Semiparametric Models for Extreme-Value Distributed Biomarkers with Measurement Error

Donglin Zeng

Department of *Biostatistics*

University of North Carolina

Joint work with Noorie Hyun and David Couper

August, 2016

Outline

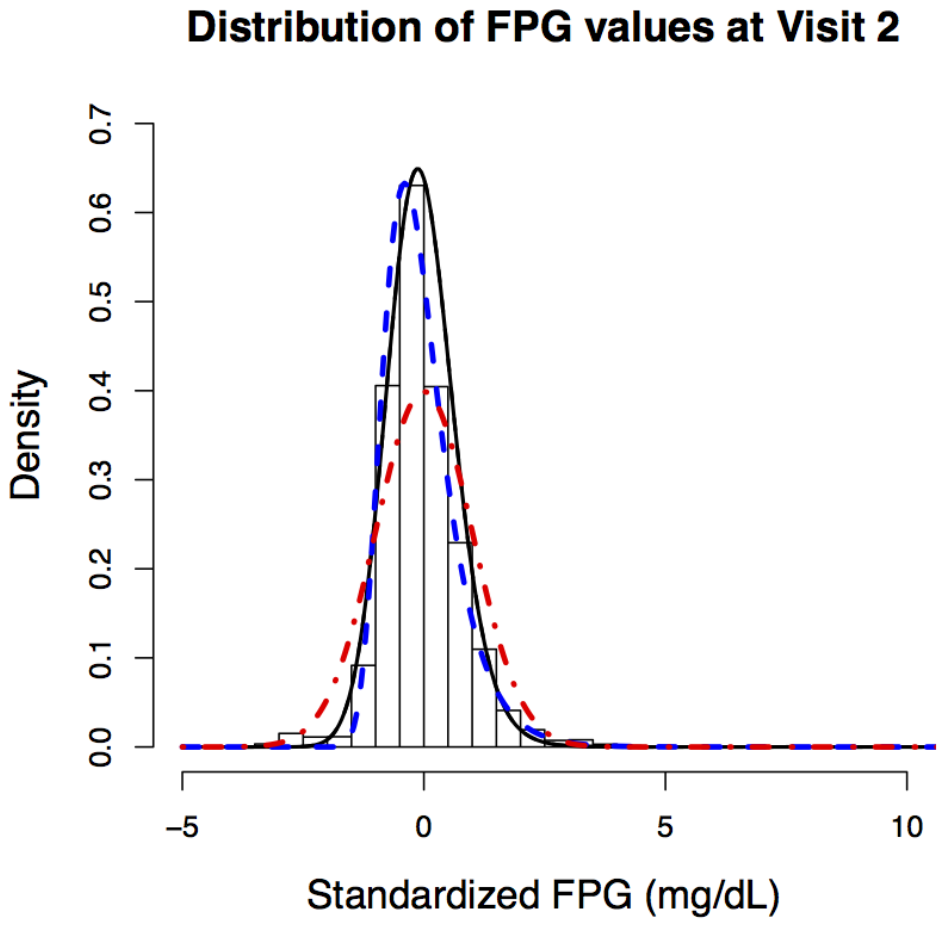
OUTLINE

DISEASE BIOMARKERS

- Many diseases are characterized or diagnosed by biomarkers.
- Diabetes: fasting plasma glucose;
  Obesity: BMI or WH ratio;
  Cancer: tumor size or imaging markers
- Only a very small fraction of populations are diagnosed as disease so their disease biomarkers are likely to be extreme compared to normal population.
- Standard normal distributions or other heavy-tail distributions may not be appropriate for modelling disease biomarker distributions.

ARIC EXAMPLE

- The Atherosclerosis Risk in Communities (ARIC) study is a prospective study of risk factors for atherosclerosis in 4 US communities.
- FPG values are used to determine diabetic status (fasting FPG$\geq 126 mg/dl$, non-fasting glucose $\geq 200 mg/dl$).
- The distribution of FPG from visit shows a long but thin tail of FGP values.


Distribution of FPG values at Visit 2
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−5
0
5
10
Standardized FPG (mg/dL)

Additional characteristics of disease biomarker measurements

- Many disease biomarkers in population tend to have a stochastically monotone trend due to natural aging processes and degrading metabolism in human bodies.
- For example, the likelihood of having higher FPG values or BMI increases with aging.
- Measurement error is inevitable: the coefficient of variation for the measurement error in laboratory glucose values is 3.5$\sim$9%.

THE GOAL OF THIS WORK

- We propose a semiparametric regression model to model extreme-value distributed biomarkers.
- The model incorporates stochastically monotone distribution of biomarkers.
- We will account for measurement error for inference.

