Susanne Rospleszcz, Silke Janitza, Anne-Laure Boulesteix

Categorical variables with many categories are preferentially selected in bootstrap-based model selection procedures for multivariable regression models

Technical Report Number 164, 2014
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Categorical variables with many categories are preferentially selected in bootstrap-based model selection procedures for multivariable regression models

Susanne Rospleszcz[*] Silke Janitza Anne-Laure Boulesteix

August 7, 2014

Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany.

Abstract

To perform model selection in the context of multivariable regression, automated variable selection procedures such as backward elimination are commonly employed. However, these procedures are known to be highly unstable. Their stability can be investigated using bootstrap-based procedures: the idea is to perform model selection on a high number of bootstrap samples successively and to examine the obtained models, for instance in terms of the inclusion of specific predictor variables. However, from the literature such bootstrap-based procedures are known to yield misleading results in some cases. In this paper we aim to thoroughly investigate a particular important facet of these problems. More precisely, we assess the behaviour of regression models—with automated variable selection procedure based on the likelihood ratio test—fitted on bootstrap samples drawn with replacement and on subsamples drawn without replacement with respect to the number and type of included predictor variables. Our study includes both extensive simulations and a real data example from the NHANES study. The results indicate that models derived from bootstrap samples include more predictor variables than models fitted on original samples and that categorical predictor variables with many categories are preferentially selected over categorical predictor variables with fewer categories and over metric predictor variables. We conclude that using bootstrap samples to select variables for multivariable regression models may lead to overly complex models with a preferential selection of categorical predictor variables with many categories. We suggest the use of subsamples instead of bootstrap samples to bypass these drawbacks.

[*] Corresponding author. Email: susanne.rospleszcz@lmu.de.

1 Introduction

In biometrical applications, multivariable regression is commonly used to model the association between an outcome and candidate predictor variables or to provide a simple interpretable prediction model for the outcome. Selection of the appropriate predictor variables is of key importance. Neglecting to include a predictor variable with a strong effect on the outcome obviously leads to a suboptimal model, while incorporating too many variables can result in serious overfitting and negatively affect model interpretability as well as prediction accuracy. To address this issue, automated selection procedures such as stepwise or backward elimination have been suggested. These methods, based either on a test procedure such as the likelihood ratio test (LR-test) or another selection criterion such as the Akaike Information Criterion (AIC), are now widely employed in biometrical applications. However, they are known to be highly unstable in the sense that a small change in the data might lead to a substantially different model (Sauerbrei et al.; 2011).

In the two last decades, the use of resampling methods has been propagated for stability investigations in this context (Altman and Andersen; 1989; Royston and Sauerbrei; 2003). Non-parametric bootstrapping, i.e. drawing with replacement from the original sample and thereby generating new data sets of the same size as the original one, is a method widely used for this purpose. Bootstrap-based procedures may allow approximate inference such as, for example, the derivation of confidence intervals in situations where parametric procedures do not exist or are computationally unfeasible.

The use of bootstrap-based methods has been proposed for model stability investigations in the context of automated selection methods such as, for instance, backward elimination (Sauerbrei and Schumacher; 1992; Austin and Tu; 2004). The considered variable selection procedure is applied to a large number of bootstrap samples successively as if they were the original sample. The inclusion frequencies over all bootstrap samples can then be calculated for each predictor variable. This percentage of inclusion can be used to assess the importance of the respective variable (Gong; 1982; Chen and George; 1985; Sauerbrei and Schumacher; 1992; Sauerbrei et al.; 2011). This method has been employed in numerous biomedical applications to investigate the stability of models (Halabi et al.; 2003; Ette; 1997) as well as for validation (Motzer et al.; 1999) and for the selection of predictor variables for prognostic models (Bruneel et al.; 2010; Heymans et al.; 2007).

However, it has been reported in various contexts within the biometrical field (Bollen and Stine; 1992; Wagenmakers et al.; 2004; Binder and Schumacher; 2008) that the distributions of test statistics derived from bootstrap samples differ from the distributions of test statistics derived from original samples; see Janitza et al. (2014) for a recent overview and theoretical considerations substantiating these ideas. Yet, to date the practical impact of these problems in bootstrap-based model selection procedures for multivariable regression is largely unknown.

In this context, the objective of the present paper is to compare regression models selected through a backward elimination procedure using the p-value of the LR-test as an elimination criterion, applied to original samples, bootstrap samples drawn with replacement and subsamples drawn without replacement, respectively. Note that similar results are expected using other variants of variable selection, as suggested both by theory (Janitza et al.; 2014) and preliminary empirical results. We aim to identify differences between models derived from the different types of samples with regard to the number and type of included variables. For this purpose we perform extensive simulation studies under different settings and illustrate our findings using real data from the 2007 cycle of the NHANES study (National Center for Health Statistics; 2012). All calculations are carried out using `R 3.0.1` (R Core Team; 2012).

2 Theoretical rationale

2.1 The bootstrap method

Let $x_i = (x_{i1}, \ldots, x_{ip})^\top$ be independent realizations of p predictor variables for an individual i and let y_i be the corresponding observed value of the response variable. The n realizations $z_i := (x_i, y_i), i = 1, \ldots, n$ are assumed to have been drawn from an unknown joint distribution F. The resulting data set $Z = (z_1, \ldots, z_n)^\top$ is referred to as the *original sample* throughout this paper. A bootstrap sample $Z^* = (z_1^*, \dots, z_n^*)^\top$ is generated by drawing n observations with replacement from the original sample Z and can thus contain duplicated observations. The estimate $\hat{\theta}$ computed from the original sample and the bootstrap estimate $\hat{\theta}^*$ computed from the bootstrap sample Z^* are estimators for θ, the parameter of interest. By repeating the bootstrap sampling and estimation of the parameter of interest B times, one obtains bootstrap samples $Z^{*1}, \dots, Z^{*B}$ and corresponding bootstrap estimates $\hat{\theta}^{*1}, \dots, \hat{\theta}^{*B}$. The bootstrap estimates can be used, for example, to derive confidence intervals for θ or to approximate the whole distribution of $\hat{\theta}$. However, this procedure fails if the considered "parameter" of interest is a test statistic or a p-value. This is briefly motivated in the next section for the LR-test statistic. Readers are referred to Janitza et al. (2014) for more details.

2.2 The LR-test statistic

Let $\beta = (\beta_1, \dots, \beta_p)^\top$ be the vector of length p that reflects the true unknown effects of $x_{i1}, \dots, x_{ip}$ on the response variable via the so-called linear predictor $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. One can test a hypothesis of the form

$$H_0 : C\beta = \zeta \text{ against } H_1 : C\beta \neq \zeta \tag{1}$$

where C is a matrix of rank $s \leq p$ and ζ is a fixed vector (see e.g. Tutz (2012)). In the simplest case one can test if the j-th parameter in β is equal to zero, i.e. $H_0 : \beta_j = 0$. In this case the matrix C reduces to a row matrix $C = (0, \dots, 0, 1, 0, \dots, 0)$ with the j-th entry taking the value 1 and all other entries taking value 0. In this case, the vector ζ takes the value 0.

If L_1 denotes the likelihood of the model where β is estimated without any constraint and L_0 denotes the likelihood of the (sub)model with the constraint that $C\beta = \zeta$, the LR-test statistic for testing the hypothesis (1) is given by

$$\Lambda = -2(log(L_0) - log(L_1)).$$

According to Wilks' theorem (Wilks; 1938) the LR-test statistic asymptotically follows a $\chi^2(s)$-distribution with $s = \text{rank}(C)$ degrees of freedom. The asymptotic expectation and variance of the LR-test statistic Λ are given by

$$\text{aE}[\Lambda] = s + \delta, \quad \text{aVar}[\Lambda] = 2s + 4\delta, \tag{2}$$

respectively, where δ represents the noncentrality parameter of the distribution, which reflects the extent to which the null hypothesis is false. If the null hypothesis holds true, the noncentrality parameter δ equals 0 and the asymptotic expectation and variance of Λ with respect to the underlying general population are given by $\text{aE}[\Lambda] = s$ and $\text{aVar}[\Lambda] = 2s$.

2.3 Computing the LR-test statistic for bootstrap samples

Bollen and Stine (1992) report that the distribution of the test statistic Λ^* computed from a bootstrap sample Z^* deviates from the distribution of the test statistic Λ computed from the original sample, leading to an increased α-level when deriving the p-value from the comparison of Λ^* to a central $\chi^2(s)$- distribution. This is due to the fact that for the bootstrap sample the null hypothesis does not hold, as the sample does not come from the theoretical distribution F but from the empirical distribution $\hat{F}$. Drawing bootstrap samples from the original sample amounts to drawing from a population where H_0 is not true (Bollen and Stine; 1992).

Following Bollen and Stine (1992), the discrepancy between the expected values and variances of Λ and Λ^* depends on the degrees of freedom of the LR-test, with increasing degrees of freedom leading to a stronger discrepancy between the distributions.

For the LR-test statistic the degrees of freedom are given by the difference of the degrees of freedom of the two models that are compared by the test. The degrees of freedom of a model are the number of model parameters that can be varied freely. A model that includes p parameters, where we assume for the moment that one parameter describes the effect of a metric or a binary predictor variable, plus an intercept has $(n - p - 1)$ degrees of freedom, with n denoting the sample size. In the following we will use the term *categorical predictor variable* to denote a nominal or ordinal predictor variable with at least three categories, while predictor variables with two categories are simply termed binary. Categorical predictor variables are usually dummy-coded such that a categorical variable with $k > 2$ categories is represented by $k - 1$ binary variables. Usually one of the k levels of the categorical variable is considered as a reference category. For each of the remaining categories, a binary variable is created that indicates whether the predictor variable takes this level or not. Hence, $k - 1$ binary variables are needed to represent the information of k possible categories in this coding scheme. Consequently, if a categorical predictor variable with k categories enters a linear regression model, it consumes $k - 1$ degrees of freedom.

Accordingly, when testing the submodel which includes only the intercept against a model including, for instance, a 3-category predictor variable, the null distribution has $s = 2$ degrees of freedom. If we test the submodel against the model including a 7-category predictor variable we have $s = 6$ degrees of freedom. We expect a higher false positive rate for LR tests performed on bootstrap samples for a test on the inclusion of a 7-category predictor variable than for a test on the inclusion of a 3-category predictor variable. That is because, according to Bollen and Stine (1992), the discrepancy between the distributions of Λ and Λ^* rises with increasing degrees of freedom. The practical impact on the results of model selection procedures in the context of multivariable regression, however, is completely unknown to date. The purpose of this paper is to systematically investigate these consequences in a quantitative manner.

