HEDG Working Paper 10/20

A Heap of Trouble? Accounting for Mismatch Bias in Retrospectively Collected Data on Smoking

Haim Bar
Dean Lillard

July 2010

york.ac.uk/res/herc/hedgwp

A Heap of Trouble? Accounting for Mismatch Bias in Retrospectively Collected Data on Smoking

Haim Y. Bar* Dean R. Lillard†

July 2010

Abstract

When event data are retrospectively reported, more temporally distal events tend to get "heaped" on even multiples of reporting units. Heaping may introduce a type of attenuation bias because it causes researchers to mismatch time-varying right-hand side variables. We develop a model-based approach to estimate the extent of heaping in the data, and how it affects regression parameter estimates. We use smoking cessation data as a motivating example to describe our approach, but the method more generally facilitates the use of retrospective data from the multitude of cross-sectional and longitudinal studies worldwide that already have and potentially could collect event data.

1 Introduction

The primary goal of this paper is to provide methods to estimate the effects of heaping in retrospective studies on parameter estimates in regression models, to quantify the degree to which heaping affects statistical inference, and to provide a method by which to recover parameter estimates of interest that are less biased.

*Cornell University

†Cornell University and DIW. This research was supported by Award # R01 HD048828 from the National Institutes of Health

As a motivating example, we use retrospectively reported data on smoking behavior from the Panel Study of Income Dynamics (PSID) and the Current Population Survey - Tobacco Use Supplements (CPS-TUS). With those data we analyze the effect of certain covariates on a smoker's decision to quit smoking. To obtain the smoker's quitting time, there are three common ways to phrase the question in the survey. The PSID survey used the question "How old were you when you quit smoking?" Other surveys (like the CPS) used one of "How long ago (in years) did you quit smoking?", while some surveys use: "In what year did you quit smoking?" These prototypical questions often result in "heaped" responses – reported answers that tend to have non-smooth distributions with peaks at multiples of five and ten years. It is conceivable that heaped responses may result in biased estimates in linear regression models. For instance, if we convert the reported ages to calendar years and let p_t be the probability that a smoker quit in Year=t, then we may be interested in fitting the logistic model

$$\log\left(\frac{p_t}{1-p_t}\right) = X_t\beta + \varepsilon_t$$

where X is a design matrix, containing covariates such as age, health status, and cigarette prices. In this example, we may be interested in testing the hypothesis that increased taxation on tobacco products causes an increase in smoking cessation rate, and we need to know how heaping might affect the estimation and significance of the coefficient on tax.

We take a model-based approach and assume that the distribution of quit-times (expressed either in terms of age, or time elapsed since quitting, or calendar years) mixes outcomes generated by two processes. The first is a stochastic process, representing the smokers who would eventually quit smoking randomly, while the second normally distributed component results from decisions of smokers who quit in response to some external conditions, such as serious health issues, changes in family or employment status, or, perhaps, due to changes in cigarette prices. We fit the mixture model using the Expectation Maximization (EM) algorithm (Dempster et al., 1977) or a Monte Carlo Markov Chain (MCMC) simulation, and obtain a smooth, parametric distribution of quit-times. To estimate whether and how heaping biases regression parameter estimates, and the misclassification rate in the responses, we sample from this mixture distribution and fit the linear model to obtain parameter estimates for the regression model. We then compare

the average estimates with the ones obtained from the observed (heaped) data, to obtain an estimate for the bias that is due to heaping.

We also use the fitted distribution to directly compute the probability of two types of misclassification errors that are (potentially) present in data on smoking behavior. These errors reflect the probability that a person reports he has quit when in fact he currently smokes, and the probability that a person erroneously reports he is currently a smoker when in fact he quit. Hausman et al. (1998) label these errors as α_0 and α_1 respectively. They show that, in models with limited dependent variables, these types of measurement error are not innocuous and they develop a method to parametrically adjust for the presence of such errors. Here we use our fitted and the empirical distributions of start and quit ages to directly estimate the average error of each type at each age at which smokers are at risk to quit. With those computed error probabilities we can then directly adjust the likelihood function in the Hausman et al. framework to recover the true parameters of interest.

We use Monte Carlo methods to simulate data with subjects that respond to a change in a single covariate of interest in a known way. We then introduce heaping into the data using (a set of) heaping rules that replicate the pattern of heaping in the observed distribution of quit ages. With these data we test whether our bias adjustment factor recovers the true underlying distribution. We also explore whether bias is mitigated by two alternative methods developed in Lillard et al. (2009) and another widely known method developed by Heitjan and Rubin (1990).

The example we study epitomizes the bias that heaping (potentially) introduces in all models of event data. Studies that use events as either dependent or independent variables data abound. For example, a Google Scholar (http://scholar.google.com/) article search yielded thousands of hits for "age at marriage," "time of marriage," "age at birth of first child," etc. Because event data are the focus of so much attention, there is great value in developing methods to reduce potential bias that heaping introduces.

Several studies identify factors associated with respondents' recall accuracy. Recall duration or time since event is a strong predictor of the quality of retrospective reports on marital history in the US Panel Study of Income Dynamics (PSID, Peters (1988)), age at first sex in the National Longitudinal Survey of Youth 1979 (NLSY79) Wu et al. (2001), and post-partum amenorrhea (the interval after a pregnancy before menstruation returns) in the Malaysian Family Life Surveys (MFLS) Beckett et al. (2001). Researchers also agree that respondents more accurately report when an event occurred

if the event is more salient to the respondent. Kenkel et al. (2004) find that smokers are more likely to report the same starting age across different waves of the NLSY79 if they are or were heavier smokers. Although marriage and divorce are both salient life events, Peters (1988) shows evidence that dates of divorce are reported less consistently than dates of marriage and conjectures that the difference may arise because divorce is less socially acceptable. Researchers have also linked with recall accuracy with demographic characteristics such as education and race/ethnicity Kenkel et al. (2004), Peters (1988), question wording Peters (1988), and even arithmetic facility Wu et al. (2001).

The paper is organized as follows. In Section 2 we review the extant literature and briefly review methods suggested there to mitigate the bias due to heaping. In Section 3 we provide graphic illustrations and several observations from the PSID data set. In Section 4 we describe two parametric mixture models for smoking cessation data and derive parameter estimates using the EM algorithm or MCMC simulations. Section 5 contains a simulation study, and in Section 6 we show how one may use our model-based approach to estimate the bias in regression parameter estimation, and the misclassification rate in the response and the set of people who are 'at risk' to quit smoking. We conclude with a discussion in Section 7.

2 Background

Our analysis follows and builds on the work of Little (1992), Torelli and Trivellato (1993), Heitjan and Rubin (1990). Each of these studies recognized the potential problem that heaping might cause. Little (1992) provides a succinct review of event history analysis and missing-data methods. Torelli and Trivellato (1993) propose solutions to heaping in data on unemployment spells of Italian youth that involve the specification of a parametric model of the errors in the reformulated likelihood function, adding a dummy variable to flag ex-smokers who heap or do not heap, and smoothing the data as recommended by Heitjan and Rubin (1990). Heitjan and Rubin (1990) attempt to solve the problem of heaping by coarsening data over broad intervals centered around the heaping unit. They use a simple framework in which survey respondents use a single heaping rule. In more recent work, Forster and Jones (2001) model smoking initiation and smoking cessation using UK data in discrete-time hazard models with and without controls for heaping.

They implement solutions proposed by Torelli and Trivellato (1993) but find little evidence that heaping biases coefficients on cigarette tax in models of smoking duration. Pudney (2007) focuses on heaping in consumption expenditure data, and changes in heaped responses between consecutive waves. Similar to our findings, he notes that in any group of survey respondents, multiple heaping rules are used. His analysis focuses on patterns of transition between heaping points for the same individual.

The working paper of Lillard et al. (2009) is the first study, of which we are aware, to compare side-by-side estimates of the coefficient on price in 4 specifications of a model of the probability of smoking cessation. Using a naive treatment of the data (OLS) as the reference the three other specifications each attempt to account for heaping with semi-parametric (sample selection) or parametric methods. They show that, relative to the OLS estimates, each specification with heaping controls yields substantially different coefficient estimates. Unfortunately Lillard et al. (2009) cannot evaluate the extent to which each method reduces bias because they do not observe the true underlying distribution of quit ages. Consequently one cannot evaluate which method is preferred or even if any of the methods reduce the apparent bias. In a new working paper, Kenkel and LeCates (2010) attempts to fill the gap by simulating the distribution of quit ages with Monte Carlo methods, generating heaping to replicate observed patterns of heaping in actual data. They then examine how well 6 different correction methods perform in recovering the parameters of interest, including variants of the three methods examined by Lillard et al. (2009).

Our model-based approach is similar to that of Wright and Bray (2003), where a hierarchical model is used to estimate the effect of heaping (via MCMC simulations). Crockett and Crockett (2006) deal with the consequences of heaping in the British religious census of 1851, and they point out that it is not plausible that the "coarsening" occurs at random, and hence, it is not 'ignorable' in the sense of Heitjan and Rubin (1990). Lambert (1992) deals with a special case of heaping, where there is excess in observations of 0 in count data. In her analysis, she shows that one has to account for heaping in Poisson regression, in the presence of 'zero-inflated' data.

