HHS Public Access

Author manuscript

Am Stat. Author manuscript; available in PMC 2017 November 01.

Published in final edited form as:

Am Stat. 2017 ; 71(2): 171–176. doi:10.1080/00031305.2017.1296375.

Efficient Computation of Reduced Regression Models

Stuart R. Lipsitz[a], **Garrett M. Fitzmaurice**[b], **Debajyoti Sinha**[c], **Nathanael Hevelone**[d], **Edward Giovannucci**[e], **Quoc-Dien Trinh**[f], and **Jim C. Hu**[g]

[a]Department of Medicine, Brigham and Women's Hospital, Boston, MA

[b]Department of Psychiatry, Harvard Medical School, Boston, MA

[c]Department of Statistics, Florida State University, Tallahassee, FL

[d]Department of Surgery, Brigham and Women's Hospital, Boston, MA

[e]Departments of Nutrition and Epidemiology, Harvard School of Public Health, Boston, MA

[f]Division of Urology, Brigham and Women's Hospital, Boston, MA

[g]Department of Urology, Weill Cornell Medical College, New York, NY

Abstract

We consider settings where it is of interest to fit and assess regression submodels that arise as various explanatory variables are excluded from a larger regression model. The larger model is referred to as the full model; the submodels are the reduced models. We show that a computationally efficient approximation to the regression estimates under any reduced model can be obtained from a simple weighted least squares (WLS) approach based on the estimated regression parameters and covariance matrix from the full model. This WLS approach can be considered an extension to unbiased estimating equations of a first-order Taylor series approach proposed by Lawless and Singhal. Using data from the 2010 Nationwide Inpatient Sample (NIS), a 20% weighted, stratified, cluster sample of approximately 8 million hospital stays from approximately 1000 hospitals, we illustrate the WLS approach when fitting interval censored regression models to estimate the effect of type of surgery (robotic versus nonrobotic surgery) on hospital length-of-stay while adjusting for three sets of covariates: patient-level characteristics, hospital characteristics, and zip-code level characteristics. Ordinarily, standard fitting of the reduced models to the NIS data takes approximately 10 hours; using the proposed WLS approach, the reduced models take seconds to fit.

Keywords

Complementary log–log regression; c survey; Weighted estimating equations; Weighted least squares

CONTACT Stuart R. Lipsitz, slipsitz@partners.org, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115.

1. Introduction

We consider settings where it is of interest to fit and assess a sequences of regression models that arise as various explanatory variables are excluded from a larger regression model. The larger regression model is referred to as the full model and the submodels to be fitted are referred to as the reduced models. In particular, we are interested in situations, where the fitting of the regression models is computationally demanding. We note that there are many "big data" settings, where the fitting of regression models is computationally demanding, even with the widespread availability of powerful computing facilities. We show that a computationally efficient approximation to the regression estimates under any reduced model can be obtained from a simple weighted least squares (WLS) approach based on the regression parameter estimates and their estimated variance matrix from the full model. This WLS approach can be considered an extension of a first-order Taylor series approach first suggested by Lawless and Singhal (1978) and later by Kowalski and Hutmacher (2001). Both Lawless and Singhal (1978) and Kowalski and Hutmacher (2001) assume that the full likelihood has been correctly specified, but we show here that this can be relaxed, and it is only necessary that we have a set of unbiased estimating equations that are robust to mis-specification of higher-order moments (other than the mean) of the response. This allows the approach of Lawless and Singhal (1978) and Kowalski and Hutmacher (2001) to be applied to a much greater class of estimating equations than the score equations from a correctly specified likelihood.

Using data from the 2010 Nationwide Inpatient Sample (NIS), a 20% weighted, stratified, cluster sample of approximately 8 million hospital stays from approximately 1000 hospitals, we illustrate the WLS approach when fitting interval censored regression models to estimate the effect of type of surgery (robotic surgery versus nonrobotic surgery) on hospital length-of-stay while adjusting for patient-level, hospital, and zip-code level characteristics. Ordinarily, standard fitting of the reduced models to the NIS data takes approximately 10 hours; using the proposed WLS approach, the reduced models take seconds to fit.

