BIOST 572 Update Talk

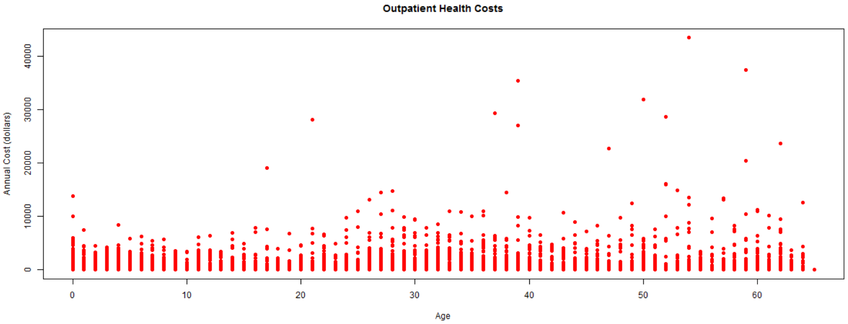
David Benkeser

University of Washington Department of Biostatistics

April 26, 2012

Few people understand a really good sandwich

James Beard (1903-1985)
American culinary artist


Outpatient Health Costs
Annual Cost (dollars)
0
10000
20000
30000
40000
0
10
20
30
40
50
60
Age

$$\boldsymbol{\beta} = \underset{\boldsymbol{\beta}}{\mathit{argmin}} \;\; \mathbb{E}_F \left[\left(\phi(\mathbf{X}) - \mathbf{X}^T \boldsymbol{\beta} \right)^2 \right]$$

If we knew $\phi(\cdot)$ and F, we could calculate β directly

Instead, embed in flexible Bayesian model and use data to derive posterior for $\phi(\cdot)$ and F and *thus* for $\boldsymbol{\beta}$

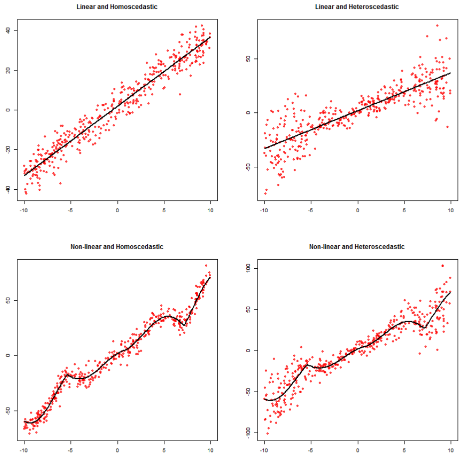
Fixed vs. Random $\mathbf{X}$
Why does it matter?

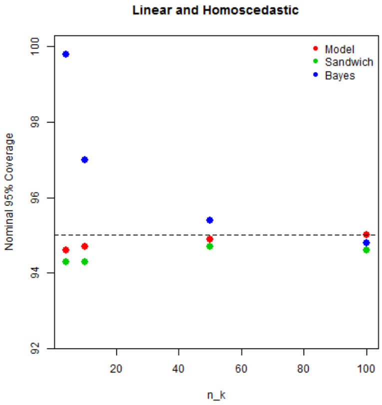
If ϕ truly is linear, it doesn't

Otherwise fixed $\mathbf{X}$ can lead to (massive) overcoverage for sandwich errors

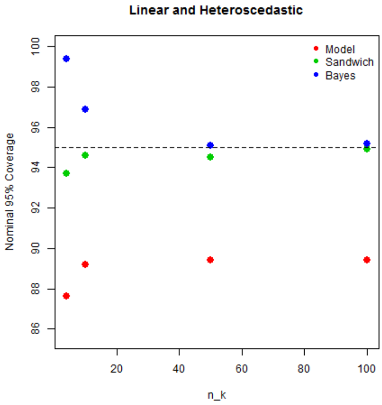
- Confuses errors from nonlinearity in ϕ with the random component