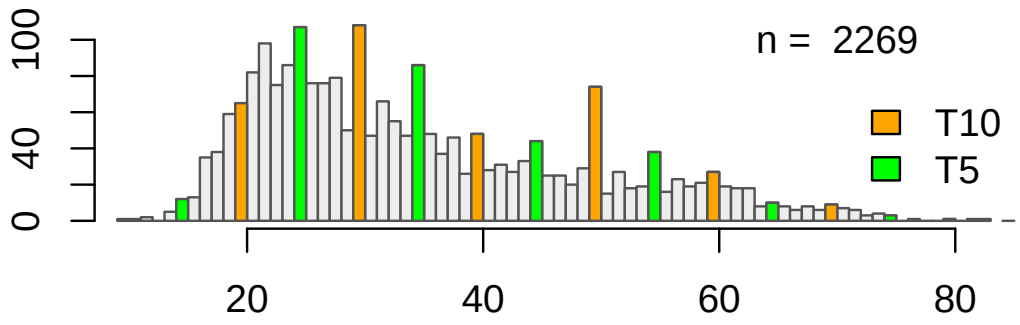

Figure 1: The distribution of reported quit ages in the 1986 PSID survey.

3 Data Explorations

Before going in detail into our model-based approach to assess the effect of heaping on parameter estimation, we begin with a number of examples.

Figure 1 shows the distribution of reported quitting age of 2,269 respondents in the 1986 wave of the PSID survey. The labels 'T5' and 'T10' correspond to ages that are multiples of five or ten years, respectively. It may be argued that at least some of the heaping observed in the data is real, and not a result of respondents rounding to the nearest multiple of five or ten years. We see evidence in the data that this is not the case (see Section 4), but our proposed model-based approach allows to account for "true heaping."

We see in Figure 2 that older respondents tend to heap their start-smoking age much more than the younger cohort. There is no obvious reason that a higher proportion of older people would have started to smoke at age 25 that similar people who happen to be younger. To check whether reporting heaped values is related to age, we fit a logistic regression model

$$\log\left(\frac{P_{I[T5]}}{1-P_{I[T5]}}\right) = \beta_0 + \beta_1 \times CurrentAge$$

where $I[T5]$ takes the value 1 if the reported start-age is a multiple of five, and 0 otherwise, and test the null hypothesis that $\beta_1 = 0$. Figure 3 shows the fitted logit function and the parameter estimates for this model. The odds for reporting a T5 age increase exponentially with age ($p < 6.1e-9$). The horizontal dashed line represents the expected proportion of people starting to smoke at an age that is a multiple of five (20%).

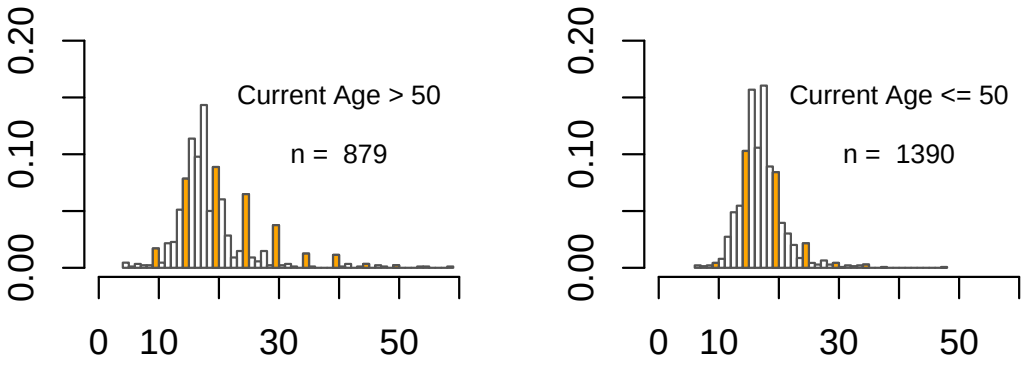

Figure 2: Reported start-smoking ages for two current-age groups.

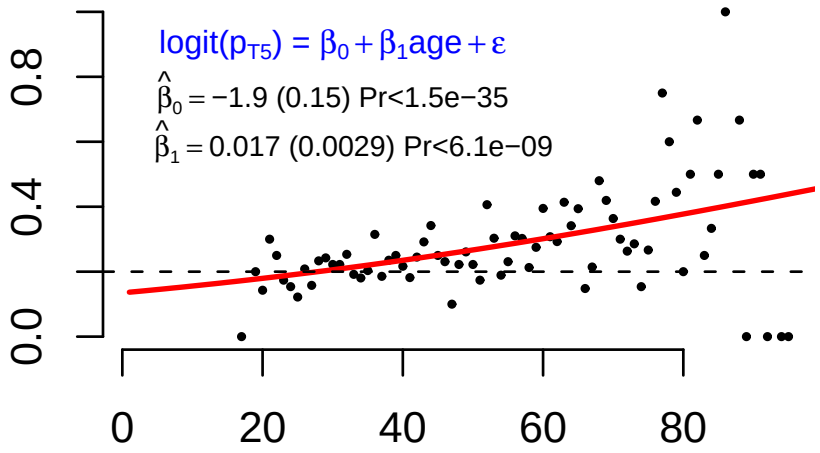

Figure 3: The probability of a heaped start-smoking age, as a function of current-age.

In the surveys we analyzed we find that among people whose age at the time of the survey was less than or equal to 50, the starting age appears to be distributed approximately normal, as in the right panel of Figure 2, with mean $\approxeq 17.5$ and variance $\approxeq 11$.

Before we proceed with parametric modeling of quitting ages in the presence of heaping, we analyze the distribution of a related quantity. Clearly, quitting age can be written as the sum of starting age and the number of years a person has been smoking. The former can be estimated empirically

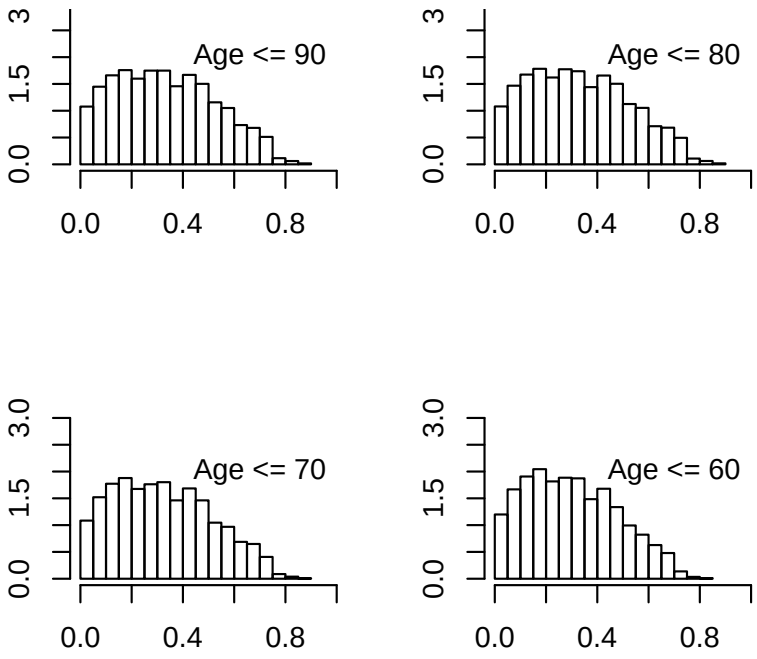

Figure 4: The distribution of $pys/100$ for different subsets from the sample.

by considering the subset of younger respondents whose starting ages will e less error prone. Because the number of years a person has been smoking depends on the person's current age, we transform it into percentage terms, i.e.

$$pys/100 = \frac{YearsSmoking}{CurrentAge},$$

where pys =Percent Years Smoking, the result is much less sensitive to the choice of the cohort. Figure 4 shows the distribution of the variable $pys/100$ for different subsets.

A natural choice for fitting this distribution is the generalized beta distribution with support (L,H) where $L = min(YearsSmoking/CurrentAge) > 0$ and $H = max(YearsSmoking/CurrentAge) < 1$, with the probability distribution function

$$f_B(x; \alpha, \beta, L, H) = \frac{(x-L)^{\alpha-1}(H-x)^{\beta-1}}{B(\alpha,\beta)(H-L)^{\alpha+\beta-1}} \quad (1)$$

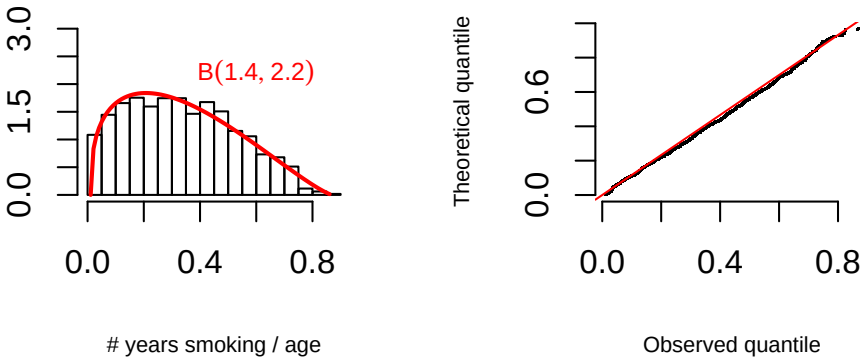

Figure 5: Fitting the distribution of $pys/100$.

where $B(,)$ is the Beta function. Figure 5 shows the fitted generalized Beta distribution (left) and a quantile-quantile plot (right).

