Interviewer Effects on Survey Questionnaire Response Times: A Hierarchical Bayesian Analysis of Multivariate Survival Paradata

Hiroaki Minato
U.S. Energy Information Administration
1000 Independence Avenue, SW, Washington, DC 20585

Abstract

Survey research often covers a range of topics in one large survey and related questions are usually grouped to form sections in the questionnaire. With a computer-assisted survey interview instrument, the time spent to complete each section of the questionnaire may be measured. Our interest is to understand the completion-rate variability due to interviewers.

With some language from biostatistics, we conceptualize multiple response times, "treatments" (e.g., interviewers, interview times), and hierarchical covariates. The multivariate survival data framework is used to model the relationship between multiple "failure" rates and survey treatments and covariates. And, we adopt the hierarchical Bayesian approach in analyzing the multivariate, multilevel data and making inference on interviewer effects on the multiple completion rates.

As an illustration, the 2009 Residential Energy Consumption Survey (RECS) data along with the paradata on survey questioning are examined.

Key Words: hierarchical Bayesian data analysis, multivariate survival analysis, paradata, survey questionnaire response time, Residential Energy Consumption Survey (RECS)

1. Overview

This paper illustrates one way to detect an unusual level of effect an interviewer might exert on time that a respondent spends to answer survey questions. We will begin by stating our motivation of why this research may be valuable. And, we will describe the general nature and structure of paradata and will provide some examples, with which we will transition to a statement of some research questions paradata can help answer. Then, we will propose a statistical model to answer the questions and a Bayesian approach to estimate the key model parameters. Finally, we will show results and will discuss their possible usage.

2. Motivations

A survey produces not only substantive data but also data about the substantive data (metadata) and data about how the substantive data were collected (paradata). Although metadata may be used for computing weights and response rates, paradata are usually treated as a retrospective evidence of data collection work. However, if paradata is analyzed during a data collection period, it can help control the data collection

operation—that is to measure and evaluate some performance and to utilize the measurement to improve the future performance.

Paradata tend to be large whenever survey processes are long and complicated. But, they are not "Big Data". Paradata have a well-organized structure, though not necessarily simple. We believe various types and levels of statisticians can find good opportunities for challenge and contribution.

3. Nature and structure of paradata

3.1 Human interactions during the field data collection period generate paradata

Paradata that we consider here are generated by or observed in human interaction processes during survey data collection. The word paradata was coined by M. Couper in his 1998 paper "Measuring survey quality in a CASIC environment," (Kreuter et al., 2010). The notion must have been existent since the beginning of surveys. We loosely define paradata to be data about survey data collection process (where a process consists of input, function or black box, and output).

3.2 A response is an outcome of interviewer's specified attempt on a person

Outcome is a reaction or non-reaction to the action. The non-reaction can be intentional or non-intentional (e.g., ring but no answer in phone interview). Attempts must be pre-specified as protocols and must follow some rules. A person may or may not be pre-selected—in a persons survey, a person is usually pre-selected before interviewing, while in a households survey, a buildings survey, or an establishments survey, a person is not usually pre-selected and only the eligibility to respond in the survey may be pre-specified. Note that a household, housing unit, building, or establishment cannot be a respondent as it could not produce a response. It is always a person who provides responses in our surveys.

3.3 A response is nominal or ordinal

A response as an outcome of each survey attempt is normally coded into a nominal or ordinal category, often called a disposition code—e.g., successful contact, hard refusal, soft refusal, and so on. Disposition codes can be as fine or detailed as they need be, but standard codes are suggested for some scenarios by AAPOR (2011).

3.4 There may be a sequence of action-reaction's between an interviewer(s) and a person(s) over time

When a person provides a response that is not an acceptable resolution at interviewer's first attempt, an additional attempt of the same or different kind may be taken; thus, a sequence of action-reaction's could be produced for each person until some resolution, defining a person-level survey contact history over a finite time period. Persons or interviewers may not be unique in each contact history.

3.5 A time from an action to a response can be measured

Dates and times of contact attempts and outcomes, lengths of interviews, and other surveying time information are usually measured quantitatively through survey instruments. These time-to-event data may be considered as metadata of the paradata.

4. Paradata Examples

4.1 Paradata example 1: Survey contacts history

Survey contact attempts and their outcomes are perhaps the most typical paradata and an example is shown in table 1. The meaning of sample case ID depends on the sampling unit, which may not correspond to the person with which a human contact was attempted. The attempt is usually completely pre-specified with respect to what, how, and when, similarly to protocols in applying a treatment in biomedical experiment. In a mixed-mode survey, interviewing modes could be used further to distinguish contact attempts. Each attempt is administered by an interviewer in an in-person interviewing survey. In some surveys, pairs of interviewers may be utilized.

In the example, an interviewer is not assigned when the mode is a self-administered questionnaire (SAQ) like a mail survey, but the person who filled in the SAQ could be considered as an interviewer of himself or herself. The times and dates of attempts are not randomly or conveniently determined but rather pre-specified by contact rules that define waiting times between attempts with specific outcomes. For example, after a soft refusal, the next attempt may have to wait for at least one day but at most two days.