Outline

Model

- Let $Y^*(t)$ denote true disease biomarker (no error) at time t and X are baseline risk factors.
- Our model assumes

$$P(Y^*(t) \leq y) = \exp\left\{-\Lambda(t)e^{-\mu y + X^T\beta}\right\}, \quad \mu > 0.$$

- Unknown parameters include μ, β and $\Lambda(t)$.
- $\Lambda(t)$ is positive and increasing.

Model interpretation

- At each fixed time t, this is an extreme-value distribution.
- Different X leads to location shift of this distribution by $X^T\beta/\mu$.
- Since $\Lambda(t)$ is increasing, $Y^*(t)$ is stochastically increasing: $Y^*(t_1) \prec Y^*(t_2)$.

Connection to threshold-defined event

- There is an interesting connection of the proposed model to a threshold-defined event.
- For any threshold value ξ, define T_ξ as the first time that biomarker value passes ξ (assuming increasing biomarker values).
- For example, in ARIC study, if $\xi = 126 mg/dl$ and $Y^*(t)$ is FPG, then T_{126} is clinically meaningful time to diabetic incidence.

Connection (cont.)

- Assume that $Y^*(t)$ has an increasing trajectory. Note

$$P(T_\xi \leq t) = P(Y^*(t) > \xi).$$

- Our model implies a proportional hazard model for each T_ξ:

$$\lambda_\xi(t) = \lambda(t) \exp\left\{-\mu\xi + X^T\beta\right\}.$$

- Thus, β can be understood as the common log-hazard ratios of risk factors on threshold-defined disease incidence.

Observed data and measurement error model

- Data are obtained cross-sectionally: $Y_i(v_i), X_i$ where v_i is the measurement time or age.
- Since $Y_i(t)$ is contaminated with measurement error, we assume
$$Y_i(t) = Y_i^*(t) + N(0, \sigma^2).$$
- Measurement error is independent and additive.
- Assume measurement time v_i to be non-informative and σ^2 known.

Likelihood function

$$\prod_{i=1}^{n} \int_{-\infty}^{\infty} \exp(-\Lambda(v_i) e^{X_i^T \beta - \mu\xi})$$

$$\times \Lambda(v_i) \mu \exp(X_i^T \beta - \mu\xi) \frac{1}{\sigma} \phi\left(\frac{Y_i(v_i) - \xi}{\sigma}\right) d\xi$$

Nonparametric maximum likelihood estimation

- We estimate Λ as a step function with positive jumps at unique values of v_i's.
- We treat the likelihood function as from missing data where ξ is missing for each subject.
- In some sense, we "pretend" each patient to have individual threshold value ξ_i and the likelihood concerns T_{ξ_i}.
- EM algorithm is adopted for maximization.

Detail of the algorithm

- In the M-step, we update μ and β using the Newton-Raphson.
- We update Λ by maxmizing

$$Q(\Lambda) = \sum_{k=1}^{K} \sum_{i=1}^{n} I(v_i = v_{(k)}) E(-\Lambda_k e^{X_i^T \beta - \mu\xi} + \log \Lambda_k \mid Y_i(v_i), \theta^{(l)}). \tag{1}$$

- The latter is a concave function over a convex cone $0 \leq \Lambda_1 \leq \cdots \leq \Lambda_K$.
- The E-step involves one-dimensional numerical integration with respect to ξ based on Gaussian quadratures.

Variance estimation

- We explicitly estimate the efficient influence functions for β and μ so the variance can be estimated using the empirical variance of this influence function.
- The unknown parameters in the influence functions can be estimated using data.
- The latter involves one-dimensional kernel density estimation.

Asymptotic results

Let $\theta = (\beta, \mu)$.

- Consistency: $|\widehat{\theta} - \theta| + \sup_v |\widehat{\Lambda}(v) - \Lambda(t)| \to_p 0$.
- Convergence rate: $d((\widehat{\theta}, \widehat{\Lambda}), (\theta, \Lambda)) = O_p(n^{-1/3})$.
- Asymptotic normality and efficiency: $\sqrt{n}(\widehat{\theta} - \theta) \to_d N(0, I(\theta)^{-1})$.
- Consistency of variance estimator.

Outline

Simulation setting

- Consider two covariates: $X_1 \sim Ber(0.5)$, $X_2 \sim N(0, 0.1)$
- We set $\Lambda(t) = 2t^{1/5}$, $\mu = 0.5$ and $\beta_1 = \beta_2 = 0.3$.
- Measurement error is from $N(0, \sigma^2)$ where $\sigma^2 = 0.25$.
- We consider time points from discrete set $\{0.1, 0.2, 0.4, 0.8\}$ or uniformly from $[0, 1]$.