Interestingly, if we consider the subset of quitters who are at least 49 years old at the time of the survey and plot the percentage of years smoking starting from age 49, i.e.

$$pys49/100 = \frac{YearsSmokingAfter48}{CurrentAge - 48},$$

then the distribution of $pys49/100$ is almost uniform, as can be seen in Figure 6. A possible explanation for the difference between the two distributions is that quitting age is actually a mixture of two distributions. This is the basic idea behind the model-based approach in the following sections, which can be summarized intuitively as follows. A certain fraction of smokers will decide to quit randomly, according to some stochastic process (or will die without ever quitting), while the rest of the population will quit smoking following a major life-changing event (such as heart attack, birth of a child, retirement, etc.), *regardless* of when they started smoking. Alternatively, the mixture distribution can be within each person so that, in a steady state, smokers quit according to a random process that is independent of external events. When a shock occurs, some smokers are pushed into the second distribution. Quit decisions for this group are distributed as described above. More generally this structure allows for the presence of a large number of smokers who are

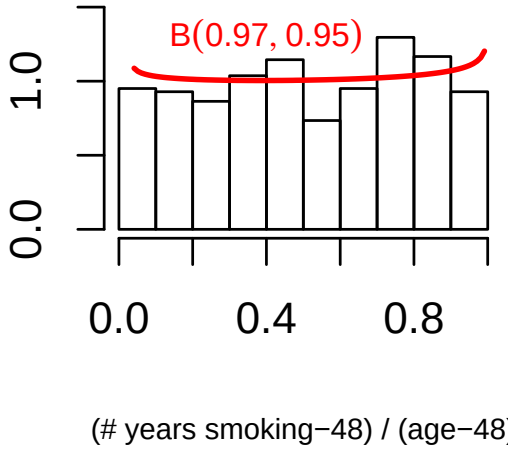

Figure 6: The distribution of $pys49/100$.

not subject to shocks but who still quit according to a random process - a point to which we return later.

4 Statistical Models

We propose a parametric model approach to fit smoking cessation data in order to assess the effect that heaping has on regression parameter estimates. We fit a mixture in which the first component represents the population of smokers who quit randomly, and the second component represents the population of smokers who quit as a response to certain events. The first component may be modeled differently, depending on the form of the smoking cessation question in the survey. The second component allows us to incorporate covariates of interest and assess their effect on people's decision to quit. We use two methods to estimate the mixture-model parameters, namely, the Expectation-Maximization (EM) algorithm, and Monte Carlo

Markov Chain (MCMC). The EM algorithm is particularly well suited for fitting parametric models to such data, since the model involves missing observations. Specifically, missing data include which mixture component each subject belongs to and quitting ages, which are censored for people who still smoke at the time of the survey. The MCMC approach, while more computationally intensive, allows us to modify the assumed underlying mixture model more easily. Here, we present both as viable approaches, and while they have different implementation considerations, they both benefit from the parsimony of our model-based approach. Here, we propose a family of parametric models for fitting quitting ages. We describe how we use these models to estimate the attenuation bias due to heaping in Section 6.

4.1 Model 1 – a Beta/Normal Mixture

The first model is motivated by the observations in Section 3 where we saw that the distribution of percentage of time smoking is different for the older cohort. In the first mixture component, in which smokers quit randomly, we use the trivial identity:

$$QuitAge = StartAge + CurrentAge \times \frac{YearsSmoking}{CurrentAge}. \tag{2}$$

As discussed in the previous section, the distribution of $StartAge$ is approximately normal, and its mean and variance can be estimated using a subset of younger respondents, and $YearsSmoking/CurrentAge$ can be estimated using (1).

For the second mixture component, we assume that quitting ages are distributed normally with mean θ and variance σ^2. Hence, given the starting age s_i and the current age c_i, the probability distribution function of the quitting age q_i, is

$$f(q_i|s_i;c_i) = b_i f_B\left(\frac{q_i - s_i}{c_i};\alpha,\beta,L,H\right) + (1-b_i)\, f_N\left(q_i;\theta,\sigma^2\right), \tag{3}$$

where the unobserved indicator variables b_i are distributed $Bernoulli\,(p)$.

Notice that the normal component allows us to incorporate covariates into the model, since it can be written in the usual form in normal linear regression, as $X\beta + \varepsilon$, where β is a vector of effects, and ε is the random error. In the next section, we describe in detail how we perform the regression

analysis while accounting for heaping. Briefly, we take a hybrid approach where in the first step (the EM algorithm) we include only the overall mean and variance of the normal component in the mixture (3), and in a second step we use a Monte Carlo approach to estimate the regression parameters. This approach has the advantage that the EM estimation is more parsimonious, and does not depend on the number of covariates in the regression model.

The complete-data likelihood for one observation, given the starting age s_i, is

$$L_C = \left(p \frac{\left(\frac{q_i - s_i}{c_i} - L\right)^{\alpha-1} \left(H - \frac{q_i - s_i}{c_i}\right)^{\beta-1}}{B(\alpha,\beta)(H-L)^{\alpha+\beta-1}} \right)^{b_i} \times \left((1-p)\sqrt{\frac{1}{2\pi\sigma^2}} \exp\left\{ -\frac{(q_i-\theta)^2}{2\sigma^2} \right\} \right)^{1-b_i},$$

so the complete-data log-likelihood for a sample of m independent subjects is

$$\begin{aligned} l_C &= \sum_{i=1}^{m} b_i \left[\begin{array}{c} \log p + (\alpha-1)\log\left(\frac{q_i-s_i}{c_i} - L\right) + (\beta-1)\log\left(H - \frac{q_i-s_i}{c_i}\right) \\ -\log B(\alpha,\beta) - (\alpha+\beta-1)\log(H-L) \end{array} \right] \\ &\quad + \sum_{i=1}^{m} (1-b_i) \left[\log(1-p) - \frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{(q_i-\theta)^2}{2\sigma^2} \right]. \end{aligned}$$

To construct the $Q\left(\Psi;\Psi^{(k)}\right)$ function for the EM algorithm, where

$$Q\left(\Psi;\Psi^{(k)}\right) \equiv E_{\Psi^{(k)}}\left[l_C(\Psi)\,|s,q\right], \Psi = \left\{p, L, H, \alpha, \beta, \theta, \sigma^2\right\}, \tag{4}$$

we need to replace the missing data variables with their expectations, given the current parameter estimates. For the Bernoulli variables, b_i, we simply use Bayes rule to find the posterior probability that subject i quitting age is distributed according to the generalized binomial distribution:

$$E(b_i) = \frac{p f_B\left(\frac{q_i-s_i}{c_i}; \alpha, \beta, L, H\right)}{p f_B\left(\frac{q_i-s_i}{c_i}; \alpha, \beta, L, H\right) + (1-p) f_N(q_i; \theta, \sigma^2)} \tag{5}$$

The parameter estimates are obtained by maximizing the Q function with respect to Ψ. We obtain the following formulas for the estimates from k-th

iteration of the EM algorithm:

$$\begin{aligned} p^{(k)} &= \frac{\sum_{i=1}^{m} b_i^{(k-1)}}{m} \\ \theta^{(k)} &= \frac{\sum_{i=1}^{m} \left(1 - b_i^{(k-1)}\right) q_i}{\sum_{i=1}^{m} \left(1 - b_i^{(k-1)}\right)} \\ \sigma^{2^{(k)}} &= \frac{\sum_{i=1}^{m} \left(1 - b_i^{(k-1)}\right) \left(q_i - \theta^{(k-1)}\right)^2}{\sum_{i=1}^{m} \left(1 - b_i^{(k-1)}\right)}, \end{aligned}$$

and for L, H, α, β we obtain the estimates by solving the following set of equations numerically:

$L^{(k)}$:

$$\sum_{i=1}^{m} b_i^{(k-1)} \left[\frac{\alpha^{(k-1)} + \beta^{(k-1)} - 1}{H^{(k-1)} - L} - \frac{\alpha^{(k-1)} - 1}{\frac{q_i - s_i}{c_i} - L} \right] = 0$$

$H^{(k)}$:

$$\sum_{i=1}^{m} b_i^{(k-1)} \left[\frac{\beta^{(k-1)} - 1}{H - \frac{q_i - s_i}{c_i}} - \frac{\alpha^{(k-1)} + \beta^{(k-1)} - 1}{H - L^{(k-1)}} \right] = 0$$

$\alpha^{(k)}$:

$$\sum_{i=1}^{m} b_i^{(k-1)} \left[\log\left(\frac{q_i - s_i}{c_i} - L^{(k-1)} \right) - \log(H^{(k-1)} - L^{(k-1)}) \right] = \psi(\alpha) - \psi(\alpha + \beta^{(k-1)})$$

$\beta^{(k)}$:

$$\sum_{i=1}^{m} b_i^{(k-1)} \left[\log\left(\frac{q_i - s_i}{c_i} - L^{(k-1)} \right) - \log(H^{(k-1)} - \frac{q_i - s_i}{c_i}) \right] = \psi(\alpha^{(k-1)}) - \psi(\alpha^{(k-1)} + \beta)$$

where $\psi()$ is the digamma function.

