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

Lei Liu

Department of Preventive Medicine
Northwestern University

Joint work with Jinsong Chen, Daowen Zhang, and Ya-Chen T. Shih
Outline

- Introduction
- Model and Estimation
- Simulation
- Application
- Discussion
Medical Cost Data

• Medical cost data have been collected routinely by hospitals, government agencies, and health insurance companies.

• Modeling medical costs is of great interest in health economics study.

• Goal: to identify the risk factors of medical costs and ascertain the most cost-effective treatment, which in turn, can assist policy makers in maximizing health benefits for individuals and society.
Rising Medical Costs

- Medical costs rise rapidly: Health care costs were projected to be $8,160 per person in 2009, $13,100 per person in 2018 (Department of Health and Human Services, February 24, 2009).

- President Obama believes that out-of-control costs are the main obstacle to securing medical coverage for all.

- Peter Orszag, director of the Office of Management and Budget, stated that “reducing the growth rate of health care costs is the single most important fiscal issue we face as a country”.
A Motivating Example

- Annual medical costs for heart failure patients
  - The only cardiac disease growing in prevalence, with 670,000 new patients diagnosed each year. Totally 5.7 million in USA.
  - It is the leading cause of hospitalization among people 65 and older in the United States.
  - One of the most expensive health care problems in the U.S.: heart diseases rank No. 1 in medical spending (NIHCM July 2011).

- Data from the clinical data repository (CDR) for the University of Virginia (UVa) Health System

- 1370 patients aged from 60 to 90 and treated first in 2004 with heart failure (ICD9 diagnosis code beginning with 428)

- Research interest: study the association between medical costs and predictors, including gender, race, hospitalization, and age
Figure 1: Histogram of Annual Medical Costs (in $10,000) for Heart Failure Patients in UVa Health System. Mean: $22,287; Median: $9,298; SD: $37,630; Max: $694,004.
Medical Cost Data

• Estimate of mean response may be quite sensitive to heteroscedasticity and severe skewness

• Conventional statistical methods for medical cost data: regress log cost on covariates
  – Re-transformation issue for log costs - from $E \log(Y)$ to $E(Y)$
  – Smearing estimate (Duan 1983): not appropriate with heteroscedasticity

• Generalized linear models, e.g., gamma distribution with log link (Manning 1998; Blough, Madden, and Hornbrook 1999; Manning and Mullahy 2001; Manning, Basu, and Mullahy 2005)
  – Parametric distribution for costs
  – Parametric covariate effects
Our Model

- Generalized semiparametric model with unknown variance function

\[
g(\mu_i) = x_i^T \beta + f_1(z_{1i}) + \ldots + f_m(z_{mi})
\]

\[
\text{Var}(Y_i) = \mathcal{V}(\mu_i)
\]

- \( \mu_i = E(Y_i) \): the medical cost for the \( i \)th subject
- \( g(\cdot) \): a known link function, \( g(\cdot) = \log(\cdot) \) for medical cost
  * Can correct the right skewness and heteroscedasticity
  * Easy to interpret the covariate effect: \( \exp(\beta) - 1 \) is the percentage change of medical cost for a unit change in \( x \).
- \( x_i \): linear covariates of length \( p \)
- \( z_i = (z_{1i}, \ldots, z_{mi})^T \): a \( m \times 1 \) vector of continuous variables
- \( f_j(\cdot) \): an unknown function
- \( \mathcal{V}(\cdot) \): an unknown but smooth function of mean
Advantages of Our Model

- Model $E(Y)$ instead of $E \log(Y)$, avoid retransformation
- Very flexible
  - Different functional forms for different covariates on the impact of the mean medical cost: linear vs. nonlinear
  - Not assuming any specific form of the variance structure: robust in fitting data with heteroscedasticity
  - No further assumption on any specific distribution (e.g., Gamma) of response variable
Estimate Unknown Functions via Penalized splines

- Linear combination of basis and coefficients

\[ f_j(z) = \tau_{j0} + \tau_{j1}z + \ldots + \tau_{jq}z^q + \sum_{k=1}^{K_j} \tau_{j(q+k)}(z - h_{jk})_+^q = B_j \tau_j \]  

- \( B_j(z) = (1, z, \ldots, z^q, (z - h_{j1})_+^q, \ldots, (z - h_{jK_j})_+^q)^T \), where

\[ x_+ = \begin{cases} 
  x & \text{if } x > 0, \\
  0 & \text{otherwise.} 
\end{cases} \]