Next, we describe the NIS dataset that motivated the application of the WLS approach to fitting reduced regression models. Robotic-assisted laparoscopic surgery has been rapidly adopted (Barbash and Glied 2010) despite the dearth of evidence demonstrating superior outcomes compared to traditional surgical approaches: non-robotic-assisted laparoscopic surgery and open surgery. One important outcome to compare among the three surgical approaches is length-of-stay, defined as time from hospital admission to discharge. To compare the three surgical approaches, we use the 2010 NIS, a 20% stratified, cluster probability sample that encompasses approximately 8 million acute hospital stays per year from approximately 1000 hospitals in 37 states. In the NIS, hospitals in the sampling frame are stratified by five key characteristics. Then, a random sample of hospitals (clusters) is chosen from each of the strata. The NIS includes all discharges from the selected hospitals. Each hospital has a different probability of being selected in the sample depending on the five characteristics that determine the strata. As a result, each hospital, and thus all discharges within the hospital, are given a weight so that any results can be extrapolated to the entire universe of hospitals in the United States. In the 2010 NIS, there are 60 strata,

1049 clusters (hospitals), and the average cluster size is 127, giving a dataset with over 8 million observations.

The NIS is an example of a complex sample survey, which is typically a stratified cluster design in which each individual in the population has a different probability of being selected into the sample. Design-based analyses usually incorporate the weighting (inverse probability of being sampled), stratification, and clustering variables. When estimating the regression parameters of a generalized linear model for any type of outcome (continuous or discrete) from complex survey data with large clusters, for reasons of computational feasibility, the most popular approach is to naively assume the observations within a cluster are independent to obtain consistent regression parameter estimates (Binder 1983; Liang and Zeger 1986); a consistent estimate of the covariance matrix of these estimates can be obtained using a "sandwich estimator" (Binder 1983; Liang and Zeger 1986). In the NIS, hospital length-of-stay is rounded to the day (integers from 0 to 365). Because of this rounding, we use interval censoring techniques (Prentice and Gloeckler 1978; Whitehead 1989) for estimating the proportional hazards models, naively assuming subjects within a cluster are independent. Using this approach requires creating "risk sets" of patients available for hospital discharge each day of the year with an indicator (outcome) equal to 1 if the patient is discharged that day and 0 otherwise, resulting in a dataset of over 45 million observations. Estimation for this interval censored proportional hazards model is then performed on this data using a complementary log–log link regression for binary outcomes. The main predictor of interest is the patient level covariate: type of surgery, with three levels: robotic-assisted laparoscopic surgery, non-robotic-assisted laparoscopic surgery and open surgery. Other predictors of length-of-stay include patient characteristics: patient race (white versus other), gender (except for prostate cancer), stage (advanced cancer stage versus other), number of patient comorbidities, and private insurance (yes, no); area (zip-code)-level covariates: the median household income of the patient's zip code of residence; and hospital characteristics: hospital bed size (categorized as small, medium, and large), hospital location (urban vs. rural), and hospital teaching status (teaching vs. not teaching).

Further, length-of-stay also depends on the procedure. We are interested in the three procedures: prostatectomy for prostate cancer, nephrectomy for kidney cancer, and partial nephrectomy for kidney cancer. We expect the model to be different for each of these diagnoses. For descriptive purposes, Table 1 gives the unadjusted estimated probability of being hospitalized more than one day by procedure and surgery type, with the probabilities obtained from a weighted interval censored Kaplan–Meier curve (Peto 1973), weighted by the complex survey weight. In general, from Table 1, the procedure prostatectomy appears to have the lowest probabilities of being hospitalized more than one day and robotic surgery has the lowest probabilities of being hospitalized more than one day. However, the results in this table do not take into account the possible confounders. We would like to fit separate interval-censored regression models for each procedure and surgery type to quantify effects and provide tests for significance. When estimating a regression model for each procedure and surgery type, one can restrict the analysis to the particular diagnoses and obtain unbiased parameter estimates. However, because of the complex survey design, restricting the analysis to the subgroup of interest does not yield correct standard errors (Graubard and Korn 1996). Unlike a simple random sample, for most complex survey designs, the

variance–covariance matrix for the parameter estimates is not block diagonal across subgroups. As a result, all subjects must be incorporated in the analysis, even when interest is only in a small subgroup. Correct standard errors of estimated regression model parameters can be obtained by analyzing data from all subjects, and including interaction terms between subgroup (procedure for NIS) and all covariates in the model. For example, using the National Health Interview Survey, Graubard and Korn (1996) showed that the 1987 death rates from digestive cancer per 100,000 individuals for whites is estimated to be 69.01; using the full sample with interactions, the estimated standard error (SE) is 3.24. However, by incorrectly restricting the analysis to whites, the estimated SE is 0.89, an almost four-fold decrease. Thus, for the NIS analysis, all 8,023,590 patients must be included in the analysis and subgroup-specific estimates are obtained through the inclusion of interactions between diagnoses and covariates. For the NIS analysis, in addition to the three procedures (subgroups) of interest, a fourth subgroup with “everybody else” (about 8 million patients) must also be created.