If the distribution of quitting ages is truly a mixture in which some fraction of the population quits randomly, then including these people in the

regression model may diminish the power to detect significant effects. Furthermore, when a survey is conducted in waves, this model makes it possible to compare the proportions of people who quit randomly. These properties are important for policy makers, since they can estimate which proportion of the population may be persuaded to quit smoking, and estimate the effect of their policy (e.g. higher cigarette taxes, advertisement campaigns, etc.).

4.2 Model 2 – an inverse-Gaussian/normal mixture

Our second model also corresponds to questions of the form "How old were you when you quit smoking?" We assume that the population of smokers is a mixture of two groups, but in this case, in the first group the response is assumed to follow an inverse Gaussian distribution, with probability distribution function

$$f_I\left(x;\mu,\lambda\right)=\sqrt{\frac{\lambda}{2\pi x^3}}\exp\left\{-\lambda\frac{\left(x-\mu\right)^2}{2x\mu^2}\right\}$$

for $x>0$. The inverse Gaussian distribution (IG) is related to 'first passage time' in Brownian motion: given a stochastic process X_t with $X_0=0$ and $X_t=\nu t+\tau W_t$ where W_t is a standard Brownian motion with a positive drift ν, the first passage time, defined as

$$T_\alpha=inf\{0<t|X_t=\alpha\}$$

is distributed $IG\left(\frac{\alpha}{\nu},\frac{\alpha^2}{\tau^2}\right)$. In the context of smoking cessation, 'first passage time' refers to a smoker's decision to quit. The definition of the IG distribution and its intuitive interpretation make is a natural candidate for modeling event occurrences in general, and smoking duration in particular. This distribution has been used to model the emptiness of dams (Hasofer, 1964), purchase incidence (Banerjee and Bhattacharyya, 1976), and duration of strikes (Lancaster, 1972). The IG model is quite popular in the field of finance, where strategies for buying or selling portfolios are often determined using a 'first passage time' rule (buy/sell when the price of a stock reaches a certain threshold). Recently, the IG distribution was also used to model time until the first substitution in soccer games (Del Corral et al., 2008). Folks and Chhikara (1978) provide a detailed description of the distribution, its origin, properties, and applications. They noted that for several data sets

which were modeled using the IG distribution, the log normal, the Weibull, and the gamma distributions seemed equally adequate. However, they recommend using the IG distribution in lifetimes applications because of "its considerable exact sampling distribution theory" and because it is preferable to base the choice on the relation to an underlying physical mechanism. In the case of smoking cessation, the "underlying physical mechanism" may, in fact, be a psychological or social one, but at any rate, the 'first passage time' interpretation seems to make the IG model a natural choice to model the stochastic process component of our data.

For the second group, we assume that quitting ages are distributed normally, as in Model 1. We assume that the population of smokers is a mixture of these two distributions, so that given the starting age s_i, the probability distribution function of the quitting age q_i, is

$$f(q_i|s_i) = b_i f_I\left(q_i - s_i; \mu, \lambda\right) + \left(1 - b_i\right) f_N\left(q_i; \theta, \sigma^2\right), \tag{6}$$

As in Model 1, b_i are unobserved indicator variables, distributed $Bernoulli\,(p)$.

Note that among those in the sample who still smoke at the time of the survey, q_i is censored, and we only observe their current age, c_i. To estimate the parameters in the model, we consider only the quitters $(i = 1, ..., m)$, and we use the EM algorithm with b_i playing the role of missing data. The complete-data likelihood for one observation, given the starting age s_i, is

$$L_C = \left(p\sqrt{\frac{\lambda}{2\pi(q_i - s_i)^3}}\exp\left\{-\lambda\frac{\left((q_i - s_i) - \mu\right)^2}{2(q_i - s_i)\mu^2}\right\}\right)^{b_i} \times \left((1-p)\sqrt{\frac{1}{2\pi\sigma^2}}\exp\left\{-\frac{(q_i - \theta)^2}{2\sigma^2}\right\}\right)^{1-b_i},$$

so the complete-data log-likelihood for a sample of m independent subjects is

$$\begin{aligned} l_C &= \sum_{i=1}^{m} b_i\left[\log p + \frac{1}{2}\log\lambda - \frac{1}{2}\log\left(2\pi(q_i - s_i)^3\right) - \lambda\frac{\left((q_i - s_i) - \mu\right)^2}{2(q_i - s_i)\mu^2}\right] \\ &\quad + \sum_{i=1}^{m}\left(1 - b_i\right)\left[\log\left(1 - p\right) - \frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{(q_i - \theta)^2}{2\sigma^2}\right]. \end{aligned}$$

Collecting the constant terms, we get

$$\begin{aligned} l_C &= \sum_{i=1}^{m} b_i \left[\log p + \frac{1}{2}\log\lambda - \frac{3}{2}\log(q_i - s_i) - \lambda\frac{\left((q_i - s_i) - \mu\right)^2}{2(q_i - s_i)\mu^2} \right] \\ &\quad + \sum_{i=1}^{m} (1 - b_i) \left[\log(1 - p) - \frac{1}{2}\log\left(\sigma^2\right) - \frac{(q_i - \theta)^2}{2\sigma^2} \right] + K. \end{aligned} \tag{7}$$

Again, we replace the missing data variables with their expectations, given the current parameter estimates:

$$E(b_i) = \frac{p\sqrt{\frac{\lambda}{2\pi(q_i - s_i)^3}}\exp\left\{-\lambda\frac{((q_i - s_i) - \mu)^2}{2(q_i - s_i)\mu^2}\right\}}{p\sqrt{\frac{\lambda}{2\pi(q_i - s_i)^3}}\exp\left\{-\lambda\frac{((q_i - s_i) - \mu)^2}{2(q_i - s_i)\mu^2}\right\} + (1 - p)\sqrt{\frac{1}{2\pi\sigma^2}}\exp\left\{-\frac{(q_i - \theta)^2}{2\sigma^2}\right\}}. \tag{8}$$

Note that the EM algorithm allows us, in principle, to include the subset of still-smokers, when fitting the model. For the censored q_i, $i = m+1, ..., n$ we can find the expected value, given that the subject still smoked at the time of the survey, which is given by

$$E(q_i - s_i | c_i, s_i) = \int_{c_i - s_i}^{\infty} x\frac{f(x)}{1 - F(c_i - s_i)}dx \tag{9}$$

where c_i is the person's age at the time of the survey, and f, F are the probability distribution function and the cumulative distribution function, respectively, obtained from (6).

The M-step of the EM algorithm involves taking the derivative of $Q\left(\Psi; \Psi^{(k)}\right)$ with respect to $\Psi = \{p, \mu, \lambda, \theta, \sigma^2\}$. The maximum likelihood estimators in the k-th step are

$$\begin{aligned} p^{(k)} &= \frac{\sum_{i=1}^{m} b_i}{m} \\ \lambda^{(k)} &= \frac{\sum_{i=1}^{m} b_i}{\sum_{i=1}^{m} b_i \frac{((q_i - s_i) - \mu)^2}{(q_i - s_i)\mu^2}} \\ \mu^{(k)} &= \frac{\sum_{i=1}^{m} b_i (q_i - s_i)}{\sum_{i=1}^{m} b_i} \\ \theta^{(k)} &= \frac{\sum_{i=1}^{m} (1 - b_i) q_i}{\sum_{i=1}^{m} (1 - b_i)} \\ \sigma^{2^{(k)}} &= \frac{\sum_{i=1}^{m} (1 - b_i)(q_i - \theta)^2}{\sum_{i=1}^{m} (1 - b_i)} \end{aligned}$$

4.3 Model 3 – an exponential/normal mixture

The third model corresponds to questions of the form "How long ago did you quit smoking?" As before, we assume that the population of smokers is a mixture of two groups, but in this case the response in the first group is assumed to follow an exponential distribution, with probability distribution function

$$f_E(x;\lambda) = \lambda e^{-\lambda x}$$

for $x > 0$. Let r_i be the response of the i-th subject. The complete data log likelihood in this case is

$$\begin{aligned} l_C &= \sum_{i=1}^{m} b_i \left[\log p + \log \lambda - \lambda r_i\right] \\ &\quad + \sum_{i=1}^{m} (1-b_i)\left[\log(1-p) - \frac{1}{2}\log\left(\sigma^2\right) - \frac{(r_i-\theta)^2}{2\sigma^2}\right] + K. \quad (10) \end{aligned}$$

Again, we order the observations so that the smokers are the $m+1, ..., n$ observations and replace the missing data variables, b_i, with their expectations, given the current parameter estimates:

$$E(b_i) = \frac{p\lambda e^{-\lambda r_i}}{p\lambda e^{-\lambda r_i} + (1-p)\sqrt{\frac{1}{2\pi\sigma^2}}\exp\left\{-\frac{(r_i-\theta)^2}{2\sigma^2}\right\}}. \quad (11)$$

The maximum likelihood estimators in the k-th step are

$$\begin{aligned} p^{(k)} &= \frac{\sum_{i=1}^{m} b_i}{m} \\ \lambda^{(k)} &= \frac{\sum_{i=1}^{m} b_i}{\sum_{i=1}^{m} b_i r_i} \\ \theta^{(k)} &= \frac{\sum_{i=1}^{m} (1-b_i) r_i}{\sum_{i=1}^{m} (1-b_i)} \\ \sigma^{2^{(k)}} &= \frac{\sum_{i=1}^{m} (1-b_i)(r_i-\theta)^2}{\sum_{i=1}^{m} (1-b_i)} \end{aligned}$$

The left panel in Figure (7) shows the distribution of 14,142 quitters in the 1995 Tobacco Use Supplement to the Current Population Survey (CPS). Notice that there seems to be significant heaping at A5 and A10 points

(corresponding to multiples of 5 and 10, respectively, in terms of number of years since quitting smoking.) Interestingly, when the response is converted to quitting ages, the heaping in T5, T10 ages is much less pronounced than in the PSID data (Figure 1).

