Motivational Example

- Data: Observational longitudinal study of obesity from birth to adulthood.
- Overall Goal: Build age-, gender-, height-specific growth charts (under 3 year) to diagnose growth abnomalities.
- Specific Aim: Estimate the reference range for age-, gender-, height-specific weight.
- A simple version: Estimate the reference range for age-, gender-specific BMI.
- Statistical Problem: Estimate covariate specific quantiles in a reference group.

New method: Semiparametric Models

- Allow the shape depend on X and do not specify specific distribution for Y. Model $\mu(X)$ and $\sigma(X)$ parametricly to gain efficiency.
- General model

$$Y_i = \mu(X_i; \boldsymbol{\theta}) + \sigma(X_i; \boldsymbol{\theta})\varepsilon(X_i),$$

$\mu(X_i; \boldsymbol{\theta})$ is location parameter, $\sigma(X_i; \boldsymbol{\theta})$ is scale parametre, i.e. $\sqrt{(}Var(Y_i|X_i))$ and $\varepsilon(X_i)$ is from baseline distribution with mean zero, unit varaince. Denote baseline distribution function by $F_0(z, X) = P(\varepsilon(X) \leq z|X)$, we have

$$Y^\alpha(X; \boldsymbol{\theta}, F_0) = \mu(X_i; \boldsymbol{\theta}) + \sigma(X_i; \boldsymbol{\theta})Z^\alpha(X),$$

$Z^\alpha(X)$ is the αth quantile of $\varepsilon(X)$, i.e.

$$F_0(Z^\alpha(X), X) = \alpha.$$

Model details

- We can use splines to model $\mu(X)$ and $\sigma(X)$. Let $\boldsymbol{\theta} = \{\beta_1, \cdots, \beta_p, \gamma_1, \cdots, \gamma_q\}$.

$$\begin{aligned} \mu(X) &= \sum_{k=1}^{p} \beta_k R_k(X) \\ \log\{\sigma(X)\} &= \sum_{k=1}^{q} \gamma_k S_k(X) \end{aligned}$$

  $R(X) = \{R_1(X), \cdots, R_p(X)\}$ and $S(X) = \{S_1(X), \cdots, S_q(X)\}$ are pre-specified regression spline basis functions.

Quasi-likelihood and GEE 2

Consider general moment restricted models as below:

$$E[f(Y, X, \theta)|X] = 0.$$

All RAL estimator for θ must be solution from

$$E[A(X; \theta)f(Y, X, \theta)] = 0.$$

The most efficient selection will be

$$A(X; \theta) = E[\frac{\partial f(Y, X, \theta)}{\partial \theta}|X] Var[f(Y, X, \theta)|X]^{-1}.$$

Quasi-likelihood and GEE 2

A special case, location and scale model, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$:

$$
\begin{aligned}
f_1(Y, X, \boldsymbol{\theta}) &= Y - \mu(X; \boldsymbol{\beta}) \\
f_2(Y, X, \boldsymbol{\theta}) &= (Y - \mu(X; \boldsymbol{\beta}))^2 - \sigma^2(X; \gamma).
\end{aligned}
$$

We have $E[\frac{\partial f_1(Y,X,\boldsymbol{\theta})}{\boldsymbol{\gamma}}|X] = E[\frac{\partial f_2(Y,X,\boldsymbol{\theta})}{\boldsymbol{\beta}}|X] = 0$, so estimating functions should be

$$
\begin{aligned}
&\frac{\partial \mu(X; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} Var(Y|X)^{-1}[Y - \mu(X; \boldsymbol{\beta})] \\
&\frac{\partial \sigma^2(X; \gamma)}{\partial \boldsymbol{\gamma}} Var[(Y - \mu(X; \boldsymbol{\beta}))^2|X]^{-1}[(Y - \mu(X; \boldsymbol{\beta}))^2 - \sigma^2(X; \gamma)]
\end{aligned}
$$