Our regression models of interest are ones that include: (1) interaction between type of surgery (robotic-assisted laparoscopic, non-robotic-assisted laparoscopic, open) and procedure; (2) Model (1) plus interaction between patient-level covariates and procedure; (3) Model (1) plus interaction between the area-level covariate and procedure; (4) Model (1) plus interaction between hospital-level covariates and procedure; (5) Model (2) plus interaction between the area-level covariate and procedure; and (6) the “full” model which is Model (5) plus interaction between hospital-level covariates and procedure. As we describe in Section 4, all models take between 10 and 11 hours to fit, so it is computationally intensive to fit all six models. We propose first fitting the “full model” with all potential confounders. Then, instead of refitting the reduced models (1, 2, 3, 4, or 5) which would take more than 10 hours each, we fit reduced models using a computationally simple weighted least-squares approach based on the regression parameter estimates and their estimated variance matrix from the “full” model.

In Section 2, we briefly describe the complex survey design and the interval censored model. In Section 3, we describe the weighted least squares approach to fitting reduced models from the “full” model. In Section 4, we apply the proposed approach to the regression analyses of the data from the NIS.

2. Estimating Equations for Interval Censored Proportional Hazards Models

For many complex sample surveys, the target population is usually thought to be of finite size N, and a total of n subjects are sampled. To indicate which n subjects are sampled from the population of N subjects, we define the indicator random variable

$$\delta_i = \begin{cases} 1 \text{ if subject } i \text{ is selected into sample} \\ 0 \text{ if subject } i \text{ is not selected into sample} \end{cases},$$

for $i = 1, \ldots, N$, where $\sum_{i=1}^{N} \delta_1 = n$. Depending on the sampling design, some of the δ_i may be correlated (e.g., for two subjects within the same cluster).We let π_i denote the probability

of subject i being selected into the survey, which is typically specified in the design of the study; π_i may depend on the variables of interest, or additional variables (screening variables, for example) not in the analytic model of interest. Each subject sampled can have a different "weight" determined by the inverse of the probability of being selected into the sample, for example, $w_i = \delta_i/\pi_i$.

Let T_i be the length-of-stay for subject i. We assume that it has been discretized to be a nonnegative integer, $T_i = 0, 1, 2, \ldots$. Also, let $\mathbf{x}_i = (x_{i1}, \ldots, x_{iK})$ be a vector of K covariates. Let $p_{ij} = pr(T_i = j | T_i > j - 1, \mathbf{x}_i)$ be the probability that subject i was discharged on day j given the patient had not yet been discharged. Assuming that the survival time was discretized from a continuous time $\mathscr{T}_i$, then

$$\mathrm{pr}(T_i=j|T_i>j-1,\mathbf{x}_i)=\mathrm{pr}(j-1<\mathscr{T}_i\le j|\mathscr{T}_i>j-1,\mathbf{x}_i)=\frac{S(j-1|\mathbf{x}_i)-S(j|\mathbf{x}_i)}{S(j-1|\mathbf{x}_i)}=1-\frac{S(j|\mathbf{x}_i)}{S(j-1|\mathbf{x}_i)},$$

where $S(\mathscr{T}_i|\mathbf{x}_i)$ is the survival distribution of the underlying continuous survival time $\mathscr{T}_i$.

If the underlying survival distribution is proportional hazards, then (Cox 1972)

$$S(\mathscr{T}|\mathbf{x}_i)=\exp\{-\Lambda_0(\mathscr{T})e^{\mathbf{x}_i'\boldsymbol{\beta}}\},$$

where $\Lambda_0(\mathscr{T})$ is the baseline cumulative hazard and $\boldsymbol{\beta}$ are the parameters of interest. Then,

$$p_{\mathrm{ij}}=\mathrm{pr}(T_i=j|T_i>j-1,\mathbf{x}_i)=1-\exp\{-[\Lambda_0(j)-\Lambda_0(j-1)]e^{\mathbf{x}_i'\boldsymbol{\beta}}\}=1-\exp\{-e^{\gamma_j+\mathbf{x}_i'\boldsymbol{\beta}}\}$$

where

$$\gamma_j=\log[\Lambda_0(j)-\Lambda_0(j-1)].$$

The probability that a subject was discharged on day j is

$$\mathrm{pr}(T_i=j|\mathbf{x}_j)=\left[\prod_{\ell=1}^{j-1}(1-p_{i\ell})\right]p_{\mathrm{ij}}.$$

and the probability that a subject was censored (transferred to a different hospital) on day j is

$$\mathrm{pr}(T_i>j|\mathbf{x}_i)=\left[\prod_{\ell=1}^{j-1}(1-p_{i\ell})\right](1-p_{\mathrm{ij}}).$$