While it is conceivable that people may have a greater tendency to quit smoking at round ages (which we may consider as 'true heaping', as opposed to misreporting), evidence and logic suggests it is not the case. There is no reason to believe that the populations in the PSID and CPS surveys will have different distribution of quitting ages, and therefore it seems that the heaping at age 50, for example, in the PSID data is mostly a result of the form of the question in the survey. In the case of data that are reported in terms of calendar time (e.g. in what calendar year did you last smoke regularly) heaping presents itself by excess reports of quits in calendar years divisible by 5. To evaluate whether there might be "true" heaping, consider the available evidence about the process smokers follow when they quit and what patterns in behavior would have to hold. In the case of calendar year heaping and age heaping a person must be able to identify years in which the calendar year or their age is evenly divisible by 5. This is plausible when quits are reported in terms of calendar year or age. It is implausible for elapsed time because, at the time a smoker is deciding to quit, he cannot know that in five or ten or fifteen years time he is going to be surveyed and asked about whether he quit or not. At the time a smoker is deciding, nothing differentiates one year from another in the distribution of elapsed time. Therefore, all heaping on T5 and T10 years is true rounding. In the case of quits reported in calendar years or age years, smoking relapse makes it unlikely that a smoker hits a quit target that is a T5 year (age). On average, smokers attempt to quit three to four times before they succeed and the period between relapses is approximately three to four months ("Center for Disease Control and Prevention [CDC]" (1993), DiClemente et al. (1991), Hatziandreu et al. (1990), Prochaska and DiClemente (1983)). To generate heaping of the magnitude shown above, smokers would have to accurately forecast how many attempts they would make and how long the periods of relapse would be in order to exactly hit the quit (years) ages that are evenly divisible by 5. Such planning is logically implausible.

Empirically one can also ask what would have to be true if the heaping in Figure 1 reflected true behavior. Figure 1 implies that a person is between 3 to 5 times more likely to have successfully quit at age 30 than at age 29 or age 31. Such large differences in quit rates would almost surely be commonly

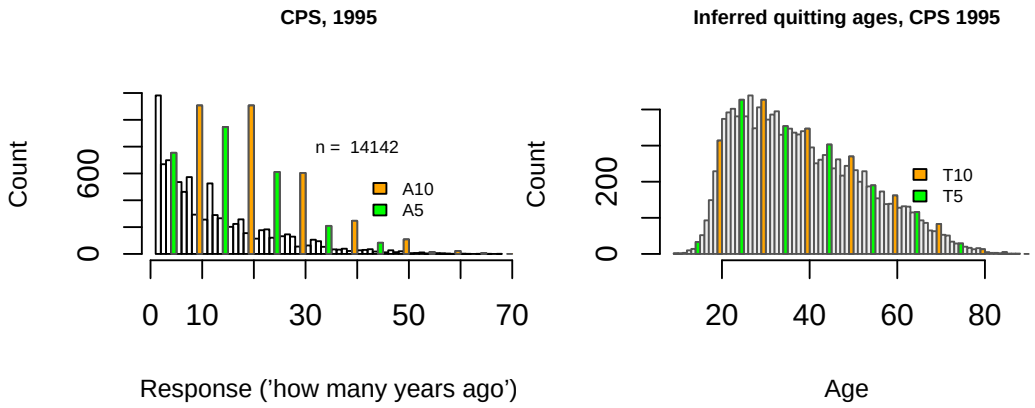

Figure 7: The distribution of the response to the question 'how long ago did you quit smoking' in the CPS survey (left), and the inferred quitting ages (right).

known. We are unaware of any such differences reported in any literature on smoking cessation.

Despite the above, it is important to note that our model-based approach allows us to incorporate a prior belief that certain ages are more likely to prompt smokers to quit. One way to do it, is to add point-mass mixture components to the model, and estimate their probabilities. A simpler way is to add these ages as dummy variables to the regression model (e.g. $I_{[is50ij]} = 1$ if subject i was 50 years old in time period j, and 0 otherwise). More details about the estimation of regression parameters are provided in the next section.

In some surveys the question about smoking cessation is asked in the form "In which calendar year did you quit smoking?" This is equivalent to the question "How long ago did you quit smoking?" since we can simply subtract the calendar year at the time of the survey from the response. Note, however, that the heaping that occurs depends on the form of the question, and more precisely, on the time units being used. So, for instance, if a survey is conducted in 2002, and the question has the form of "in which calendar year", then after the transformation to the form "how long ago" we expect the heaping points to be 2,7,12, etc.

4.4 The EM Algorithm – Estimating the Heaping

In the previous subsections we described three possible models for retrospective event data and the EM framework for fitting such data. However, in the presence of heaping the parameter estimates are expected to be biased, and indeed, our simulations show that this is the case. For example, when we fit data generated according to the exponential/normal model, the parameter estimates obtained by the EM algorithm are very accurate as long as the total percentage of heapers is less than $10 - 15\%$, and when the percentage of heapers increases, the estimate of the (normal component) variance, σ^2, as well as the estimate of the mixture parameter, p, are inflated. In these simulations, by the way, the rate (λ) and the mean of the normal component (θ) are estimated quite accurately even in the presence of significant heaping.

The EM algorithm framework allows us to account for heaping by incorporating other unobserved variables, in a way similar to the inclusion of the Bernoulli variables in the previous subsections. In this subsection, we discuss the idea in detail.

The hierarchical nature of our model-based approach is depicted in Figure 8: as we described in the previous subsections, subject i may quit smoking according to a stochastic process with probability p, and we denote the probability distribution function generically by $f_R(r_i; \psi)$ where ψ is a set of hyper-parameters; or, it is distributed normally with mean μ and standard deviation σ with probability $1 - p$. Note that b_i is a Bernoulli random variable, so either $b_i = 1$ (in which case the subject belongs to the group of people who quit randomly), or $1 - b_i = 1$ (which corresponds to the group of people who quit in response to certain conditions whose overall effect has mean μ and variance σ^2.) Hence, r_i represents the true response of subject i.

However, there is some probability that this subject is a "heaper" and instead of reporting r_i he reports $H_c(r_i)$, where c is a multiple of the time units (e.g., in Model 2 we denote the heaped ages by T5 or T10 for multiples of 5 or 10 years, respectively). The probability that a subject heaps is distributed according to a generic function, F, with a set of hyper parameters ϕ. Note that F may depend, for instance, on the the current age of the subject. As we have shown in Section 3, there is evidence that the current age affects the probability that the subject will report a heaped quitting age.

This framework can be extended to allow for multiple heaping rules, and each one can be modeled with a different probability distribution function, F_c

(e.g. T5 and T10 may have different probabilities in Model 2.) Furthermore, the distribution function F of the indicator variables h_i is not assumed to have certain properties. For instance, it does not have to be symmetrical around the heaping points. In the simplest case, we might assume that h_i are drawn from a Bernoulli distribution such that person i heaps his response ($h_i = 1$) with probability q (regardless of their age, or any other covariates). Alternatively, we can assume that q_i depends on a person's age, and use a logistic regression as we did in Section 3 (Eq. 3), to estimate a subject-specific heaping probability.

In the above discussion we assumed that the covariates that affect a person's decision to quit are independent of the covariates that affect their probability to heap. We recognize that this assumption may not always be true. For example, a person may have quit smoking following a heart attack, the timing of which he recalls perfectly, and as a result he will also report the correct quitting age, and does not heap, regardless of his current age. However, having analyzed multiple data sets, we believe that our assumption is reasonable, since a person's current age (or the amount of time since quitting) is the strongest predictor for heaping, and other covariates, like major life events, seem to have a much smaller effect the probability of heaping.