Before considering multivariable regression models in Sections 3 and 4, however, we first perform a simple simulation to examine the distribution of the LR-test statistics under H_0 with varying degrees of freedom and compare the original distributions to the distributions on bootstrap samples for different types of predictors. We generate $n = 1000$ independent observations of a standard normally distributed response variable Y and a binary predictor variable, as well as categorical predictor variables with 3, 4, 5, 6 and 7 categories which are all independent of Y (i.e., the null hypothesis is true) and independent of each other. For each predictor variable in turn, a LR-test is conducted for the linear submodel containing only an intercept (null hypothesis) and the linear model containing an intercept and the considered binary or categorical variable (alternative hypothesis). Subsequently, a bootstrap sample is drawn from the simulated data set and the LR-tests described above are conducted again based on this bootstrap sample. The whole procedure (data generation, bootstrapping and performing LR-tests) is repeated 10000 times to obtain reliable approximations of the distributions of interest, resulting in a total of 10000 LR-test statistics computed from original samples and 10000 LR-test statistics computed from bootstrap samples for each type of predictor variable. Figure 1 shows the corresponding empirical distributions of the LR-test statistic for original samples (solid lines) and bootstrap samples (dashed lines).

These results corroborate the considerations outlined above. The distribution of the LR-test statistic on bootstrap samples is increasingly distorted with increasing degrees of freedom, i.e., in our case with an increasing number of categories. To assess the impact of this distortion on type I error, we calculate the empirical α-levels (i.e., the percentage of LR-test statistics out of the 10000 samples that exceed the critical value of the χ^2-distribution with the respective degrees of freedom) from the derived empirical distributions. Table 1 shows to which extent the discrepancy between the distributions of Λ and Λ^* leads to a type I error for increasing degrees of freedom, when the LR-tests are performed at the 5% level. For the categorical predictor variable with 3 categories, α amounts to 22.3% and for the categorical predictor variable with 7 categories, α increases to 39.3%. The latter means that H_0 is falsely rejected in almost 40% of the cases when performing the LR-test on a bootstrap sample under the null hypothesis at a nominal level of 5%.

2.4 Bootstrap inclusion frequencies

In automated iterative variable selection methods such as backward elimination, forward selection or stepwise selection, the LR-test is often used at each step to select variables to be included in or excluded from the model (see e.g. Sauerbrei et al. (2011)). In this paper we focus on the backward elimination procedure. It consists of starting with the full model, i.e, that which includes all predictor variables, and eliminating at each step the variable yielding the largest p-value from the LR-test. This process is iterated until no predictor yields a p-value larger than the threshold fixed by the user (5% in our paper). This procedure outputs a submodel that most often includes

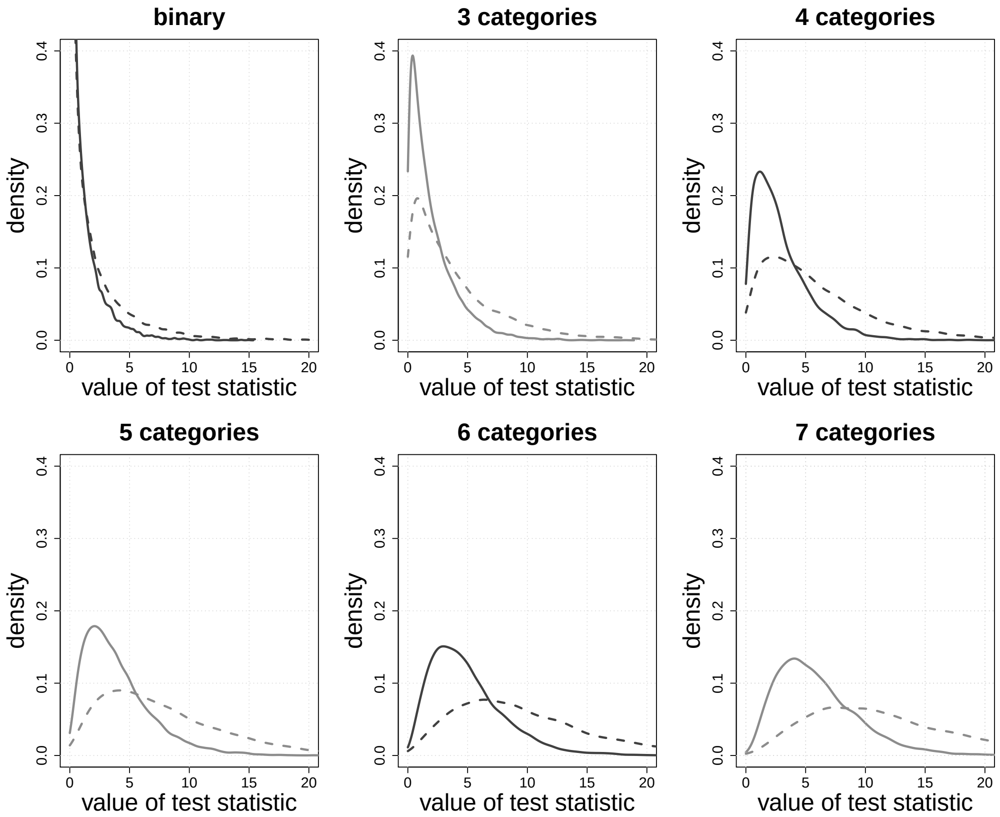


Figure 1: Distribution of LR-test statistic for the comparison of the linear regression model with only the intercept to the model including intercept and one categorical predictor variable, for original samples (solid lines) and bootstrap samples (dashed lines).

fewer predictor variables than the full model. Note that throughout this paper we always assume that a categorical predictor variable is included or eliminated from the model as a whole. This is done by performing a global test that tests if the effect of at least one of the dummy variables significantly differs from 0.

If the procedure is repeated for a large number of bootstrap samples, one can compute the so-called bootstrap inclusion frequency for each predictor variable, i.e., the proportion of bootstrap samples for which this variable is included in the submodel selected by the automated procedure. Readers are referred to Sauerbrei et al. (2011) for more details on this approach.

The distortion between the null-distribution of the LR-test statistic for original samples and for bootstrap samples displayed in Figure 1 is expected to affect the results of the automated selection procedure outlined above. More precisely, (i) the models selected based on bootstrap samples are expected to be more complex, i.e. include more predictor variables, and (ii) under the null hypothesis that no predictor variable has an effect on the response, predictor variables with many categories are expected to be preferentially selected over binary or metric predictor variables.

In Section 3 we present an extensive simulation study to provide empirical evidence for these conjectures; we further assess their practical impact on multivariable model selection based on bootstrap samples in a quantitative manner. More precisely, we simulate data under different settings and perform backward elimination based on the original samples, bootstrap samples (drawing with replacement) and subsamples (drawing without replacement) successively.

Table 1: Empirical α-levels when performing the LR-test on 10000 original samples and bootstrap samples.

	critical value of central χ^2-distribution for $\alpha = 5\%$					
	df=1 3.84	df=2 5.99	df=3 7.81	df=4 9.49	df=5 11.07	df=6 12.59
corresponding α in original samples (in %)	5.46	5.06	4.83	5.23	4.85	4.87
corresponding α in bootstrap samples (in %)	16.91	22.31	27.18	31.97	35.33	39.28

3 Simulation design

For each simulation setting a total of 5000 original *i. i. d.* samples are drawn from the considered distribution. Subsequently, one bootstrap sample is drawn with replacement and one subsample including approximately 63.2% of the original sample is drawn without replacement from each of these original samples. The value 63.2% is chosen to obtain the same number of unique data points in the subsample and (on average) in the bootstrap sample. A regression model is fitted to each of the original samples, bootstrap samples and subsamples by backward elimination, based on the LR-test with a threshold of 5%. After backward elimination, the percentage of resulting models (out of 5000) that include a specific predictor variable and the distribution of the number of selected predictor variables (over 5000 models) are derived successively for the original samples, bootstrap samples and subsamples.

In the remainder of this section we describe all simulation settings in detail. In our main simulation we set up a series of linear regression models which include several uncorrelated predictor variables with varying effects (no effect, strong effect, moderate effect). This simulation setting will be examined in detail and is described in subsection 3.1. In three subsequent analyses we explore if results are comparable when (i) predictor variables are mutually correlated (subsection 3.2), (ii) sample size is reduced (subsection 3.3) and (iii) the response variable is a censored time-to-event in a multivariable Cox proportional hazard model (Cox; 1972) (subsection 3.4).

3.1 Main simulation

We simulate and analyze data from three underlying linear regression models which differ in the size of predictor effects. In our main simulation design each sample comprises $n = 1000$ observations, which corresponds to "asymptotic settings", and a total of 17 mutually independent predictor variables. The 17 predictor variables consist of five metric variables $X_1, \dots, X_5$, two binary predictors X_6, X_7, and 10 categorical predictors $X_8, \dots, X_{17}$. Metric variables are sampled from a standard normal distribution, binary predictor variables are sampled from a Bernoulli distribution with probability $p = 0.5$ and categorical predictors are sampled from a multinomial distribution with values in $\{1, \dots, k\}$—where k denotes the number of categories of the predictor variable—and equal probabilities for all k categories. We have two predictor variables for each $k \in \{3, 4, 5, 6, 7\}$, yielding a total of 10 categorical predictor variables. Categorical variables are dummy-coded using the first category as a reference category. For example, the 3-category predictor variable X_8 is coded as two dummy variables X_{8_2} and X_{8_3} which take value 1 if $X_8 = 2$ or $X_8 = 3$, respectively, and 0 otherwise.

The regression coefficients reflecting the effects of the predictor variables are shown in Table 2. The null model, in which no predictor variable has any effect on the response, is denoted by Null-LM-n1000. For the first model with non-zero effects, which is denoted by LM1-n1000, the effects are stronger than for the second model: an effect of 0.2 is assumed for three of the five metric variables. For the informative 4-category predictor variable X_{10}, the coefficients of two of four categories are set to 0.2. The coefficients for three of the six categories of the informative 6-category predictor variable X_{14} are set to 0.1. The same predictor variables that have an effect in model LM1-n1000 also have an effect in the second model with non-zero effects, LM2-n1000, with the difference that effects are now smaller for the categorical predictor variables.

Table 2: Effects of (uncorrelated) predictor variables for the linear models Null-LM-n1000, LM1-n1000 and LM2-n1000 with $n = 1000$. For categorical predictor variables with k categories, effects for the corresponding $(k-1)$ dummy variables are shown.

Predictor	Scale	Effect(s)		
		Null-LM-n1000	LM1-n1000	LM2-n1000
X_1	metric	0	0.2	0.01
X_2	metric	0	0.2	0.02
X_3	metric	0	0.2	0.03
X_4	metric	0	0	0
X_5	metric	0	0	0
X_6	binary	0	0	0
X_7	binary	0	0	0
	categorical with			
X_8	3 categories	0, 0	0, 0	0, 0
X_9	3 categories	0, 0	0, 0	0, 0
X_{10}	4 categories	0, 0, 0	0.2, 0.2 ,0	0.08, 0.08, 0
X_{11}	4 categories	0, 0, 0	0, 0, 0	0, 0, 0
X_{12}	5 categories	0, 0, 0, 0	0, 0, 0, 0	0, 0, 0, 0
X_{13}	5 categories	0, 0, 0, 0	0, 0, 0, 0	0, 0, 0, 0
X_{14}	6 categories	0, 0, 0, 0, 0	0.1, 0.1, 0.1, 0, 0	0.04, 0.04, 0.04, 0, 0
X_{15}	6 categories	0, 0, 0, 0, 0	0, 0, 0, 0, 0	0, 0, 0, 0, 0
X_{16}	7 categories	0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0
X_{17}	7 categories	0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0	0, 0, 0, 0, 0, 0

With $x = (x_1, \ldots, x_7, x_{8_2}, x_{8_3}, \ldots, x_{17_7})^\top$ now also including dummy-coded variables derived from the multicategorical predictor variables, the response variable is generated according to the linear model

$$y = x^\top \beta + \epsilon, \tag{3}$$

where β is the vector of predictor effects given in Table 2 and $\epsilon \sim N(0,1)$ is the error term.

