Correcting the Regression Bias from Variable Selection : An Example of Mixture Model

Longhai Li

Department of Statistics

University of Toronto

Joint work with Radford Neal and Jianguo Zhang

Outline

- The Regression Bias from Variable Selection
- The Idea to Correct the Bias
- Correcting the Bias in A Mixture Model
- Simulation Studies
- Discussion and Conclusion

The Regression Bias from Variable Selection

- We want to predict X_1(response) using $X_2, \cdots, X_D$(explanatory variables)
- When D is very large a common practice is to select part of variables to work with. But such variable selection will make the relationship between the response and the selected variables stronger than it actually is. An extreme case to understand this problem is that when all of the variables are unrelated to the response. If we look at only the selected variables some degree of relation will be shown by the selected data.
- People in practice may be unaware of such possible bias. We need a safer inference method on selected data set to protect us from such bias.

The General Idea to Correct the Regression Bias

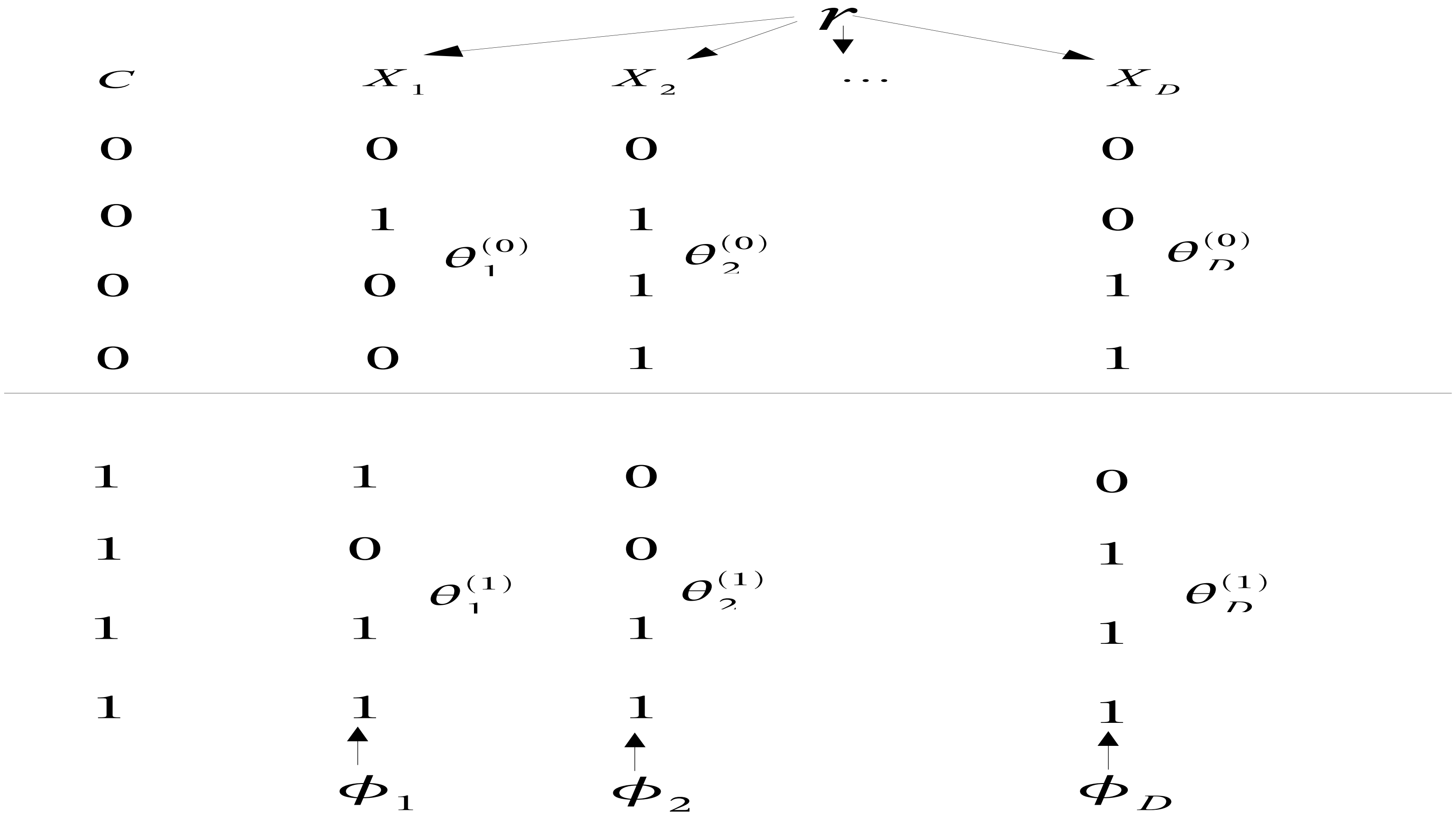
- Suppose $X_{:,1}, \cdots, X_{:,d}$ are included and $X_{:,d+1}, \cdots, X_{:,D}$ are found to have absolute correlations less than γ with the response.
- The idea to correct the bias is **adding some partial information from the un-selected variables**. We need devise a regression model in which some parameters,denoted R, are used to control the overall degree of the relationship between the response and the explanatory variables. Then the partial information from un-selected variables can be used to modify the likelihood of R. I.e. the likelihood is modified to be