To obtain the parameter estimates for the generic model in Figure 8, we write the complete data likelihood as before, except in this case it contains two sets of unobserved variables, $\{b_i\}, \{h_i\}$. Let $F_R(H_C(r_i)), F_N(H_C(r_i))$ be the (discrete) probabilities of observing the heaped value $y_i = H_C(r_i)$ rather than the true value r_i. Let H be the set of all the possible heaped responses and H^C is its complement.

$$\begin{aligned} L_C &= \prod_{i \in H^C} \left\{ [p f_R(y_i)]^{b_i} \left[(1-p) f_N(y_i) \right]^{1-b_i} \right\} \\ &\quad \times \prod_{i \in H} \left((1-q) \left[p f_R(y_i) \right]^{b_i} \left[(1-p) f_N(y_i) \right]^{1-b_i} \right)^{1-h_i} \\ &\quad \left(q \left[p F_R(y_i) \right]^{b_i} \left[(1-p) F_N(y_i) \right]^{1-b_i} \right)^{h_i} \end{aligned} \tag{12}$$

We proceed with the EM algorithm as before, but the main difference is that now we also plug-in the posterior probability that subject i is a heaper, $E(h_i)$, in the E-step. Given the parameter estimates in the k-th step of

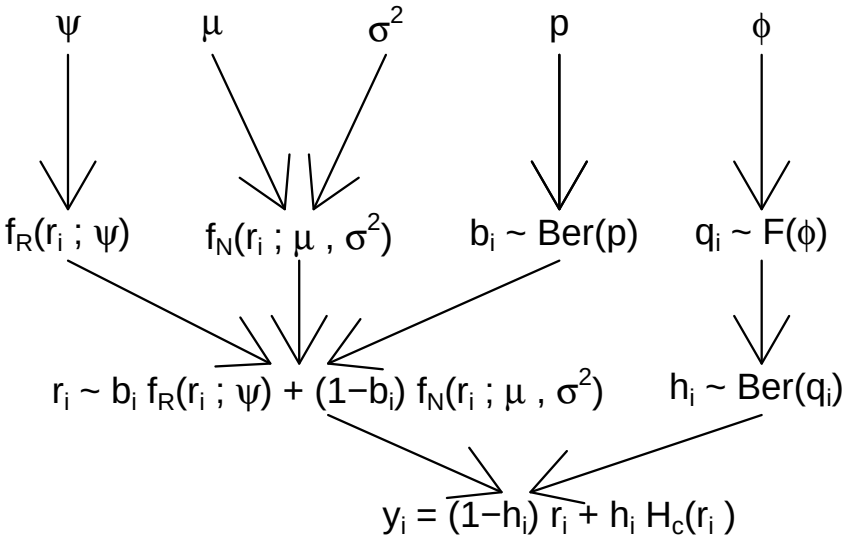

Figure 8: Graphical representation of process generating observed smoking cessation ages

the algorithm, we can express F_N, F_R in terms of the continuous cumulative distribution functions, and take the derivatives in order to find the current iteration's maximum likelihood estimates.

For example, under Model 2, f_I is the exponential p.d.f, and if we assume that the probability of heaping to, say, 25 is symmetrical in the range $[22.5, 27.5)$, then we can write

$$F_I(25) = \int_{22.5}^{27.5} \lambda \exp\{-\lambda x\} dx = e^{-22.5\lambda} - e^{-27.5\lambda}$$

We can similarly express $F_N(y)$ using the normal cumulative distribution function.

4.5 Using MCMC to Fit Parametric Models to Quitting Ages

The model-based approach presented in this paper lends itself quite naturally to a fully-Bayesian framework. While the EM algorithm provides tractable expressions for parameter estimates, and is therefore computationally efficient, it requires one to obtain the derivatives with respect to each parameter of the complete data likelihood function. Hence, when a different underlying mixture model is assumed, a new program is needed, and as we demonstrated in the previous subsection, when we account for heaping (and in particular when we allow for multiple heaping rules), the derivations become more tedious. The Bayesian approach is similar to the one in Wright and Bray (2003), except that in our case the response is not only heaped, but it is also a mixture of two components.

In the fully Bayesian approach which employs Monte Carlo Markov Chain (MCMC) simulations, changing the specification of the distribution functions in the model (which is displayed generically in Figure 8), or implementing different heaping rules, is much simpler. In fact, the specification of the model is determined by the directional acyclic graph in Figure 8. However, choosing the MCMC estimation approach trades off simplicity for speed. Although the models described here are parsimonious, an MCMC sampling approach tend to be slow since it involves a random sampling part, whereas the EM algorithm with its analytically derived estimates tends to converge much faster. Furthermore, MCMC sampling may require multiple runs to assess convergence and to tune up initial values or prior distributions.

Both methods (EM and MCMC) are viable options for recovering the assumed ("true") underlying distribution of responses, which is a critical step toward bias estimation, which we describe in the next section.

5 Simulation Study

We stated earlier that two factors can contribute to biased estimates in regression models. First, we assume that some of the population quits smoking according to a stochastic process that does not depend on the covariates in the regression. It is important to separate out the two groups, in order to assess what impact certain policies can have on smoking habits. For example, in the extreme case in which all the subjects quit at random, we cannot

Parameter	True Value	Estimate
p	0.9	0.857(0.024)
q	0.34	0.336(0.023)
μ	14	14.235(0.444)
σ	2	2.425(0.451)
λ	0.1	0.092(0.003)

Table 1: Posterior means and standard errors of the simulation parameters

expect any policy to affect people's decision to quit smoking. The second factor, which is the focus of this paper is heaping. Heitjan and Rubin (1990) coined the term 'ignorable coarsening', but as stated here and other places (e.g., (Wright and Bray, 2003)), heaping cannot always be ignored. Our goal is to estimate the extent of heaping in data sets, and use the smooth parametric model to quantify to what degree regression parameters are biased as a result of heaping.

In Section 4 we described two approaches to recovering the underlying mixture model, one using the EM algorithm, and the other using a Bayesian approach (MCMC simulations). We simulated 1000 subjects who reported their quitting time in terms "how long ago". We set $p = 0.9$, meaning that 900 subjects quit according to a stochastic process (exponential distribution with rate $\lambda = 0.1$), and 100 subjects quit because of external conditions, according to a normal distribution with mean $\mu = 14$ and variance $\sigma^2 = 4$. We also set $q_5 = 4$, that is, 40% of the people reported a heaped response (rounded to the nearest multiple of 5 years). Note that in this simulation we set it so that only people who quit more than three years ago can heap, so the actual number of heapers in this simulation is 295 (out of 867 whose response is greater than 2), so effectively $q = 0.34$.

Using the WinBUGS MCMC sampler (Spiegelhalter et al., 2003), we were able to obtain very good estimates for the simulation parameters. Figure 9 shows the simulated data, with substantial heaping at multiples of 5 years.

Table 5 summarizes the results of the MCMC simulation. The posterior means of the five parameters in the model are very close to the true values, even in the presence of significant heaping, and a relatively small proportion of the normal component in the mixture.

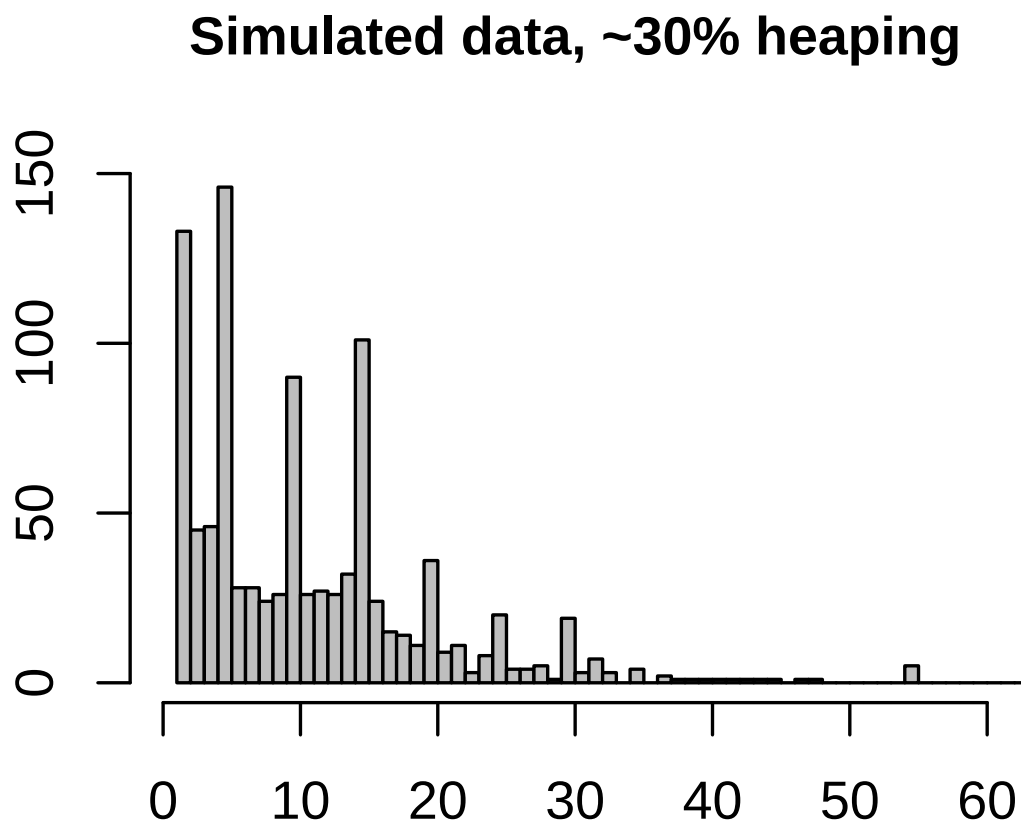

Figure 9: Simulated data with a mixture of 0.9 exponential(0.1) and 0.1 normal(14,4), with 34% heaping.