3.2 Variant 1: Correlated predictors

Additional simulations are conducted with mutually dependent predictor variables. More precisely, the simulated data sets now include five pairs of predictor variables: one binary variable and one categorical variable with five categories. The four first pairs of predictor variables $(X_1, X_6), (X_2, X_7), (X_3, X_8), (X_4, X_9)$ are pairwise dependent, while X_5 and X_{10} are independent.

Pairs of mutually dependent binary and 5-category variables are generated using the `ordsample` function from the R-package `GenOrd` (Barbiero and Ferrari; 2012). This function first generates correlated metric variables using the multivariate standard normal distribution with an appropriate covariance matrix and subsequently categorizes the resulting metric variables into categorical variables with an ordering in the categories. In our case we have two correlated variables where one metric variable is categorized into 5 ordered categories and the other into 2 categories. The Spearman correlation between the binary and the ordered categorical variable is used in the following to quantify the strength of the correlation between the two variables.

Effect sizes for the underlying models are displayed in Table 3. The effect sizes for the model with informative predictor variables and with the same Spearman correlation ρ among all correlated variable pairs are set up in such a way that for the first correlated pair of predictor variables (X_1, X_6), both variables have an effect. For the second pair (X_2, X_7) only the binary variable X_2 has an effect and for the third pair (X_3, X_8) only the categorical variable X_8 has an effect. For the fourth pair (X_4, X_9), neither of the variables has an effect. The fifth pair (X_5, X_{10}) consists of two uncorrelated predictor variables with non-zero effects: the effect of the binary variable is set to 0.1 and two of the five categories of the 5-category variable have a regression coefficient of 0.1; see Table 3. We set the correlation coefficient ρ to $\rho = 0.3, 0.5, 0.7$, successively. The term corr(ρ)-Null-LM-n1000 (with ρ taking values $0.3, 0.5, 0.7$) refers to the linear model, in which no predictor variable has an effect, while corr(ρ)-LM-n1000 refers to the models with informative predictor variables.

The response variable is generated according to Eq. (3) with $x = (x_1, \ldots, x_5, x_{6_2}, x_{6_3}, x_{6_4}, x_{6_5}, x_{7_2}, \ldots, x_{10_5})^\top$.

Table 3: Effects of correlated predictor variables for the linear regression models with $n = 1000$ and Spearman correlation of $\rho \in \{0.3, 0.5, 0.7\}$. For categorical predictor variables with 5 categories regression coefficients for 4 dummy variables are shown.

Predictor	Scale	correlated to	Effect(s)	
			corr(ρ)-Null-LM-n1000	corr(ρ)-LM-n1000
X_1	binary	X_6	0	0.1
X_2	binary	X_7	0	0.1
X_3	binary	X_8	0	0
X_4	binary	X_9	0	0
X_5	binary	-	0	0.1
	categorical with			
X_6	5 categories	X_1	0, 0, 0, 0	0.1, 0.1, 0, 0
X_7	5 categories	X_2	0, 0, 0, 0	0, 0, 0, 0
X_8	5 categories	X_3	0, 0, 0, 0	0.1, 0.1, 0, 0
X_9	5 categories	X_4	0, 0, 0, 0	0, 0, 0, 0
X_{10}	5 categories	-	0, 0, 0, 0	0.1, 0.1, 0, 0

Table 4: Effects of the three predictor variables for the linear regression models Null-LM-n100, LM1-n100 and LM2-n100, with $n = 100$.

Predictor	Scale	Effect(s)		
		Null-LM-n100	LM1-n100	LM2-n100
X_1	binary	0	0.1	0.05
	categorical with			
X_2	4 categories	0	0.1, 0.1, 0	0.05, 0.05, 0
X_3	7 categories	0	0.1, 0.1, 0, 0, 0, 0	0.05, 0.05, 0, 0, 0, 0

3.3 Variant 2: Smaller sample size

We now consider three linear regression settings with independent predictor variables but a smaller sample size of $n = 100$. The three considered settings differ in the effects of the predictor variables. Considering the smaller sample size, the number of predictor variables is set to only three instead of 17. The binary variable X_1 is drawn from a Bernoulli distribution with $p = 0.5$. The categorical predictor variables are drawn from multinomial distributions with equal probabilities for all categories and values in $\{1, \ldots, 4\}$ for the 4-category variable X_2 and values in $\{1, \ldots, 7\}$ for the 7-category variable X_3.

The effects of the predictor variables are presented in Table 4. The first model (termed Null-LM-n100) is a null model, in which none of the three predictor variables has an effect. In the other two models, LM1-n100 and LM2-n100, all three predictor variables have an effect which is larger in LM1-n100 than in LM2-n100 for all three variables. The response variable is generated according to Eq. (3) with $x = (x_1, x_{2_2}, x_{2_3}, x_{2_4}, x_{3_2}, x_{3_3}, \ldots, x_{3_7})^\top$.

3.4 Variant 3: Survival response

In an additional study we again consider uncorrelated predictor variables with the same effects as in the main simulation setting (presented in Table 2) and sample size $n = 1000$, but with a censored time-to-event as the response variable. Times-to-event are simulated according to the Cox proportional hazard model where the hazard is described by

$$h(t|x) = h_0(t) \exp(x^\top \beta), \tag{4}$$

where $h_0(t)$ is the baseline hazard and β is the vector of regression coefficients representing the effects of the predictor variables $X_1, \ldots, X_{17}$ (some of them dummy-coded) on time-to-event; see Table 2. The survival function is given by

$$S(t|x) = \exp\left(-\int_0^t h(u|x)du\right). \tag{5}$$

For the generation of times-to-event a constant baseline hazard of 0.025 is assumed. The censoring process is considered to be independent of the predictor variables and a constant censoring hazard of 0.05 is chosen. As with the linear regression setting described within the main simulation design, the settings with a survival response are referred to as Null-Cox-n1000, Cox1-n1000 and Cox2-n1000 (see Table 2). Multivariable Cox regression is used in place of multivariable linear regression for the analyses, also including backward elimination based on the LR-test with a threshold of 5%.

4 Results

4.1 Simulation studies

In this section we first describe the results for the main simulation design (linear regression settings with uncorrelated predictor variables and a large sample size of $n = 1000$) and point out the differences observed in the additional studies with correlations between binary and categorical predictors (Variant 1), a reduced sample size (Variant 2) and a time-to-event as response (Variant 3).

4.1.1 Main simulation design

Model complexity
First we look at the model complexity in terms of the total number of predictor variables being included in a model. Figure 2 depicts the total numbers of included predictor variables for the three analyzed models. For all models, the range of the number of included predictor variables is wider for bootstrap samples than for original samples or subsamples. It is clear that on average with the bootstrap more predictors are included in a model. For Null-LM-n1000, many of the models of original samples and subsamples (about 40% of models) yield the true model, which includes no predictor variable at all. In contrast, very few models on bootstrap samples yield the true model. Many of the bootstrap models (about 20%) include four predictor variables. There are even bootstrap models that include up to 12 of the 17 predictor variables. For original samples, in contrast, not more than 5 predictor variables are ever chosen in a model.
A similar behavior of the bootstrap towards overly complex models – including too many predictor variables – is observed for models where some of the predictor variables are associated with the response. For LM1-n1000, the true model includes 5 predictor variables. Around 30% of the models derived from original samples or subsamples include exactly that number of predictors and are thus of the same complexity. Another 52% of the models from original samples and 60% of the models from subsamples include fewer predictor variables. On bootstrap samples in contrast, only about 10% of the models include 5 or fewer predictor variables. The remaining 90% of the models include 6 to 15, with the most common number being 7. The tendency toward too complex models for the bootstrap can also be observed for LM2-n1000 (lower panel in Figure 2). These findings are a direct consequence of the theoretical considerations on the increased α level when performing tests on a bootstrap sample, leading to a higher inclusion of predictor variables that are actually not of importance.

Type of predictor variables in a model
Figure 3 displays the results for Null-LM-n1000, the model in which none of the predictor variables has an effect. Depicted are the mean inclusion frequencies for all predictors; the "metric" inclusion frequency is averaged over the five metric predictors, and the "binary" is the average of the two binary predictors. There is a systematic preference for categorical predictor variables on bootstrap samples when compared to metric or binary variables. In addition, among the categorical predictor variables there is a clear preference for those with many categories. Inclusion frequencies rise distinctly with the predictor variable's number of categories. While metric and binary predictor variables are included with a frequency of around 17.5%, categorical predictor variables with three categories are included in 23.4% of the models. This percentage increases approximately linearly with an increasing number of categories of the categorical variable. A categorical predictor variable with seven categories is included in 43.3% of the models. Notably, this effect cannot be observed for original samples or subsamples. The inclusion frequency in models derived on original samples and subsamples averages about 5% for every predictor variable and does not systematically vary