Quasi-likelihood and GEE 2

Using independent working correlation and use normal distribution as working model for $Y|X$, we obtain quasi-likelihood score equation.

$$Var[(Y - \mu(X;\boldsymbol{\beta}))^2|X] = 2Var(Y|X)^2$$

$$\begin{aligned}
0 &= \sum_i \frac{\partial\mu(X_i;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} \frac{Y_i - \mu(X_i;\boldsymbol{\beta})}{\sigma^2(X_i;\gamma)} \\
0 &= \sum_i \frac{\partial\sigma^2(X_i;\gamma)}{\partial\boldsymbol{\gamma}} \frac{(Y_i - \mu(X_i;\boldsymbol{\beta}))^2 - \sigma^2(X_i;\gamma)}{2\sigma^4(X_i;\gamma)}
\end{aligned}$$

Quasi-likelihood and GEE 2

For our model specification, we have

$$\frac{\partial \mu(X_i; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = R(X_i)^T$$

$$\frac{\partial \sigma^2(X_i; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2S(X_i)^T \sigma^2(X_i; \boldsymbol{\beta})$$

$$\begin{aligned} 0 &= \sum_i R(X_i)^T \frac{Y_i - \mu(X_i; \boldsymbol{\beta})}{\sigma^2(X_i; \gamma)} \\ 0 &= \sum_i S(X_i)^T \frac{(Y_i - \mu(X_i; \boldsymbol{\beta}))^2 - \sigma^2(X_i; \gamma)}{\sigma^2(X_i; \gamma)} \end{aligned}$$

Estimate Baseline Function

- Obtain consistent estimates of $\mu(X)$ and $\sigma(X)$ as above
- Obtain the estimated residual

$$\hat{e}_i(X_i) = \frac{Y_i - \mu(X_i; \hat{\boldsymbol{\theta}})}{\sigma(X_i; \hat{\boldsymbol{\theta}})},$$

then estimate the baseline function $F_0(z, X)$ from $\hat{e}_i(X_i)$.

Estimate Baseline Function

- Speical case: F_0 does not depend on X.

$$\hat{F}_0(z, X) = n^{-1} \sum_{i=1}^{n} I(\hat{e}_i \leq z).$$

- Step function $\Rightarrow$ continuous function.

$$\hat{F}_0(z, X) = n^{-1} \sum_{i=1}^{n} K_{\lambda_2}(z, \hat{e}_i),$$

where $K_{\lambda_2}(z, \hat{e}_i) = K\{(z - \hat{e}_i)/\lambda_2\}$ and $K(\cdot)$ is any continuous distribution function.

$$\lim_{\lambda_2 \to 0} K\{(z - \hat{e}_i)/\lambda_2\} = I(\hat{e}_i < z) + K(0)I(\hat{e}_i = z).$$

$\hat{F}_0(z, X)$ is monotonicly increasing in z.

Estimate Baseline Function

- Speical case: F_0 does not depend on X.

$$\hat{F}_0(z, X) = \frac{\sum_{i=1}^{n} w_{\lambda_1}(X, X_i) I(\hat{e}_i \leq z)}{\sum_{i=1}^{n} w_{\lambda_1}(X, X_i)},$$

where $w_{\lambda_1}(X, X_i) = W((X - X_i)/\lambda_1)$ and $W(\cdot)$ can be any kernel function satisfy

$$W(x) \geq 0$$
$$\int W(x)dx = 1$$
$$\sigma_W^2 = \int x^2 W(x)dx < \infty$$

- Continuous version:

$$\hat{F}_0(z, X; \lambda_1, \lambda_2) = \frac{\sum_{i=1}^{n} w_{\lambda_1}(X, X_i) K_{\lambda_2}(z, \hat{e}_i)}{\sum_{i=1}^{n} w_{\lambda_1}(X, X_i)},$$

Bandwith and Kernel Function

- We will use following kernels (for simplicity, not most efficient):

$$\begin{aligned} w_{\lambda_1}(X, X_i) &= \phi((X - X_i)/\lambda_1) \\ K_{\lambda_2}(z, \hat{e}_i) &= \Phi((z - \hat{e}_i)/\lambda_2) \end{aligned}$$

- For λ_1 and λ_2, we can use either fixed value or allow they depend on X.

