A METHOD FOR COMPUTING STATE PROBABILITIES UNDER COMPETING RISKS

Lawrence L. Wu
New York University

Steven P. Martin
Christopher Newport University

September 2012

Research funding from NICHD (R01 HD 29550) is gratefully acknowledged. Direct all correspondence to Lawrence L. Wu, Department of Sociology, Puck Building, 295 Lafayette Street, 4th floor, New York University, New York, NY 10012-9605, lawrence.wu@nyu.edu.

ABSTRACT

This paper outlines a computationally intensive method for computing state probabilities in problems involving competing risks. Consider individuals who begin life in a single origin state and who are subsequently exposed to $j = 1, \ldots, J$ competing risks. We discuss a computationally intensive method that provides, for any arbitrary time t, the $J + 1$ quantities $\{p_0(t), p_1(t), \ldots, p_J(t)\}$, where $p_0(t)$ denotes the probability of occupying the origin state at time t and $p_j(t)$ denotes the probability of having experienced transition j by time t. The resulting state probabilities provide a more easily interpretable set of quantities that can be useful to those analyzing competing risks.

Competing risk and multistate life table models are used frequently in demography, in part because they represent straightforward extensions of single-transition hazard or single-decrement life table methods. It is, however, considerably less straightforward to interpret the resulting competing risk survivor or multistate life table probabilities. Students encountering estimates are often confused when probabilities across transitions do not sum to one or when a synthetically estimated probability differs markedly from the corresponding descriptive proportion. Their confusion may not be clarified by statements such as "this should interpreted as the probability of this transition under a counterfactual in which individuals are thought to remain exposed to risk." A more serious issue is that the subtleties contained in such counterfactual statements could lead readers and analysts to misinterpret findings.

To fix ideas, consider a competing risk problem in which there is a single origin state and J potential destination states and let $T_1, T_2, \ldots, T_J$ denote the random variables for the event times associated with the J competing risks. Within this framework, what is observable to the analyst is the random variable $T = \min(T_1, T_2, \ldots, T_J)$ and the transition j for which $T = T_j$.

The "state probabilities" that are the subject of this paper are conceptualy straightforward and thus easy to interpret. For concreteness, consider J distinct causes of death and let τ denote some prespecified time chosen by the analyst. Then the state probabilities we seek are the $J + 1$ quantities $\{p_0(\tau)$, $p_1(\tau)$, $\ldots$, $p_J(\tau)\}$, where $p_0(\tau)$ denotes the probability that members in a population remain alive at age τ, $p_1(\tau)$ denotes the probability of dying from cause 1 by age τ, and so forth. It is well known how to estimate $p_0(\tau)$, but the literature to date, to the best of our knowledge, is silent on how to estimate the quantities $p_1(\tau), p_2(\tau), \ldots, p_J(\tau)$.

We outline a computationally intensive method for obtaining these $J + 1$ state probabilities. This method can be applied in both nonparametric and parametric settings, that is, for nonparametric estimates obtained from a Kaplan-Meier or life table estimator as well as when particular parametric forms are assumed for the distribution of event times. We further sketch and apply bootstrap methods to obtain standard errors and confidence intervals for the $J + 1$ state probabilities.

We have previously used this method to compute state probabilities in a paper examining cohort trends in premarital first birth (England, Wu, and Shafer 2012); this paper also contains a very brief and somewhat terse description of the method in an appendix. The goal of that paper was substantive, whereas the object of this paper is didactic and methodological—to discuss in fuller detail the theory behind the method.

BACKGROUND

Many demographic processes involve so-called *competing risks.* Demographers studying mortality often distinguish between distinct causes of death and the corresponding cause-specific mortality risks. Pregnant women can take their pregnancy to term, but some may terminate a pregnancy via an elective abortion, while others may suffer a miscarriage. Individuals can exit unemployment by finding and being hired into a new job, by exiting the labor force temporarily, or by retiring, thus exiting permanently. Although most students at a given grade level will, at the end of the school year, proceed to the next grade level, some may may change schools while others may be held back or drop out of school altogether.

Demographers encountering such problems often employ multistate life table or competing risk event history methods. At one level, these methods involve relatively straightforward extensions of classic single-decrement life table or single-transition hazard models. However, interpreting the resulting estimates from such models is far less straightforward.

