2015

The number of subjects per variable required in linear regression analyses

Peter Austin, *Institute for Clinical Evaluative Sciences*
Ewout Steyerberg

Journal of Clinical Epidemiology 68 (2015) 627–636

Journal of Clinical Epidemiology

The number of subjects per variable required in linear regression analyses

Peter C. Austin[a,b,c,*], Ewout W. Steyerberg[d]

[a]Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ontario, Canada M4N 3M5
[b]Institute of Health Policy, Management and Evaluation, University of Toronto, 155 College Street, Suite 425 Toronto, ON M5T 3M6, Canada
[c]Schulich Heart Research Program, Sunnybrook Research Institute, 2075 Bayview Avenue, Toronto, ON M4N 3M5, Canada
[d]Department of Public Health, Erasmus MC—University Medical Center Rotterdam, 's-Gravendijkwal 230 3015 CE, Rotterdam, The Netherlands

Accepted 24 December 2014; Published online 22 January 2015

Abstract

Objectives: To determine the number of independent variables that can be included in a linear regression model.

Study Design and Setting: We used a series of Monte Carlo simulations to examine the impact of the number of subjects per variable (SPV) on the accuracy of estimated regression coefficients and standard errors, on the empirical coverage of estimated confidence intervals, and on the accuracy of the estimated R^2 of the fitted model.

Results: A minimum of approximately two SPV tended to result in estimation of regression coefficients with relative bias of less than 10%. Furthermore, with this minimum number of SPV, the standard errors of the regression coefficients were accurately estimated and estimated confidence intervals had approximately the advertised coverage rates. A much higher number of SPV were necessary to minimize bias in estimating the model R^2, although adjusted R^2 estimates behaved well. The bias in estimating the model R^2 statistic was inversely proportional to the magnitude of the proportion of variation explained by the population regression model.

Conclusion: Linear regression models require only two SPV for adequate estimation of regression coefficients, standard errors, and confidence intervals.

Keywords: Regression; Linear regression; Bias; Monte Carlo simulations; Explained variation; Statistical methods

1. Introduction

The question of how many independent predictor variables may be included in a multivariable regression model is one that is faced by statistical analysts and applied researchers in diverse fields of research. Overfitting of regression models arises when a regression model includes more predictor variables or incorporates more analytic steps (e.g., univariate prescreening of variables, stepwise selection of variables, searching for nonlinear transformations and statistical interactions) than are warranted by the amount of data available [1–3]. A consequence of overfitting a regression model is that the model may predict poorly in subsequent subjects who were not used for model derivation. This arises because the systematic component of the fitted model has incorporated idiosyncrasies of the sample in which it was developed. Furthermore, the estimated R^2 statistic of a linear regression model can be artificially inflated as the number of subjects per variable (SPV) decreases [when outcomes are binary or time-to-event in nature, the corresponding quantities are the effective sample size, denoting the number of observed events and the number of events per variable (EPV)] [4]. Although the effects of overfitting in the context of logistic regression or survival analysis have been examined in greater detail, the number of subjects required to achieve accurate estimation of regression coefficients in the context of linear regression has not been explored to the same extent.

Conflict of interest: None.

Funding: This study was supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results, and conclusions reported in this article are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. This research was supported by an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508). P.C.A. is supported in part by a Career Investigator award from the Heart and Stroke Foundation. The Enhanced Feedback for Effective Cardiac Treatment (EFFECT) data used in the study were funded by a CIHR Team Grant (CRT 43823 and CTP 79847) in Cardiovascular Outcomes Research. These data sets were linked using unique, encoded identifiers and analyzed at the Institute for Clinical Evaluative Sciences (ICES).

* Corresponding author. Tel.: 416-480-6131; fax: 416-480-6048.
E-mail address: peter.austin@ices.on.ca (P.C. Austin).

http://dx.doi.org/10.1016/j.jclinepi.2014.12.014

What is new?

Key findings

- Two subjects per variable (SPV) tends to permit accurate estimation of regression coefficients in a linear regression model estimated using ordinary least squares.
- When the number of SPV is low, the adjusted R^2 is to be preferred over the conventional R^2 for quantifying the proportion of variance explained by the model.

What this adds to what was known?

- Prior studies have examined the minimum number of events per variable required for estimation of logistic regression models and Cox proportional hazards regression models.
- The required number of SPV for linear regression appears much smaller than in logistic or Cox regression.

What is the implication and what should change now?

- When fitting multivariable/multiple linear regression models, analysts should require a minimum of only two SPV in the model to guarantee unbiased estimation of coefficients and adjusted R^2 values but higher numbers for adequate statistical power.

The primary objective of the present study was to investigate the number of SPV required for accurate estimation of a linear regression model. We examined estimation of regression coefficients, confidence intervals, standard errors of the estimated regression coefficients (i.e., whether the estimated standard errors of the estimated regression coefficients approximate the variability of the sampling distribution of the estimated regression coefficients), and the R^2 of the fitted model. The article is structured as follows: in Section 2, we provide a brief summary of previous approaches to the required sample size for reliable regression modeling. In Section 3, we describe the design of Monte Carlo simulations and we report the results in Section 4. Finally, in Section 5, we summarize our results and place them in the context of the existing literature.