Comparing to kernel estimator from Y_i

- A nonparametric way will be use

$$\hat{F}(z, X) = \frac{\sum_{i=1}^{n} w_{\lambda_1}(X, X_i) I(Y_i \leq z)}{\sum_{i=1}^{n} w_{\lambda_1}(X, X_i)},$$

- For certain z, X and λ_1, assuming we have a uniform distribution for X, the bias will be

$$\begin{aligned} & \frac{\int W(u/\lambda_1)[F(z, X+u) - F(z, X)]du}{\int W(u/\lambda_1)du} \\ \approx & \frac{\int W(u/\lambda_1)[\dot{F}(z, X)u + \frac{1}{2}\ddot{F}(z, X)u^2]du}{\int W(u/\lambda_1)du} \\ = & \frac{1}{4}\ddot{F}(z, X)\lambda_1^2\sigma_W^2 \end{aligned}$$

- So whether semiparametric method gain depend on comparison of $\ddot{F}(z, X)$ and $\ddot{F}_0(z, X)$ or $\int \ddot{F}(z, X)^2 dF_X$ and $\int \ddot{F}_0(z, X)^2 dF_X$.

Optimal band width

- For certain z, X and λ_1, the variance is approximate

$$(n\lambda_1)^{-1}F(z,X)\int W(u)^2du.$$

- The mean square error (MSE) will be

$$\frac{1}{4}\ddot{F}(z,X)^2\lambda_1^4\sigma_W^4 + (n\lambda_1)^{-1}F(z,X)\int W(u)^2du$$

$$\lambda_1^* = \left(\frac{n^{-1}F(z,X)\int W(u)^2du}{\ddot{F}(z,X)^2\sigma_W^4}\right)^{1/5}$$

The estimator is $n^{2/5}$ consistent for both $\hat{F}$ and $\hat{F}_0$. Semiparametric method and nonparametric one has same convergence rate!

Multiple covariates

- Estimating equation part: Use two set of spline basis $R(X_1)$, $R(X_2)$ and $S(X_1)$, $S(X_2)$. Use their tensor product to generate the new basis $R(X) = R(X_1) \otimes R(X_2)$, $S(X) = [S(X_1), S(X_2)]$.
- Baseline estimator: Not approximate indicator function by continuous one. For kernel $W(X)$, can use any multivariate kernel, for example the tensor product of two univariate kernel $W_1(X_1) \times W_2(X_2)$. In application, they use $W(X) = W_1(X_1)$.