THEORY

Let

$$S_j(t) = \exp\Big[-\int_0^t r_j(u)\,du\Big] \qquad j = 1,\ldots,J\,, \tag{2}$$

denote the cause-specific survivor probability for states $j = 1,\ldots,J$. The textbook interpretation of $S_j(t)$ (see, e.g., Cox and Oakes 1984; Crowder 2001; Preston, Heuveline, and Guillot 2001; Wu 2003) involves a counterfactual, often difficult to convey to students and readers, in which $1 - S_j(t)$ denotes the expected proportion of the population who has experienced the jth cause of death by age t *were all other causes of death to be eliminated.*

When $J = 1$, the quantities $S(t)$ and $1 - S(t)$ provide the probability of remaining in the origin state and the probability of having moved to the single destination state, with these two quantities necessarily summing to one. What often confuses students and lay readers is that no such relationship holds when $J \geq 2$ for the quantities $S_0(t)$ and the $1 - S_j(t)$, with the sum of these quantities in general differing from one at all times t.

One standard textbook discussion of this phenomena asks one to imagine a counterfactual world in which one cause of death, j', has been eliminated, but in which mortality from all other causes remains unchanged. If so, by construction we would have a zero risk of death from cause j' with the remaining $r_j(t)$

and $S_j(t)$ unchanged. This then impies that the *probability* of death from other causes will necessarily rise, despite no change in the remaining $J-1$ cause-specific mortality risks. This is then used to motivate the classical counterfactual interpretation of $S_j(t)$.

Let $T_1, T_2, \ldots, T_J$ denote the random variables for the event times associated with the J competing risks; then what is observed by the analyst is the random variable $T = \min(T_1, T_2, \ldots, T_J)$ and the transition j associated with T, i.e., the j for which $T = T_j$. The state probabilities we seek are the $J+1$ quantities $\{p_0(\tau), p_1(\tau), \ldots, p_J(\tau)\}$, where $p_0(\tau)$ denotes the probability that members in a population remain alive at age τ, $p_1(\tau)$ denotes the probability of dying from cause 1 by age τ, and so forth.

The probability at time t of not having experienced any of the J possible events (and thus the probability of continuing to reside in the origin state) is well known and is given by

$$S_0(t) = \exp\Big[-\sum_{j=1}^{J} \int_0^t r_j(u)\, du\Big]$$

however, the literature to date, to the best of our knowledge, is silent on how to estimate the quantities $p_1(\tau), p_2(\tau), \ldots, p_J(\tau)$.

[Paragraph, not yet written, on why getting an analytic expression for $p_j(t)$ is difficult. Core problem lies in $T = \min(T_1, T_2, \ldots, T_J)$. To do things analytically would first require obtaining an expression for the joint distribution of the J event times $f(t_1, t_2, \ldots, t_J)$. Then obtaining $p_j(t)$ would presumably require integrating the joint distribution over all $t_j < t_k, k \neq j$. In an nonparametric setting, this would be really difficult!]

It is, however, possible to approach this problem using computationally intensive methods. The key idea is that both nonparametric and parametric estimates of $S_j(t)$, although interpretable only under seemingly restrictive counterfactuals, are asymptotically consistent under the usual conditions. This further implies that $F_j = 1 - S_j$ yields an asymptotically consistent estimate of the cumulative distribution of the event times for the jth competing risk. Sampling with replacement from $F_j = 1 - S_j$ then allows one to generate draws from the jth event time distribution.

Let t_j^* denote a simulated event time for the jth competing risk, with the superscript $*$ denoting that these are simulated event times, and let τ denote the time at which to evaluate the state probabilities $\{p_0(\tau), p_1(\tau), \ldots, p_J(\tau)\}$. Then given estimates of the J cumulative distributions, $\widehat{F}_j(t) = 1 - \widehat{S}_j(t)$,

1. Generate J simulated event times $t_1^*, \ldots, t_J^*$ from the $F_j(t)$.

2. Let $t^* = \min(t_1^*, \ldots, t_J^*)$ denote the smallest of the J simulated times from step (1) and let j^* denote the transition associated with t^*, as derived from $t^* = t_j$. (When doing this nonparametrically, one can break possible "ties" by, for example, a coin flip.)
3. If $t^* \leq \tau$, then the simulation in (1) implies a transition to j^* during the period of interest; otherwise, all simulated times are greater than τ, in which case the simulations imply that this case has not transitioned out of the origin state by τ. Thus, steps (1)–(3) yield one of $J + 1$ possibilities: the origin state or a transition to one of the J destinations states.
4. Repeating steps (1)–(3) M times yields frequencies summing to M, with these counts in turn yielding simulated probabilities $\{p_0^*(\tau), p_1^*(\tau), \ldots, p_J^*(\tau)\}$ for the $J + 1$ possible states.
5. We then repeat step (4) until the simulated probabilities satisfy a convergence criteria.