In a phone interview, sample cases may be randomly assigned to interviewers, although in an in-person interview, interviewers are often recruited locally. In order to complete difficult cases (e.g., those that complain about the survey), special interviewers (e.g., senior or supervisory interviewers) could be assigned to them after some failed attempts.

Table 1: Paradata Example—Survey Contacts History

Case ID	*Attempt*	*Outcome*	*Time and Date*	*Mode*	*Interviewer ID*
1	1	Ring, no answer	09:00 01-Aug-2014	Phone	1
1	2	Soft refusal	15:00 01-Aug-2014	Phone	1
1	3	Hard refusal	10:00 03-Aug-2014	Phone	1
2	1	Complete	09:30 01-Aug-2014	Phone	1
3	1	Disconnected	11:00 01-Aug-2014	Phone	1
3	2	Complete	20:00 07-Aug-2014	Mail/SAQ	NA
…	…	…	…	…	…

4.2 Paradata example 2: Survey questionnaire administration outcomes

Another common example of paradata is found when a respondent is answering questions in a survey. (Earlier we characterized such answering times as metadata of paradata, but the characterization is not important.) An answering time can be measured for an entire questionnaire or for each section (i.e., each set of questions) as well as each question. The example in Table 2 shows the elapsed time in each section of a questionnaire. A survey

instrument often does not allow skipping questions or sections unless it is legitimate to do so, but a respondent can break off from a survey, making all remaining questions or sections incomplete. In the survival analysis language term, those cases are right-censored. Mode and interviewer are also identified in the example. (Again, for a SAQ mode, the respondent can be thought as self-interviewer.)

Table 2: Paradata Example 2—Survey Questionnaire Administration Outcomes

Case ID	*Section*	*Outcome*	*Time duration (seconds)*	*Mode*	*Interviewer ID*
1	A	Complete	300	Face-to-face	1
1	B	Complete	600	Face-to-face	1
1	C	Complete	420	Face-to-face	1
2	A	Complete	100	Web/SAQ	NA
2	B	Complete	200	Web/SAQ	NA
2	C	Incomplete	0	Web/SAQ	NA
…	…	…	…	…	…

An example to illustrate our application is a special case of Example 2. Specifically, we have a uni-mode face-to-face survey and look at only cases that completed all sections in questionnaire. (This later constraint may not be applicable if we are analyzing select sections in real time. We choose only the first two sections of a questionnaire in our illustration, but the constraint is kept for simplicity.) Further, we select those interviewers who completed a significant number of cases, at least 100 cases.

Our paradata come from the 2009 Residential Energy Consumption Survey (RECS), which is a multi-phase national multi-stage area sampling survey by computer-assisted personal interviewing (CAPI) for residential housing unit's energy consumption and characteristics. The questionnaire has fourteen sections or components: A. Housing unit characteristics; B. Kitchen appliances; C. Home appliances and electronics; D. Space heating; E. Water heating; F. Air conditioning; G. Miscellaneous; H. Fuels used; I. Housing unit measurements; J. Fuel bills; K. Residential transportation; L. Household characteristics; M. Energy assistance; and N. Scanning of fuel bills. We select Sections A and B for our illustrative analysis.

5. Research questions

5.1 Do interviewers differentially affect survey answering times?

An interviewer performance can be defined and measured in various ways. One simple measure may be an interview completion rate by each interviewer. Given completed (or partially completed) interviews, we go further to examine a more complex aspect of interviewing that can still be easily measured. Our research addresses the question of whether interviewers differentially affect the lengths of time respondents use to answer survey questions.

We analyze in our illustration some of the paradata collected for the 2009 RECS, which achieved the household sample size of 12,083. Our analysis focuses on nine interviewers

who completed at least one hundred interviews, and we try to detect any sign of unusual (either too good or too bad) interviewer performance with respect to the survey answering times. The first two sections of the 2009 RECS questionnaire are selected for our analysis. It is not our goal to generalize or infer to some population of interviewers. Rather, we demonstrate how we might help a data collection operator statistically control one possible sign of interviewer performance in a simple, timely, and yet effective manner.

An interviewer effect either exists or does not exist and an existing effect can be positive or negative. Small effects are not substantively or practically important. Also, in our problem, interviewers always exist as "treatments". That is, we cannot conduct our survey without interviewers. So, an interviewer effect is not due to existence vs. absence of an interviewer but due to a relative difference among interviewers. We care about an interviewer's effect that is uniquely and significantly different from the other interviewers' effects.

Once we detect a unique and unusual interviewer effect, i.e., a sign of possible performance problem by an interviewer, we can advise a data collection operator to search for a potential cause for or association with the detected effect. Then, the data collection operator might investigate the interviewer and/or the data that interviewer has collected. If any problems are found, they should be corrected.