2. Previous approaches to sample size for adequate regression modeling

Green [5] used statistical power analysis to compare the performance of different rules-of-thumb for how many subjects were required for linear regression analysis. These rules-of-thumb can be classified into two different classes. The first class consists of those rules-of-thumb that specify a fixed sample size, regardless of the number of predictor variables in the regression model, whereas the second class consists of rules-of-thumb that incorporate the number of SPV. In the former class, Green described a rule, attributable to Marks, that specifies a minimum of 200 subjects for any regression analysis. In the latter class, Green described a rule, attributable to Tabachnick and Fidell, who suggested (with what Green described as some hesitancy) that although 20 SPV would be preferable, the minimum required SPV should be five. Another rule attributed to Harris is that the number of subjects should exceed the sum of 50 and the number of predictor variables. Schmidt [6] determined that, in a variety of settings, the minimum number of SPV lies in the range of 15 to 20. In a similar vein, Harrell [2] suggested that 10 SPV was the minimum required sample size for linear regression models to ensure accurate prediction in subsequent subjects.

In the epidemiologic and clinical literature, dichotomous outcomes and time-to-event outcomes that can be subject to right censoring are more common than are continuous outcomes [7]. Peduzzi et al. [8,9] conducted a series of simulation studies to examine the effective sample size that was required to allow logistic regression models and Cox proportional hazards models to be reliably estimated. Their focus was on the number of events, which defines the effective sample size. For dichotomous outcomes, the number of events (i.e., the effective sample size) was defined to be the smaller of the number of outcomes and the number of nonoutcomes. For survival outcomes, the number of events was the observed number of events that occurred during follow-up (i.e., the number of noncensored observations). They found that if the number of EPV was at least 10, then logistic regression models and Cox proportional hazards models could be estimated accurately, with an expected relative bias of less than 10%. In a more recent study, Vittinghoff and McCulloch [10] found that this requirement of 10 EPV could be relaxed in the context of confounder adjustment, to requiring five to nine EPV. Courvoisier et al. [11] suggested that there was not a single value of the number of EPV that would be sufficient for all contexts. Instead, the number of EPV would depend on the number of predictors, the anticipated magnitude of the regression coefficients, and the correlations between the predictor variables. Steyerberg suggested three thresholds for the number of EPV in the context of accurate prediction of binary outcomes: 10 for any prediction modeling, 20 would remove the need for shrinkage of estimated regression coefficients in prespecified models, and 50 EPV would be required to permit reliable variable selection from a set of candidate predictors (where the number of variables is equal to the total number of candidate variables considered) [12].

3. Methods

We used a modification of the methods used by Peduzzi et al. [8,9] to examine the number of events per predictor variable when estimating logistic regression or Cox proportional hazards regression models. To do so, we used a real data set and estimated two prespecified linear regression models using ordinary least squares. Subsets of increasing size were then randomly drawn from the data set, and the regression coefficients estimated in the full sample were used to simulate a continuous outcome for each subject in each randomly drawn data set. The prespecified linear regression models were then estimated in each subset, and the estimated regression coefficients were compared with those used in simulating outcomes.

3.1. Data sources

The Enhanced Feedback for Effective Cardiac Treatment (EFFECT) study was a cluster randomized trial intended to assess the effect of public reporting of hospital performance on an array of quality indicators for patients with cardiovascular disease in Ontario, Canada [13,14]. Detailed clinical data on patients hospitalized with heart failure between April 1, 1999, and March 31, 2001 (Phase 1), and April 1, 2004, and March 31, 2005 (Phase 2), at 103 hospitals in Ontario, Canada, were obtained by retrospective chart review. Data on patient demographics, vital signs and physical examination at presentation and at discharge, medical history, results of laboratory tests, and medications administered during the hospital stay were collected. The present study was restricted to those subjects discharged alive from hospital. Furthermore, subjects with missing data on baseline covariates necessary to estimate the prespecified regression model were excluded from the analyses, leaving 6,982 patients for analysis.

For this study, we considered two different continuous outcome variables: systolic blood pressure at discharge (measured in mm Hg) and heart rate discharge (measured in beats per minute). For the regression of discharge blood pressure, we considered twelve predictor variables: age, sex, presence of hypertension, ischemic heart failure (vs. nonischemic etiology), systolic blood pressure at hospital admission, and left ventricular ejection fraction (LVEF) [categorized as low ($\leq$20%) vs. medium (20% to 40%) vs. high ($>$40%), with high being the reference category]. We also included use of the following medications during hospital stay: angiotensin converting enzyme inhibitors or angiotensin receptor blockers, beta-blockers, digoxin, calcium channel antagonists, vasodilators, and diuretics. The regression of discharge heart rate modified the first regression model by replacing two predictor variables (systolic blood pressure at admission and history of hypertension) with two different predictor variables (heart rate at admission and history of chronic obstructive pulmonary disease). Summary statistics for the continuous outcome variable and the predictor variables are reported in Table 1.