Comparing with current-status analysis

- Recall that our model is equivalent to Cox PHM for threshold-defined time event.
- One alternative is to consider a fixed threshold and its corresponding time-to-event; then data will reduce to current status data for this event which can be analyzed using Cox model for current status data (e.g., ICM).
- We compare the results for threshold values chosen to be 90-, 80- and 70-quantile of data.

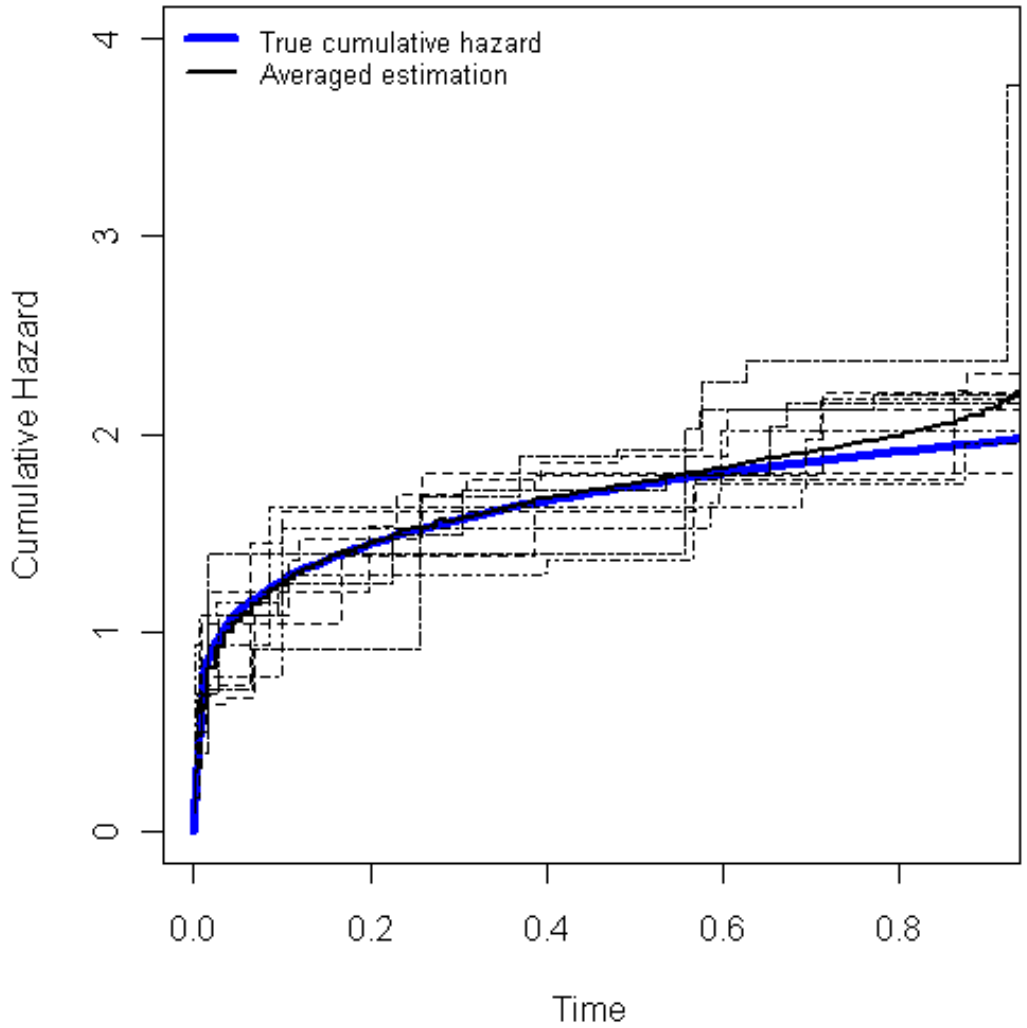
Our results

Sample size	Variance ratio	Par	True Value	Bias	SE	ASE	CP
400	0.04	μ	0.5	0.013	0.024	0.023	0.930
		β_1	0.3	0.010	0.104	0.122	0.977
		β_2	0.3	0.000	0.174	0.195	0.981
	0.16	μ	1.0	0.028	0.054	0.053	0.924
		β_1	0.3	0.009	0.118	0.135	0.974
		β_2	0.3	0.002	0.213	0.214	0.951
800	0.04	μ	0.5	0.007	0.016	0.016	0.940
		β_1	0.3	0.001	0.081	0.084	0.955
		β_2	0.3	0.002	0.121	0.132	0.966
	0.16	μ	1.0	0.014	0.040	0.036	0.920
		β_1	0.3	0.001	0.086	0.092	0.960
		β_2	0.3	0.000	0.148	0.146	0.941