- \( \{h_{jk}\}_{k=1}^{K_j} \) are spline knots
- \( (z - h_{kj})_+^q \) piecewise polynomial: useful when the data have some special feature locally
- \( \tau_j = (\tau_{j0}, \tau_{j1}, \ldots, \tau_{j(q+K_j)})^T \)
- We use \( q = 2 \), quadratic splines
Matrix form of the mean function

\[ g(\mu) = \mathbf{x}^T \beta + \mathbf{B}_1 \mathbf{\tau}_1 + \ldots + \mathbf{B}_m \mathbf{\tau}_m = \mathbf{X} \theta \]

- \( \mathbf{X} = (\mathbf{x}, \mathbf{B}_1, \ldots, \mathbf{B}_m) \)
- \( \theta = (\beta^T, \mathbf{\tau}_1^T, \ldots, \mathbf{\tau}_m^T)^T \)
Choosing knots

• The knots can be placed at equally-spaced sample quantiles of the continuous predictor variable (Yu and Ruppert, 2002).
  – Example: if there are 9 knots, then they would be at the 10th percentile, 20th percentile, ... of the values of predictor variable.

• More knots, better model fitting, but less smooth (more wiggly).
Penalized splines

- Add a penalty to control the trade-off between fidelity to the data and smoothness of the fitted spline

\[ L - \sum_{j=1}^{m} \lambda_j \theta^T K_j \theta \]  \hspace{1cm} (4)

- \( K_j \): penalty matrix, e.g., identity matrix
- \( \lambda_j \): a smoothing parameter selected by generalized cross validation (GCV) score (Craven and Wahba 1979)
A similar matrix form can be obtained for the variance function $V(\cdot)$.

Can be nonparametrically estimated from a penalized least square

$$\sum_i \{\hat{\varepsilon}_i^2 - V(\hat{\mu}_i)\}^2 - J(V)$$

where $\hat{\varepsilon}_i^2 = (Y_i - \hat{\mu}_i)^2$, and $J(\cdot)$ is a penalty function.
Quasi-likelihood

- Quasi-likelihood (Wedderburn 1974; McCullagh and Nelder 1989)

\[ Q(\mu; y) = \sum_{i=1}^{n} \int_{y}^{\mu} \frac{Y_i - t}{\sigma^2V(t)} dt, \]

where \( V(t) \) is known.

- Instead of specifying a probability distribution for the data, only a relationship between the mean and the variance is specified in \( V(\cdot) \)

- Quasi-likelihood is not the true likelihood - it may not correspond to any known probability distribution

- Often used to allow over-dispersion.
Nonparametric Penalized Quasi-likelihood (NPQL)

- Nonparametric quasi-likelihood for mean

\[ \tilde{Q}(\mu; y) = \sum_{i=1}^{n} \int_{y}^{\mu} \frac{Y_i - t}{\hat{V}(t)} dt, \]  

(5)

- Given \( \hat{V}(\cdot) \), nonparametric penalized quasi-likelihood for \( f(\cdot) \)

\[ \tilde{Q}(\mu; y) - \sum_{j=1}^{m} \lambda_j \theta^T K_j \theta \]  

(6)

- Score function for \( \theta \)

\[ \sum_{i=1}^{n} D_i^T \hat{V}_i^{-1}(Y_i - \mu_i) - \sum_{j=1}^{m} \lambda_j K_j \theta \]  

(7)
Estimation Procedure

- **Step 0**: Initial values of $\hat{\beta}$ and $\hat{\tau}_j$ are estimated by quasi-likelihood using constant variance, $\mathcal{V}(\cdot) = 1$.