If we let C_i be the censoring (transfer) time, then we observe $U_i = \min(T_i, C_i)$. Further, if we let $D_{i\ell}$ equal 1 if the subject was discharged at time ℓ, ($\ell = 0, 1, 2, \ldots$.) and equal 0 otherwise, then the likelihood for subject i is

$$L(U_i|\boldsymbol{\gamma},\boldsymbol{\beta})=\left[\prod_{\ell=1}^{U_i-1}(1-p_{i\ell})\right]p_{\mathrm{iU}_i}^{D_{\mathrm{iU}_i}}(1-p_{\mathrm{iU}_i})^{1-D_{\mathrm{iU}_i}}. \quad (2.1)$$

Note further, without loss of generality, we can write (2.1) as

$$L(U_i)=\left[\prod_{\ell=1}^{U_i}p_{i\ell}^{D_{i\ell}}(1-p_{i\ell})^{1-D_{i\ell}},\right] \quad (2.2)$$

where $D_{i\ell}=0$ for $\ell < U_i$. Then (2.2) is equivalent to a product of independent Bernoulli observations, where the probability of "success" for each binary observation follows a complementary log–log regression model.

One possibility for estimation of the proportional hazards model for complex survey data is the Cox regression approach proposed by Binder (1992) which incorporates the weights for each subject; unfortunately, for discrete times (many "ties"), this approach can give badly biased estimates. Thus, a far better approach is to use the weights in a weighted pseudo-likelihood for the product of binary observations for each subject using any complex survey generalized linear model program that allows for weights and a complementary log–log link. Note, for the complementary log–log regression model, we will need to fit a different intercept for each failure time (with the large samples of complex surveys like NIS, adequate sample size to estimate the intercepts is not an issue).

To obtain a consistent estimate of $(\boldsymbol{\gamma}, \boldsymbol{\beta})$, one can use a weighted estimating equation for independent Bernoulli observations with a complementary log–log link, which is the solution to $\mathbf{U}_{\mathrm{wee}}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}) = \mathbf{0}$, where

$$\mathbf{U}_{\mathrm{wee}}(\hat{\boldsymbol{\gamma}},\hat{\beta})=\sum_{i=1}^{N}\frac{\delta_i}{\pi_i}\sum_{\ell=1}^{U_i}\mathbf{x}_{i\ell}[\exp\{\gamma_j+\mathbf{x}_i'\boldsymbol{\beta}-e^{\gamma_j+\mathbf{x}_i'\boldsymbol{\beta}}\}]\times(D_{i\ell}-\hat{\mathrm{p}}_{i\ell})/[\hat{\mathrm{p}}_{i\ell}(1-\hat{\mathrm{p}}_{i\ell})], \quad (2.3)$$

where $\mathbf{x}_{i\ell}$ is a vector containing indicators for the intercept and $\mathbf{x}_i$. Here, the "weights" are $w_i=\frac{\delta_i}{\pi_i}(w_i=\frac{1}{\pi_i}$ if sampled $\delta_i=1)$. Note, also, these are weighted likelihood score equations under a working "independence" assumption for the N subjects (disregarding any clustering). Heuristically, using method-of-moments ideas (Binder 1983), $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ is consistent because (2.3) is an unbiased estimating equation with $E[\mathbf{U}_{\mathrm{wee}}(\boldsymbol{\gamma}, \boldsymbol{\beta})] = 0$ and $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ obtained as the solution to $\mathbf{U}_{\mathrm{wee}}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}) = 0$.