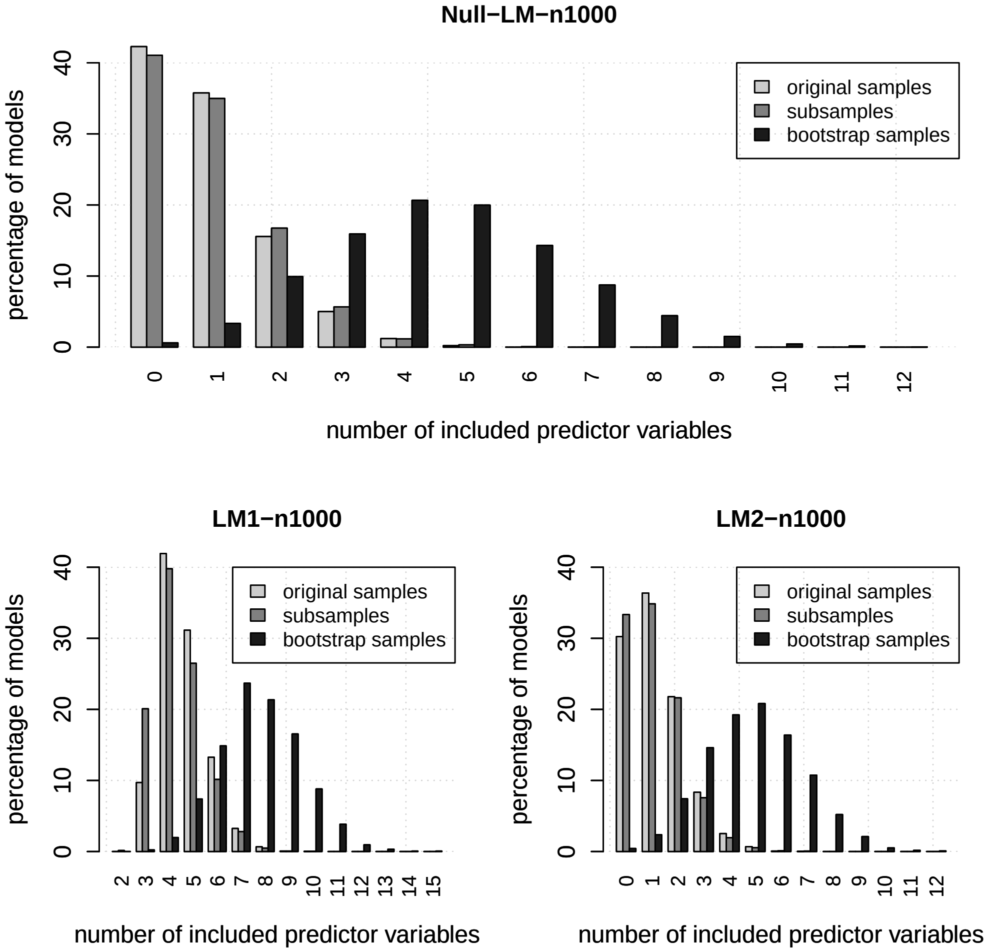


Figure 2: Percentage of models including displayed number of predictor variables in original, subsamples and bootstrap samples: 5000 models, $n = 1000$.

between the different variable types.
This result corroborates our considerations from the section Theoretical rationale. As expected, given the true model where no predictor variable has an effect on the response, categorical predictor variables with many categories are preferentially selected in bootstrap samples compared to binary or metric predictor variables.

Results for LM1-n1000 are displayed in Table 5. In general inclusion frequencies for variables with a non-zero effect are similar for original samples and subsamples, but tend to be smaller for subsamples. This finding of smaller inclusion frequencies for subsamples can be explained by the lower power of the LR-test to detect these effects, as the sample size of the subsamples is smaller than that of original samples. Again, for original samples or subsamples no preference for any particular type of predictor can be observed. For bootstrap samples in contrast we again observe a preferential selection of categorical predictors without any effect over metric and binary predictors without effect, as well as over categorical predictors with fewer categories and no effect. When considering inclusion frequencies for predictors with effect one can see that the metric predictors with a large effect of 0.2 are included in nearly all models, irrespective of whether derived from original samples, bootstrap samples or subsamples. For categorical predictors with effect there are differences among the sampling approaches. However, all sampling approaches come to the same conclusion regarding the relative importance of the predictors according to their inclusion frequencies: the most important predictors are the metric predictors X_1, X_2, X_3. After that the 4-

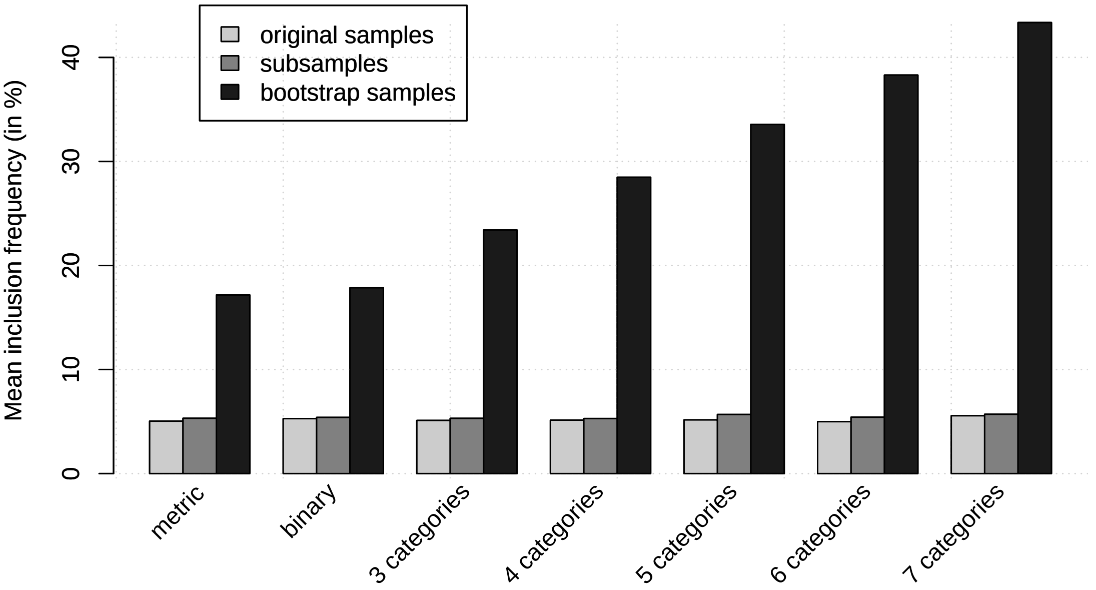


Figure 3: Null-LM-n1000: average inclusion frequencies over 5000 models for predictor variables in original samples, subsamples and bootstrap samples.

category predictor X_{10} is the fourth most important predictor and finally the predictor X_{14} with 6 categories is the predictor among the informative predictors which is least important. Considering the relative sizes of the inclusion frequencies of the two categorical informative predictors, it can be seen that for the bootstrap scheme the discrepancy between the inclusion frequency of the informative 4-category predictor is much closer to the inclusion frequency of the (less important) 6-category predictor. This finding is attributable to the preference for categorical predictors with many categories: the less important predictor (X_{14}) with more categories is preferred over the more important predictor (X_{10}) with fewer categories. However, here the difference in effects of the two predictors is large enough to ensure that the more important predictor with fewer categories indeed receives a higher inclusion frequency in bootstrap samples. This is however not the case for LM2-n1000, in which predictor effects are smaller (Table 6). For this model the consequences of the preferential selection are much worse than for LM1-n1000. On bootstrap samples, the categorical variable X_{14} with six categories is included almost as often (in 39.9% of the models) as the variable X_{10} with four categories (in 40.7% of the models), though the 4-category variable is actually more strongly associated with the response than the 6-category variable. Even more problematic may be the finding that both the 4-category and the 6-category variable are less frequently included in models derived from bootstrap samples than variables X_{16} and X_{17} with seven categories and no effect. This is not the case for LM1-n1000, and is attributable to the smaller effect sizes in LM2-n1000, which lead to the overriding of effect size by the number of categories. For this model with weaker effects, bootstrap inclusion frequencies give completely misleading conclusions regarding the importance of predictors.

4.1.2 Variant 1: Correlated predictors

We now turn to the case of pairwise correlated binary and 5-category predictor variables. Again, the mean number of included variables is always substantially higher when models are derived from bootstrap samples (data not shown). In the following we focus on the type of predictor variables included in the models and if this is influenced by strong correlations among predictors

Table 5: LM1-n1000: Inclusion frequencies of variables in 5000 models from original samples, subsamples and bootstrap samples of size $n = 1000$.

Predictor	Scale	Effect(s)	Inclusion Freq. (in %)		
			original sample	subsample	bootstrap sample
X_1	metric	0.2	100	99.82	99.98
X_2	metric	0.2	100	99.86	99.88
X_3	metric	0.2	100	99.86	99.86
X_4	metric	0	5.36	5.74	18.00
X_5	metric	0	5.30	5.86	17.80
X_6	binary	0	5.54	5.48	17.58
X_7	binary	0	5.34	5.36	18.06
	categorical with				
X_8	3 categories	0, 0	5.46	5.36	23.28
X_9	3 categories	0, 0	5.04	5.26	23.48
X_{10}	4 categories	0.2, 0.2, 0	76.22	55.58	76.90
X_{11}	4 categories	0, 0, 0	5.56	5.54	28.66
X_{12}	5 categories	0, 0, 0, 0	5.46	6.04	33.64
X_{13}	5 categories	0, 0, 0, 0	5.28	5.72	33.46
X_{14}	6 categories	0.1, 0.1, 0.1, 0, 0	19.54	14.62	52.62
X_{15}	6 categories	0, 0, 0, 0, 0	5.20	5.68	38.40
X_{16}	7 categories	0, 0, 0, 0, 0, 0	5.78	5.86	43.06
X_{17}	7 categories	0, 0, 0, 0, 0, 0	5.66	5.62	42.66

Table 6: LM2-n1000: Inclusion frequencies of predictor variables in 5000 models from original samples, subsamples and bootstrap samples of size $n = 1000$. For categorical predictor variables with k categories effect sizes for $(k - 1)$ dummy variables are shown.

Predictor	Scale	Effect(s)	Inclusion Freq. (in %)		
			original sample	subsample	bootstrap sample
X_1	metric	0.01	6.24	5.74	17.92
X_2	metric	0.02	10.18	9.10	21.50
X_3	metric	0.03	14.94	11.90	25.74
X_4	metric	0	5.28	5.54	17.84
X_5	metric	0	5.20	5.78	17.88
X_6	binary	0	5.38	5.28	17.54
X_7	binary	0	5.26	5.38	17.86
	categorical with				
X_8	3 categories	0, 0	5.36	5.26	23.00
X_9	3 categories	0, 0	4.8	5.28	23.80
X_{10}	4 categories	0.08, 0.08, 0	16.06	12.04	39.86
X_{11}	4 categories	0, 0, 0	5.56	5.5	28.36
X_{12}	5 categories	0, 0, 0, 0	5.42	5.84	33.62
X_{13}	5 categories	0, 0, 0, 0	5.30	5.72	33.22
X_{14}	6 categories	0.04, 0.04, 0.04, 0, 0	7.50	6.90	40.74
X_{15}	6 categories	0, 0, 0, 0, 0	5.22	5.48	38.48
X_{16}	7 categories	0, 0, 0, 0, 0, 0	5.58	6.08	43.36
X_{17}	7 categories	0, 0, 0, 0, 0, 0	5.64	5.42	42.94

Table 7: corr(ρ)-LM-n1000 with $\rho \in \{0.3, 0.5, 0.7\}$: Inclusion frequencies of predictor variables in 5000 models from original samples, subsamples and bootstrap samples of size $n = 1000$.

