.

Appendix 1: Trend Decomposition based on stratified nonparametric regression

Xiaoyu Song, Ying Wei, Sanyukta Mathur and John Santelli.

October 7, 2014

1.1 Background

Understanding the change of prevalence and incidence of a health outcome or risk factor in a population has been a long-standing research question in epidemiology and population research. Various analyses have been conducted to understand what are the underlying driving forces behind the change to observed prevalence and incidence. When there exist multiple factors that simultaneously associate with the disease prevalence and incidence, it is of natual interest to decompose the total effect into individual ones. For example, in Santelli et al (2004) and Santelli et al (2007) used population level data to decompose the the change of the pregnancy rate in a population into the changes in sexual activity and contraceptive use. Prithwis Das Gupta (1993) also described an effect decomposition formulation to quantify the individual contributions towards the change of prevalence between two time points.

In this article, we used a model-based trend decomposition approach. It quantifies to what extend the change of a risk factor contributes to the change of prevalence/incidence over the time. Specifically, we performed two decompositions; we examined the contribution of the declining prevalence of sexual experience among young women to the decline in HIV incidence; we also examined the contribution of increasing school enrollment among young women to the decline in sexual experience. The method fully utilizes the longitudinal follow up data in RCCS, and models the trend decomposition as a function of time, hence it is not restrict to two time points.

1.2 The model based trend decomposition

To illustrate the idea behind the model based trend decomposition, we start with a simple scenario where the risk factor is binary. For example, whether the respondent is sexually experienced is expressed as the prevalence of sexual experience. And whether the respondent is in a school is expressed as school enrollment. We first make the following notations.

- $I_Y(t)$ is the HIV incidence rate at year t among the teen women at Rakai.
- $P_X(t)$ is the prevalence of a binary risk factor at year t. If X is the indicator of being sexual active, then $P_X(t)$ is the percent of teen women who are sexually experienced. If X is the indicator of being in school, then $P_X(t)$ is the school attendance rate in among the teen women at Rakai.
- $I_{Y|1}(t)$ is the HIV incidence rate at year t conditioning on the presence of X, and $I_{Y|0}(t)$ is the HIV incidence rate at year t among those $X = 0$. Again, if X is the school indicator, then $I_{Y|1}(t)$ is the HIV incidence rate among those who are currently enrolled in schools, and $I_{Y|0}(t)$ is that among those who are not.

By definition, the HIV incidence rate $I_Y(t)$ can be written as

$$I_Y(t) = I_{Y|1}(t)P_X(t) + I_{Y|0}(t)\{1 - P_X(t)\}. \tag{1}$$

From the equation above, we see that the change of HIV incidence rate $I_Y(t)$ is driven by two factors, the prevalence of X and the conditional incidence rates $I_{Y|1}$ and $I_{Y|0}$. Since the latter are conditioning at a fixed level of X, they are unrelated to X.

If the reduced incidence rate $Y(t)$ is purely due to X, such as fewer sexually active teen women, then we could expect both $I_{Y|1}(t)$ and $I_{Y|0}(t)$ remain unchanged over the time. In other words, if we denote t_0 as the starting date of the cohort study, $I_{Y|1}(t) = I_{Y|1}(t_0)$ and $I_{Y|0}(t) = I_{Y|0}(t_0)$ for all t. Following this logic, we define an intermediate outcome

$$\widetilde{I}_Y(t) = I_{Y|1}(t_0) * P_X(t) + I_{Y|0}(t_0) * (1 - P_X(t)),$$

which is the expected incidence rate assuming the declined sexual active rate in the population is the only contributing factor. Consequently, we could decompose the change of HIV incidence from baseline $D(t) = I_Y(t) - I_Y(t_0)$ by

$$D(t) = \{I_Y(t) - \widetilde{I}_Y(t)\} + \{\widetilde{I}_Y(t) - I_Y(t_0)\}.$$

If we further define

$$P_2(t) = \frac{\widetilde{Y}(t) - Y(t_0)}{D(t)}, \text{ and } P_1(t) = \frac{Y(t) - \widetilde{Y}(t)}{D(t)}$$

then the numerate of $P_2(t)$ is the expected change of Incidence rate if the change of X is the only contributing factor, hence $P_2(t)$ **is the proportion of the change due to the change of X.** $P_1(t) = 1 - P_2(t)$ is then the proportion of the change due to the other reasons than X.