Using a first-order Taylor series expansion and a suitable central limit theorem for sample survey data (Binder 1983), $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ has an asymptotic multivariate normal distribution with mean $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ and covariance matrix

$$V_{\gamma,\beta}=\mathrm{var}[(\hat{\gamma},\hat{\beta})]=\left[E\left(\frac{d\mathbf{U}_{\mathrm{wee}}(\boldsymbol{\gamma},\boldsymbol{\beta})}{d\boldsymbol{\theta}}\right)\right]^{-1}\times\mathrm{var}[\mathbf{U}_{\mathrm{wee}}(\boldsymbol{\gamma},\beta)]\left[E\left(\frac{d\mathbf{U}_{\mathrm{wee}}(\boldsymbol{\gamma},\boldsymbol{\beta})}{d\boldsymbol{\theta}}\right)\right]^{-1}, \quad (2.4)$$

where $\boldsymbol{\theta}=(\boldsymbol{\gamma}',\boldsymbol{\beta})'$. Note, var[$\mathbf{U}_{\mathrm{wee}}$ $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$)] depends on the sample design (stratification, clustering, sampling with or without replacement) as well as the finite population correction factor. Empirically, (2.4) is estimated via the "sandwich variance estimator." Further, we denote

$$V_0=\left[E\left(\frac{d\mathbf{U}_{\mathrm{wee}}(\boldsymbol{\gamma},\boldsymbol{\beta})}{d\boldsymbol{\theta}}\right)\right]^{-1} \quad (2.5)$$

as the variance of ($\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\beta}}$) under the working 'independence' assumption for the N subjects (disregarding any clustering).

3. One-Step Weighted Least Squares Estimator

With large complex surveys such as NIS with 8million observation, fitting the complementary log–log model in the previous section can be very time consuming. Even the simple unadjusted model with only the main covariate of interest (type of surgery) takes 10 hours to run on a high-end PC workstation, so it is computationally intensive to fit a sequence of model with the different sets of covariates of interest. To avoid this computational burden, instead of fully iterating to fit each model of interest (which takes more than 10 hours each), we propose use of a first-order Taylor series approximation similar to that discussed by Lawless and Singhal (1978). Given the parameter estimates and their estimated covariance matrix from the "full" model, one would only need to use noniterative weighted least squares (WLS) to fit all of the reduced models.

To use WLS to fit any submodel, we fit the largest possible model (the "full model").

Suppose we partition $\boldsymbol{\theta}'=[\boldsymbol{\theta}_1',\boldsymbol{\theta}_2']$ and want to fit the submodel with $\boldsymbol{\theta}_2 = 0$. Let $\hat{\boldsymbol{\theta}}'=[\hat{\boldsymbol{\theta}}_1',\hat{\boldsymbol{\theta}}_2']$ be the estimate of $\boldsymbol{\theta}$ from the full model. Suppose $\boldsymbol{\theta}_1$ has dimension ($a \times 1$) and $\boldsymbol{\theta}_2$ has dimension ($b \times 1$). When $\boldsymbol{\theta}_2 = 0$, the asymptotic expectation (denoted E_A) of $\hat{\boldsymbol{\theta}}$ is

$$E_A\begin{bmatrix}\hat{\boldsymbol{\theta}}_1\\ \hat{\boldsymbol{\theta}}_2\end{bmatrix}=Z\boldsymbol{\theta}_1=\begin{bmatrix}I_a\\ 0_{\{b\times a\}}\end{bmatrix}\boldsymbol{\theta}_1, \quad (3.6)$$

where I_a is and ($a \times a$) identity matrix and $0_{\{b\times a\}}$ is a ($b \times a$) matrix of 0's.With $\hat{\boldsymbol{\theta}}$ as the "outcome"' vector, we propose using WLS to estimate $\boldsymbol{\theta}_1$ (assuming $\boldsymbol{\theta}_2 = 0$), with weight matrix equal to the inverse of $\hat{V}_0$ given in (2.5). Since we are using the estimating equations under a working "independence" assumption for all models, we must use the estimated covariance matrix under this independence assumption in the WLS algorithm. Without loss of generality, we assume that $\hat{V}_0$ is partitioned in the correct order of the vector ordering $[\boldsymbol{\theta}_1',\boldsymbol{\theta}_2']$.