- **Step 1**: Estimate $\mathcal{V}(\cdot)$ by minimizing the penalized least square (e.g., using package mgcv in R)

$$\sum_i \{\hat{\varepsilon}_i^2 - \mathcal{V}(\hat{\mu}_i)\}^2 - J(\mathcal{V})$$

- **Step 2**: Given $\hat{\mathcal{V}}(\cdot)$, estimate $\beta$ and $\tau_j$ by solving the nonparametric penalized quasi-score function:

$$\sum_{i=1}^n D_i^T \hat{\mathcal{V}}^{-1}_i \{y_i - \mu(x_i^T \beta + B_1 \tau_1 + \ldots + B_m \tau_m)\} - \sum_{j=1}^m \lambda_j K_j \theta = 0.$$ 

- **Step 3**: Iterate between Step 1 and Step 2 till convergence
• Covariance matrix of $\hat{\theta}$

$$V(\theta) = \left(\sum_{i=1}^{n} D_i^T \hat{\gamma}_i^{-1} D_i + \sum_{j=1}^{m} \lambda_j K_j \right)^{-1} \left(\sum_{i=1}^{n} D_i^T \hat{\gamma}_i^{-1} (y_i - \hat{\mu}_i)^2 \hat{\gamma}_i^{-1} D_i \right) \left(\sum_{i=1}^{n} D_i^T \hat{\gamma}_i^{-1} D_i + \sum_{j=1}^{m} \lambda_j K_j \right)^{-1}.$$ 

• Variance of $\hat{f}_j(\cdot)$

$$\hat{\text{Var}}\{\hat{f}_j(\cdot)\} = B_j^T V(\hat{\theta}) \tau_j \tau_j B_j$$

where $V(\hat{\theta}) \tau_j \tau_j$ is the $(1 + q + K_j) \times (1 + q + K_j)$ block matrix of $V(\hat{\theta})$. 
Simulation: Example 1

• Gamma data

\[ f(y) = \frac{1}{\lambda^\alpha \Gamma(\alpha)} y^{\alpha-1} \exp(-y/\lambda) \]

• \( \alpha = \mu^3, \lambda = 1/\mu^2, \ E(Y) = \alpha \lambda = \mu \)

• Log link: \( \log(\mu) = \beta_0 + \beta_1 x + f(z) \)
  - \( \beta_0 = 0, \beta_1 = 1 \)
  - \( f(z) = \sin(2\pi z) + 1 \)
  - \( Z \sim \text{unif}[0, 1] \)
  - \( X \sim \text{N}(0, 1) \)

• \( \text{Var}(Y) = \alpha \lambda^2 = 1/\mu \)
Simulation: Example 2

- Overdispersed Poisson generated via Gamma-Poisson mixture.

- Log link: \( \log(\mu) = \beta_0 + \beta_1 x + f(z) \)
  - \( \beta_0 = 0, \beta_1 = 1 \)
  - \( f(z) = \{\sin(2\pi z)\}/2 + 5 \)
  - \( Z \sim \text{unif}[0, 1] \)
  - \( X \sim N(0, 1) \)

- \( \text{Var}(Y) = \mu(1 + 6\mu) \)

- 1000 replicates, each with sample size 200.
Simulation: Fitting Methods

- All three methods have the same mean function \( \log(\mu) = \beta_0 + \beta_1 x + f(z) \)

- PQL method assuming correct parametric variance structure: PQL-cv (gold standard)
  - Gamma data: \( \text{Var}(Y) = 1/\mu \)
  - Overdispersed Poisson data: \( \text{Var}(Y) = \mu(1 + 6\mu) \)

- Our method: NPQL method assuming unknown variance function

- PQL method with misspecified parametric variance structure: PQL-icv
  - Gamma data: \( \text{Var}(Y) = \mu \)
  - Overdispersed Poisson data: \( \text{Var}(Y) = \mu \)
## Simulation Results