Bayesian approach allows distinction between fixed and random sampling

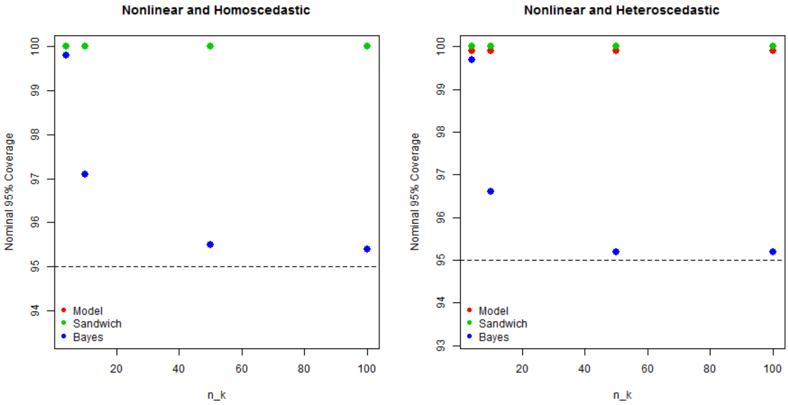

Linear and Homoscedastic
Linear and Heteroscedastic
Non-linear and Homoscedastic
Non-linear and Heteroscedastic



Linear and Homoscedastic
Nominal 95% Coverage
92
94
96
98
100
20
40
60
80
100
n_k
Model
Sandwich
Bayes



Linear and Heteroscedastic
Nominal 95% Coverage
86
88
90
92
94
96
98
100
20
40
60
80
100
n_k
Model
Sandwich
Bayes



Nonlinear and Homoscedastic
Nominal 95% Coverage
100
99
98
97
96
95
94
Model
Sandwich
Bayes
20
40
60
80
100
n_k
Nonlinear and Heteroscedastic
Nominal 95% Coverage
100
99
98
97
96
95
94
93
Model
Sandwich
Bayes
20
40
60
80
100
n_k

Theorem: *For a discrete covariate space,* $\hat{\beta}_{fixed} = \mathbb{E}_\pi(\beta_{fixed}|\mathbf{X}, Y)$ *takes the form*

$$\hat{\beta}_{fixed} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$$

and assuming at least four samples for each covariate value, the corresponding uncertainty estimate has form

$$\hat{\sigma}_\beta = [(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\Sigma'\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}]^{1/2}$$

where Σ' *is diagonal matrix*

$$\Sigma'_{ij} = \begin{cases} \frac{1}{n_k-3}\sum_{i:X_i=\xi_k}(Y_i - \bar{y}_k)^2 & \textit{if } i{=}j \\ 0 & \textit{else} \end{cases}$$

Theorem: *Assuming* $Y|\mathbf{X}$ *has bounded first and second moments,* $\hat{\beta} = \mathbb{E}_\pi(\beta|\mathbf{X}, Y)$ *takes the asymptotic form*

$$\hat{\beta} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y \rightarrow_{as} 0$$

and assuming at least four samples for each covariate value, the corresponding uncertainty estimate has asymptotic form

$$\hat{\sigma}_\beta - diag[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\Sigma\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}]^{1/2} = o(n^{-1})$$

where Σ *is diagonal matrix*

$$\Sigma_{ij} = \begin{cases} (Y_i - X_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y)^2 & \textit{if } i{=}j \\ 0 & \textit{else} \end{cases}$$

Let $\xi = (\xi_1, ..., \xi_K)$ be K m-vectors of values covariates X can take

Recall our assumed likelihood

$$Y|(X = \xi_k, \phi_k, \sigma_k^2) \sim N\left(\phi_k, \sigma_k^2\right)$$

We use noninformative, improper priors for ϕ, σ^2

$$p_{\phi,\sigma^2} \propto \prod_{k=1}^{K} \sigma_k^{-2}$$

Posterior $\pi(\phi_k) \sim t(\bar{y}_k, \frac{1}{n_k}s_k^2, n_k - 1)$, where

$$s_k^2 = \frac{1}{n_k - 1} \sum_{i:X_i=\xi_k}^{n_k} (Y_i - \bar{y}_k)^2$$

which can be rewritten as

$$\pi(\phi_k) = \bar{y}_k + \epsilon_k$$

i.e. deterministic component: $\bar{y}_k$ + random component: $\epsilon_k \sim t(0, \frac{1}{n_k}s_k^2, n_k - 1)$