3.2. Simulation methods

A linear regression model, estimated using ordinary least squares, was used to regress each continuous dependent variable on the 12 predictor variables described previously. Each model was estimated in the full sample described previously, consisting of 6,982 subjects. From each of the two fitted regression models, we extracted the estimated regression coefficients for each of the predictor

Table 1. Description of study sample and regression coefficients relating predictors to discharge blood pressure

Variable	Summary of distribution [median (IQR) or *N* (%)]	Regression coefficient for modeling discharge systolic blood pressure	Regression coefficient for modeling discharge heart rate
Discharge systolic blood pressure	122 (110–140)	Outcome variable	
Admission systolic blood pressure	147 (127–170)	0.258	
Discharge heart rate	76 (66–85)		Outcome variable
Admission heart rate	93 (76–111)		0.144
Age	76 (67–83)	0.055	−0.122
Female sex	3,486 (48.5)	1.162	1.225
Hypertension	4,197 (60.1)	5.832	
COPD	1,216 (17.4)		0.855
Ischemic heart disease (vs. nonischemic heart disease)	2,360 (33.8)	−0.470	−0.101
LVEF low	501 (7.2)	−6.123	1.969
LVEF medium	1,388 (19.9)	−2.584	0.509
Angiotensin-converting enzyme inhibitor/angiotensin receptor blocker	5,742 (82.2)	−1.802	−1.149
Beta-blocker	3,488 (50.0)	−0.358	−3.366
Calcium channel antagonist	2,376 (34.0)	3.486	−1.286
Digoxin	2,630 (37.7)	−0.281	−1.229
Diuretic	6,647 (95.2)	−0.773	0.438
Vasodilators	347 (5.0)	3.794	−0.855

Abbreviations: IQR, interquartile range; COPD, chronic obstructive pulmonary disease; LVEF, left ventricular ejection fraction.

variables and the estimate of the variance of the residual distribution (Table 1). For each of the two regression models, we denote the vector of estimated regression coefficients and the estimated variance of the residual distribution by β_{full} and σ^2_{full}, respectively. The R^2 statistics of the fitted models were 0.24 and 0.10, for the regression of systolic blood pressure at discharge and heart rate at discharge, respectively. The adjusted R^2 statistics of the two fitted models were also 0.24 and 0.10, respectively. The variance inflation factors for the predictor variables were all less than 1.1 in each of the two regression models, indicating that multicollinearity was not an issue. Each fitted regression model consisted of 12 predictor variables; however, LVEF was a three-level categorical variable that required two indicator variables for inclusion in the regression model. Thus, the estimated model used 13 degrees of freedom (df). The estimated regression coefficients will serve as the true population parameters that will be used to simulate outcomes in subsequent simulated samples.

For a given number of SPV, we sampled, with replacement, from the original data set of 6,982 subjects, a sample consisting of 13 × SPV subjects (because there were 13 df in each of the two regression models). For each subject in the randomly selected sample, we simulated a continuous outcome from the following model: $y_i \sim N(\mathrm{x}\beta_{\text{full}}, \sigma^2_{\text{full}})$. Thus, regression coefficients estimated in the full sample (β_{full}) were used as the true or population regression parameters that described the true linear relationship between the covariates and the continuous outcome. Then, in the randomly selected sample of size 13 × SPV, linear regression was used to regress the simulated outcome variable, Y, on the 13 predictor variables. This process was repeated 1,000,000 times per value of the number of SPV. This procedure was conducted for SPV ranging from 2 to 50, in increments of 1. This entire process was done twice. First, using the regression coefficients for the systolic blood pressure at discharge regression model and second using the regression coefficients for the heart rate at discharge regression model.

Note that the parameters used in the data-generating process were obtained from analyses conducted in the EFFECT study data set, which was a cluster randomized trial. Thus, it would be reasonable to anticipate a degree of within-hospital correlation of outcomes and subjects' baseline covariates. However, the estimated regression coefficients and the estimated residual variance were only used to simulate outcomes in our data-generating process. Outcomes were simulated for subjects in samples drawn from the EFFECT data set in a way that simulated independent observations. Therefore, analytic methods that assume independent observations can reasonably be applied to the simulated data sets.

3.3. Statistical analyses

For a given value of the number of SPV, in some of the simulated samples, the fitted regression model consisted of fewer than 13 covariates. This was due to low prevalence of some of the dichotomous predictor variables. In some of the randomly selected samples (particularly when the number of SPV was low), all the sampled subjects had the same value of one of the predictor variables. Because this variable was eliminated from the fitted regression model in the randomly selected sample, there was no regression coefficient associated with this predictor variable in this randomly selected sample. When this occurred, the randomly selected sample was discarded. Thus, some of the subsequent analyses used fewer than the number of randomly selected samples. We refer to the retained samples as the useable samples: samples in which the fitted regression model consisted of 13 predictor variables. For a given value of the number of SPV, we let N denote the number of useable samples.

From the linear regression model fit in each randomly selected sample, we extracted the estimated regression coefficients for the 13 predictor variables along with the estimated standard error of each regression coefficient. We estimated 95% confidence intervals for each of the estimated regression coefficients. In the ith randomly selected sample, we followed the approach of Peduzzi et al. [8,9] and determined the relative bias of the kth estimated regression coefficient as $\mathrm{RB}_{i,k} = 100 \times (\beta_{\text{estimate},k} - \beta_{\text{true},k})/\beta_{\text{true},k}$. Mean relative bias for the kth regression coefficient was determined as $\mathrm{RB}_k = (1/N)\sum_{i=1}^{N} \mathrm{RB}_{i,k}$, where N is the number of randomly selected samples that were useable (as described previously). For each of the 13 predictor variables, we determined the proportion of estimated 95% confidence intervals that contained the true value of the regression parameter. To determine whether the estimated standard errors of the regression coefficients correctly approximated the sampling variability of the estimated regression coefficients, we did the following for each predictor variable: first, we determined the mean standard error of the estimated regression coefficient across the N useable samples; second, we determined the standard deviation of the estimated regression coefficients across the N useable samples; third, we computed the ratio of the first quantity to the second quantity. If the estimated standard error is correctly approximating the sampling variability of the estimated regression coefficients, then this ratio should be close to one. Finally, we determined both the R^2 and the adjusted R^2 of the fitted regression model in each of the useable samples. We then calculated the mean R^2 and the mean adjusted R^2 across the useable samples for each value of the number of SPV. We determined the relative bias in the estimated (adjusted) R^2 by comparing this quantity to the (adjusted) R^2 of the model fit in the full sample of 6,982 subjects (R^2 = 0.24 and 0.10 for two regression models, respectively). All analyses were done separately for the two sets of simulations. The first set of simulations were those based on the regression of systolic blood pressure at discharge regression model, whereas the second set of simulations were those based on the regression of heart rate at discharge.