We have not yet settled on a specific illustrative application, but as noted above, we have previously used this method in analyses of cohort trends in premarital first births (England, Wu, and Shafer 2012). In that paper, we posited two competing risks: (1) a premarital conception that was taken to term and that yielded a first birth, and (2) a first marriage not preceded by such a premarital conception. We then specified a proportional hazard model using a parametric specification for the baseline hazard (a splined piecewise Gompertz model) for the two competing risks. Proportional effects of dummy variables for five-year birth cohorts (1925-29, 1930-34, . . . , 1960-64) were used to operationalize cohort trends in the two competing risks. Data analyzed were obtained by pooling the marital and fertility histories contained in the marital and fertility supplements to the June 1980, 1985, 1990, and 1995 Current Population Surveys.

Table 1, drawn from England, Wu, and Shafer, compares the observed proportions in the origin and two destination states with state probabilities as estimated from the procedure outlined above. These results show a close fit between observed and estimated state probabilities for successive cohorts of both white and black women. (Agreement in this case will depend both on the asymptotic properties of the above procedure and on fit of the parametric model specified for these processes.)

[Table 1 about here]

Although Table 1 shows good fit between observed and predicted percentages in the three states, the point of the exercise may seem somewhat academic—the observed percentages are easily computed. Thus, the real utility of the above procedure is that it can be easily adapted to obtain inferences for state probabilities under conditions that deviate from what is observed—for example, the implications of the model for state probabilities were values of specific covariates to be altered or if the overall composition of a population were to change over time. The above procedure can also be used in decompositions that are frequently used in demographic research. For example, England, Wu, and Shafer use these methods to pose counterfactuals questions concerning trends, for example, "what would our model say about trends in premarital conceptions were there to have been no cohort trend in women's age-specific first marriage risks?" See Figure 1.

[Figure 1 about here]

REFERENCES

Cox, D. R., and D. Oakes. 1984. *Analysis of Survival Data*. London: Chapman and Hall.

Crowder, Martin. 2001. *Classical Competing Risks.* Boca Raton, FL: Chapman and Hall/CRC.