Pred.	Scale	correla- ted to	Effect(s)	Inclusion Frequency (in %) original sample corr 0.3 / 0.5 / 0.7	 subsample corr 0.3 / 0.5 / 0.7	 bootstrap sample corr 0.3 / 0.5 / 0.7
X_1	binary	X_6	0.1	30.82 / 27.66 / 21.56	20.86 / 18.56 / 15.06	37.94 / 35.60 / 29.86
X_2	binary	X_7	0.1	35.56 / 34.46 / 31.62	24.52 / 23.70 / 22.84	40.62 / 38.14 / 34.00
X_3	binary	X_8	0	5.46 / 5.52 / 5.96	5.18 / 5.18 / 5.70	18.18 / 18.04 / 17.64
X_4	binary	X_9	0	5.10 / 4.62 / 4.96	4.78 / 5.02 / 5.16	17.22 / 17.80 / 17.56
X_5	binary	-	0.1	34.66 / 34.52 / 33.96	23.24 / 23.14 / 23.28	39.18 / 39.70 / 38.88
X_6	categorical with 5 categories	X_1	0.1, 0.1, 0, 0	20.86 / 22.04 / 22.92	14.36 / 15.22 / 16.10	47.86 / 49.62 / 49.82
X_7	5 categories	X_2	0, 0, 0, 0	5.86 / 7.28 / 8.92	5.62 / 6.76 / 8.40	34.18 / 36.04 / 38.40
X_8	5 categories	X_3	0.1, 0.1, 0, 0	20.42 / 20.48 / 20.86	13.96 / 14.04 / 14.08	46.26 / 47.88 / 49.86
X_9	5 categories	X_4	0, 0, 0, 0	5.64 / 5.76 / 5.56	5.52 / 5.66 / 5.60	32.72 / 33.62 / 34.84
X_{10}	5 categories	-	0.1, 0.1, 0, 0	21.06 / 20.96 / 21.04	14.90 / 14.96 / 15.06	48.00 / 47.54 / 47.88

of different scales.

The results for the null model are comparable to those obtained for Null-LM-n1000 and are shown in the appendix. Results for models corr(0.3)-LM-n1000, corr(0.5)-LM-n1000 and corr(0.7)-LM-n1000 are shown in Table 7. In these models, both the binary variable and the 5-category variable in the first pair of variables (X_1, X_6) have an effect. For the second pair (X_2, X_7) only the binary predictor variable has an effect and for the third pair (X_3, X_8) only the categorical predictor variable has an effect. For the fourth pair (X_4, X_9), neither variable has an effect. The remaining—independent—predictor variables X_5 (binary) and X_{10} (5-category) both have an effect on the response. The results of our studies illustrate that the presence of correlations among predictor variables of different scales intensifies the problem of preferential selection of categorical predictor variables for bootstrap samples. We first look at the results for the model corr(0.3)-LM-n1000, in which correlations among binary and categorical predictor variables are moderate. When looking at the correlated pair (X_2, X_7) for which only the binary predictor variable X_2 has an effect, it is obvious that the inclusion frequency is higher for the truly informative binary predictor variable X_2 than for the dependent (non-informative) categorical predictor variable X_7. This is also the case when deriving inclusion frequencies from bootstrap samples, but here the difference in inclusion frequencies between the informative and the non-informative predictor is smaller, giving a worse discrimination. If we now look at the model corr(0.7)-LM-n1000 which has a higher dependence value, however, we can see that the non-informative 5-category predictor variable X_7 now has a higher inclusion frequency than the binary variable X_2 when using bootstrap samples.

This is attributable to the preferential selection of categorical predictor variables in combination with the spurious effect that is induced by the high association with another variable (here X_2) which has an effect on the response. This was however not the case when the dependence between the predictor variables was lower ($\rho = 0.5$ or $\rho = 0.3$). Here the spurious effect of X_7 was not high enough to select the categorical variable X_7 more often into bootstrap models than the binary variable X_2 .

When looking at inclusion frequencies in original samples and in subsamples it is noticeable that the inclusion frequencies of mutually dependent binary and categorical predictor variables get closer to each other with rising correlations. This behavior would be expected because if two variables are highly correlated, both of them contain almost the same information and are therefore interchangeable. However, this is not observed for bootstrap samples, probably because the preference for categorical predictor variables prevents the inclusion frequencies of dependent predictor variables from getting closer.

4.1.3 Variant 2: Smaller sample size

The results for Null-LM-n100 are shown in Table 8. Overall, the results are comparable with the results for models with larger sample sizes in the sense of model complexity and preference for categorical predictor variables in bootstrap samples. However, here original samples and subsamples show a very small preference of the 4-category predictor variable X_2 over the binary predictor variable X_1 and a preference of the 7-category predictor variable X_3 over the binary and the 4-category predictor variables, though none of the predictors has an effect. This difference in inclusion frequencies – though very marginal – can also be observed for much higher numbers of original samples or subsamples, respectively, indicating that it is not due to randomness. One possible explanation for the different inclusion frequencies in original samples and in subsamples might be that the sample size of $n = 100$ is not sufficiently large to guarantee the reliability of asymptotic theory. This is also supported by the fact that in none of our other simulations (all with a much larger sample size of $n = 1000$) such an effect is observed.

Table 8: Null-LM-n100: Inclusion frequencies of predictor variables in 5000 models from original samples, subsamples and bootstrap samples of size $n = 100$.

Predictor	Scale	Effect(s)	Inclusion Freq. (in %)		
			original sample	subsample	bootstrap sample
X_1	binary	0	5.32	5.84	20.54
	categorical with				
X_2	4 categories	0, 0, 0	5.94	6.88	32.70
X_3	7 categories	0, 0, 0, 0, 0, 0	6.74	7.88	45.60

Results for models LM1-n100 and LM2-n100 are presented in the appendix and are consistent with the results for models with larger sample sizes.

4.1.4 Variant 3: Survival response

The results for the analyses of Cox regression models are consistent with those for the linear regression models from the main simulation and are thus only briefly summarized here. Detailed results are given in the appendix. Models selected based on bootstrap samples contain on average more predictor variables in total than models selected based on original samples or subsamples. This is apparent for the null model (Null-Cox-n1000) as well as for the two models with informative predictors (Cox1-n1000 and Cox2-n1000). In addition, for the latter two models a clear preference for categorical predictor variables in bootstrap sample-based models, resulting in systematically too high inclusion frequencies, can be observed, as with the linear regression setting. In some cases this systematic preferential selection even overrides the advantage of predictor variables with effect over predictor variables without effect. A preferential selection of categorical predictor variables over metric and binary predictor variables and over categorical predictor variables with fewer categories is thus not specific to the linear regression model but is also present for other models such as the Cox model.

4.2 Real data study

We also use a real data set to more deeply investigate the findings of our simulation studies. We consider data from the 2007-2008 cycle of the National Health and Nutrition Examination Survey (NHANES) (National Center for Health Statistics; 2012) which is maintained by the Centers for Disease Control and Prevention. NHANES is designed as a series of cross-sectional surveys and uses a stratified multistage sampling method to obtain a representative sample of the US population. The data are freely available from the institution's homepage or from the Interuniversity Consortium for Political and Social Research (ICPSR; 2012). We analyze the level of high-sensitive C-reactive protein (CRP) as response variable, a plasma protein involved in the acute phase response during inflammatory states (Black et al.; 2004). The considered data set comprises a total of $n = 1914$ subjects. Table 5 in the appendix shows descriptive statistics for the predictor variables from the NHANES data set considered in our application. A more detailed description of the predictor variables is given in the appendix.

The backward elimination procedure based on the LR-test used in the simulations is employed again here for model selection in the linear regression framework for (i) the original data set, (ii) 5000 bootstrap samples drawn with replacement, and (iii) 5000 subsamples drawn without replacement. Since the true data generating process is unknown it is not possible to compare the results obtained from bootstrap samples to those from several original samples. Instead we compare the results of bootstrap samples to the results of subsamples, since we have evidence

Table 9: Inclusion frequencies for predictor variables in the NHANES linear regression model for 5000 subsamples and 5000 bootstrap samples.

Predictor	Scale	Inclusion Freq. (in %)	
		subsample	bootstrap sample
age	metric	30.12	51.08
alcohol	metric	0.00	4.88
BMI	metric	99.98	99.90
BPdias	metric	7.46	22.36
BPsys	metric	20.12	45.38
cholesterol	metric	4.46	21.56
waistcircum	metric	6.98	17.06
WBCcount	metric	100.00	99.96
100cig	binary	12.36	28.68
AcuteIllness	binary	82.58	87.10
asthma	binary	6.46	17.94
chronicBronchitis	binary	0.68	11.96
diabetes	binary	9.24	35.60
heartFailure	binary	0.08	4.02
heavyDrinker	binary	0.48	5.84
sex	binary	72.54	71.12
stroke	binary	17.86	30.86
	categorical with		
country_of_birth	4 categories	0.48	16.70
depression_screening	4 categories	7.60	45.98
education	5 categories	2.28	31.26
HealthStatus	5 categories	41.56	69.50
medicalPlaceToGo	5 categories	0.02	7.56
race	5 categories	39.62	73.78
sleepTrouble	5 categories	12.46	41.20
ToothCond	5 categories	32.86	68.72
wakeUp	5 categories	38.26	68.38
marital_status	6 categories	51.98	74.90
income	12 categories	46.00	87.52

from our simulation studies that backward elimination can reliably recover the true ordering of predictor variables. For the original sample, our backward elimination procedure yields the model *CRP* $\sim$ *WBCcount* + *BPSys*+ *age*+ *BMI* + *race* + *ToothCond* + *wakeUp* + *income* + *sex* + *AcuteIllness*.

Figure 4 depicts the percentage of models that included a specific number of predictor variables. The results are very similar to those of the simulation studies. On average more predictor variables are included in models when using the bootstrap approach than with subsampling. However, as noted in the preceding section, due to the smaller statistical power with subsampling one tends to select fewer predictor variables than for the original sample. This property is evident when we examine the model obtained from the original sample, which includes 10 predictors. Examination of Figure 4 shows that there are only few subsampling models which include 10 or more predictor variables and conversely, for the bootstrap there are only few models which have 10 or fewer predictor variables. This indicates that for this data set models selected from bootstrap samples may be too complex.

We now further investigate the type of included predictor variables for the two sampling approaches. A special focus is laid on cases where binary or categorical predictors show different inclusion frequencies in bootstrap samples and subsamples. Inclusion frequencies for subsample and bootstrap sample-based models of the NHANES data set are displayed in Table 9. In general the inclusion frequency for a variable for bootstrap samples is higher than its counterpart for subsamples for all predictor variables. There are a few cases in which binary predictor variables yield higher inclusion frequencies than a categorical predictor variable for subsamples but the categorical predictor variable is more frequently included than the binary for the bootstrap samples.