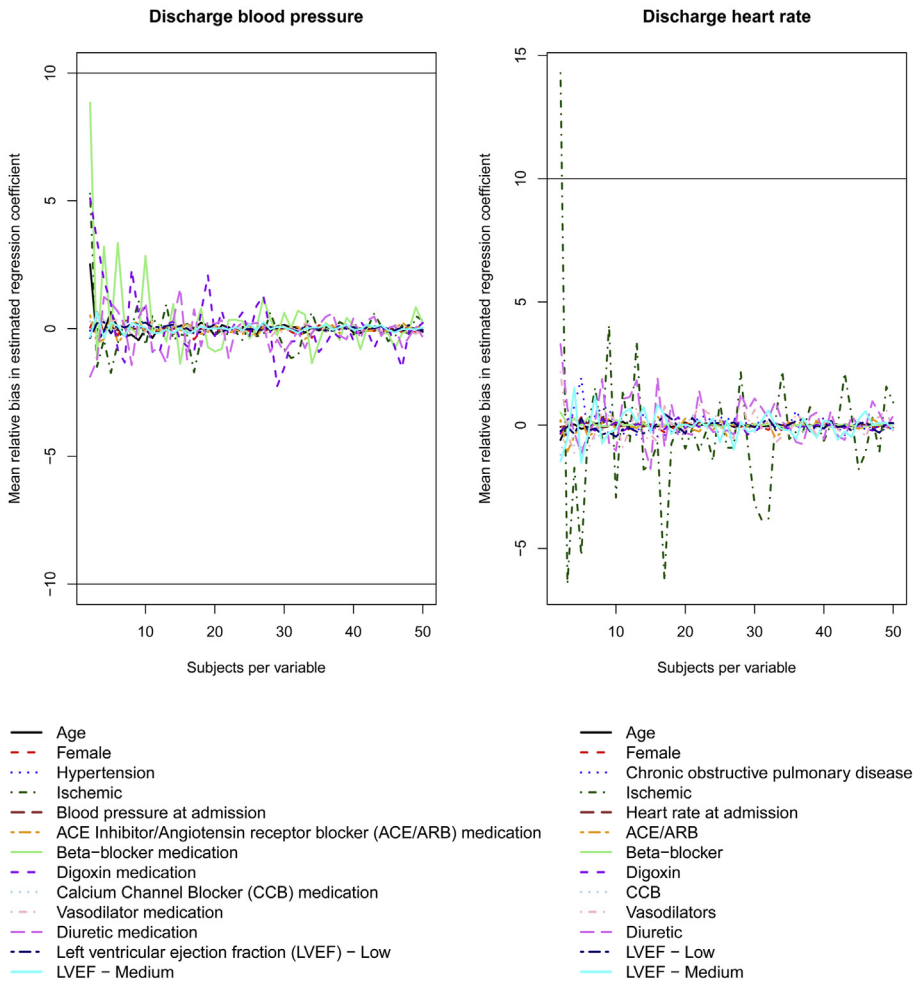

Fig. 1. Mean relative bias in estimating regression coefficients. (For interpretation of references to color in this figure, the reader is referred to the web version of this article.)

4. Results

When the simulations were based on discharge blood pressure being the outcome variable, the percentage of simulated samples that were not useable were 56%, 31%, 16%, 8%, and 4% when the number of SPV were equal to 2, 3, 4, 5, and 6, respectively. Once the number of SPV was at least 25, then all the 1,000,000 samples were useable. Comparable numbers were observed when the simulations were based on discharge heart rate as the outcome variable.

The relationship between mean relative bias and the number of SPV is reported in Fig. 1. The left panel describes results for the simulations based on the regression of systolic blood pressure at discharge, whereas the right panel describes the results for the simulations based on the regression of heart rate at discharge. On each panel, we have superimposed two horizontal lines denoting relative bias of ±10%, which we consider, admittedly arbitrarily, as denoting "limited bias." In the first set of simulations, based on blood pressure as the outcome variable, the mean relative bias was less than 10% for all variables and across all values of the number of SPV (range −2.2% to 8.8%). In the second set of simulations, based on heart rate as the outcome variable, the mean relative bias was less than 14.3% for all values of the number of SPV (and this value occurred for only one variable) (range −6.4% to 14.3%). Apart from this variable, the relative bias was less than 4% across the range of the number of SPV considered.

The relationship between the empirical coverage rates of 95% confidence intervals and SPV is reported in Fig. 2. On each of the two panels, we have superimposed curves denoting the lower and upper range of empirical coverage rates that would not be statistically significantly different from the advertised rate of 0.95, based on a standard normal-theory test and the number of useable samples (note that these superimposed curves are not horizontal due to not all the samples being useable). The empirical coverage rates were approximately equal to the advertised rates of 95% across all values of number of SPV. Even when the number of SPV was very low, empirical coverage rates were very close to 95%.