Let Φ be n-vector, with $\Phi_i = \phi(X_i)$ and define

$$\bar{Y} = \mathbb{E}_\pi(\Phi|\mathbf{X}, Y) = (\bar{y}(X_1), ..., \bar{y}(X_n))$$

and note,

$$\begin{aligned}
\mathrm{Var}_\pi(\phi_k|\mathbf{X}, Y) &= \mathrm{Var}_\pi(\epsilon_k|\mathbf{X}, Y) \\
&= \frac{1}{n_k} s_k^2 \frac{n_k - 1}{(n_k - 1) - 2} \\
&= \frac{1}{n_k(n_k - 3)} \sum_{i:X_i=\xi_k}^{n_k} (Y_i - \bar{y}_k)^2
\end{aligned}$$

Let $\lambda(\cdot)$ be density with mass restricted to $\xi \subset \mathbb{R}^m$ such that

$$\lambda(\cdot) = \sum_{k=1}^{K} \lambda_k \delta_{\xi_k}(\cdot)$$

In other words,

$$\Pr\left(X = \xi_k | \lambda(\cdot)\right) = \lambda_k$$

$$\sum_{k=1}^{K} \lambda_k = 1$$

In fixed setting, let λ_{fixed} be deterministic density of actual sampled values

$$\lambda_{fixed}(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(\cdot)$$

$\mathbb{E}_{x;\lambda_{fixed}}$ refers to integration w.r.t. known distribution (not ϕ or posterior of estimated parameters)

$$\begin{aligned}\boldsymbol{\beta}_{fixed} &\equiv \underset{\boldsymbol{\beta}}{argmin}\ \mathbb{E}_{x;\lambda_{fixed}}\left[\left(\phi(\mathbf{x}) - \mathbf{x}^T\boldsymbol{\beta}\right)^2\right] \\ &= \mathbb{E}_{x;\lambda_{fixed}}\left[\mathbf{x}^T\mathbf{x}\right]^{-1}\mathbb{E}_{x;\lambda_{fixed}}\left[\mathbf{x}^T\phi(\mathbf{x})\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\Phi\end{aligned}$$

Thus,

$$\hat{\boldsymbol{\beta}}_{fixed} = \mathbb{E}_{\pi}\left(\boldsymbol{\beta}_{fixed}|\mathbf{X}, Y\right) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^TY$$

and

$$\hat{\sigma}^2_{\beta_{fixed}} = \mathrm{Cov}_{\pi}\left(\boldsymbol{\beta}_{fixed}|\mathbf{X}, Y\right) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\Sigma'\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

Now, we need a prior on $\lambda(\cdot)$

$$p_\lambda \propto \prod_{k=1}^{K} \lambda_k^{-1}$$

which yields posterior

$$p_{\lambda|X}(\lambda(\cdot)) \propto \prod_{k=1}^{K} \lambda_k^{-1+n_k}$$

$\mathbb{E}_{x;\lambda}$ thus refers to integrating w.r.t. posterior probability measure
$\mathbb{E}_\pi$ refers to integrating over posterior of λ, ϕ, ϵ

First note,

$$
\begin{aligned}
(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y &= \left(\sum_{k=1}^{K} \frac{n_k}{n} \xi_k^T \xi\right)^{-1} \left(\sum_{k=1}^{K} \frac{n_k}{n} \xi^T \bar{y}(\xi_k)\right) \\
&\rightarrow_{as} \left(\sum_{k=1}^{K} \lambda_k^* \xi_k^T \xi\right)^{-1} \left(\sum_{k=1}^{K} \lambda_k^* \xi^T \phi^*(\xi_k)\right)
\end{aligned}
$$

Now,

$$\begin{aligned}
\hat{\beta} &= \mathbb{E}_{\pi}\left[\mathbb{E}_{x;\lambda}\left(\mathbf{x}^T\mathbf{x}\right)^{-1}\mathbb{E}_{x;\lambda}\left(\mathbf{x}^T\phi(x)\right)|\mathbf{X},Y\right] \\
&= \mathbb{E}_{\pi}\left[\mathbb{E}_{x;\lambda}\left(\mathbf{x}^T\mathbf{x}\right)^{-1}\mathbb{E}_{x;\lambda}\left(\mathbf{x}^T\bar{y}(x)\right)|\mathbf{X},Y\right] \\
&= \mathbb{E}_{\pi}\left[\left(\sum_{k=1}^{K}\lambda_k\xi_k^T\xi\right)^{-1}\left(\sum_{k=1}^{K}\lambda_k\xi^T\bar{y}(\xi_k)\right)|\mathbf{X},Y\right] \\
&\rightarrow_{as} \left(\sum_{k=1}^{K}\lambda_k^*\xi_k^T\xi\right)^{-1}\left(\sum_{k=1}^{K}\lambda_k^*\xi^T\phi^*(\xi_k)\right)
\end{aligned}$$

