcal{V}(\mu)$	Bias	SD	SE	CP	$\mathcal{V}(\mu)$	Bias	SD	SE	CP
Method										
PQL-cv	$1/\mu$	0.009	0.173	0.166	94.9%	$\mu(1 + 6\mu)$	-0.055	0.948	0.907	95.2%
PNQL	$\sigma^2(\mu)$	0.005	0.184	0.174	95.3%	$\sigma^2(\mu)$	-0.062	0.981	0.948	95.9%
PQL-icv	μ	0.047	0.285	0.269	96.9%	μ	-0.094	1.265	1.246	97.1%

Table 2

Summary of the Heart Failure Data

Covariate	Mean (Percent)	SD
Age (years)	72.2	7.7
Male	54.2%	
White	73.6%	
Inpatient	37.7 %	
Death	5.9 %	
Cost (U.S. dollar)	22,287	37,630

Table 3

Estimates for UVa Medical Cost Data using GSUV with Linear Covariates: Gender, White, Inpatient, and Death. GSUV(N): GSUV with an unknown function of age; GSUV(Q): GSUV with a parametric quadratic association for age.

	GSUV(N)			GSUV(Q)		
	Estimate	SE	P-value	Estimate	SE	P-value
Gender	0.013	0.074	0.858	0.011	0.074	0.892
White	-0.226	0.080	0.005	-0.239	0.078	0.002
Inpatient	1.360	0.072	< 0.001	1.356	0.071	< 0.001
Death	0.371	0.185	0.045	0.383	0.170	0.025
Age	—	—	—	1.486	0.581	0.010
Age ²	—	—	—	-1.756	0.655	0.007