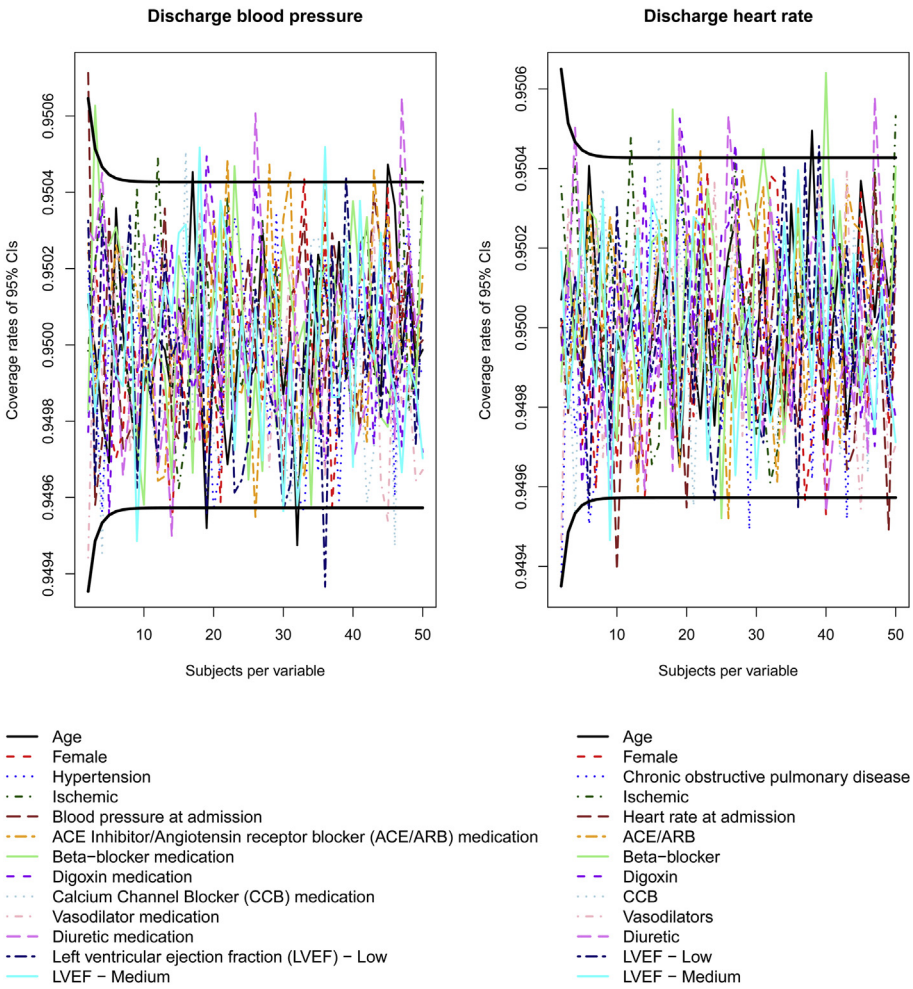

Fig. 2. Empirical coverage rates of 95% confidence intervals. (For interpretation of references to color in this figure, the reader is referred to the web version of this article.)

The relationship between the ratio of the mean estimated standard error for the regression coefficient and the standard deviation of the estimated regression coefficients and the number of SPV is described in Fig. 3. Once the number of SPV was greater than approximately 10, this ratio tended to be between 0.97 and 1.01, indicating that the estimated standard error closely approximated the sampling distribution of the estimated regression coefficient. Even when the number of SPV was very low, the ratio of these two quantities was still relatively close to one.

The relationship between the mean (adjusted) R^2 of the estimated linear model and the number of SPV is described in Fig. 4. Four curves are depicted in each of the two panels. Two, whose scale is depicted on the left axis, denote the mean R^2 and adjusted R^2 of the estimated linear model, whereas two, whose scale is depicted on the right axis, denote the relative bias of the R^2 and adjusted R^2 of estimated linear model when compared with the R^2 and adjusted R^2 of the model fit in the full sample (0.24 and 0.10 for the two simulation scenarios). Two horizontal lines have been superimposed on each panel: the first denoting the R^2 in the full sample and the second denoting a relative bias of 0%. When the number of SPV was 20, the relative bias in the estimated R^2 was 15% for the simulations based on blood pressure and 43% for the simulations based on heart rate. The number of SPV had to be greater than 30 before the relative bias in the estimated R^2 was less than 10% in the first set of simulations, whereas the relative bias still exceeded 10% even when the number of SPV was equal to 50 in the second set of simulations. In contrast, the bias in the adjusted R^2 was minimal, regardless of the number of SPV.