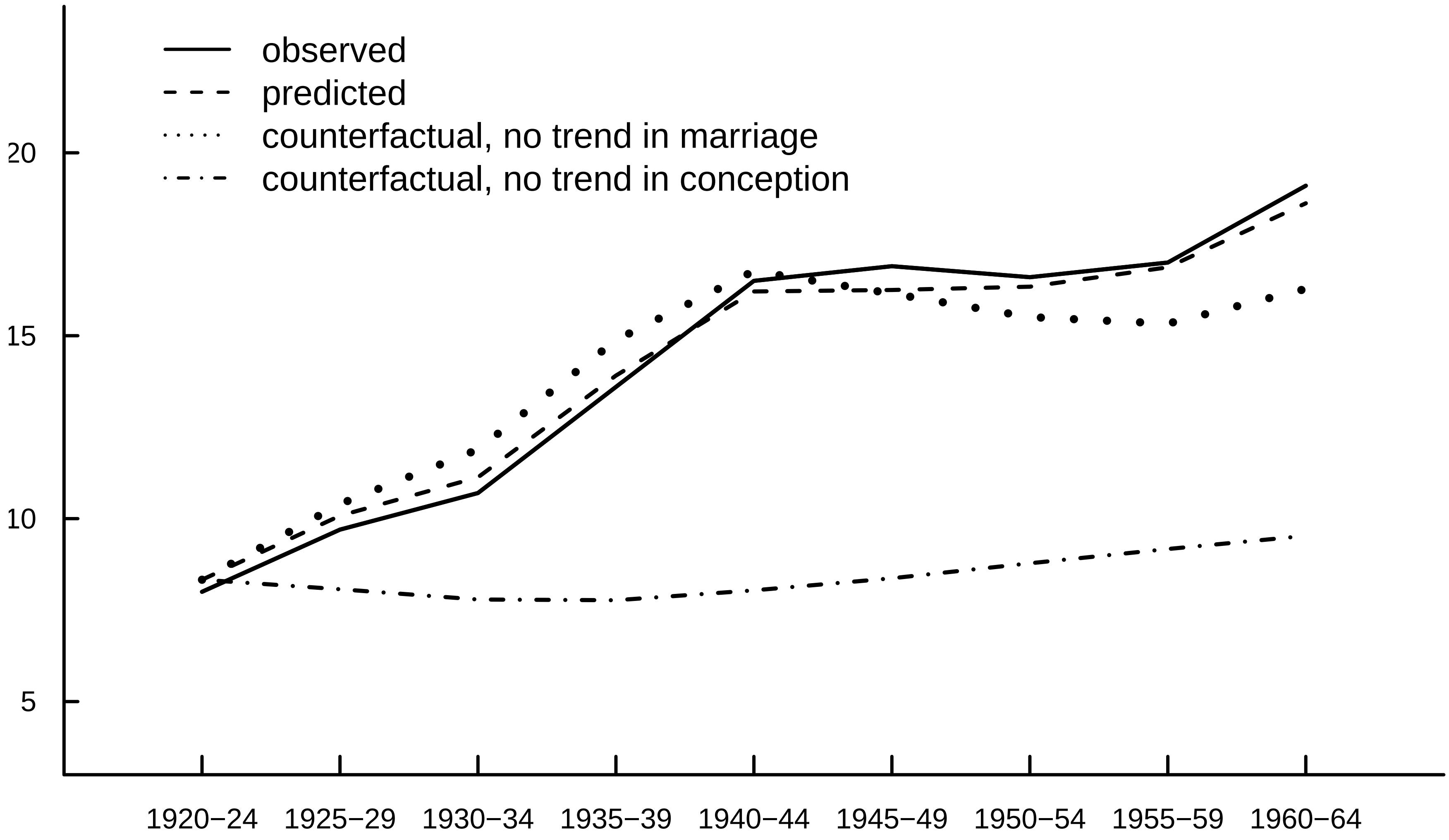
England, Paula, Lawrence L. Wu, and Emily Fitzgibbons Shafer. 2012. "Cohort Trends in Premarital First Births: Premarital Conceptions vs. the Retreat from Marriage?" Unpublished manuscript, Department of Sociology, New York University.