For example, the inclusion frequency for the binary variable *stroke* is 17.86% for subsamples. Inclusion frequencies for the categorical variables *depression screening* ($k = 4$) and *sleep trouble* ($k = 5$) are lower, with 7.60% and 12.46%, respectively. Since in our simulation studies inclusion frequencies obtained by subsampling have been shown to reliably reflect the relative importance

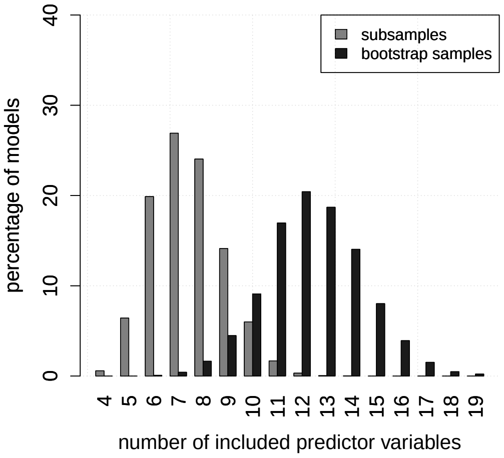


Figure 4: Linear regression model for NHANES data: percentage of models (of 5000) against number of included predictor variables in subsamples (gray) and bootstrap samples (black). $n = 1914$.

of predictors, we therefore assume that the variable *stroke* is more strongly associated with CRP than the variables *depression screening* ($k = 4$) and *sleep trouble* ($k = 5$). For bootstrap samples, however, the inclusion frequency for *stroke* is 30.86% (which is higher than that for the subsamples, as one would expect) but the inclusion frequencies for *depression screening* ($k = 4$) and *sleep trouble* ($k = 5$) are even higher, with 45.98% and 41.2%, respectively. In other words, if the importances of these variables were to be assessed based on bootstrap samples, the association between predictor variables *depression screening* ($k = 4$) and *sleep trouble* ($k = 5$) and the response would be incorrectly estimated to be higher than the association between the predictor variable *stroke* and the response.

This also occurs for the binary variable *100cigarettes*, which describes whether the proband has smoked at least 100 cigarettes in his/her life. For subsamples, the inclusion frequency for this variable is 12.36%. This exceeds the inclusion frequency for *depression screening* ($k = 4$), which is 7.60%. For bootstrap samples, the variable *100cigarettes* is included in 28.68% of the models, which is less than the inclusion frequency of 45.98% for *depression screening* ($k = 4$). Again, the importance of *depression screening* is (likely incorrectly) estimated to be higher when bootstrap inclusion frequencies are used. We note that none of the variables *stroke*, *100cigarettes*, *sleep trouble* (k =5) or *depression screening* ($k = 4$) are selected for the model on the original data set.

Our results also indicate that the preferential selection of categorical predictor variables for bootstrap samples might be present even for stronger true effect sizes. Two further examples which indicate a preferential selection of a categorical predictor variable over a binary predictor variable can be seen with the variables *sex* and *marital status* ($k = 6$), and *sex* and *race* ($k = 5$). For subsamples, the binary variable *sex* is included in 72.54% of the models, which suggests a rather strong effect of *sex* on CRP in our sample. The variable is also selected in the model derived from the original data set. This inclusion frequency is higher than those of the categorical variables *marital status* ($k = 6$, 51.98%) and *race* ($k = 5$, 39.62%). However, for bootstrap samples the opposite occurs: the inclusion frequencies for the two categorical variables are slightly higher than the inclusion frequency for *sex* (74.9% and 73.78% vs 71.12%).

Furthermore, the binary variable *acute illness* has an inclusion frequency of 82.58% for subsamples. After the metric variables *BMI* and *white blood cell count*, *acute illness* has the highest inclusion frequency of all variables for subsamples. We would therefore assume that the true effect of *acute illness* on CRP is substantial. However, for bootstrap samples, the inclusion frequency of 87.10% of *acute illness* is slightly exceeded by the inclusion frequency of 87.52% of the categorical variable *income* (k =12). *Income* has an inclusion frequency of only 46% for subsamples, suggesting that its true effect on CRP is considerably smaller than the effect of *acute illness*. Nevertheless *income* (k =12) is preferentially selected, showing the highest overall inclusion frequency for bootstrap samples. It is worth noting that *income* is the categorical variable with the most categories ($k = 12$) in this analysis.

It is important to note that in the present analysis there is no case where a categorical predictor variable shows a higher inclusion frequency than a metric or binary predictor variable for the subsamples, but is less frequently included than the metric or binary variable on bootstrap samples. Overall, these findings obtained from the real data application are in line with those obtained from the simulation studies. They illustrate that the effect of categorical predictor variables might be considerably overestimated when performing model selection on bootstrap samples, and that this issue is of high practical relevance.

5 Discussion

In this paper we performed extensive simulations to compare multivariable regression models selected from (nonparametric) bootstrap samples to those selected from original samples and subsamples, using backward elimination based on LR-test with linear and Cox regression models as examples. Theoretical considerations of the distribution of the LR-test statistic computed from bootstrap samples suggest that models selected from bootstrap samples include a higher number of predictor variables and that categorical variables are preferentially selected with preferential selection increasing with the number of categories. These conjectures were confirmed and quantitatively assessed in our simulation study, which also further demonstrated that these mechanisms can have a substantial impact on practical results.

The number of categories may even override the effect size, especially in the case of small effect sizes, so that in some settings non-informative categorical predictor variables are selected on average more often than informative binary or metric predictor variables. This also became manifest in the analysis of a real data set from the NHANES study. Binary predictor variables which received higher inclusion frequencies than categorical predictor variables for models based on subsamples were often excluded in favor of categorical predictor variables with many categories when model selection was conducted on bootstrap samples.

We emphasize that this problem only occurs if candidate predictor variables include categorical predictors with different numbers of categories or categorical predictors together with metric/binary predictors. If the variable selection procedure is applied to a set of metric and binary predictor variables only, no preferential selection is observed. Furthermore, in our analyses categorical predictor variables were entered as a whole in the model based on a global LR-test, the null hypothesis being that the coefficients of all the dummy variables derived from this variable are zero. As an alternative procedure, the dummy variables of categorical predictors could be tested separately. This would not lead to the here described preferential selection of categorical predictor variables, but yield a multiple testing problem instead. We also want to stress that we did not analyze the predictive abilities of models when these are derived from a bootstrap sample. This should be examined in future research.

We considered a backward elimination procedure based on the LR-test, but other methods such as forward or stepwise selection are applicable as well. In principle, they are also affected by the bias investigated in our paper. Moreover, there exists a wide range of selection criteria like AIC, BIC or the Wald test that can be used in automated variable selection procedures. From preliminary analyses (data not shown) we strongly expect similar behaviour when using these criteria, however, it remains to be investigated in future studies how the choice of the selection procedure and criterion affects results.

All our investigations were performed in a low-dimensional setting, i.e., in cases where the sample size greatly exceeds the number of predictor variables. Consequences of variable selection on bootstrap samples for high-dimensional data, for instance a mixture of clinical and genetic risk factors, remain to be investigated.

For specific applications, solutions to avoid the overcomplexity of models derived from bootstrap samples have already been suggested. Bollen and Stine (1992) proposed a corrected test statistic for bootstrapping in structural equation models and Steck and Jaakkola (2003) suggested a bias correction term to avoid the overcomplexity in graphical models. However, neither of these are directly applicable to the problem of preferential selection of categorical predictor variables considered here. However, we can corroborate the findings of Strobl et al. (2007) that subsampling might be a promising solution, since the preferential selection of categorical predictor variables does not occur for subsamples. In all of our simulation settings original samples and subsamples had identical inclusion frequency patterns with regard to the type of predictor variable included. This is due to the fact that the original properties of the LR statistic are maintained if the subsample is drawn without replacement. As the subsample is a proper subset of the original sample, it is valid to assume that both the subsample and the original sample have been drawn from the same underlying population where H_0 holds. Therefore, the distribution of the LR statistic for subsamples is the same as in original samples under the null hypothesis. As a consequence, the LR-test does not favor categorical predictor variables with many categories under the null hypothesis, which states that no predictor variable is associated with the response. If the possibility of drawing reasonably sized subsamples exists for a data set to be analyzed, subsampling presents a simple but effective alternative to the biases introduced by bootstrapping.

Acknowledgements
The authors thank Rory Wilson for helpful comments.

Appendix

Table 10: corr(ρ)-Null-LM-n1000 with $\rho \in \{0.3, 0.5, 0.7\}$: Inclusion frequencies of predictor variables in 5000 models from original samples, subsamples and bootstrap samples of size $n = 1000$.

Predictor	Scale	correla- ted to	Effect(s)	Inclusion Frequency (in %) original sample corr 0.3 / 0.5 / 0.7	 subsample corr 0.3 / 0.5 / 0.7	 bootstrap sample corr 0.3 / 0.5 / 0.7
X_1	binary	X_6	0	5.04 / 4.90 / 4.70	5.44 / 5.58 / 5.24	17.66 / 17.72 / 17.34
X_2	binary	X_7	0	4.82 / 4.92 / 4.70	4.66 / 4.74 / 4.28	16.74 / 16.90 / 18.08
X_3	binary	X_8	0	5.54 / 5.04 / 4.68	5.26 / 4.76 / 5.22	17.78 / 18.06 / 17.54
X_4	binary	X_9	0	4.84 / 4.44 / 4.80	4.70 / 5.02 / 4.78	17.08 / 17.14 / 17.48
X_5	binary	-	0	5.16 / 5.14 / 5.24	5.56 / 5.56 / 5.66	17.12 / 17.40 / 17.06
	categorical with					
X_6	5 categories	X_1	0, 0, 0, 0	5.18 / 5.70 / 5.56	5.06 / 5.88 / 6.10	33.24 / 33.90 / 35.10
X_7	5 categories	X_2	0, 0, 0, 0	5.20 / 5.40 / 5.76	4.96 / 5.70 / 5.60	33.88 / 35.24 / 35.50
X_8	5 categories	X_3	0, 0, 0, 0	5.46 / 5.26 / 5.26	5.08 / 5.26 / 5.76	33.72 / 33.58 / 35.04
X_9	5 categories	X_4	0, 0, 0, 0	5.30 / 5.66 / 5.42	5.24 / 5.54 / 5.60	32.68 / 33.76 / 34.54
X_{10}	5 categories	-	0, 0, 0, 0	5.02 / 4.98 / 5.02	5.12 / 5.34 / 5.36	31.82 / 31.78 / 31.88