5. The effect of model R^2 and SPV on estimating R^2

The findings mentioned previously may be surprising to analysts in medical and epidemiologic research who are more familiar with generalized linear models and likelihood-based inference. We conducted a second set of simulations for two reasons. First, to confirm the low number of SPV required for accurate estimation of regression coefficients in a different setting. Second, in the simulations mentioned previously, the only quantity that was poorly

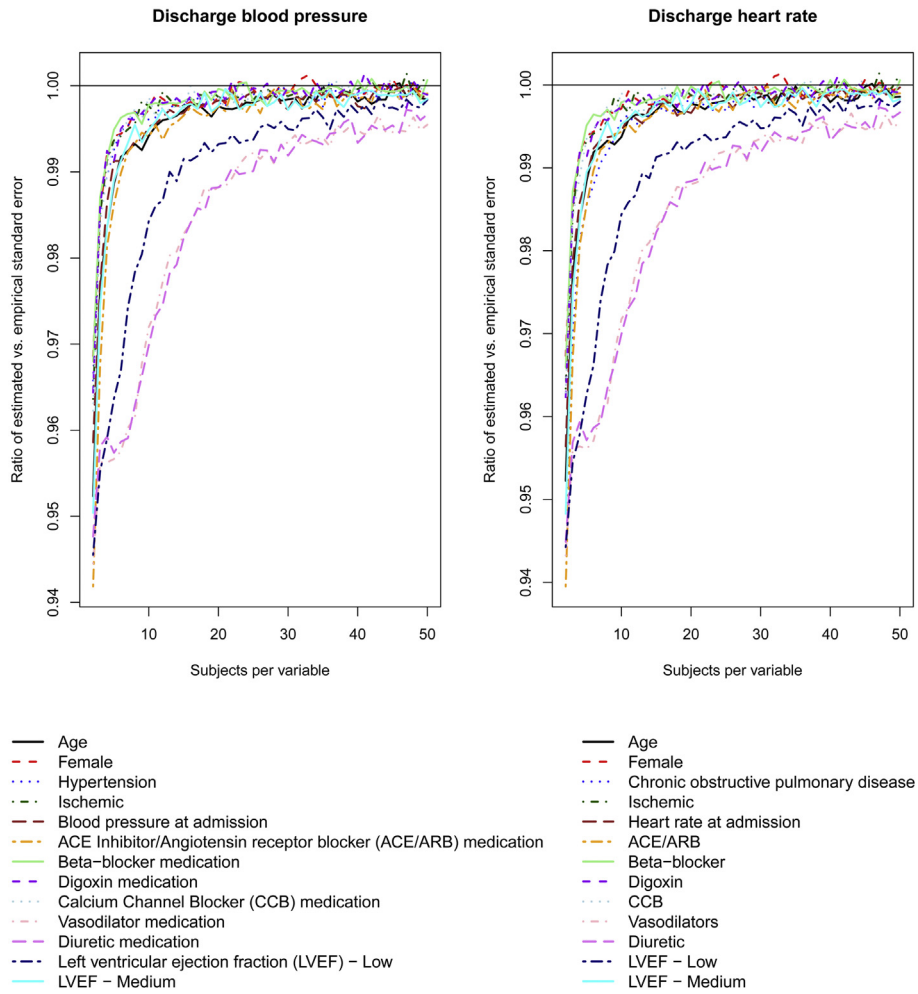

Fig. 3. Ratio of mean estimated standard error to standard deviation of estimated coefficients. (For interpretation of references to color in this figure, the reader is referred to the web version of this article.)

estimated in the presence of a low number of SPV was the model R^2. In this second set of simulations, we sought to examine this issue in greater depth.

5.1. Simulations—methods

For each subject, we simulated five predictor variables from independent standard normal random variables: $X_{ij} \sim N(0,1), j=1,\ldots,5$, and $i=1,\ldots,5 \times \text{SPV}$. We simulated data sets of size 5 × SPV. For each subject, a continuous outcome was simulated from the following model: $Y_i = 0.2X_{i1} + 0.4X_{i2} + 0.6X_{i3} + 0.8X_{i4} + 1X_{i5} + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$. The variance of the residual distribution (σ^2) was selected to induce the desired model R^2. Data were simulated with the following values for the model R^2: 0.05, 0.10, 0.25, and 0.50. For each of the scenarios, 100,000 simulated data sets were generated. The number of SPV was allowed to range from 2 to 4 in increments of one and then from 5 to 100 in increments of five.

In each simulated data set, a linear regression model was fit in which the continuous outcome was regressed on the five predictor variables. The R^2 and adjusted R^2 statistic of the fitted regression model was determined. The relative bias of each regression coefficient and of each of the R^2 statistics was determined, and the relative biases were averaged across the 100,000 simulated data sets in each of the four scenarios.

5.2. Simulations—results

Across the four scenarios defined by the four different values of the true population R^2 and across all values of the number of SPV, the mean relative biases for the estimated regression coefficients ranged from −2.5% to 4.0%. Thus, even when the number of SPV was very low (i.e., 2 or 3), the mean relative bias for the estimated regression coefficients was minimal. There is one panel in Fig. 5 for each of the true values of R^2 (0.05, 0.10, 0.25, and 0.50). In each of the four scenarios, the adjusted R^2 was essentially unbiased across the range of the number of SPV. However, the relationship between the number of SPV and the relative bias of the estimated R^2 statistic