Then, the weighted least squares estimate of $\boldsymbol{\theta}_1$, using the weight matrix $\hat{V}_0^{-1}$, is

$$\tilde{\boldsymbol{\theta}}_1=\left[Z'\hat{V}_0^{-1}Z\right]^{-1}\left[Z'\hat{V}_0^{-1}\hat{\boldsymbol{\theta}}\right], \quad (3.7)$$

The large sample covariance matrix of $\tilde{\boldsymbol{\theta}}_1$ equals

$$\widehat{\mathrm{var}}(\tilde{\boldsymbol{\theta}}_1)=C\hat{V}_{\gamma,\beta}C', \quad (3.8)$$

where

$$C=\left[Z'\hat{V}_0^{-1}Z\right]^{-1}Z'\hat{V}_0^{-1}$$

and $\hat{V}_{\gamma,\beta}$ is the robust sandwich estimator given in (2.4). Given $\tilde{\boldsymbol{\theta}}_1$ and $\widehat{\mathrm{var}}(\tilde{\boldsymbol{\theta}}_1)$, from the particular submodel, one can use the Wald statistic for all the elements of $\tilde{\boldsymbol{\theta}}_1$. For the pth element of $\boldsymbol{\theta}_1$, the Wald statistic is

$$Z_p=\frac{\tilde{\boldsymbol{\theta}}_{1p}}{\widehat{\mathrm{var}}(\tilde{\boldsymbol{\theta}}_{1p})}.$$

The approach here outlined for deriving $\tilde{\boldsymbol{\theta}}_1$ and $\widehat{\mathrm{var}}(\tilde{\boldsymbol{\theta}}_1)$ is similar to the first-order Taylor series estimate of $\boldsymbol{\theta}_1$ given in Lawless and Singhal (1978) and Kowalski and Hutmacher (2001). Unlike their likelihood-based approach which assumes that the full likelihood has been correctly specified, we use an estimating equations approach for clustered data that only requires an unbiased set of estimating equations for $\boldsymbol{\theta}$. In particular, to get a set of unbiased estimated equations, we are estimating the parameters under a working independence assumption. Because we are using estimating equations, the covariance estimate (3.8) is based on a sandwich estimator. Thus, for the clustered data in our example, the novelty of this approach is that we can use the naive independence assumption in the first order Taylor series approximation for the reduced models, thereby avoiding assumptions about the higher order moments (other than the mean) of the response.

4. Application to 2010 Nationwide Inpatient Sample

In this section, we present results for analyses of data from the NIS study discussed in the Introduction. The main outcome of interest is the patient's length-of-stay in days (1–365) in 2010. We fit an interval censored proportional hazards model for this outcome, which first requires creating "risk sets" of patients available for hospital discharge each day of the year with an indicator (outcome) equal to 1 if the patient is discharged that day and 0 otherwise, resulting in a dataset of over 45 million observations. Estimation for this interval censored proportional hazards model is then performed using complementary log-log link regression for binary outcomes. The main predictor of interest is the patient-level covariate type of

surgery, classified as robotic-assisted laparoscopic surgery, non-robotic-assisted laparoscopic surgery and open surgery, with open as the reference group. Other patient-level covariates used for adjustment are: Caucasian (1 if Caucasian, 0 if otherwise), comorbidities (number of patient comorbidities), age (in years), gender (1 if female, 0 if male), stage (1 if advanced cancer stage, 0 if otherwise), insurance (1 if private insurance, 0 if otherwise). The hospital-level covariates of interest are: medium bedsize (1 if medium bed size, 0 if otherwise), large bedsize (1 if large bed size, 0 if otherwise), urban (1 if urban, 0 if otherwise), and teaching (1 if teaching, 0 if otherwise). The one zip code level covariate median household income of the patient's zip code of residence, with 4 levels (<$41,000; $41,000-$50,999; $51,000-$66,999; and ≥ $67,000). This income variable is a proxy for patient income to at least account for some of the possible confounding due to patient's income.

We are particularly interested in length-of-stay for urological conditions: prostate, kidney, and bladder cancer. Table 2 gives the fully iterated and WLS estimates of $\boldsymbol{\beta}$ for the three surgical procedures. The computation time (real, not CPU) isbetween10 and 11 hours for all fully iterated models (even unadjusted) as implemented using the complementary log-log link in SAS Proc SurveyLogistic (SAS Institute 2015). The striking result from Table 2 is that there is very little evidence of confounding, that is, all models yield very similar results. Note that the WLS estimates are similar to the fully iterated estimates. Instead of fitting all models, with each taking at least 10 hours to fit, one could have seen there is little evidence of confounding by simply fitting the "full-model," and then using the WLS approach to obtain estimates from the reduced models.