ICM RESULT

n			q(90%)		q(80%)		q(70%)	
			Bias	SE	Bias	SE	Bias	SE
400	0.04	β_1	-0.004	0.189	-0.002	0.160	0.000	0.148
		β_2	0.004	0.281	0.010	0.260	0.009	0.240
	0.16	β_1	-0.005	0.352	-0.004	0.247	-0.015	0.193
		β_2	-0.003	0.540	-0.011	0.387	-0.010	0.300
800	0.04	β_1	0.002	0.124	-0.002	0.112	-0.004	0.102
		β_2	-0.006	0.201	-0.009	0.183	-0.008	0.167
	0.16	β_1	-0.001	0.235	-0.005	0.168	-0.009	0.132
		β_2	-0.031	0.377	-0.016	0.274	-0.015	0.212

ESTIMATE OF $\Lambda(t)$



Conclusion from the simulation study

- Our method performs well and is always more efficient than ICM method.
- ICM is biased if measurement error is not small.

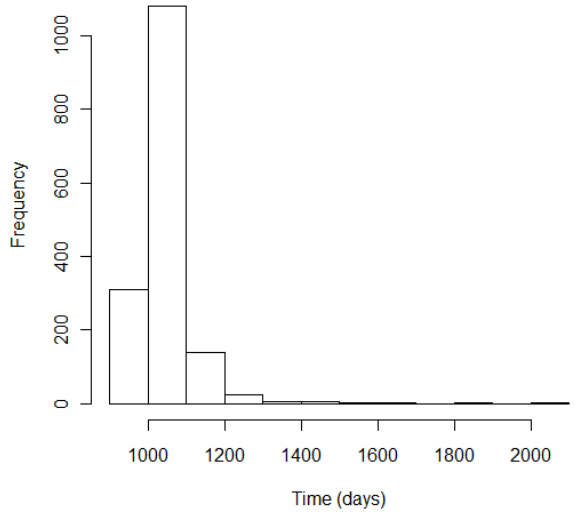
Outline

DATA DESCRIPTION

- The whole data consist of about 12,000 subjects from 4 counties.
- Due to computation burden, we restrict analysis to 1,560 Caucasian females from Forsyth County, North Carolina.
- Each subject had up to 4 visits; however, since participants were instructed to take medicine or prevention (dietary change) after diagnosis of diabetes after visit 2, the follow-up FPG values could be changed especially for extreme-tail patients.
- We thus restrict analysis to visit 2 data.

MORE DATA INFORMATION

- Visit times are random:



- The covariates include age, BMI, current smoking status, and hypertension.
- FPG values below 75 mg/dl were winsorized to reduce the influence of outliers in the lower tail of the distribution, because our interest is in crossing a threshold towards the upper end of the distribution.

Analysis result

Label	Our method Estimate	ASE.	p-value
Results using all the data			
Threshold effect	1.35	0.028	<0.001
Current smoker	0.203	0.055	<0.001
Hypertension	0.382	0.062	<0.001
Age (years)	0.016	0.004	<0.001
BMI (kg/m^2)	0.035	0.006	<0.001
Results using the data with 6 outliers excluded			
Threshold effect	1.35	0.028	<0.001
Current smoker	0.167	0.050	<0.001
Hypertension	0.403	0.060	<0.001
Age (years)	0.015	0.004	<0.001
BMI (kg/m^2)	0.032	0.005	<0.001

