DISCUSSION OF THE PAPER BY LIN AND YING

Xihong Lin
University of Michigan

Raymond J. Carroll
Texas A&M University

http://stat.tamu.edu/~carroll
OUTLINE

- L–Y address a hard problem
  - Computationally convenient general method
  - With some tweaking, there is the potential to greatly increase the method’s efficiency
- A Special Case: the partially linear model
  - Optimal estimation
  - Single nearest neighbor estimation
  - Efficiency
- If covariate process is not completely observed
  - Inconsistency of Single nearest neighbor
- Efficiency of semiparametric versus parametric modeling
**POINT #1: A SPECIAL CASE**

- Covariate process is non–time varying: $X_i(t) \equiv X_i$
- Each individual observed once, at $T_i$
- $X_i$ and $T_i$ are independent
- Non-varying coefficient: $\beta(t) \equiv \beta$
- Residual error has constant variance
- Then the L–Y model takes the form

$$Y_i = \alpha(T_i) + X_i\beta + \epsilon_i$$

- This is known as the **partially linear model**, and can be fit in Splus using **gam**
PARTIALLY LINEAR MODEL

\[ Y_i = \alpha(T_i) + X_i\beta + \epsilon_i \]

- **Studied extensively**, e.g., Severini & Staniswalis (1994)
- **No iteration**:
  - Regress \( Y \) on \( T \) nonparametrically (e.g., spline, kernel): \( \bar{m}_y(t) \)
  - Regress \( Y - \bar{m}_y(T) \) on \( X - \bar{X} \)
  - This yields \( \hat{\beta} \)
- L–Y use this algorithm too, **except** that they do not use a standard nonparametric regression
  - If we understand it correctly (?), they use \( \bar{m}_y(t) = \bar{Y} \) by approximating \( Y_i(t) \) using its nearest neighbor observation.
- Extremely computationally convenient
- What cost computational convenience?
The partially linear model is given by:

\[ Y_i = \alpha(T_i) + X_i \beta + \epsilon_i \]

- Thus, L–Y use a computationally convenient but non-standard nonparametric regression.
- The efficiency of their method can be arbitrarily low:

\[
\text{efficiency} = \frac{\text{var}(\epsilon)}{\text{var}(\epsilon) + \text{var}\{\alpha(T)\}}
\]

- Thus, if there is a major time effect on the intercept, then the L–Y method can have near-zero efficiency.
- The efficiency can be corrected by using splines (say)
POINT #2: COVARIATE PROCESS NOT OBSERVED

- Pretend $X(T)$ is time–varying and only a few observations per subject are available

- Common in longitudinal models

- Not common that the entire covariate history $X(t), t \leq T$ is observed

- Not discussed by L-Y how to handle this case. One possibility is to approximate $X(t)$ by its nearest neighbor observation $X^*(t)$.

- **Conjecture:** The modified L–Y method is inconsistent unless $X^*(t)$ is replaced by a nonparametric regression of $X$ on $T$ at $t$, e.g., splines.
POINT #3: SEMIPARAMETRIC EFFICIENCY

\[ Y_i = \alpha(T_i) + X_i\beta + \epsilon_i \]

- **Semiparametric** means that you make **no assumptions** about \( \alpha(t) \)
- One could instead assume \( \alpha(t) \) is a parametric function and estimate the parameters.
- How much efficiency do you lose for estimating \( \beta \) by not making **correct parametric** assumptions?
- **NONE** if \( X \) and \( T \) are independent (as in L–Y simulations)
- **LOTS** otherwise, at least **potentially** if \( E(X|T) \) varies substantially and nonlinearly.
CONCLUSIONS

- We discussed in detail the partially linear model, which has a rich history.

- We suggest that the computational convenience of the L–Y methods carries costs:
  - Loss of efficiency, potentially great loss
  - Inconsistency if the covariate process is not observed and estimated using the nearest neighbor method.
  - Both points make a case for using spline/loess methods for the nonparametric regressions.

- One needs to be somewhat careful not to ignore parametric models
  - Optimal semiparametric methods can have near–zero efficiency compared to parametric methods.