In our example, we look for a potential sign of interviewer effect in the questionnaire responding time, more specifically in the rate of questionnaire section completion conditional on elapsed time. Two rates are calculated because we have selected the first two sections of the 2009 RECS questionnaire: A. Housing unit characteristics and B. Kitchen appliances. The selection was made in order to control the effect of responses on the responding times and the other effects that may not be attributable to interviewers. In each section, some respondents might legitimately skip some of the questions, which certainly could add variation in the time duration to complete the section. But, we assume this variation is sufficiently controlled by the covariates we introduce to the model and that the remaining variation is more or less a random noise.

6. A statistical model

A model is a set of assumptions.

How do we statistically approach to detect any interviewer effects in survey answering times?

6.1 Why the Cox proportional hazards model?

Why not just compute simple means? Means are computed for descriptive purposes along with the other distributional statistics. But, for inference on differences of time duration, we must control for covariates, i.e., we need some models for comparisons. One could use a linear regression model possibly without the intercept. If the time duration is log-transformed, the intercept may be put back in. However, the reasonableness of the normal error assumption could remain dubious. Try a nonlinear model? Further, to analyze multiple response times, one would need multivariate models. And, a big problem comes up when some time response data are censored data.

Meanwhile, the Cox model that we suggest here is very natural for the current data and problem (Cox, 1972). It has been well established like a regression model and is simple (as simple as a regression model). This model choice has nothing to do with being Bayesian or frequentist, although there seems some advantage for Bayesians using the Cox model (Kalbfleisch, 1978; Sinha et al., 2003).

6.2 Completion rate depends on elapsed time: $h(t)$

A completion occurs as an outcome if and when a respondent completes a given section of a survey questionnaire. A time is measured in terms of elapsed time, not calendar time in our example. Completion rate at a given time *t* is a rate of completing a section at *t* by those respondents who had not completed the section yet. More specifically, a completion rate at *t* is the ratio of the number of respondents who completed at *t* over the number of respondents who have not completed before *t*. In survival analysis literature, it is called a hazard function or a conditional failure rate.

It is worth mentioning that in our data there is no censoring, because we are using only those cases that completed the survey, i.e., survey respondents. However, break-offs within a section could be modeled as censored, as long as they happen independently of the responding time to complete the section. (An open problem is handling of dependent censoring.) If some questions in a section are skipped in non-legitimate ways, we can no longer analyze the data by section and we would need first to tackle the missing data problem.

6.3 The rate also depends on the person: $h_i(t) = h_0(t)\exp(f_i)$

The person means everything about and around the person, including the interviewer. Denote by f_i a function of person *i* and express h_i as a multiplicative function with a base term that only depends on time *t* and with a (positive) multiplicative term $\exp(f_i)$ that depends only on person *i*. The exponential function simply keeps the multiplicative term positive. The 0 in $h_0(t)$ indicates that the term is a base rate term or a baseline hazard function. If f_i is a linear function of some variables like a multiple linear regression but without the intercept, then the model is Cox's (1972) proportional hazards model.

The proportional hazards (PH) mean that the hazard ratio (or ratio of two hazards) is constant over time or does not depend on time. It is an assumption but can be checked with data. For us, it is a good enough model (the semi-parametric approach with the partial likelihood) because we do not care about the baseline hazard function (which could be modeled with, e.g., exponential, Weibull, or gamma distribution for $t \geq 0$) and also because we are interested only in the relative comparisons of interviewer parameters. The simplicity also helps reduce the number of parameters for prior specification in the later Bayesian analysis.

6.4 There are multiple completion rates per person: $h_{s(i)}(t) = h_{0s}(t)\exp(f_{s(i)})$

We have two survey questionnaire sections and thus two response times per respondent. Our model is multivariate. But, again, we have made a simplifying assumption, ignoring the order of sections and the possible dependency between sections, and we formulate our model as the stratified Cox model, which stratifies the data/cases so that the PH assumption is reasonable within each stratum (Kleinbaum and Klein, 2005). Stacking the two section-level data sets to have one time response variable, we specify the section variable as the stratification variable:

$$\begin{cases} h_{A(i)}(t) = h_{0A}(t)\exp\left(f_{A(i)}\right) \\ h_{B(i)}(t) = h_{0B}(t)\exp\left(f_{B(i)}\right) \end{cases},$$

where *A* and *B* signify the two sections in the questionnaire we analyze.

6.5 $f_{s(i)}$ is a linear function without the constant term

Let $f_{s(i)}$ be a linear function of multiple variables defined for person *i* in section *s*. When all the variables in $f_{s(i)}$ are zero, exp(0) = 1. The exponential of a non-zero value, positive or negative, is an increasing or decreasing multiplicative factor, respectively. The intercept, if included, would be cancelled out in a hazard ratio. For our model, we could consider respondent-specific log-frailties (say, θ_i), which do not depend on *s*. But, we do not do this in order to avoid over-parameterization. We believe the regression model is sufficient for our purpose. Further, we assume that $f_{A(i)} = f_{B(i)} = f_i$, i.e., the linear regression specification does not depend on section *s*. It is reasonable for us, because we are interested in the overall effect of an interviewer over all sections, here two sections. Thus, as mentioned before, we have the stratified Cox proportional hazards model, where the questionnaire section ID is the stratifying variable and does not interact with covariates in f_i:

$$\begin{cases} h_{A(i)}(t) = h_{0A}(t)\exp(f_i) \\ h_{B(i)}(t) = h_{0B}(t)\exp(f_i) \end{cases}.$$

With the 2009 RECS data and paradata, f_i contains the variables in Table 3 as covariates. In the table, covariates are grouped by their substantive information purpose.