Example

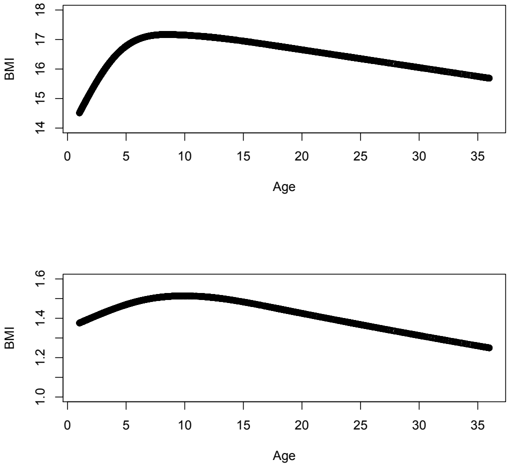


Figure: Fitted mean (top) and standard deviation (bottom) function

Example

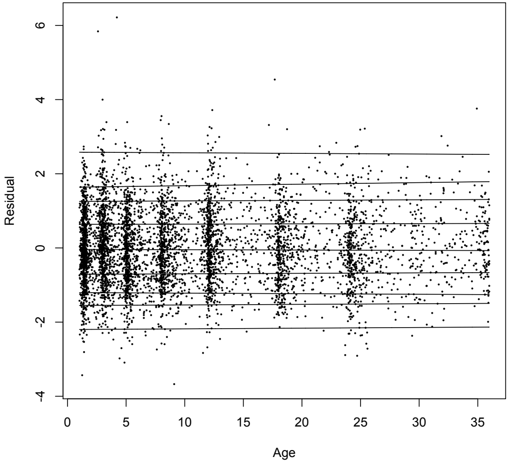


Figure: Residual quantile regression for fitted 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 99th quantiles

Example

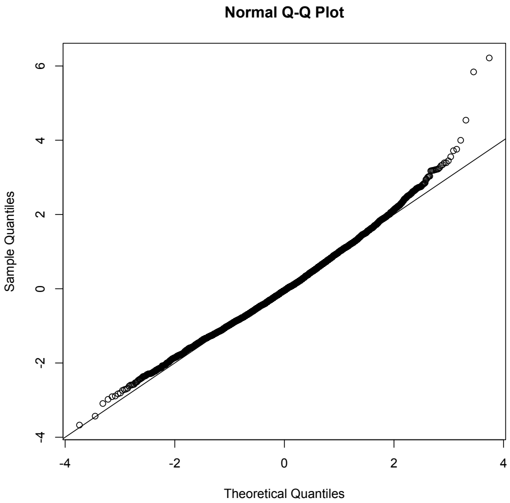
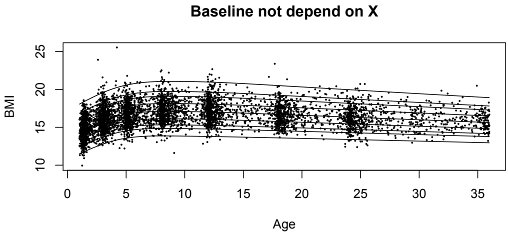


Figure: Residual Q-Q plot

Example



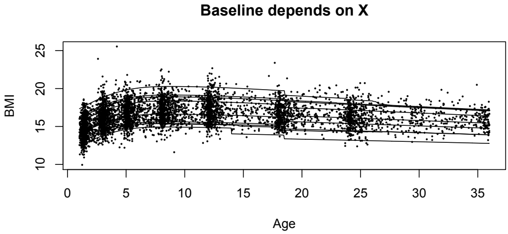


Figure: Fitted 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 99th quantiles with (top) or without (bottom) allow baseline depend on X

Simulation: Methods Comparing

- We will compare following methods in univariate covariate setting:
  1. Bin and smooth quantile
  2. Nonparametric quantile regression
  3. Locally weighted method (empirical one)
  4. Parametric method (LMS model)
  5. Semiparametric method assuming same baseline (empirical one)
  6. Semiparametric method allowing baseline depend on X (empirical one)
- For methods using bandwidth, we will choose several bandwidths $(0.05, 0.1, 0.2)$.

Simulation: Methods Comparing

- For simplicity, we use a linear term for all parametric parts.
- True model is generated with semiparametric model with $n = 5000$, where

$$\begin{aligned} \mu(X) &= 50 + 10X \\ \log(\sigma(X)) &= 3 + 2X, \end{aligned}$$

  X follow standard uniform distribution and the $e(X)$ from following distributions

  1. Standard normal $N(0, 1)$.
  2. Standardized log normal distribution.
  3. Mixture normal $N(-0.5, 1)$, $N(0.5, 1)$.

Simulation: Methods Comparing

- We are interested in estimating following things from $M = 1000$ simulations
  1. Bias, variance and MSE of the estimator for specific quantiles (90%, 95%, 99%) and specific covariate values (0.05, 0.25, 0.5, 0.75, 0.95)
  2. Integrated mean square error of the estimator for specific quantiles