Finally, to calculate the overall contribution from X, we can integrate $P_2(t)$ over time.

Estimation from the RCCS data The defined statistics above depends on unknown functions, $I_Y(t)$, $P_X(t)$, $I_{Y|1}(t)$ and $I_{Y|0}(t)$. They can be estimated from RCCS data, which consist of the following observations.

- $Y_{i,j}$ is whether the i-th subject is a new HIV case in his j's round interview at time $t_{i,j}$
- $X_{i,j}$ is the measured risk factor of the he i-th subject and j's round interview
- $D_{i,j}$ is the time gap between the ith subject's j-the interview and his previous interview.

To estimate incidence rate $I_Y(t)$, we apply nonparametric poison regression (Bowman and Azzalini, 1997)

$$\ln\{E(Y_{i,j}/D_{i,j})\} = g(t_{i,j}), \quad I_Y(t) = \exp\{g(t_{i,j})\}. \tag{2}$$

To the prevalence $P_X(t)$, nonparametric logistic regression can be used to regress $X_{i,j}$ against $t_{i,j}$. Finally, we consider the following nonparametric coefficient-varying log-linear model to estimate $I_{Y|1}(t)$ and $I_{Y|0}(t)$

$$\ln\{E(Y_{i,j}/D_{i,j})\} = \beta_0(t_{i,j}) + \boldsymbol{\beta}_1(t_{i,j})^T X_{i,j},$$

where $\beta_0(t)$ and $\beta_1(t)$ are coefficient functions and estimated non-parametrically. Following the model, $I_{Y|0}(t) = \beta_0(t)$ and $I_{Y|1}(t) = \beta_0(t) + \beta_1(t)$. All the coefficient functions in the models will be estimated non parametrically using spline functions. We select the optimal splines using AIC.

Figure 1 shows the estimated HIV incidence rate from Model (2), which has been decreasing gradually from 2000 to 2004.

Using whether the respondent ever had sex as an example, Figure 2 shows the proportion of sexually experienced teen women (EverSex) that is estimated from nonparametric logistic regression(Bowman and Azzalini, 1997). We observed a

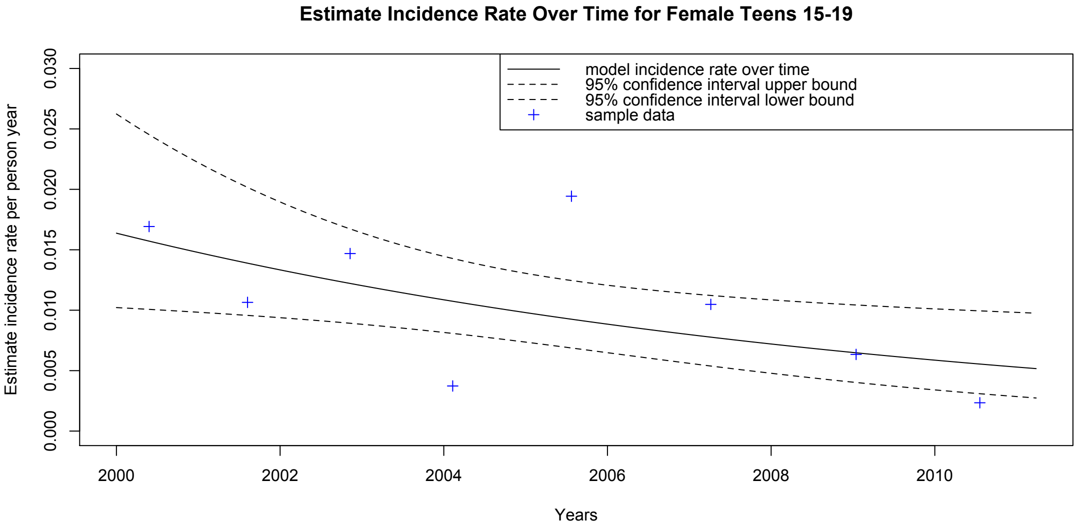

Figure 1:

rapid decline after 2004. Moreover, Figure 3 shows the estimated HIV incidence rate among the sexually experienced teen women only. No HIV cases were observed among the youth women who never had sex, hence we assume the incidence rate is zero among this subpopulation. We used the same approach to explain the change in sexual experienced based on change in school enrollment.