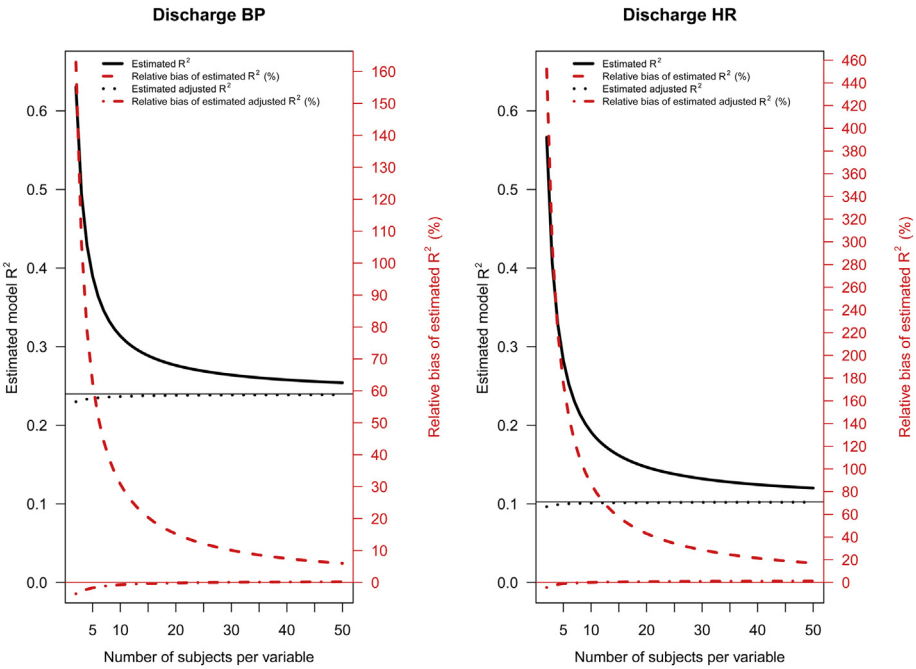

Fig. 4. R^2 of estimated regression model (first set of simulations).

depended on the magnitude of the true R^2 statistic. For a fixed value of the number of SPV, the relative bias of the estimated R^2 statistic decreased as the true population R^2 increased. Thus, a greater relative bias was observed when the true population R^2 was low, compared with when it was high.

6. Discussion

We conducted a series of simulations, based on an empirical analysis of existing data, to determine the minimum number of SPV to permit accurate estimation of linear regression models using ordinary least squares. We

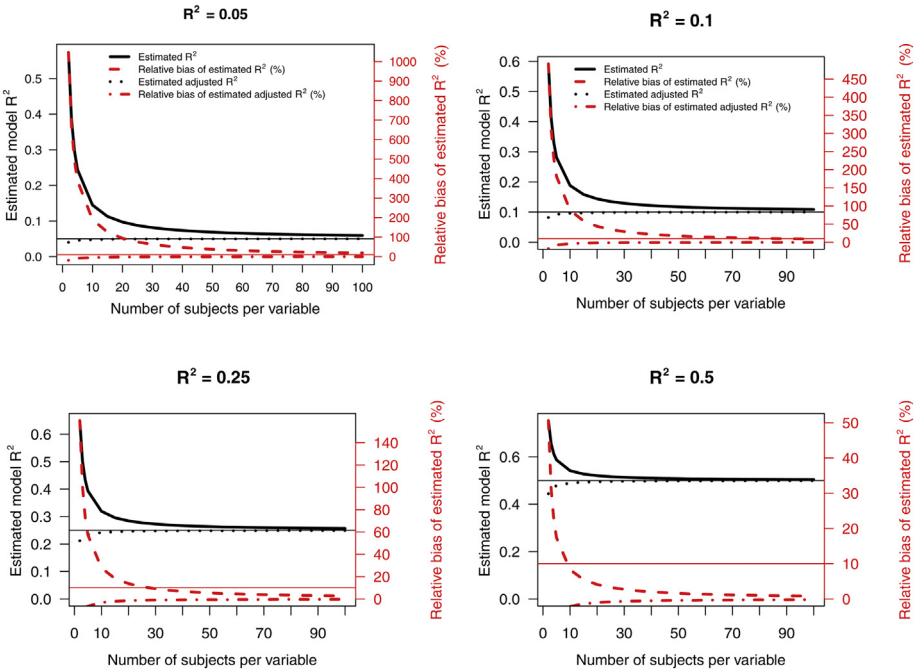

Fig. 5. R^2 of estimated regression model (second set of simulations).

observed that accurate estimation of regression coefficients was possible even when the number of SPV was as low as two. Furthermore, even when the number of SPV is equal to two, the estimated standard errors of the regression coefficients will accurately estimate the sampling variability of the estimated regression coefficients and the coverage rates of the estimated confidence intervals for the regression coefficients will have coverage rates as advertised. Minimal bias in the adjusted R^2 was observed over the entire range of number of SPV, whereas substantial bias was observed in the conventional R^2 when the number of SPV was low. Hence, the adjusted R^2 is clearly preferable over the conventional R^2 when quantifying the predictive ability of a linear regression model when the number of SPV is low.

These observations differ somewhat from those of Schmidt [6], who used simulations in the early 1970s to examine the effect of differing numbers of SPV when estimating linear regression models. The design of his simulations was informed by prior empirical studies in the psychological and education literature. In summarizing the findings by Schmidt, Green [5] suggested that the minimum number of SPV ranges from 15 to 25. As our study arrived at more liberal guidelines than those of Schmidt, it is important to note that the two studies used very different criteria with which to evaluate model performance. The present study focused on accurate estimation of regression coefficients, standard errors, and confidence intervals. In contrast, Schmidt compared the R^2 from an estimated linear regression model with the R^2 that arose from simply summing up the predictor variables (equivalent to assuming that all the regression coefficients were equal to one). Schmidt's thresholds on the number of SPV reflect the criterion necessary for the R^2 of the estimated linear model to exceed that of the R^2 obtained by summing up the predictor variables. We believe that our criteria for identifying the minimum number of SPV are of greater utility when the focus is on model estimation. Apart from the present study and that of Schmidt, there is a lack of guidelines derived from theoretical derivations or Monte Carlo simulations on the minimum SPV required for estimating linear regression models.