Based on the results from the "full-model," for all three surgical procedures, robotic surgery significantly ($P < 0.05$) reduces length-of-stay when compared to open surgery. In particular, for prostatectomy, nephrectomy, and partial nephrectomy, when compared to open surgery, robotic surgery increases the hazard rate of "discharge" by exp(.567) ≈ 1.76, exp(.421) ≈ 1.52, and exp(.603) ≈ 1.83, respectively. When compared to open surgery, laparoscopic surgery increases the hazard rate of "discharge" by exp(.428) ≈ 1.53, exp(.287) ≈ 1.33, and exp(.406) ≈ 1.50, respectively; the latter two hazard ratios are significant at the 5% level, but the first is not. Both unadjusted and adjusted for other factors, robotic surgery (versus either laparoscopic or open surgery) appears to lead to the greatest increase in the hazard rate of discharge for all three procedures.

5. Conclusion

In this article, we show that a computationally efficient approximation to the regression estimates under any reduced model can be obtained from a simple weighted least squares (WLS) approach based on the regression parameter estimates and their estimated variance matrix from the full model. We note that this one-step weighted least squares approach for reduced models can be used for any type of regression model, and will be particularly useful when the fitting of all reduced models is computationally intensive. For example, Cox regression models with time-varying covariates, many Bayesian approaches, generalized estimating equations with very large clusters are all computationally intensive approaches. The novelty of our approach is that, when using the generalized estimating equations approach (or any method-of-moments approach), the one-step WLS for the reduced model

can be based on a "working correlation model," with standard errors estimated by a sandwich variance estimator. Both Lawless and Singhal (1978) and Kowalski and Hutmacher (2001) assume that the full likelihood has been correctly specified, but we show here that this can be relaxed, and it is only necessary that we have a set of unbiased estimating equations. This allows the approach of Lawless and Singhal (1978) and Kowalski and Hutmacher (2001) to be applied to a much greater class of estimating equations than the score equations from a correctly specified likelihood.

When applied to data from the 2010 NIS we found that, after fitting the "fully iterated" model, it was transparent that there was little evidence of confounding for submodels, without having to wait 10 hours to fit each reduced model. Our results also suggest that a one-step WLS approach using the fully iterated estimates from the "full model" can lead to a huge computational saving when there is interest in examining various submodels.

Acknowledgments

Funding

The authors gratefully acknowledge the support provided by grant CA 160679 from the U.S. National Institutes of Health.

References

Barbash GI, Glied SA. New Technology and Health Care Costs–The Case of Robot-Assisted Surgery. New England Journal of Medicine. 2010; 363:701–704. [PubMed: 20818872]

Binder DA. On the Variances of Asymptotically Normal Estimators From Complex Surveys. International Statistical Review/Revue Internationale de Statistique. 1983; 51:279–292.

Binder DA. Fitting Cox's Proportional Hazards Models From Survey Data. Biometrika. 1992; 79:139–147.

Cox DR. "Regression Models and Life Tables" (with discussion). Journal of the Royal Statistical Society. 1972; 34:187–220.

Graubard BI, Korn EL. Survey Inference for Subpopulations. American Journal of Epidemiology. 1996; 144:102–106. [PubMed: 8659480]

Kowalski KG, Hutmacher MM. Efficient Screening of Covariates in Population Models Using Wald's Approximation to the Likelihood Ratio Test. Journal of Pharmacokinetics and Pharmacodynamics. 2001; 28:253–275. [PubMed: 11468940]

Lawless JF, Singhal K. Efficient Screening of Nonnormal Regression Models. Biometrics. 1978; 143:318–327.

Liang KY, Zeger SL. Longitudinal Data Analysis Using Generalized Linear Models. Biometrika. 1986; 73:13–22.

Peto R. Experimental Survival Curves for Interval-Censored Data. Applied Statistics. 1973; 22:86–91.

Prentice RL, Gloeckler LA. Regression Analysis of Grouped Survival Data With Application to Breast Cancer Data. Biometrics. 1978; 34:57–67. [PubMed: 630037]

SAS Institute. SAS/STAT Software, Version 9.4. Cary, NC: SAS Institute; 2015.

Whitehead J. The Analysis of Relapse Clinical Trials, With Application to a Comparison of Two Ulcer Treatments. Statistics in Medicine. 1989; 8:1439–1454. [PubMed: 2616934]

Table 1

Estimated probability of being hospitalized more than one day by procedure and surgery type from weighted interval censored Kaplan–Meier curves.

	Type of surgery		
Procedure	**Robotic**	**Laparoscopic**	**Open**
Prostatectomy	0.147	0.134	0.138
Nephrectomy	0.154	0.135	0.378
Partial Nephrectomy	0.135	0.383	0.377

Table 2

WLS and fully iterated complementary log–log regression estimates for different models.