For posterior variance, note

$$\begin{aligned}
\text{Cov}_\pi(\boldsymbol{\beta}) &= \text{Cov}_\pi\left(\mathbb{E}_{x;\lambda}[\mathbf{x}^T\mathbf{x}]^{-1}\mathbb{E}_{x;\lambda}[\mathbf{x}^T\phi(\mathbf{x})]|\mathbf{X}, Y\right) \\
&= \text{Cov}_\pi\left(\mathbb{E}_{x;\lambda}[\mathbf{x}^T\mathbf{x}]^{-1}\mathbb{E}_{x;\lambda}[\mathbf{x}^T(\bar{y}(\mathbf{x}) + \epsilon(\mathbf{x}))]|\mathbf{X}, Y\right) \\
&= \text{Cov}_\pi\left(\mathbb{E}_{x;\lambda}[\mathbf{x}^T\mathbf{x}]^{-1}\mathbb{E}_{x;\lambda}[\mathbf{x}^T\bar{y}(\mathbf{x})]|\mathbf{X}, Y\right) \\
&\quad + \text{Cov}_\pi\left(\mathbb{E}_{x;\lambda}[\mathbf{x}^T\mathbf{x}]^{-1}\mathbb{E}_{x;\lambda}[\mathbf{x}^T\epsilon(\mathbf{x})]|\mathbf{X}, Y\right)
\end{aligned}$$

We can calculate sandwich form for each of these terms

$$\text{Cov}_\pi\left(\mathbb{E}_{x;\lambda}[\mathbf{x}^T\mathbf{x}]^{-1}\mathbb{E}_{x;\lambda}[\mathbf{x}^T\bar{y}(\mathbf{x})]|\mathbf{X},Y\right) \rightarrow_{as} diag[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\tilde{\Sigma}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}]$$

where $\tilde{\Sigma}$ is diagonal matrix

$$\tilde{\Sigma}_{ij} = \begin{cases} (\bar{Y}_i - X_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\bar{Y})^2 & \text{if i=j} \\ 0 & \text{else} \end{cases}$$

and

$$\text{Cov}_\pi\left(\mathbb{E}_{x;\lambda}[\mathbf{x}^T\mathbf{x}]^{-1}\mathbb{E}_{x;\lambda}[\mathbf{x}^T\epsilon(\mathbf{x})]|\mathbf{X},Y\right) \rightarrow_{as} diag[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\Sigma'\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}]$$

Three sandwiches:

$$\begin{aligned}\Sigma'_{ii} &= \frac{1}{n_k - 3} \sum_{i:X_i=\xi_k} (Y_i - \bar{y}_k)^2 \\ \tilde{\Sigma}_{ii} &= (\bar{Y}_i - X_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\bar{Y})^2 \\ \Sigma_{ii} &= (Y_i - X_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y)^2\end{aligned}$$

Finally, we can show that

$$\sum_{i:X_i=\xi_k} \Sigma_{ii} = \sum_{i:X_i=\xi_k} \left(\tilde{\Sigma}_{ii} + \Sigma'_{ii}\right)$$

Classic sandwich (Σ) accounts for residual errors (Σ') *as well as* the errors due to the linear fit to non-linear ϕ ($\tilde{\Sigma}$)

Conclusions and Future Work

Standard sandwich errors lead to over-coverage in fixed $\mathbf{X}$ situations

Bayesian methods allow for distinction between fixed and random sampling and inform us about what the sandwich is actually doing

Future Work

- Derivation of random $\mathbf{X}$ asymptotic results
- Extension to continuous covariates (simulations)

Questions?