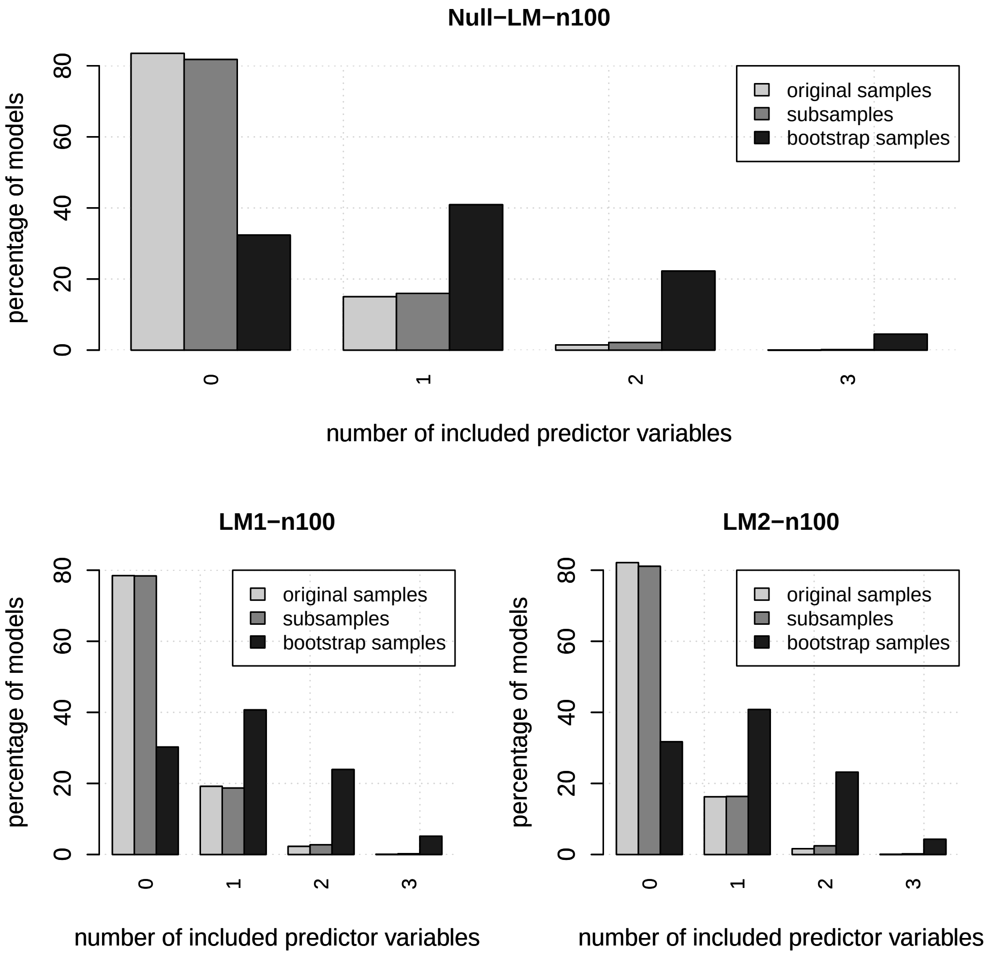


Figure 5: Percentage of models including displayed number of predictor variables in original, subsamples and bootstrap samples: 5000 models, $n = 100$.

Table 11: Null-LM-n100: Inclusion frequencies of predictor variables in 5000 models from original samples, subsamples and bootstrap samples of size $n = 100$.

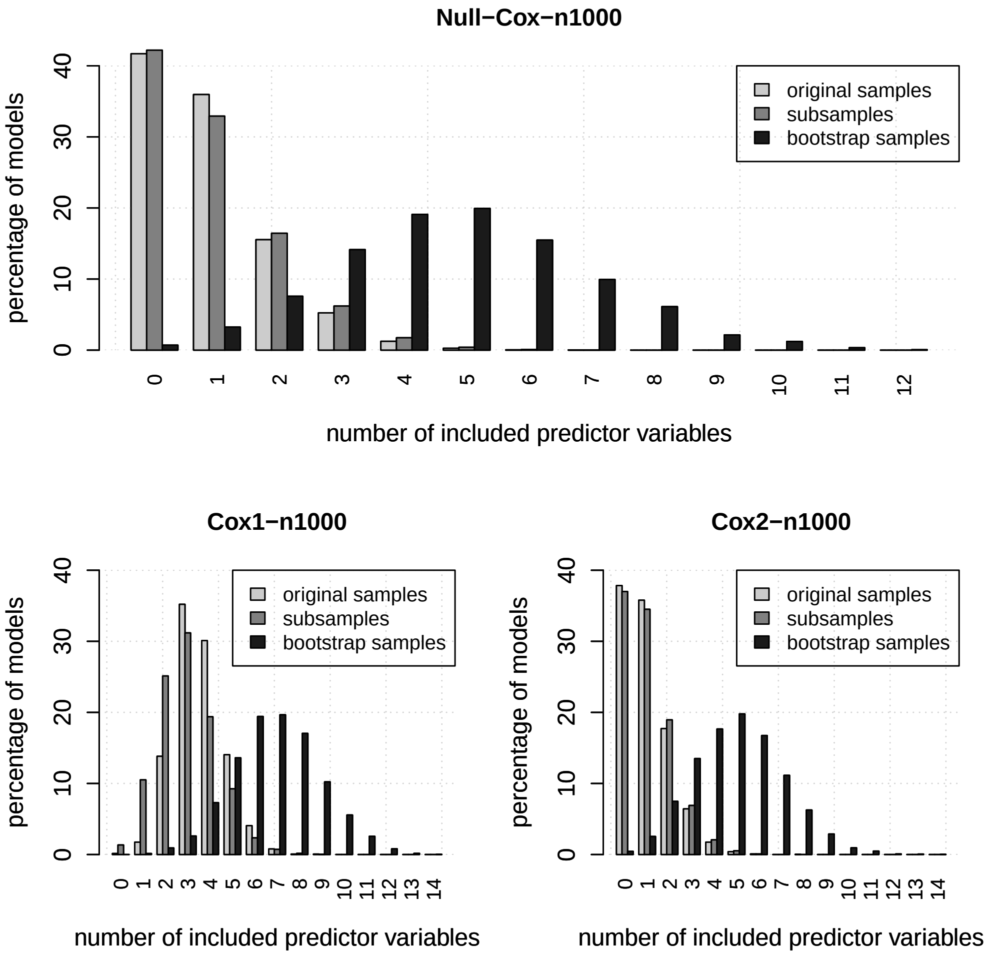
Predictor	Scale	Effect(s)	Inclusion Freq. (in %)		
			original sample	subsample	bootstrap sample
X_1	binary	0	5.32	5.84	20.54
X_2	categorical with 4 categories	0, 0, 0	5.94	6.88	32.70
X_3	7 categories	0, 0, 0, 0, 0, 0	6.74	7.88	45.60

Table 12: LM1-n100: Inclusion frequencies of predictor variables in 5000 models from original samples, subsamples and bootstrap samples of size $n = 100$.

Predictor	Scale	Effect(s)	Inclusion Freq. (in %)		
			original sample	subsample	bootstrap sample
X_1	binary	0.1	8.58	7.74	22.82
X_2	categorical with 4 categories	0.1, 0.1, 0	8.04	8.18	34.12
X_3	7 categories	0.1, 0.1, 0, 0, 0, 0	7.36	8.82	47.06

Table 13: LM2-n100: Inclusion frequencies of predictor variables in 5000 models from original samples, subsamples and bootstrap samples of size $n = 100$.

Predictor	Scale	Effect(s)	Inclusion Freq. (in %)		
			original sample	subsample	bootstrap sample
X_1	binary	0.05	6.36	6.24	20.92
X_2	categorical with 4 categories	0.05, 0.05, 0	6.20	7.10	33.04
X_3	7 categories	0.05, 0.05, 0, 0, 0, 0	7.02	8.30	46.10



Null−Cox−n1000

Cox1−n1000

Cox2−n1000

Figure 6: Percentage of Cox models including displayed number of predictor variables in original, subsamples and bootstrap samples: 5000 models, $n = 1000$.

Table 14: Null-Cox-n1000: Inclusion frequencies of predictor variables in 5000 models from original samples, subsamples and bootstrap samples of size $n = 1000$.

Predictor	Scale	Effect(s)	Inclusion Freq. (in %)		
			original sample	subsample	bootstrap sample
X_1	metric	0	5.62	5.24	18.12
X_2	metric	0	5.28	5.24	18.14
X_3	metric	0	4.90	5.12	18.92
X_4	metric	0	4.78	5.22	17.20
X_5	metric	0	5.18	5.26	18.22
X_6	binary	0	5.34	5.3	18.58
X_7	binary	0	5.72	5.88	18.98
	categorical with				
X_8	3 categories	0, 0	4.90	5.30	24.448
X_9	3 categories	0, 0	4.76	5.30	25.32
X_{10}	4 categories	0, 0.2, 0	5.58	5.28	32.54
X_{11}	4 categories	0, 0, 0	4.50	5.34	30.36
X_{12}	5 categories	0, 0, 0, 0	6.04	5.88	36.60
X_{13}	5 categories	0, 0, 0, 0	5.60	6.06	35.46
X_{14}	6 categories	0, 0, 0, 0, 0	5.04	5.76	41.08
X_{15}	6 categories	0, 0, 0, 0, 0	5.20	6.02	40.52
X_{16}	7 categories	0, 0, 0, 0, 0, 0	5.76	6.26	44.52
X_{17}	7 categories	0, 0, 0, 0, 0, 0	5.20	5.52	45.26

Table 15: Cox1-n1000: Inclusion frequencies of predictor variables in 5000 models from original samples, subsamples and bootstrap samples of size $n = 1000$.

Predictor	Scale	Effect(s)	Inclusion Freq. (in %)		
			original sample	subsample	bootstrap sample
X_1	metric	0.2	86.88	67.72	77.46
X_2	metric	0.2	85.52	66.44	77.54
X_3	metric	0.2	85.80	67.60	77.90
X_4	metric	0	5.08	5.5	17.76
X_5	metric	0	5.14	5.5	18.34
X_6	binary	0	5.54	5.76	19.18
X_7	binary	0	4.82	5.24	18.36
	categorical with				
X_8	3 categories	0, 0	5.66	6.36	26.28
X_9	3 categories	0, 0	5.20	5.26	24.60
X_{10}	4 categories	0.2, 0.2, 0	22.78	15.82	45.76
X_{11}	4 categories	0, 0, 0	5.74	5.86	31.92
X_{12}	5 categories	0, 0, 0, 0	6.04	7.04	37.10
X_{13}	5 categories	0, 0, 0, 0	4.86	5.92	35.92
X_{14}	6 categories	0.1, 0.1, 0.1, 0, 0	9.46	8.70	44.58
X_{15}	6 categories	0, 0, 0, 0, 0	5.78	6.04	41.06
X_{16}	7 categories	0, 0, 0, 0, 0, 0	6.10	7.42	46.04
X_{17}	7 categories	0, 0, 0, 0, 0, 0	6.02	6.36	45.34

Table 16: Cox2-n1000: Inclusion frequencies of predictor variables in 5000 models from original samples, subsamples and bootstrap samples of size $n = 1000$. For categorical predictor variables with k categories effect sizes for $(k-1)$ dummy variables are shown.