In a similar simulation, we generated a data set as in the first example, but with $\sigma^2 = 0.25$, which represents a "one-time shock", which we think of (simplistically) as the effect of tax. The (Bayesian) 95% Highest Probability Density Intervals for the tax effect parameters, μ, σ^2 are: $\hat{\mu} \in (13.96, 14.4)$ and $\hat{\sigma}^2 \in (0.197, 0.32)$ where the true values are $\mu = 14, \sigma^2 = 0.25$. The 95% HPD Interval for the proportion of people who quit due to the tax is $\hat{p} \in (0.079, 0.1276)$ and the true value is 0.1.

Figure 10 shows the sampling properties of the posterior distribution of

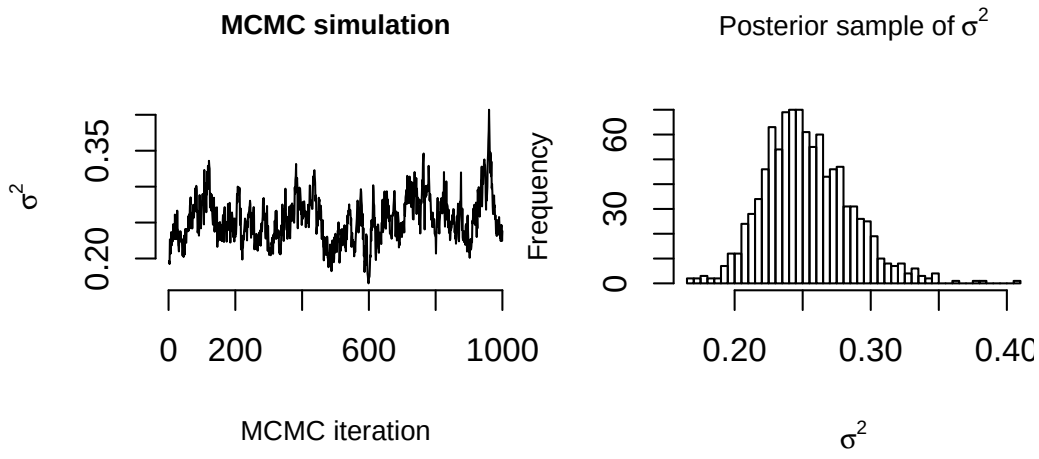

Figure 10: Sampling from the posterior distribution of σ^2

σ^2. The left panel shows a trace plot of the last 1000 MCMC iterations (suggesting good mixing and convergence), and the right panel shows a histogram of the sampled values, which demonstrates the accuracy of the estimation procedure, as was also summarized in the HPD Interval above. Similarly, Figure 11 shows the sampling properties of the posterior distribution of the mixture parameter p.

6 Bias Estimation

Our main goal in this paper is to estimate to what extent heaping in surveys can affect parameter estimates in regression models. We begin this section with a brief review of discrete time survival analysis regression. Our notation below follows that in Singer and Willett (1993), which we summarize here for completeness.

6.1 Discrete-Time Survival Analysis – Brief Review

Let $g_{i,j} = Pr\{T_i = j | T_i \geq j, Z_{1ij} = z_{1ij}, \ldots, Z_{Pij} = z_{Pij}\}$ be the discrete time hazard function, which is defined as the conditional probability that person i quits smoking in time j, given that he did not quit before time j, and given

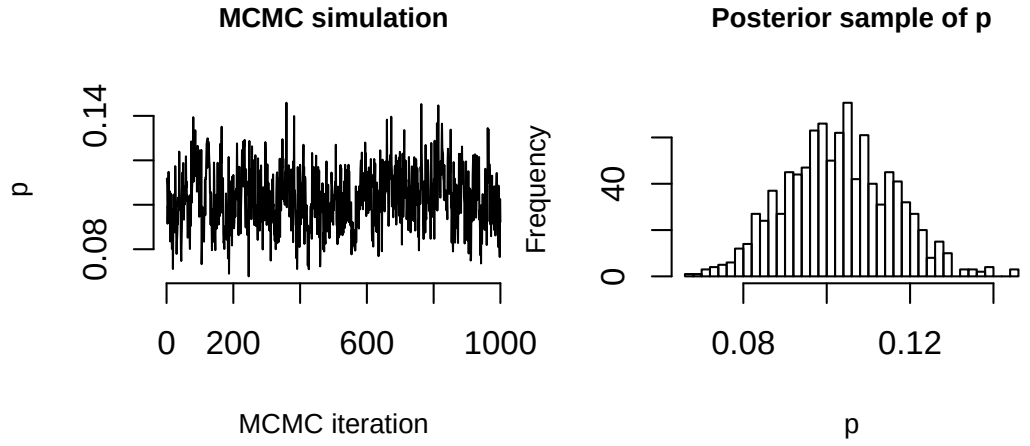

Figure 11: Sampling from the posterior distribution of σ^2

some covariates Z_{kij}. We use the model proposed by Cox (1972) and assume that the log-odds of quitting follow a linear model:

$$log\left(\frac{g_{i,j}}{1-g_{i,j}}\right) = \alpha_1 D_{1ij} + \ldots + \alpha_J D_{Jij} + \beta_1 Z_{1ij} + \ldots + \beta_P Z_{Pij}, \qquad (13)$$

where the data is stored on person-years format: suppose that the earliest year reported in the survey is $t_{(1)}$ and the last one is $t_{(J)}$. For example, $t_{(1)}$ can be the earliest birth year in the survey, and $t_{(J)}$ can be the year in which the survey was conducted. The first index in D_{tij} represents the range of time periods in the sample, $1, \ldots, J$. The second index, i, represents the subject. For each subject, the data set contains s_i rows, where s_i is the number of years the subject reported to have been smoking. The variable D_{tij} is set to 1 if subject i has been smoking for j years at time period t, and 0 otherwise. Note that for a fixed t, for each pair ij at most one dummy variable D_{tij} can be 1. The parameters α_t represent the baseline hazard in each time period.

The variables Z_{pij} record the values of P covariates for each subject i, in each time period j in which he was 'at risk' for quitting. These covariates may be fixed for all time period (e.g., sex, race, etc.) or time-varying (e.g., cigarette price, or major events such as marriage, heart attack, etc.) The parameters β_p describe the effect of the P predictors on the baseline hazard function (on the logit scale).

The response Y_{ij} is either 1, if subject i quit in his j-th year as a smoker, or 0 if he was still smoking.

Hence, the design matrix X for the logistic regression consists of $J + P$ columns where the first J correspond to the smoking duration indicators, D_{tij}, and the last P correspond to the linear predictors. The number of rows in X is $S = \sum_{i=1}^{n} s_i$, the total number of person-years. To estimate the parameters $\varphi \equiv \{\alpha_1, \ldots, \alpha_J, \beta_1, \ldots, \beta_P\}$, we maximize the likelihood function

$$L = \prod_{i=1}^{n} \prod_{j=1}^{s_i} g_{ij}^{y_{ij}} (1 - g_{ij})^{1-y_{ij}}. \tag{14}$$

where g_{ij} is obtained from (13).

Assessing the significance of the predictors is typically done by comparing $-2LL$ of the complete and reduced models, where the latter includes only the intercept parameters, α_t, and $-2LL$ is -2 times the log-likelihood. The drop in $-2LL$ is compared with a Chi-square distribution with P degrees of freedom.

6.2 Estimating Heaping-Induced Bias and Misclassification Probabilities

Using the notation in 6.1, it is obvious that heaping (or for that matter, any type of error in reported ages in retrospective surveys) will result in a different design matrix X and response vector Y, and hence may result in biased estimates for β_p.

To estimate the heaping-induced bias we use Monte Carlo simulations: recall that in Section 4 we provided a model-based approach to estimate the distribution of quitting ages. We use these estimates to construct random design matrices and response vectors, $X^{(m)}, Y^{(m)}$, respectively, and for each such pair we obtain the regression parameter estimates, $\hat{\varphi}_m$ (for $m = 1, \ldots, M$). Specifically, we take the birth years of subjects in the survey, and draw start- and quit-smoking ages according to the fitted distribution for the appropriate model from Section 4, and convert them to calendar years, which we then use to construct $X^{(m)}, Y^{(m)}$ as described in the previous subsection.

For each predictor $p = 1, \ldots, P$ we then estimate the bias by

$$Bias_p = \frac{1}{M} \sum_{m=1}^{M} (\hat{\beta}_p - \hat{\beta}_{p,m})$$

where $\hat{\beta}_p$ is the maximum likelihood estimator of β_p obtained from the survey data (without accounting for heaping), and $\beta_{p,m}$ is the estimator from the m-th random pair $X^{(m)}, Y^{(m)}$.