Table 3: Classification of Covariates by Substantive Information Purpose

Purpose	*Covariates*
Treatment	Interviewer_ID DayEvening
Householder	OccupyAge Race_White TotalBtu
Housing unit	YearBuilt SquareFootage HUType
Geography	Census_Division
Sample	W4_PostStrat

The interviewer identification variable (Interviewer_ID) is recoded in this analysis in order to prevent any linking to or identification of the interviewers. The DayEvening variable indicates whether the interview ended before or after 5 p.m.: 1 = Day (interview ended before 5 p.m.) and 2 = Evening (interview ended after 5 p.m.). This can be considered as another treatment. The OccupyAge variable measured the number of years the household had occupied the housing unit, for which the respondents who did not

know the answer or refused to answer are excluded from the current analysis. (There were no refusals in the data.) The Race_White variable indicates the race of the householder: 1 = white and 0 = non-white. The TotalBtu variable measured the total annual energy consumption in Btu (the British thermal unit) by the householder at the housing unit—the values include some imputed values. The YearBuilt variable records the year in which the housing unit was built—respondents with Don't Know or Refusal answers are excluded from the analysis. (There were no refusals in the data.) The SquareFootage variable measured the area in square footage of the housing unit—some of the values are imputed values. The HUType variable classifies the housing units by: 1 = Mobile Home; 2 = Single-Family Detached; 3 = Single-Family Attached; 4 = Apartment in Building with 2 - 4 Units; 5 = Apartment in Building with 5+ Units. The Census_Division variable identifies the geographical location by: 1 = New England Census Division (CT, MA, ME, NH, RI, VT); 2 = Middle Atlantic Census Division (NJ, NY, PA); 3 = East North Central Census Division (IL, IN, MI, OH, WI); 4 = West North Central Census Division (IA, KS, MN, MO, ND, NE, SD); 5 = South Atlantic Census Division (DC, DE, FL, GA, MD, NC, SC, VA, WV); 6 = East South Central Census Division (AL, KY, MS, TN); 7 = West South Central Census Division (AR, LA, OK, TX); 8 = Mountain North Sub-Division (CO, ID, MT, UT, WY) and Mountain South Sub-Division (AZ, NM, NV); and 9 = Pacific Census Division (AK, CA, HI, OR, WA). And, finally, W4_PostStrat is the final sampling weights variable, which was adjusted by unit nonresponse rates and post-stratified to the U.S. Census Bureau's 2009 American Community Survey totals.

Our current statistical objectives do not include hypothesis testing or model selection. Instead, we are interested in understanding the given model for its reasonableness and usefulness for the current detection purpose. There is no optimization in terms of goodness of model fit. After all, a model is a set of assumptions and is wrong. Particularly in a large and complex survey like RECS, there does not pre-exist a substantive data model nor could we acquire perfect data—perfect in the sense of measurable, accurate, and precise.

With that said, model selection/building/fitting and model assessment/checking could be conducted with the frequentist methods or general statistical methods such as least squares and AIC/BIC, if finding a more correct model is important. In our analysis, only the PH assumption and the interaction effects were checked with the frequentist methods (significance tests and graphical diagnostics) available in SAS PROC PHREG without the BAYES specification.

All the variables except Interviewer_ID are there to control. Specifically, we are controlling the variation in completion rate possibly due to the variation in interviewing time of the day, householder, housing unit, geography, and sample characteristics. This is necessary because interviewers were not randomly assigned to the respondent-to-be's (or the times of the day). Some of the covariates may be more strongly associated with one questionnaire section than the other, e.g., the housing unit type variable may be more closely related to the housing unit characteristic section than the kitchen appliances section. But, in our current analysis we focus on the overall relationship between each covariate and the two select sections together.

Compared to a model with respondent-specific log-frailties, e.g., that of Gustafson (1997), our model has the much less number of parameters, because we assume that a certain homogeneity among the respondents exists and because we replace respondent-

specific parameters with higher-than-respondent-level or respondent-classifying parameters such as householder characteristics (i.e., TotalBtu, OccupyAge, and Race_White) and housing unit characteristics (i.e., YearBuilt, SquareFootage, and HUType). Above the interviewer level, we include parameters to control for the sampling design variation in W4_PostStrat and for the geographic variation in Census_Division.

Note that there are no interviewer-classifying or -level variables in our analysis. No covariates besides identification variables for interviewers are considered, as we think they are unnecessary for our analysis. Meanwhile, characteristics traits of interviewers could be useful if one is to conduct analysis to help screen or select interviewers for a new survey. Similarly, there are no section-level variables besides the identifier, with which the nature of sections or questions could be examined in developing or testing a new survey questionnaire. But, these questions are out of our current scope.