Predictor	Scale	Effect(s)	Inclusion Freq. (in %)		
			original sample	subsample	bootstrap sample
X_1	metric	0.01	5.32	5.72	18.12
X_2	metric	0.02	6.82	6.60	19.66
X_3	metric	0.03	8.84	7.94	20.78
X_4	metric	0	4.82	5.32	18.78
X_5	metric	0	5.46	5.44	17.58
X_6	binary	0	5.58	5.46	19.18
X_7	binary	0	5.34	5.34	18.98
	categorical with				
X_8	3 categories	0, 0	4.92	5.24	25.24
X_9	3 categories	0, 0	4.84	5.22	25.08
X_{10}	4 categories	0.08, 0.08, 0	8.92	8.06	35.60
X_{11}	4 categories	0, 0, 0	5.88	6.12	31.12
X_{12}	5 categories	0, 0, 0, 0	5.18	5.82	35.36
X_{13}	5 categories	0, 0, 0, 0	5.20	5.76	35.70
X_{14}	6 categories	0.04, 0.04, 0.04, 0, 0	6.38	7.12	42.76
X_{15}	6 categories	0, 0, 0, 0, 0	5.44	6.40	42.42
X_{16}	7 categories	0, 0, 0, 0, 0, 0	5.40	6.10	46.36
X_{17}	7 categories	0, 0, 0, 0, 0, 0	5.74	6.74	45.94

Table 17: NHANES: Original interview question or description of selected variables

Abbreviation	Interview question / description	Values
race	Recode of reported race and ethnicity information	Mexican American Other Hispanic Non-Hispanic White Non-Hispanic Black Other Race - Including Multi-Racial
country of birth	In what country (were you/was SP) born?	50 US States or Washington, DC Mexico Other Spanish Speaking Country Other Non-Spanish Speaking Country
education	What is the highest grade or level of school (you have/SP has) completed or the highest degree (you have/she/he has) received?	less than 9th up to 11th high school some college graduate
marital status	Marital staus	married widowed divorced separated never married living with partner
HealthStatus	Would you say (your/SP's) health in general is . . .	excellent very good good fair poor
depression screening	Over the last 2 weeks, how often have you been bothered by the following problems: little interest or pleasure in doing things? Would you say...	not at all several days over half the days nearly every day
ToothCond	Now I have some questions about the condition of your teeth and gums. How would you describe the condition of (your/SP?s) teeth? Would you say . . .	excellent very good good fair poor
sleepTrouble	In the past month, how often did (you/SP) have trouble falling asleep?	never rarely sometimes often almost always
wakeUp	In the past month, how often did (you/SP) wake up during the night and had trouble getting back to sleep?	never rarely sometimes often almost always
medicalPlaceToGo	What kind of place (do you/does SP) go to most often: is it a clinic, doctor's office, emergency room, or some other place?	clinic doctor's office hospital emergency hospital outpatient other
income	Total household income (reported as a range value in dollars)	under \$5k \$5k - under \$10k \$10k - under \$15k \$15k - under \$20k \$20k - under \$25k \$25k - under \$35k \$35k - under \$45k \$45k - under \$55k \$55k - under \$65k \$65k - under \$75k \$75k - under \$100k over \$100k
AcuteIllness	Did (you/SP) have a head cold or chest cold that started during the last 30 days? *or* Did (you/SP) have flu, pneumonia, or ear infections that started during those 30 days? *or* Did (you/SP) have a stomach or intestinal illness with vomiting or diarrhea that started during those 30 days?	no yes
100cig	Have you/Has SP smoked at least 100 cigarettes in (your/his/her) entire life?	yes no
diabetes	(Other than during pregnancy, (have you/has SP)/(Have you/Has SP)) ever been told by a doctor or health professional that (you have/(he/she/SP) has) diabetes or sugar diabetes?	yes no

Continued on next page

Continued from previous page

asthma	Has a doctor or other health professional ever told (you/SP) that (you/she/he) have/has asthma?	yes no
heartFailure	Has a doctor or other health professional ever told (you/SP) that (you/she/he) had congestive heart failure?	yes no
stroke	Has a doctor or other health professional ever told (you/SP) that (you/she/he) had a stroke?	yes no
chronicBronchitis	Has a doctor or other health professional ever told (you/SP) that (you/she/he) had chronic bronchitis?	yes no
heavyDrinker	Was there ever a time or times in (your/SP's) life when (you/he/she) drank 5 or more drinks of any kind of alcoholic beverage almost every day?	yes no

		Overall (n = 1914)
race	Mexican American	277 (14%)
	Other Hispanic	195 (10%)
	Non-Hispanic White	1005 (53%)
	Non-Hispanic Black	370 (19%)
	Other	67 (4%)
country of birth	US	1559 (81%)
	MEX	138 (7%)
	other spanish-speaking	135 (7%)
	other non-spanish-speaking	82 (4%)
education	less than 9th	179 (9%)
	up to 11th	304 (16%)
	high school	480 (25%)
	some college	541 (28%)
	graduate	410 (21%)
marital status	married	1091 (57%)
	widowed	117 (6%)
	divorced	252 (13%)
	separated	70 (4%)
	never married	253 (13%)
	living with partner	131 (7%)
HealthStatus	excellent	198 (10%)
	very good	565 (30%)
	good	722 (38%)
	fair	346 (18%)
	poor	83 (4%)
depression screening	not at all	1403 (73%)
	several days	345 (18%)
	over half the days	87 (5%)
	nearly every day	79 (4%)
ToothCond	excellent	267 (14%)
	very good	337 (18%)
	good	625 (33%)
	fair	404 (21%)
	poor	281 (15%)
sleepTrouble	never	742 (39%)
	rarely	397 (21%)
	sometimes	419 (22%)
	often	215 (11%)
	almost always	141 (7%)
wakeUp	never	665 (35%)
	rarely	373 (19%)
	sometimes	467 (24%)
	often	252 (13%)
	almost always	157 (8%)
medicalPlaceToGo	clinic	384 (20%)
	doctor's office	1396 (73%)

		Overall (n = 1914)
	hospital emergency	71 (4%)
	hospital outpatient	32 (2%)
	other	31 (2%)
income	under \$5k	33 (2%)
	\$5k - under \$10k	69 (4%)
	\$10k - under \$15k	125 (7%)
	\$15k - under \$20k	139 (7%)
	\$20k - under \$25k	160 (8%)
	\$25k - under \$35k	246 (13%)
	\$35k - under \$45k	184 (10%)
	\$45k - under \$55k	176 (9%)
	\$55k - under \$65k	129 (7%)
	\$65k - under \$75k	134 (7%)
	\$75k - under \$100k	206 (11%)
	over \$100k	313 (16%)
sex	male	967 (51%)
	female	947 (49%)
AcuteIllness	no	1437 (75%)
	yes	477 (25%)
100cig	yes	984 (51%)
	no	930 (49%)
diabetes	yes	260 (14%)
	no	1654 (86%)
asthma	yes	287 (15%)
	no	1627 (85%)
heartFailure	yes	52 (3%)
	no	1862 (97%)
stroke	yes	63 (3%)
	no	1851 (97%)
chronicBronchitis	yes	142 (7%)
	no	1772 (93%)
heavyDrinker	yes	317 (17%)
	no	1597 (83%)
waistcircum in cm	Mean ± SD — Median	$100.4 \pm 16.37 \vert 99.4$
Cholesterol in md/dl		$196.9 \pm 41.59 \vert 193.0$
WBCcount in (1k cells/μl)		$7.3 \pm 2.88 \vert 6.9$
BPsys in mmHg		$124.4 \pm 18.62 \vert 122.0$
BPdias in mmHg		$71.2 \pm 11.84 \vert 72.0$
age in years		$50.0 \pm 16.68 \vert 50.0$
BMI in kg/m^2		$29.3 \pm 6.66 \vert 28.3$
alcohol in units		$3.9 \pm 20.18 \vert 2.0$
CRP in mg/dl		$0.4 \pm 0.61 \vert 0.2$

Table 18: NHANES sample: Characteristics of $n = 1914$ participants for considered predictor variables and response variable CRP.

References

Altman, D. and Andersen, P. (1989). Bootstrap investigation of the stability of a Cox regression model, *Statistics in Medicine* **8**: 771–783.

Austin, P. and Tu, J. (2004). Bootstrap methods for developing predictive models, *The American Statistician* **58**: 131–137.

Barbiero, A. and Ferrari, P. A. (2012). *GenOrd: Simulation of ordinal and discrete variables with given correlation matrix and marginal distributions.*, R package version 1.0.1. http://CRAN.R-project.org/package=GenOrd.

Binder, H. and Schumacher, M. (2008). Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples., *Statistical Applications in Genetics and Molecular Biology* **7**: Article 12.

Black, S., Kushner, I. and Samols, D. (2004). C-reactive protein., *Journal of Biological Chemistry* **279**: 48487–48490.

Bollen, K. A. and Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models., *Sociological Methods & Research* **21**: 205–229.

Bruneel, F., Tubach, F., Corne, P., Megarbane, B., Mira, J.-P., Peytel, E., Camus, C., Schortgen, F., Azoulay, E., Cohen, Y., Georges, H., Meybeck, A., Hyvernat, H., Trouillet, J.-L., Frenoy, E., Nicolet, L., Roy, C., Durand, R., Le Bras, J., Wolff, M. and Group, S. S. (2010). Severe imported falciparum malaria: A cohort study in 400 critically ill adults., *PLoS ONE* **5**: e13236.

Chen, C. H. and George, S. (1985). The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model, *Statistics in Medicine* **4**: 39–46.

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

Ette, E. I. (1997). Stability and performance of a population pharmacokinetic model., *Journal of Clinical Pharmacology* **37**: 486–495.

Gong, G. (1982). *Some ideas on using the bootstrap in assessing model variability.*, Springer, New York.

Halabi, S., Small, E. J., Kantoff, P. W., Kattan, M. W., Kaplan, E. B., Dawson, N. A., Levine, E. G., Blumenstein, B. A. and Vogelzang, N. J. (2003). Prognostic model for predicting survival in men with hormone-refractory metastatic prostate cancer., *Journal of Clinical Oncology* **21**: 1232–1237.

Heymans, M., Buuren, S., Knol, D., Mechelen, W. and de Vet, H. (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study., *BMC Medical Research Methodology* **7**: 1–10.

ICPSR (2012). National health and nutrition examination survey (NHANES), 2007-2008., Inter-university Consortium for Political and Social Research.

Janitza, S., Binder, H. and Boulesteix, A.-L. (2014). Pitfalls of hypothesis tests and model selection on bootstrap samples: causes and consequences in biometrical applications., *Technical Report 163*, Department of Statistics, University of Munich. https://epub.ub.uni-muenchen.de/21038/.

Motzer, R. J., Mazumdar, M., Bacik, J., Berg, W., Amsterdam, A. and Ferrara, J. (1999). Survival and prognostic stratification of 670 patients with advanced renal cell carcinoma., *Journal of Clinical Oncology* **17**: 2530–2540.

National Center for Health Statistics (2012). NHANES 2007 to 2008 public data general release file documentation, http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/generaldoc_e.htm.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Royston, P. and Sauerbrei, W. (2003). Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation., *Statistics in Medicine* **22**: 639–659.

Sauerbrei, W., Boulesteix, A.-L. and Binder, H. (2011). Stability investigations of multivariable regression models derived from low- and high-dimensional data., *Journal of Biopharmaceutical Statistics* **21**: 1206–1231.

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building: application to the Cox regression model., *Statistics in Medicine* **11**: 2093–2109.

Steck, H. and Jaakkola, T. S. (2003). Bias-corrected bootstrap and model uncertainty., *Advances in Neural Information Processing Systems* **16**.

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution., *BMC Bioinformatics* **8**: 25.

Tutz, G. (2012). *Regression for categorical data*, Cambridge University Press, Cambridge.

Wagenmakers, E.-J., Farrell, S. and Ratcliff, R. (2004). Naive nonparametric bootstrap model weights are biased., *Biometrics* **60**: 281–283.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses., *The Annals of Mathematical Statistics* **9**: 60–62.