Green [5] evaluated different rules-of-thumb for the number of SPV in the context of statistical power analysis. He found that many such rules resulted in samples sizes that provided inadequate statistical power to detect meaningful effect sizes. The focus of the present study was on accuracy of estimation of a linear regression model, rather than on statistical hypothesis testing. Statistical power analysis is an appropriate lens through which to evaluate sample size when the focus is on detecting as statistically significant a regression coefficient of a given magnitude. However, not all linear regression models are estimated with the objective of testing hypotheses about the magnitude of certain regression coefficients. Instead, there is often an interest in the magnitude of estimated regression coefficients and/or their associated confidence intervals.

In the present study, we considered only the number of SPV for fitting a prespecified linear regression model. We did not consider the effect of variable reduction algorithms, such as backward or stepwise variable selection, on the required number of SPV. Prior research in the context of logistic regression found that backward variable selection resulted in increased bias compared with fitting a prespecified regression model [15]. Based on these observations, one would anticipate that a higher threshold for the number of SPV would be required if one were to account for data-based variable selection. Development of thresholds for the number of SPV that account for variable selection merits examination in subsequent research.

The primary limitation of the present study is that our conclusions are based on Monte Carlo simulations from a specific setting and thus are restricted to situations that resemble that described by our data-generating process. Our data-generating processes were based on patients hospitalized with congestive heart failure. Accordingly, the distribution and correlation of covariates were based on that of these patients. Thus, our simulations reflect the multivariate complexity of patients with an acute medical condition. Furthermore, our data included dichotomous variables that occurred very frequently, dichotomous variables that occurred for approximately half of the sample, and dichotomous variables that occurred rarely. In all, our simulations may well reflect a real-world complexity that is not observed in all studies that use Monte Carlo simulations. It should be noted that the covariates on which our simulations were based displayed a low degree of multicollinearity. It is possible that divergent results would be obtained in a setting with substantial multicollinearity. We would also note that we replicated our observations and conclusions in a second set of Monte Carlo simulations that were not based on actual clinical data and thus which may be more representative of data-generating processes in other studies that used Monte Carlo simulations.

Our primary set of simulations found that the number of SPV required to result in estimation of regression coefficients and standard errors that displayed minimal bias was lower than the corresponding number of EPV required in logistic regression models or for Cox proportional hazard regression models. This observation is, in one aspect, unsurprising, as pointed out by a reviewer of this article. In ordinary least squared regression, closed-form expressions exist for estimating these quantities [16]. Furthermore, these estimates will be unbiased provided the model assumptions are satisfied (as they were in our simulations). This is in contrast to estimation methods for logistic or survival analysis, in which likelihood-based methods are used, and methods to approximate the likelihood and determine the parameter values that maximize the likelihood must be used [8,9]. Thus, it was not surprising that we observed essentially unbiased estimation of regression coefficients (and their associated standard errors) even when the number of SPV was very low. In contrast to these observations, we also observed

that biased estimation of the model R^2 occurred when the number of SPV was low and that this bias was amplified when the true population R^2 was low compared with when it was high. This biased estimation of the model R^2 in the presence of a low number of SPV is a consequence of overfitting, in which the model that is fit has greater complexity than is permitted by the available data. When overfitting has occurred, the systematic component of the fitted model has incorporated idiosyncrasies of the sample in which it was estimated. This results in the fitted model appearing to explain a greater proportion of the variation than is explained by the population model. Our findings suggest that although estimation of regression coefficients are relatively unaffected by the number of SPV, the estimation of the model R^2 is susceptible to a low number of SPV.

In summary, prior research in the context of logistic regression and Cox proportional hazards models suggests that a minimum of 10 EPV is required for estimating these regression models [8,9]. Our findings suggest that in the context of linear regression estimated using ordinary least squares, a minimum of only two SPV is required for adequate estimation of regression coefficients.

Acknowledgments

The authors would like to thank three anonymous reviewers whose comments let to improvements in the manuscript.

References

[1] Steyerberg EW. Clinical prediction models. New York, NY: Springer-Verlag; 2009.
[2] Harrell FE Jr. Regression modeling strategies. New York, NY: Springer-Verlag; 2001.
[3] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Data mining, inference, and prediction. New York, NY: Springer-Verlag; 2001.
[4] Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med 2004;66(3):411−21.
[5] Green S. How many subjects does it take to do a regression analysis. Multivariate Behav Res 1991;26(3):499−510.
[6] Schmidt FL. The relative efficiency of regression and simple unit predictor weights in applied differential psychology. Educ Psychol Meas 1971;31(3):699−714.
[7] Austin PC, Manca A, Zwarenstein M, Juurlink DN, Stanbrook MB. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. J Clin Epidemiol 2010;63:142−53.
[8] Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 1995;48:1503−10.
[9] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373−9.
[10] Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 2007;165:710−8.
[11] Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol 2011;64:993−1000.
[12] Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making 2001;21:45−56.
[13] Tu JV, Donovan LR, Lee DS, Wang JT, Austin PC, Alter DA, et al. Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial. JAMA 2009;302:2330−7.
[14] Tu JV, Donovan LR, Lee DS, Austin PC, Ko DT, Wang JT, et al. Quality of cardiac care in Ontario. Toronto, Ontario: Institute for Clinical Evaluative Sciences; 2004: [Ref Type: Report].
[15] Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 1999;52:935−42.
[16] Myers RH. Classical and modern regression with applications. Belmont, California: Duxbury Press; 1990.