Using the Monte Carlo simulation approach we can also estimate the response misclassification probabilities, defined for each time period j, as in Hausman et al (1998), by

$$\gamma_{0,j} = Pr(Y_{ij} = 1|\tilde{Y}_{ij} = 0)$$
$$\gamma_{1,j} = Pr(Y_{ij} = 0|\tilde{Y}_{ij} = 1)$$

where $\tilde{Y}_{ij}$ is the true response (subject i quit in time period j), and Y_{ij} is the reported response. Similarly, we can estimate the misclassification probabilities of the 'at-risk' set at time t,

$$\delta_{0,t} = Pr(D_{tij} = 1|\tilde{D}_{tij} = 0)$$
$$\delta_{1,t} = Pr(D_{tij} = 0|\tilde{D}_{tij} = 1).$$

7 Discussion

The above exercise shows that our method recovers the parameters of the underlying distribution in our simulated data. With these estimated values of the underlying parameters, we can then estimate the bias. Our method is similar to the one developed by Wright and Bray 2003 with two important extensions. Those authors assume there is one underlying distribution and that respondents all use a single heaping rule. Here we assume a mixture of two distributions (that replicate the observed data very well) and we allow for multiple heaping rules. Our approach also relaxes two very strong assumptions of Heitjan and Rubin 1990 - that respondents only use a single heaping rule and that reported quits are coarsened at random (they term this "ignorability"). In most data both assumptions probably do not hold.

As of this writing we have made significant progress toward our goal of developing a method for measuring and correcting for the attentuation bias that heaping introduces to models of the probability that events occur. This problem arises principally when the timing of the event is related to factors that also vary over time. In such cases, misreported outcome data will be less correlated with the time-varying explanatory factor.

Our method depends on two key assumptions. First, we assume that observed data result from the mixture of two distributions that reflect distinct

and separate processes. In one we assume that the behavior of interest occurs stochastically. In the other we assume that observed behavior responds to external shocks. We also assume that respondents to surveys fall into two or more groups, each of which uses a different heaping rule. We fit distributions, show that we can replicate the observed data, and we recover the parameters of an underlying distribution in data with heaping that we simulated.

Our next steps are to extend the simulation to the distribution of quits, compute the bias correction factor for different underlying distributions, and then apply those correction factors to actual data.[1] We will also develop formal tests of our fitted parameters from the simulated data.

Acknowledgement: We thank Hua Wang for her work on an earlier (empirical) version of this paper, Donald Kenkel, George Jakubson, and Alan Mathios for comments, Robert Strawderman for his statistical advice, and Eamon Molloy for his programming assistance. We also thank the ASHE 2010 conference discussant, Anna Sommers, for her comments. All errors are our own.

References

Banerjee, A. K. and Bhattacharyya, G. K. (1976). A purchase incidence model with inverse gaussian interpurchase times. *Journal of American Statistical Association*, **71** 823–829.

Beckett, M., DeVanzo, J., Sastry, N., Panis, C. and Peterson, C. (2001). The quality of retrospective data - an examination of long-term recall in a developing country. *Journal of Human Resources*, **36** 593–625.

"Center for Disease Control and Prevention [CDC]" (1993). Smoking cessation during previous year among adults united states, 1990 and 1991. *Morbidity and Mortality Weekly Report*, **42** 504–505.

Crockett, A. and Crockett, R. (2006). Consequences of data heaping in the british religious census of 1851. *Historical Methods*, **39** 24–39.

[1] An additional advantage of our method is that it easily accommodates any distribution. If the analyst suspects the underlying distribution takes some other form, he can use that distribution and compare results.

Del Corral, J., Barros, C. P. and Prieto-Rodriguez, J. (2008). The determinants of soccer player substitutions - a survival analysis of the spanish soccer league. *Journal of Sports Economics*, **9** 160–172.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, **39** 1–38.

DiClemente, C., Prochaska, J., Fairhurst, S., Velicer, W., Valesquez, M. and Rossi, J. (1991). The process of smoking cessation: An analysis of precontemplation, contemplation, and preparation stages of change. *Journal of Consulting and Clinical Psychology*, **59** 295–304.

Folks, J. L. and Chhikara, R. S. (1978). The inverse gaussian distribution and its statistical application - a review. *Journal of Royal Statistical Society, B.*, **40** 263–289.

Forster, M. and Jones, A. M. (2001). The role of tobacco taxes in starting and quitting smoking: duration analysis of british data. *Journal of the Royal Statistical Society Series a-Statistics in Society*, **164** 517–547.

Hasofer, A. M. (1964). A dam with inverse gaussian input. *Proc. Camb. Phil. Soc.*, **60** 931–933.

Hatziandreu, E., Pierce, J., Lefkopoulou, M., Fiore, M., Mills, S., Novotny, T., Giovino, G. and Davis, R. (1990). Quitting smoking in the united states in 1986. *Journal of the National Cancer Institute*, **82** 1402–1406.

Hausman, J. A., Abrevaya, J. and Scott-Morton, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, **87** 239–269.

Heitjan, D. F. and Rubin, D. B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, **85** 304–314.

Kenkel, D. and LeCates, J. (2010). Errors in retrospective data on smoking: Comparing maximum likelihood and ad hoc approaches. Presented at the 3rd Biennial Meetings of the American Society of Health Economists. June 20-23, 2010. Ithaca, NY.

KENKEL, D., LILLARD, D. R. and MATHIOS, A. (2004). Accounting for measurement error in retrospective smoking data. *Health Economics*, **13** 1031–1044.

LAMBERT, D. (1992). Zero-inflated poisson regression, with an application to defetcs in manufacturing. *Technometrics*, **34** 1–14.

LANCASTER, A. (1972). A stochastic model for the duration of a strike. *Journal of Royal Statistical Society, A.*, **135** 257–271.

LILLARD, D. R., BAR, H. and WANG, H. (2009). A heap of trouble? accounting for mismatch bias in retrospectively collected data. Working paper.

LITTLE, R. (1992). Incomplete data in event history analysis. In *Demographic applications of event history analysis* (J. Trussell, R. Hankinson and J. Tilton, eds.). Clarendon Press, Oxford, England, 209–230.

PETERS, H. (1988). Retrospective versus panel data in analyzing lifecycle events. *Journal of Human Resources*, **23** 488–513.

PROCHASKA, J. and DICLEMENTE, C. (1983). Stages and process of self-change of smoking: Toward an integrative model of change. *Journal of Consulting and Clinical Psychology*, **51** 390–395.

PUDNEY, S. (2007). Heaping and leaping: Survey response behaviour and the dynamics of self-reported consumption expenditure. Manuscript, Institute for Social and Economic Research, University of Essex.

SINGER, J. D. and WILLETT, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and timing of events. *Journal of Educational Statistics*, **18** 155–195.

SPIEGELHALTER, D., THOMAS, A., BEST, N. and LUNN, D. (2003). *WinBUGS User Manual Version 1.4*.

TORELLI, N. and TRIVELLATO, U. (1993). Modeling inaccuracies in job-search duration data. *Journal of Econometrics*, **59** 187–211.

WRIGHT, D. and BRAY, I. (2003). A mixture model for rounded data. *The Statistician*, **52** 3–13.

Wu, L. L., Martin, S. P. and Long, D. A. (2001). Comparing data quality of fertility and first sexual intercourse histories. *Journal of Human Resources*, **36** 520–555.

A The PSID and CPS Data Sets

The Panel Study of Income Dynamics (PSID) (United States)

The PSID began in 1968 with a sample of about 5,000 households, representing a disproportionate number of low-income individuals. All current PSID families contain at least one member who was either part of the original 5,000 families or born to a member of one of these families. Although the original sampling scheme disproportionately selected individuals from low-income families, a representative sample of the United States population can be obtained by excluding the original over-sample from the data or by applying sample weights. Starting in 1997 the PSID administers its survey every other year. Only the head and "wife" (a PSID term designating the household member with whom the head has a "significant" relationship) are asked about their cigarette consumption. Retrospective smoking questions were asked in 1986, 1990 (for those age 65+), 1999, 2001, 2003, 2005, and 2007 and are scheduled to be asked in all currently planned surveys (2009 and 2011). The PSID data is the world's longest running panel study. As of 2007, it has followed individuals for up to 38 years. It also has data on up to three generations.

Current Population Survey - Tobacco Use Supplements (CPS-TUS)

The Tobacco Use Supplements to the Current Population Survey, sponsored by the National Cancer Institute and administered as part of the U.S. Census Bureau's continuing labor force survey, have been collected since 1955 (Haenszel, Shimkin, Miller 1956, Hartman et al. 2002). In the more recent CPS-TUS surveys, data on smoking behavior of a large, nationally representative sample of about 240,000 individuals 15 years of age and older is collected in a three-month survey cycle. These cycles were conducted in September 1992, January and May 1993; September 1995, January and May 1996; September 1998, January, and May 1999; and June and November 2001 and February 2002. A separate TUS "Special Topics supplement" was administered in 2003. The CPS-TUS are supplemental surveys given on top of the monthly survey the CPS is administering.