7. Bayesian estimation

7.1 Hierarchical and bivariate model structure

Our data are structured by respondents per interviewer and by sections per respondent.

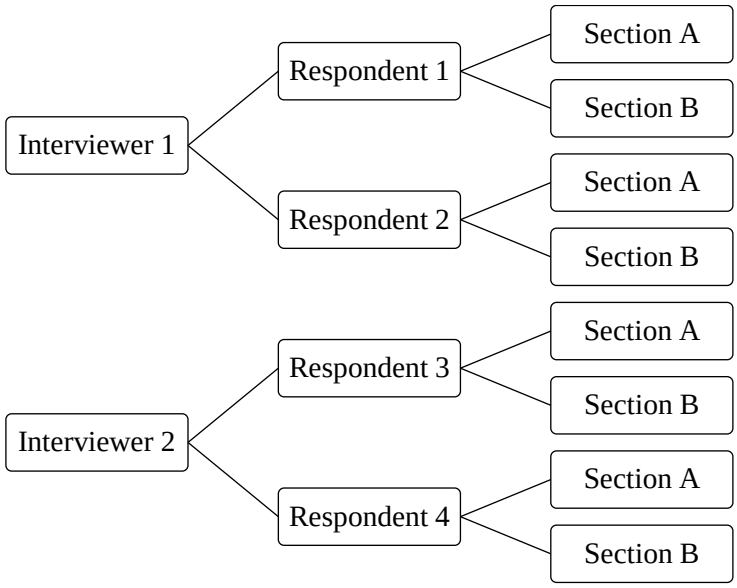

Figure 1: Hierarchical data structure

In our data, there are nine interviewers who completed at least 100 interviews each—the total number of completed interviews or respondents is 1,189. Note, however, that the number of respondents analyzed was reduced to 868, because some of the respondents had missing values at least in OccupyAge or YearBuilt, which are two of the covariates in our model. Since we look at two sections, A and B, of the 2009 RECS questionnaire, the number of time responses is: $868 \times 2 = 1{,}736$.

Sampling design variables (e.g., stratification variables) and geographies reside over interviewers and they further group interviewers. Note geographies themselves could be hierarchical (e.g., state, division, and region). We use the final weight variable

(W4_PostStrat) to extract and summarize all the sampling design information. For the geographical control, we use the Census_Division variable, as described before.

The 2009 RECS is a households survey, and at the respondent level, we have the following householder-level and housing-unit-level covariates, respectively, classified in Table 3: OccupyAge; Race_White; TotalBtu and YearBuilt; SquareFootage; HUType. They control or classify the 2009 RECS respondents.

As discussed earlier, our illustration specifies the following stratified Cox proportional hazards model with the questionnaire sections as strata:

$$\begin{cases} h_{A(i)}(t) = h_{0A}(t)\exp(f_i) \\ h_{B(i)}(t) = h_{0B}(t)\exp(f_i) \end{cases}.$$

To estimate the regression parameters, we require only a partial likelihood function, actually a product of two partial likelihood functions given by the two strata or questionnaire sections. Recall that we compute the exact likelihood, not Breslow likelihood or Efron likelihood, because of many ties in the elapsed time measurements we have.

Let $\boldsymbol{\beta}$ be the vector of regression parameters and $\boldsymbol{x}_{j(t_i)}$ be the vector of covariates for the *j*th respondent at time t_i, where $t_1 < \cdots < t_i < \cdots < t_k$ denote the *k* distinct, ordered completion times (in our special case $\boldsymbol{x}$ does not depend on time). For the questionnaire section *s*, if we let $\Omega_{s,i}$ denote the set of respondents who are yet to complete before the *i*th ordered completion time t_i and let $\Omega^*_{s,i}$ denote the set of respondents whose completion or censored times exceed t_i or whose censored times equal t_i, then we can write the exact likelihood function for the questionnaire section *s* as:

$$\mathcal{L}_s(\boldsymbol{\beta}) = \prod_{i=1}^{k} \left\{ \int_0^{\infty} \prod_{j\in\Omega_{s,i}} \left[1 - e^{-\frac{e^{\boldsymbol{\beta}'\boldsymbol{x}_{j(t_i)}}}{\sum_{l\in\Omega^*_{s,i}} e^{\boldsymbol{\beta}'\boldsymbol{x}_{l(t_i)}}} t} \right] e^{-t}\, dt \right\}.$$

Our likelihood is: $\mathcal{L}(\boldsymbol{\beta}) = \mathcal{L}_A(\boldsymbol{\beta}) \times \mathcal{L}_B(\boldsymbol{\beta})$.