With this estimated functions, Figure 4 displays the estimated $P_2(t)$, which steadily increase over time. That suggests having fewer sexually experienced women plays more significant role in reducing HIV incidence in later years comparing to the earlier years. Overall, it contributes to 71.3% decline of HIV incidence rate. The remainder (28.7%) is explained by a decline in HIV incidence among those women who are sexually active.

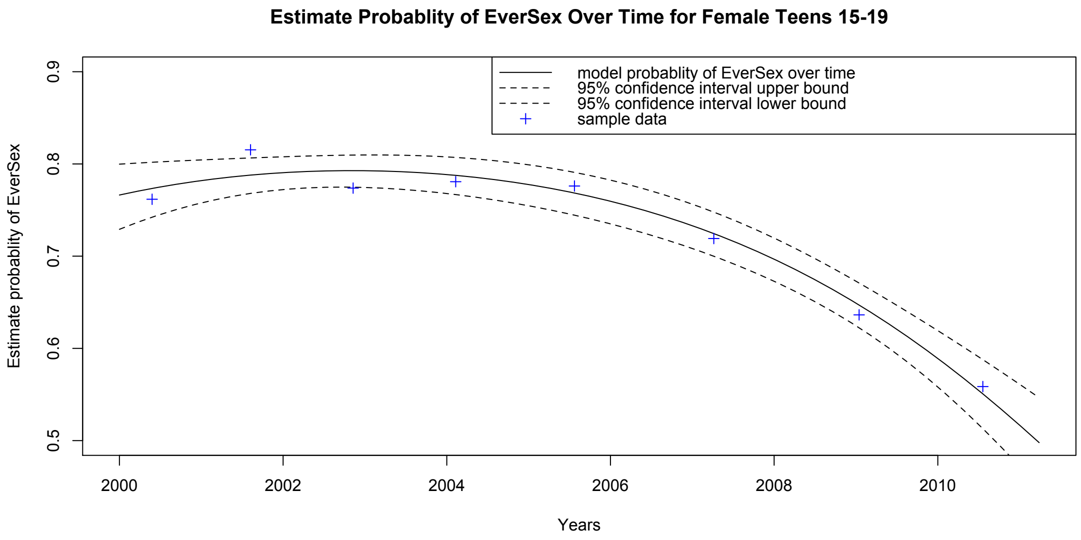

Figure 2:

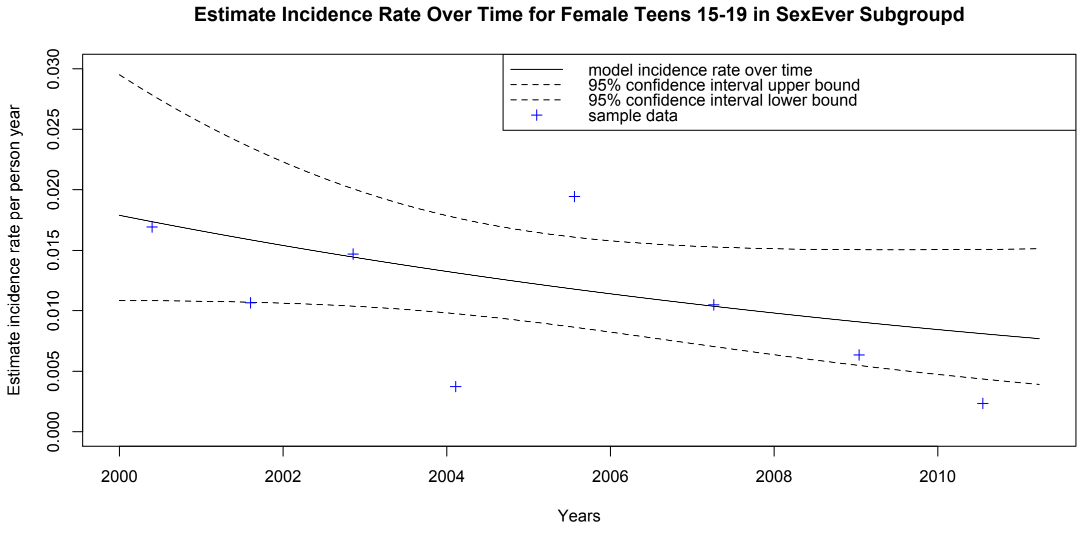

Figure 3:

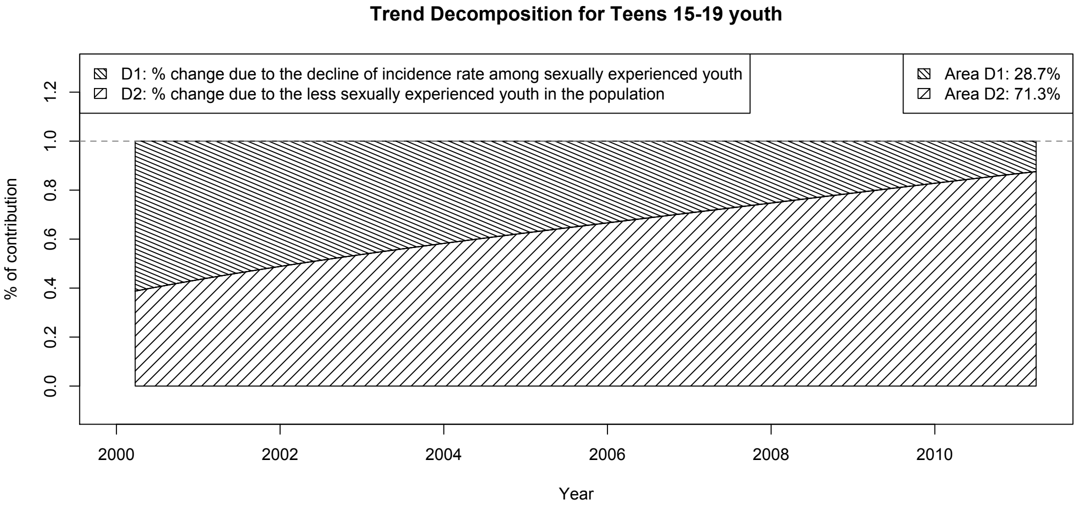

Figure 4:

1.3 Extension to other type of covariates

Categorical X If a risk factor X is a categorical variable with K levels. We define $P_{x,k}(t)$ as the proportion of k-level X at time t, where $\sum_{k=1}^{K} P_{x,k}(t) = 1$. We can also define $I_{Y|k}(t)$ as the HIV incidence rate given the k-th level of X. With these notations, the HIV incidence rate $I_Y(t)$ can be written as

$$I_Y(t) = \sum_{k=1}^{K} I_{Y|k}(t) P_{x,k}(t)$$

Same as before, if the reduced incidence rate $Y(t)$ is purely due to X, then we could expect both $I_{Y|k}(t)$'s remain unchanged over the time. Consequently, we define the intermediate outcome

$$\widetilde{I}_Y(t) = \sum_{k=1}^{K} I_{Y|k}(t_0) * P_X(t).$$

Once we have defined $\widetilde{I}_Y(t)$, we could decompose $D(t)$ in the same way as in previous section.

Continuous X For continuous risk factor X, we can define $F_X(x, t)$ as the distribution function of X at time t, and define $I_{Y|x}(t)$ as the HIV incidence rate given the $X = x$. Consequently, we could define the intermediate outcome by

$$\widetilde{I}_Y(t) = \int_x I_{Y|x}(t_0) F_X(x, t)\, dx.$$

Since contributing factors are often correlated with each other, there is no unique decomposition over multiple factors. For example, results of ANOVA often relies on the order of the covariates in the model. For this reason, we do not extend the method to the multiple factors.

2 Reference

[1] John S Santelli, Joyce Abma, Stephanie Ventura, Laura Lindberg, Brian Morrow, John E Anderson, Sheryl Lyss. Can Changes in Sexual Behaviors among High School Students Explain the Decline in Teen Pregnancy Rates in the 1990s? J Adolescent Health. 2004; 35 (2):80-90. and 1.

[2] JS Santelli, LD Lindberg, LB Finer, S Singh. Explaining Recent Declines in Adolescent Pregnancy in the United States: The Contributions of Abstinence and Improved Contraceptive Use. Am J Public Health. 2007; 97:150-156.

[3] Need reference for Prithwis Das Gupta (1995)

[4] Bowman, A.W. and Azzalini, A. (1997). Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations. Oxford University Press, Oxford.