$$
\begin{aligned}
& P(X_{:,1} = x_{:,1}, X_{:,2} = x_{:,2}, \cdots X_{:,d} = x_{:,d} \mid \theta) \cdot \\
& P(|\mathrm{Cor}(X_{:,k}, X_{:,1})| \leq \gamma \text{ for } k = d+1, \cdots, D \mid x_{:,1}, x_{:,2}, \cdots x_{:,d}, R)
\end{aligned}
$$

  where θ denotes the all parameters of the models.

- To make this correction useful we also require the explanatory variables are independent given R.

The Picture of A Binary Mixture Model



r

C	X_1	X_2	$\cdots$	X_D
0	0	0		0
0	1	1		0
0	0	1		1
0	0	1		1
1	1	0		0
1	0	0		1
1	1	1		1
1	1	1		1

$\theta_1^{(0)}$ $\theta_2^{(0)}$ $\theta_D^{(0)}$

$\theta_1^{(1)}$ $\theta_2^{(1)}$ $\theta_D^{(1)}$

ϕ_1 ϕ_2 ϕ_D

Mathematical description of the model

$$\begin{aligned}
p^{(0)} &\sim \text{Beta}(\alpha_0, \alpha_1)\\
C_1, C_2, \cdots, C_n | p^{(0)}, p^{(1)} \text{ i.i.d.} &\sim \text{Discrete}(p^{(0)}, p^{(1)})\\
r &\sim \text{Uniform}(r, R_2, \cdots, R_{n_r})\\
\phi_1, \phi_2, \cdots, \phi_D \text{ i.i.d.} &\sim \text{Uniform}(a_0, b_0)\\
(\theta_j^{(0)}, \theta_j^{(1)}) | r, \phi_j \text{ i.i.d.} &\sim \text{Beta}(r\phi_j, r(1-\phi_j))\\
P(X_{ij} = x_{ij}, j = 1, \cdots, D | C_i = c_i, \theta^{(c_i)}) &= \prod_{j=1}^{D} \text{Bernolli } (x_{ij}; \theta_j^{(c_i)}), \text{ for } i = 1, 2 \cdots, n
\end{aligned}$$

Note that $\text{Var}(\theta_j^{(c)}) = \frac{\phi_j(1-\phi_j)}{r+1}$. When r is bigger the relationship between the response and the explanatory variables is weaker. The labels $C_1, \cdots, C_n$ also affect the degree of the regression.

Integrating away θ and $p^{(0)}$

Using Polya Urn Scheme, θ's in the model can be integrated away. The parameters are now only r, $\phi_1, \cdots, \phi_D$ and the labels $C_1, \cdots, C_n$. The density function of $X_{:,1}, \cdots, X_{:,D}$ given the parameters is :

$$\begin{aligned} & P(X_{:,1}, \cdots, X_{:,D}|c_1, \cdots, c_n, r, \phi_1, \phi_D) \\ = & \prod_{c=0}^{1} \prod_{j=1}^{D} \frac{\prod_{k=0}^{I_j^{(c)}} (k + r\phi_j) \prod_{k=0}^{O_j^{(c)}} (k + r(1-\phi_j))}{\prod_{k=0}^{N^{(c)}} (k + r)} \end{aligned}$$

where $I_j^{(c)}$,$O_j^{(c)}$ is the number of 1's and 0's respectively in group c for variable j,$N^{(c)}$ is the number of cases in group c . We see that given C_i's, ϕ_j's and r, the $X_{:,j}$'s are independent.