7.2 Bayesian model assumptions, i.e., prior specifications

In our Bayesian analysis, we specify a flat prior for the *K* regression parameters in $f_{A(i)}$: $p(\beta_{A1}, \ldots, \beta_{AK}) \propto 1$, where $-\infty < \beta_{Ak} < \infty$ for each k = 1, …, *K*. Similarly for $f_{B(i)}$. But, in our example, we just have f_i: $p(\beta_1, \ldots, \beta_K) \propto 1$, where $-\infty < \beta_k < \infty$ for each k = 1, …, *K*.

The flat prior is improper but the posterior would be proper. Alternatively we can use the diffuse prior of normal distribution with the zero mean vector of the length *K* and the identity covariance matrix of the dimension *K* or the diagonal covariance matrix with diagonal elements being equal to the maximum likelihood estimates of the corresponding variances. We have tried each, and the results were similar; thus, we take the flat prior.

As we assume the Cox proportional hazards model, the baseline hazard functions $h_{0A}(t)$ and $h_{0B}(t)$ are left unspecified (and infinite-dimensional), requiring no prior distributions to be assigned.

7.3 Calculation of the posterior distributions

To calculate the posterior distributions of the regression parameters, we use the Markov chain Monte Carlo (MCMC) simulation, specifically, the Gibbs sampling with the maximum likelihood estimates (MLE's) as the initial values. The MLE's are shown in Table 4.

Table 4: Maximum Likelihood Estimates of the Interviewer Parameters

Parameter	*DF*	*Estimate*	*Standard Error*	*95% Confidence Limits*	
interviewer_id362702		0			
interviewer_id204537	1	0.074	0.2477	-0.4114	0.5594
interviewer_id269307	1	0.1443	0.2285	-0.3036	0.5923
interviewer_id212865	1	0.3355	0.2736	-0.2007	0.8718
interviewer_id244205	1	0.6576	0.2511	0.1654	1.1498
interviewer_id204297	1	0.7085	0.1157	0.4816	0.9353
interviewer_id245813	1	0.753	0.2487	0.2654	1.2405
interviewer_id259937	1	0.9365	0.2222	0.501	1.3721
interviewer_id245832	1	1.6738	0.2376	1.2081	2.1396

Specifying flat priors on the regression parameters and using their MLE's as the initial values, we depend more on data than on priors and our Bayesian analysis starts from where a frequentist analysis has ended. (Computing the MLE's, by the way, took only eight seconds of SAS time.)

After 3,000 or so samples, the posterior distribution seemed to converge to a stable distribution with small auto correlation, and we then drew 8,000 samples to form our posterior distribution for our inference. We did not "thin" the draws, that is, we did not throw any samples away systematically, e.g., every other sample, in order to tame sample auto correlations.

To determine if MCMC has converged or been stabilized to give us reasonable posterior distributions, we have utilized the following convergence diagnostics, available in SAS PROC PHREG: (a) autocorrelation, (b) the standard error of the posterior mean estimate, (c) the stationarity (the Heidelberger and Welch tests), (d) the convergence (the Gelman and Rubin statistic and the Geweke statistic), and (e) the accuracy of the estimated quantile of a chain (the Raftery and Lewis statistic).

To describe the diagnostics briefly: (a) autocorrelation literally computes autocorrelations of various lags for each parameter; (b) the effective sample size of Kass et al. (1998) measures the efficiency of the chain for each parameter and the Monte Carlo standard error measures the simulation accuracy and is the standard error of the posterior mean estimate, calculated as the posterior standard deviation divided by the square root of the effective sample size; (c) the Heidelberger and Welch (1981, 1983) tests are a stationary test and a halfwidth test for each parameter; (d) the Gelman and Rubin (1992) statistic compares convergence of two or more parallel chains and the Geweke (1992) statistic compares the first portion of the chain and the last portion of the chain; and (e) the

Raftery and Lewis (1992, 1996) statistic measures the accuracy of the estimated quantile of a chain.

We can say that in our chain: (a) dependency among Markov chain samples is low; (b) the effective sample size and the Monte Carlo standard error suggest good mixing; (c) the Heidelberger Welch tests suggest the chain has become stationary with enough samples to estimate the mean accurately; (d) the Gelman Rubin statistic suggest different starting values converge to the same value and the Geweke statistic indicates mean estimates are stabilized; (e) the Raftery and Lewis statistic shows we have sufficient samples to estimate 0.025 percentile within +/− 0.005 accuracy (Ibrahim et al., 2005).

8. Results

8.1 Assessment and validation of the Bayesian model

The deviance information criterion (DIC) and the effective number of parameters (ENP), available in SAS PROC PHREG, are not utilized in assessing our Bayesian model's goodness of fit. However, we report that the DIC was 17220.60 and ENP was 26.949.

The posterior predictive check (PPS) and the cross validation (CV) are popular model validation techniques Bayesians use. However, PPS is not particularly useful for our current analysis, as seeking predictive accuracy is not our main objective. Also, we did not conduct the model validation at this time.