### Table 1: Estimates of Linear Coefficients in the Simulation Studies

<table>
<thead>
<tr>
<th>Methods</th>
<th>(\nu(\mu))</th>
<th>Bias</th>
<th>SD</th>
<th>SE</th>
<th>CP</th>
<th>(\nu(\mu))</th>
<th>Bias</th>
<th>SD</th>
<th>SE</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta_0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PQL-cv</td>
<td>(1/\mu)</td>
<td>0.014</td>
<td>0.201</td>
<td>0.189</td>
<td>95.3%</td>
<td>(\mu(1 + 6\mu))</td>
<td>0.053</td>
<td>0.959</td>
<td>0.925</td>
<td>94.8%</td>
</tr>
<tr>
<td>NPQL</td>
<td>(\sigma^2(\mu))</td>
<td>0.012</td>
<td>0.208</td>
<td>0.194</td>
<td>95.1%</td>
<td>(\sigma^2(\mu))</td>
<td>0.051</td>
<td>0.993</td>
<td>0.961</td>
<td>95.7%</td>
</tr>
<tr>
<td>PQL-icv</td>
<td>(\mu)</td>
<td>0.069</td>
<td>0.298</td>
<td>0.280</td>
<td>96.5%</td>
<td>(\mu)</td>
<td>0.059</td>
<td>1.268</td>
<td>1.257</td>
<td>97.3%</td>
</tr>
<tr>
<td>(\beta_1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PQL-cv</td>
<td>(1/\mu)</td>
<td>0.009</td>
<td>0.173</td>
<td>0.166</td>
<td>94.9%</td>
<td>(\mu(1 + 6\mu))</td>
<td>-0.055</td>
<td>0.948</td>
<td>0.907</td>
<td>95.2%</td>
</tr>
<tr>
<td>NPQL</td>
<td>(\sigma^2(\mu))</td>
<td>0.005</td>
<td>0.184</td>
<td>0.174</td>
<td>95.3%</td>
<td>(\sigma^2(\mu))</td>
<td>-0.062</td>
<td>0.981</td>
<td>0.948</td>
<td>95.9%</td>
</tr>
<tr>
<td>PQL-icv</td>
<td>(\mu)</td>
<td>0.047</td>
<td>0.285</td>
<td>0.269</td>
<td>96.9%</td>
<td>(\mu)</td>
<td>-0.094</td>
<td>1.265</td>
<td>1.246</td>
<td>97.1%</td>
</tr>
</tbody>
</table>
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. NPQL estimate: dashed line; PQL estimate with the correct variance function: dash dotted 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. NPQL estimate: dashed line; PQL estimate with the correct variance function: dash dotted line; PQL estimate with a misspecified variance function: dotted line.
Other distributions

- We also fit data from other distribution, e.g., binomial, with logistic link, and the results are similar.
What we have learned?

• The performance of our NPQL compares well with the PQL with correct variance function.

• Mis-specification of the variance structure mostly affects efficiency, not consistency:
  – PQL method with misspecified variance function does not have substantial impact on the consistency of the estimates
  – But there is loss of efficiency in estimating both linear coefficients and smoothing functions.
Application

• Annual medical costs for heart failure patients in the clinical data repository (CDR) at University of Virginia (UVa) Health System

• Patients over 60 years old who were first diagnosed and treated in 2004 with heart failure, with at least one year follow-up.

• We exclude patients who died within one year because the high end-of-life cost could complicate the comparison.

• A total of 1370 patients.

• Outcome: UVa health system costs (actual monetary expenses of the hospital) incurred within one year follow-up.
Table 2: Summary of the Heart Failure Data

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Mean (Percent)</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>72.2</td>
<td>7.7</td>
</tr>
<tr>
<td>Male</td>
<td>54.2%</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>73.6%</td>
<td></td>
</tr>
<tr>
<td>Inpatient</td>
<td>37.7 %</td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td>$22,287</td>
<td>$37,630</td>
</tr>
</tbody>
</table>
Our model

\[ \log(\mu_i) = \beta_0 + \beta_1 \text{gender} + \beta_2 \text{white} + \beta_3 \text{inpatient} + f(\text{age}) \quad (8) \]

\[ \text{Var}(Y_i) = \mathcal{V}(\mu_i) \quad (9) \]
Table 3: Estimates for UVa Medical Cost Data. Left: model with an unknown function of age; Right: model with a parametric quadratic effect of age.

<table>
<thead>
<tr>
<th></th>
<th>Unknown age function</th>
<th>Quadratic age function</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>s.e.</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.002</td>
<td>0.006</td>
</tr>
<tr>
<td>White</td>
<td>-0.209</td>
<td>0.007</td>
</tr>
<tr>
<td>Inpatient</td>
<td>1.386</td>
<td>0.006</td>
</tr>
<tr>
<td>Age</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Age²</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
Interpretation of parametric coefficients

• The mean annual medical cost for white patients is 20% less than other races, indicating the possibility of racial disparity in this heart failure cohort.

• Being hospitalized increases the annual medical cost by 4 times.

• No gender difference
Figure 4: Curve Estimation for UVa Heart Failure Data. The plot on the left is the estimated curve for age with 95% point-wise confidence interval. The plot on the right is the estimated variance function.
Curvature in Age Effect

• Starting from age 60, the medical costs increase since the physical conditions of patients deteriorated as age increases.