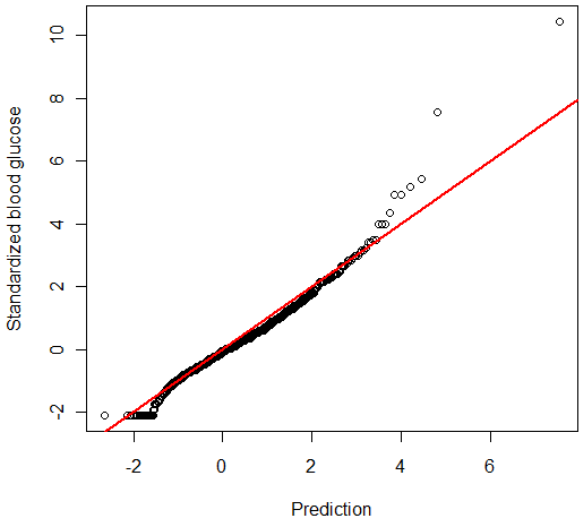
SUMMARY OF THE RESULTS

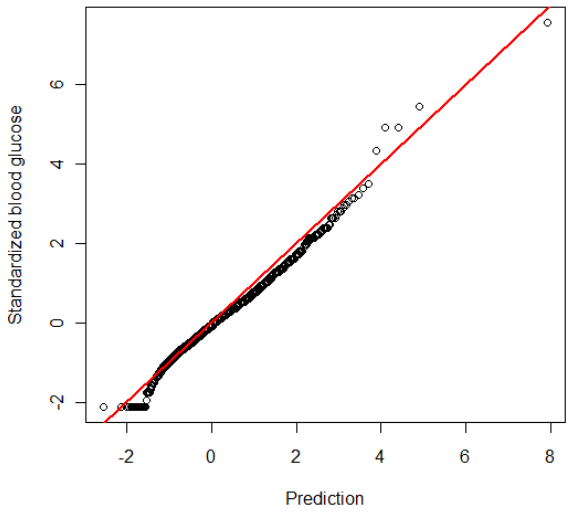
- The covariates current smoking, hypertension, higher age, and higher BMI have strongly significant associations with FPG level so diabetes.
- Smokers and subjects with hypertension have 1.22 times and 1.46 times greater hazard of diabetes than non-smokers and normotensive subjects, respectively.
- For each 1-year increase in age, the hazard of diabetes increases by a factor of 1.02; when BMI increases by 1 kg/m^2, the hazard of diabetes increases by a factor of 1.04.
- Comparatively, the analysis of the ICM method gives very different results in effect size and significance due to the lack of events.

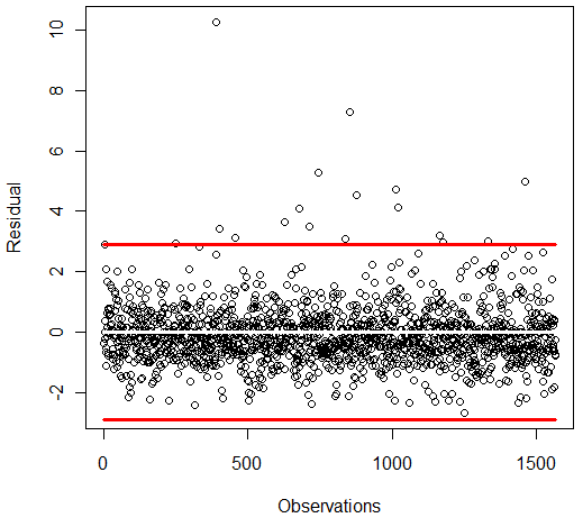
CHECK MODEL FIT

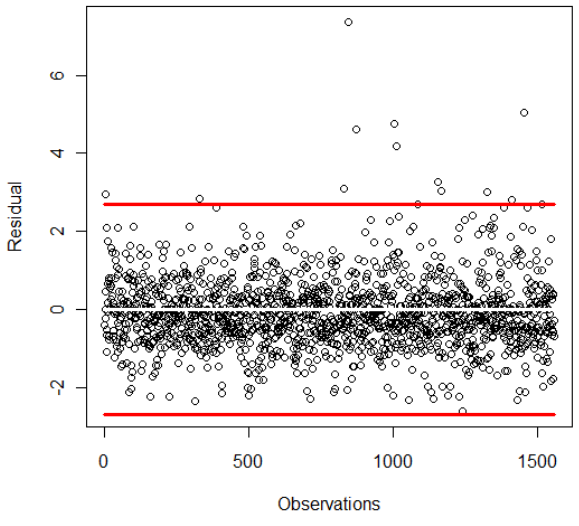
- We generated predicted glucose values based on the parameter estimation and covariate information and measurement error randomly generated from the normal distribution with mean 0 and variance σ^2.
- Using the predicted values, Quantile-Quantile (QQ) plots are generated to compare the distribution of the real observed glucose values with the predicted distribution.
- We calculated the residuals by subtracting the predicted means from the real observed glucose values and made residual plot.

MODEL CHECKING PLOT









Outline

Conclusion

- We proposed a semparametric extreme-value model for modelling disease biomarkers.
- The model implies a proportional hazards model for threshold-defined disease incidence.
- We proposed semiparametrically efficient inference using data with measurement errors.
- The proposed method works well in real application.

Extension

- The model and method can be extended to modelling longitudinal disease biomarkers.
- The inference can be extended to incorporate exact observation of disease incidence for some fixed threshold values.
- Further development can be to incorporate multivariate or even high-dimensional biomarkers for disease diagnosis.