Guided by the analytical objectives compelled by a particular research question, we should always balance theoretical model accuracy and practical computation time. Our exact likelihood model with a reasonable convergence, including the convergence diagnostics, required about 18 hours of CPU as well as real times in SAS PROC PHREG. (Note in particular that the Gelman-Rubin diagnostic multiplies the number of chains to produce. Without the diagnostic, the times are reduced to about six hours.) This was rather prohibiting, but we could not use Breslow's or Efron's approximate likelihood because they lead to completely different posterior distributions for the same number of iterations even though only a few minutes are required in those computations.

8.2 Inference on the interviewer parameters

Some summaries of the posterior distributions for the interviewer effect parameters are given in Table 5.

Table 5: Posterior Distributions Summary

Parameter	*N*	*Mean*	*Standard Deviation*	*Percentiles*			*95% HPD Credible Interval*	
				25%	*50%*	*75%*	*Lower Bound*	*Upper Bound*
interviewer_id362702	0							
interviewer_id204537	8000	0.06	0.242	-0.10	0.06	0.23	-0.40	0.54
interviewer_id269307	8000	0.14	0.225	-0.02	0.13	0.29	-0.29	0.58
interviewer_id212865	8000	0.33	0.271	0.15	0.32	0.51	-0.16	0.90
interviewer_id244205	8000	0.66	0.246	0.48	0.66	0.83	0.17	1.12
interviewer_id204297	8000	0.71	0.115	0.63	0.71	0.79	0.48	0.93
interviewer_id245813	8000	0.75	0.245	0.58	0.75	0.92	0.29	1.25
interviewer_id259937	8000	0.93	0.219	0.78	0.93	1.08	0.50	1.35
interviewer_id245832	8000	1.67	0.232	1.51	1.68	1.83	1.25	2.13

First, we note that the posterior means, sorted in the ascending order here, happen to be all positive against the baseline interviewer's zero mean value. This is by chance. All of the posterior distributions are almost bell-shaped. HPD stands for the highest posterior density, and their credible intervals are displayed in the table. The equal-tail intervals (not displayed) give similar results because the posterior distributions are highly symmetric.

The first four interviewers, including the baseline interviewer, seem to have the similar level of interviewer effects on the completion times of Sections A and B. The next four interviewers seem similarly different from the baseline interviewer in terms of their effects. The ninth or last interview appears distinctively different from the baseline interviewer and all the other interviewers, as this interviewer's completion rate is exp(1.67) = 5.3 times higher than that of the baseline interviewer (and the three others if their interwar effects are considered identical to the baseline interviewer's). Based on the observation, the data collection operator might investigate the interviewer and the data collected by the interviewer and might take some corrective actions in order to maintain a required data quality. In Figure 2, the normal kernel density of the posterior probability is plotted for each interviewer parameter. They estimate the posterior marginal distributions for the interviewer parameters.

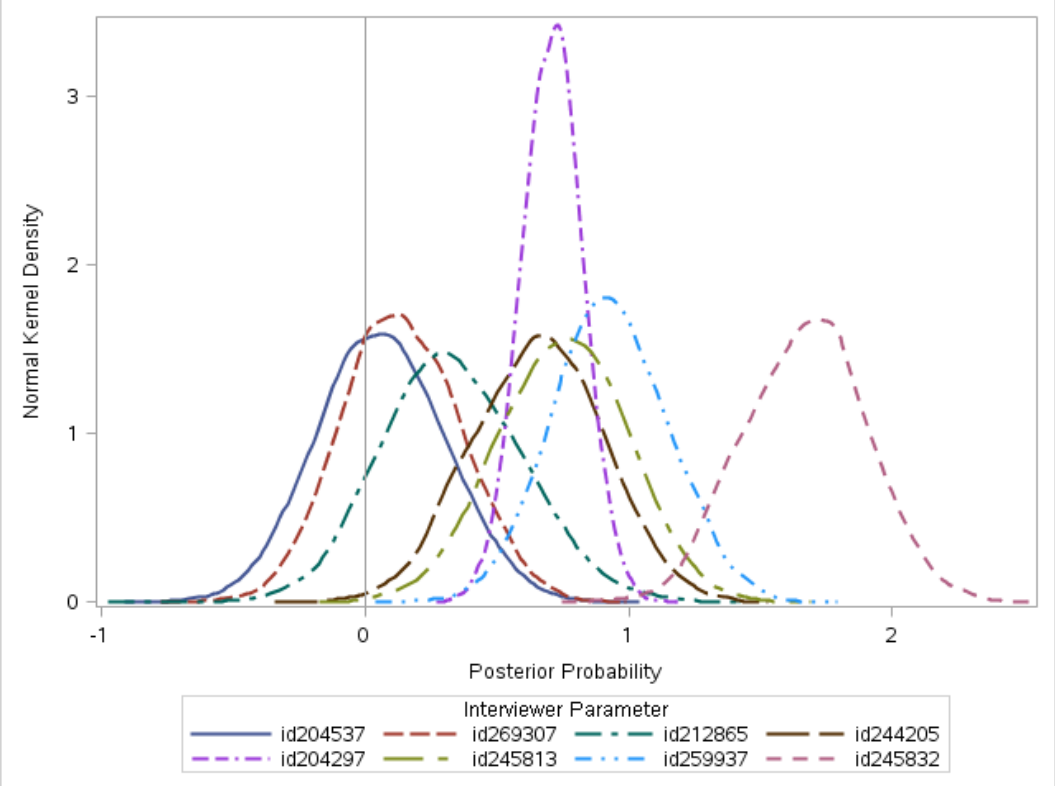

Figure 2: Normal Kernel Density of Posterior Probability of Each Interviewer Parameter

In principle, time and cost available and allocated for controlling data or operation quality as well as the level of quality required in data or operation should determine where to draw a line for flagging. Without those constraints and requirements, the determination of

the line could rely on a statistical or graphical examination of the data. Such an examination would be rather simple and straightforward, and it could be automated.