• After age 70, older patients with heart failure were often treated less aggressively, resulting in lower medical costs.
  – Stukel et al. (2005, 2007): younger patients with heart diseases were more likely to receive invasive treatment and medical therapy.
Discussion

• Comparison to Alternatives I:
  – Same model for the mean
  – Model the dispersion as a semiparametric function of multiple covariates, e.g., Yau and Kohn (2003), Rigby and Stasinopoulos (2005), Nott (2006), Leng et al. (2010), and Gijbels et al. (2010).

\[
\text{dispersion} = \mathbf{x}_i^T \boldsymbol{\alpha} + g_1(z_{1i}) + \ldots + g_m(z_{mi})
\]

  – Our model is simpler in that there is only a single smooth function (of \(\mu\)) in the variance function. Thus, we can avoid the variable selection and model averaging for the variance function (e.g., Yau and Kohn, 2003).
• Comparison to Alternatives II (Chiou and Muller, 1999):
  – Unknown link function $h(\cdot)$ for mean and unknown variance function $\mathcal{V}(\mu)$.
    $h(\mu_i) = \mathbf{x}_i^T \beta$
    $\text{Var}(Y_i) = \mathcal{V}(\mu_i)$
  – Interpretation of $\beta$ depends on the unknown link function $h(\cdot)$
  – All covariates are restricted to having the same functional relation
Comparison to Alternatives III: Extended estimating equations in GLM (Basu and Rathouz, 2005)

- Box-Cox transformation for link function \( v = x_i^T \beta \), where

\[
v = \begin{cases} 
  \omega^{-1}(\mu^\omega - 1) & \omega \neq 0, \\
  \log \mu & \omega = 0.
\end{cases}
\]

- Power variance form: \( \text{Var}(Y_i) = \theta_1 \mu_i^{\theta_2} \), or quadratic variance form: \( \text{Var}(Y_i) = \theta_1 \mu_i + \theta_2 \mu_i^2 \).

- Interpretation of \( \beta \) depends on the parameter \( \omega \) in the Box-Cox transformation

- The variance function is not flexible enough.
Other Topics in Medical Cost Analysis


- Liu and colleagues (2008a, 2008b, 2009) and Yabroff et al. (2009) were interested in the temporal trend of longitudinal medical costs, e.g., monthly medical costs.
  - Also considered joint models of longitudinal medical costs and survival
  - Particularly attractive in cost effectiveness study: when both costs and survival are of interest simultaneously (Pullenayegum and Willan 2007).

- Methods to handle the end-of-life cost have been developed by Stearns and Norton (2004), Liu, Wolfe and Kalbfleisch (2007), and Chan and Wang (2010).
Ongoing work

- Analysis of medical costs and social costs of the COMBINE alcohol treatment trial, collaborating with health economists at RTI
- Extend our model to longitudinal medical cost data
Collaboration between Health Economists and Statisticians

• Organized invited sessions in JSM and ENAR - Analysis of Medical Cost Data: Joint Venture between Health Economists and Statisticians.

• Topic contributed session in JSM 2012

• R01 grant proposal on innovative methods in longitudinal medical cost data (PIs: Liu and Shih)
  – Joint with health economists Drs. Tina Shih and Anirban Basu.
  – Very positive review for 1st submission: Percentile 2.0. Funded Sept 2011!!!
Acknowledgements

• 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 AA019274 (PIs: Liu, Johnson and O’Quigley) and AHRQ grant R01 HS020263 (PIs: Liu and Shih).
References

- Supplemental issue of Medical Care 47, No. 7, Supp 1, 2009, and references therein.


• Liu et al. (2012). *Statistical Methods in Medical Research* in press.


• Tooze et al. (2002). *Statistical Methods in Medical Research* 11, 341-355


Thank you!
Selection of smoothing parameter

- Minimize approximate generalized cross validation (AGCV) score for quasi-likelihood

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

- \( W = \text{diag}(1/V_i) \)
- \( A = X(X^T W X + \sum_{j=1}^{m} \lambda_j K_j)^{-1} X^T W \)
- \( A \) is the influence (hat) matrix for the model, i.e. \( \hat{\mu} = Ay \).
- For example, we use the 20-point grid where the values of \( \log_{10}(\lambda_j) \) is equally spaced between -6 and 3.