		Fully iterated			Weighted least squares		
Effect	**Model**	**Estimate**	**SE**	***P*-value**	**Estimate**	**SE**	***P*-value**
Prostatectomy:							
Robotic	Unadjusted	0.581	0.096	<0.001	0574	0.092	<0.001
	Patient	0.580	0.096	<0.001	0579	0.091	<0.001
	Hospital	0.562	0.093	<0.001	0.561	0.092	<0.001
	Zip	0.570	0.094	<0.001	0.561	0.092	<0.001
	Patient + Hospital	0.564	0.094	<0.001	0570	0.092	<0.001
	Patient + Zip	0.577	0.095	<0.001	0573	0.092	<0.001
	Hospital + Zip	0.559	0.091	<0.001	0554	0.093	<0.001
	Full	0.567	0.093	<0.001	0567	0.093	<0.001
Laparoscopic	Unadjusted	0.416	0.298	0.163	0.383	0.303	0.206
	Patient	0.422	0.304	0.164	0.409	0.303	0.178
	Hospital	0.428	0.292	0.143	0.402	0.302	0.184
	Zip	0.398	0.298	0.182	0.376	0.302	0.214
	Patient + Hospital	0.439	0.301	0.145	0.428	0.303	0.157
	Patient + Zip	0.408	0.305	0.182	0.406	0.303	0.181
	Hospital + Zip	0.414	0.293	0.157	0.396	0.302	0.189
	Full	0.428	0.302	0.156	0.428	0.302	0.156
Nephrectomy:							
Robotic	Unadjusted	0.367	0.042	<0.001	0.360	0.042	<0.001
	Patient	0.421	0.042	<0.001	0.423	0.041	<0.001
	Hospital	0.368	0.041	<0.001	0.362	0.041	<0.001
	Zip	0.363	0.043	<0.001	0.354	0.042	<0.001
	Patient + Hospital	0.422	0.041	<0.001	0.424	0.041	<0.001
	Patient + Zip	0.419	0.043	<0.001	0.419	0.041	<0.001
	Hospital + Zip	0.364	0.041	<0.001	0.356	0.041	<0.001
	Full	0.421	0.041	<0.001	0.421	0.041	<0.001
Laparoscopic	Unadjusted	0.288	0.038	<0.001	0.275	0.032	<0.001

		Fully iterated			Weighted least squares		
Effect	**Model**	**Estimate**	**SE**	***P*-value**	**Estimate**	**SE**	***P*-value**
	Patient	0.289	0.032	<0.001	0.286	0.032	<0.001
	Hospital	0.285	0.037	<0.001	0.273	0.032	<0.001
	Zip	0.280	0.037	<0.001	0.269	0.032	<0.001
	Patient + Hospital	0.293	0.032	<0.001	0.290	0.032	<0.001
	Patient + Zip	0.283	0.031	<0.001	0.283	0.032	<0.001
	Hospital + Zip	0.277	0.036	<0.001	0.267	0.031	<0.001
	Full	0.287	0.031	<0.001	0.287	0.031	<0.001
Partial Nephrectomy:							
Robotic	Unadjusted	0.563	0.062	<0.001	0564	0.052	<0.001
	Patient	0.585	0.063	<0.001	0.591	0.052	<0.001
	Hospital	0.574	0.050	<0.001	0573	0.050	<0.001
	Zip	0.563	0.063	<0.001	0559	0.052	<0.001
	Patient + Hospital	0.597	0.049	<0.001	0.603	0.050	<0.001
	Patient + Zip	0.590	0.064	<0.001	0.591	0.052	<0.001
	Hospital + Zip	0.576	0.051	<0.001	0570	0.050	<0.001
	Full	0.603	0.050	<0.001	0.603	0.050	<0.001
Laparoscopic	Unadjusted	0.395	0.062	<0.001	0.392	0.072	<0.001
	Patient	0.404	0.077	<0.001	0.404	0.073	<0.001
	Hospital	0.399	0.058	<0.001	0.395	0.072	<0.001
	Zip	0.377	0.061	<0.001	0.376	0.072	<0.001
	Patient + Hospital	0.410	0.073	<0.001	0.410	0.073	<0.001
	Patient + Zip	0.397	0.076	<0.001	0.398	0.073	<0.001
	Hospital + Zip	0.384	0.056	<0.001	0.382	0.071	<0.001
	Full	0.406	0.072	<0.001	0.406	0.072	<0.001