Acknowledgements

I would like to thank Thomas Leckey in the Office of Energy Consumption and Efficiency Statistics at the U.S. Energy Information Administration for his support and suggestions to improve this paper.

References

American Association for Public Opinion Research (AAPOR) (2011). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys [online]. Available at http://www.aapor.org/Standard_Definitions_New_and_Improved1.htm#.U_OLkfldWIh.

Bouman, P., Meng, X.-L., Dignam, J., and Dukić, V. (2007). A multiresolution hazard model for multicenter survival studies: Application to tamoxifen treatment in early stage breast cancer. *Journal of the American Statistical Association* **102** 1145-1157.

Breslow, N. E. (1972). Discussion of Professor Cox's paper. Journal of Royal Statistical Society: Series B (Statistical Methodology) **34** 216–217.

Brooks, S. P., and Gelman, A. (1997). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7** 434–455.

Brooks, S. P., and Roberts, G. O. (1998). Assessing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing* **8** 319–335.

Brooks, S. P., and Roberts, G. O. (1999). On quantile estimation and Markov chain Monte Carlo convergence. *Biometrika* **86** 710–717.

Couper, M. (1998). Measuring survey quality in a CASIC environment. In American Statistical Association Proceedings of the Survey Research Methods Section.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **34** 187-220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62** 269-79.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* **72** 557-565.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. Chapman and Hall, Boca Raton, FL.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7** 457–472.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting* (J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds.). Oxford University Press.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling with Gibbs sampling. *Applied Statistics* **44** 455–472.

Gilks, W. R., and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41** 337–348.

Glidden, D. V., and Vittinghoff, E. (2004). Modelling clustered survival data from multicenter clinical trials. *Statistics in Medicine* **23** 369-388.

Gustafson, P. (1997). Large hierarchical analysis of multivariate survival data. *Biometrics* **53** 230-242.

Gustafson, P., Aeschliman, D., and Levy, A. R. (2003). A simple approach to fitting Bayesian survival models. *Lifetime Data Analysis* **9** 5-19.

Heidelberger, P., and Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Communication of the ACM* **24** 233–245.

Heidelberger, P., and Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research* **31** 1109–1144.

Hougaard, P. (2000). Analysis of Multivariate Survival Data. Springer, New York.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2005). *Bayesian Survival Analysis.* Springer, New York.

Kalbfleisch, J. D. (1978). Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **40** 214-221.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician* **52** 93–100.

Keiding, N. (2014). Event history analysis. Annual Review of Statistics and its Application 1 333-360.

Kleinbaum, D. G., and Klein, M. (2005). *Survival Analysis*, 2nd ed. Springer, New York.

Kreuter, F., ED. (2013). *Improving Surveys with Paradata: Analytic Uses of Process Information*. Wiley, Hoboken, NJ.

Kreuter, F., Couper, M., and Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. In *American Statistical Association Proceedings of the Survey Research Methods Section*.

Lee, K. H., Chakraborty, S., and Sun, J. (2011). Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data. *The International Journal of Biostatistics* **7** 1-32 [online]. DOI: 10.2202/1557-4679.1301. Available at http://www.degruyter.com/view/j/ijb.2011.7.issue-1/ijb.2011.7.1.1301/ijb.2011.7.1.1301.xml?format=INT.

Lu, J., and Shen, D. (2014). Survival analysis approaches and new developments using SAS. In *Pharmaceutical Industry SAS Users Group Proceedings* [online]. Available at http://www.pharmasug.org/proceedings/2014/PO/PharmaSUG-2014-PO02.pdf.

Raftery, A. E., and Lewis, S. M. (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science* **7** 493–497.

Raftery, A. E., and Lewis, S. M. (1996). The number of iterations, convergence diagnostics and generic Metropolis algorithms. In *Markov Chain Monte Carlo in Practice* (W. R. Gilks, D. J. Spiegelhalter, and S. Richardson, eds.). Chapman & Hall, London.

Sinha, D., Ibrahim, J. G., and Chen, M.-H. (2003). A Bayesian justification of Cox's partial likelihood. *Biometrika* **90** 629-641.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 583–616.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16** 385-395.

U. S. Department of Energy Energy Information Administration (2010). 2009 Residential Energy Consumption Survey Household Questionnaire [online]. Available at http://www.eia.gov/survey/form/eia_457/form.pdf.