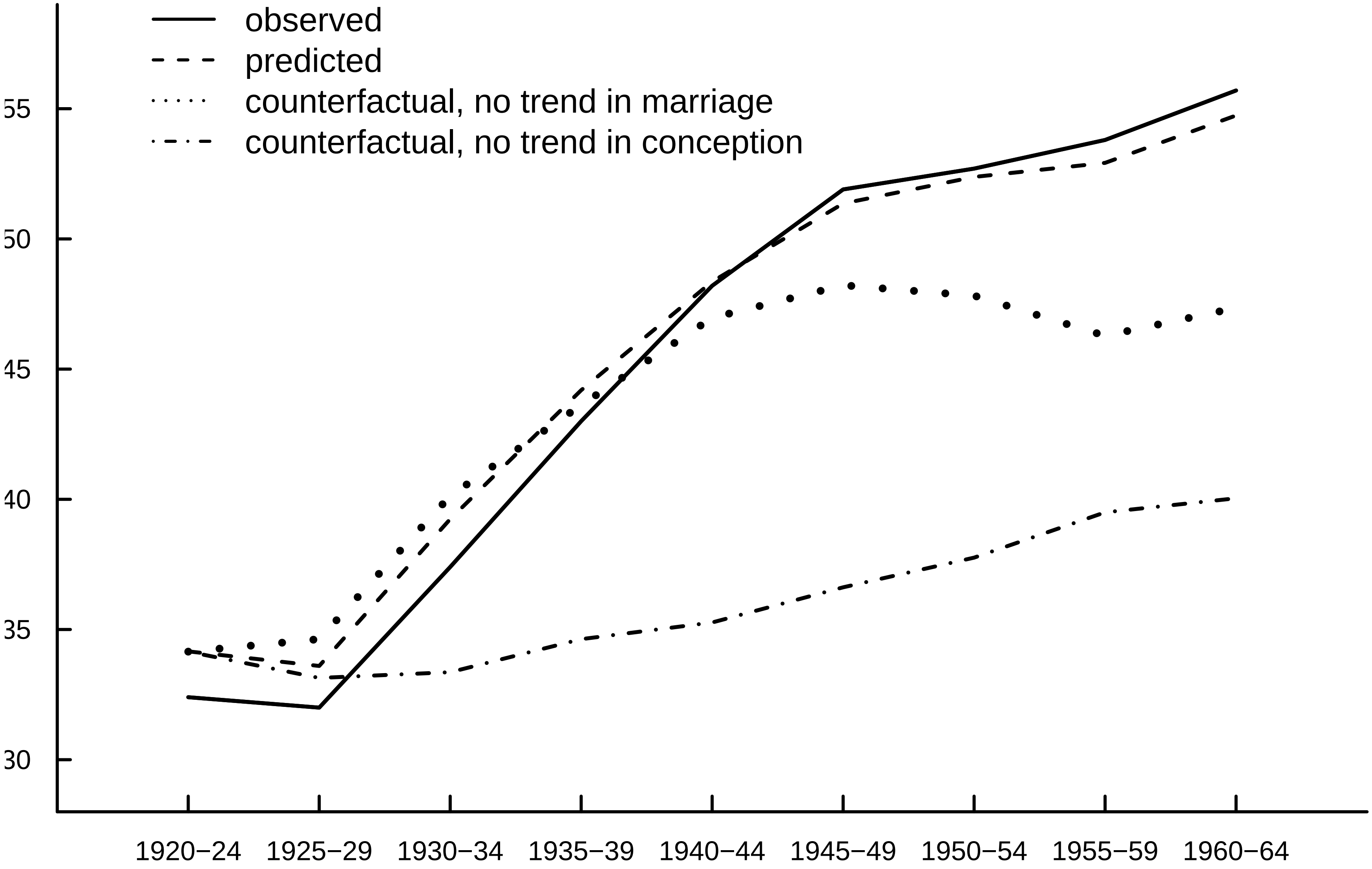
Preston, Samuel H., Patrick Heuveline, and Michel Guillot. 2001. *Demography: Measuring and Modeling Population Processes.* Oxford: Blackwell.

Wu, Lawrence L. 2003. "Event History Models for Life Course Analysis." Pp. 477–502 in Jeylan Mortimer and Michael Shanahan (Eds.), *Handbook of the Life Course,* New York: Plenum.

Table 1: Percent by age 25 experiencing: (1) a premarital conception taken to term, (2) a first marriage not preceded by such a conception, or (3) neither. Observed and predicted percentages for white and black women by birth cohort.

	Neither		First Marriage		Premarital Conception	
	observed	predicted	observed	predicted	observed	predicted
White women						
1920-24	21.0	20.9	71.0	70.7	8.0	8.3
1925-29	17.3	17.7	73.0	72.2	9.7	10.1
1930-34	14.6	14.9	74.8	74.0	10.7	11.1
1935-39	13.5	13.9	72.8	72.2	13.6	13.9
1940-44	14.5	15.5	69.0	68.3	16.5	16.2
1945-49	16.8	18.6	66.3	65.2	16.9	16.2
1950-54	22.8	22.6	60.6	61.1	16.6	16.3
1955-59	28.0	26.9	55.0	56.2	17.0	16.9
1960-64	31.6	30.3	49.2	51.1	19.1	18.6
Black women						
1920-24	21.4	19.5	46.1	46.4	32.4	34.2
1925-29	17.1	17.0	51.0	49.4	32.0	33.6
1930-34	16.1	15.3	46.5	45.5	37.4	39.2
1935-39	16.3	16.2	40.6	39.6	43.0	44.2
1940-44	16.1	15.9	35.7	35.8	48.2	48.4
1945-49	16.3	17.8	31.8	30.9	51.9	51.4
1950-54	20.4	20.2	27.0	27.4	52.7	52.4
1955-59	25.6	24.4	20.5	22.7	53.8	52.9
1960-64	26.6	23.6	17.7	21.7	55.7	54.7


observed
predicted
counterfactual, no trend in marriage
counterfactual, no trend in conception
20
15
10
5
1920−24
1925−29
1930−34
1935−39
1940−44
1945−49
1950−54
1955−59
1960−64


observed
predicted
counterfactual, no trend in marriage
counterfactual, no trend in conception
55
50
45
40
35
30
1920−24
1925−29
1930−34
1935−39
1940−44
1945−49
1950−54
1955−59
1960−64