Integrating away $p^{(0)}$, the prior of $C_1, \cdots, C_n$ is

$$P(C_1, C_2, \cdots, C_n) = \frac{\prod_{k=0}^{N^{(0)}-1} (\alpha_0 + k) \prod_{k=0}^{N^{(1)}-1} (\alpha_1 + k)}{\prod_{k=0}^{n-1} (k + \alpha_1 + \alpha_0))}$$

The Correction factor

$$
\begin{aligned}
& P(|\mathrm{Cor}(X_{:,k}, X_{:,1})| \leq \gamma \text{ for } k = d+1, \cdots, D \mid X_{:,1}, C_1, \cdots, C_n, \phi_{d+1}, \cdots, \phi_D, r) \\
= & \prod_{k=d+1}^{D} P(|\mathrm{Cor}(X_{:,k}, X_{:,1})| \leq \gamma \mid X_{:,1}, r, \phi_k, C_1, \cdots, C_n)
\end{aligned}
$$

Integrating away $\phi_{d+1}, \cdots, \phi_D$ in the joint distribution, the correction factor becomes:

$$
\begin{aligned}
& \left(\int_{a_0}^{b_0} P(|\mathrm{Cor}(X_{:,d+1}, X_{:,1})| \leq \gamma \mid X_{:,1}, r, \phi_{d+1}, C_1, \cdots, C_n) d\phi_{d+1} \right)^{D-d} \\
= & \; Cor(r, C_1, \cdots, C_n)
\end{aligned}
$$

Training the model with sub data sets

Uncorrected Method

The posterior of $r, \phi_1, \cdots, \phi_d, C_1, \cdots, C_n$ is proportional to

$$\prod_{c=0}^{1} \prod_{j=1}^{d} \frac{\prod_{k=0}^{I_j^{(c)}} (k + r\phi_j) \prod_{k=0}^{O_j^{(c)}} (k + r(1-\phi_j))}{\prod_{k=0}^{N^{(c)}} (k + r)} \cdot \text{Prior}(r, \phi_1, \cdots, \phi_d, C_1, \cdots, C_n)$$

- Using Metropolis-Hasting update r, and ϕ_j, the posterior of which are proportional to the joint distribution
- $P(C_i = c | \cdots) \propto \frac{\alpha_0 + n_{-i}^{(c)}}{\alpha_0 + \alpha_1 + n - 1} \cdot \prod_{j=1}^{d} \text{Bernolli}(X_{ij}; \frac{r\phi_j + I_{j,-i}^{(c)}}{r + n_{-i}^{(c)}})$, where, the definitions of $I_{j,-i}^{(c)}$ and $n_{-i}^{(c)}$ are as before, except that ith observation is not counted in.

Corrected Method

The posterior of r and $C_1, \cdots, C_n$ is multiplied by $Cor(r, C_1, \cdots, C_n)$.

Predicting the response

First we can obtain that

$$P(X_{*1},\cdots,X_{*d}\mid X_{1,:},\cdots,X_{n:},\phi_j's,C_i's,r,r,C_*=c)=\prod_{j=1}^{d}\text{Bernolli}(X_{*j},\frac{r\phi_j+I_j^{(c)}}{r+n^{(c)}})$$

Then it is easy to get the predictive probability with above equation:

$$P(X_{*1}=1\mid X_{*2},\cdots,X_{*d},X_{1,:},\cdots,X_n)=\frac{P(X_{*1}=1,X_{*2},\cdots,X_{*d}\mid X_{1,:},\cdots,X_{n,:})}{P(X_{*2},\cdots,X_{*d}\mid X_{1,:},\cdots,X_{n,:})}$$

To calculate the numerator,

$$\begin{aligned}&P(X_{*1}=1,X_{*2},\cdots,X_{*d}\mid X_{:,1},\cdots,X_n)\\ =\ &\sum_{c=0}^{1}\int P(X_{*1}=1,X_{*2},\cdots,X_{*d}\mid X_{1,:},\cdots,X_{n,:},\Phi_d,C_*=c)\frac{n^{(c)}+\alpha_c}{n+\alpha_0+\alpha_1}\cdot P(\Phi_d\mid X_{1,:},\cdots,X_{n,:})\,d\Phi_d\\ =\ &\sum_{c=0}^{1}\int\prod_{j=1}^{d}\text{Bernolli}(X_{*j},\frac{r\phi_j+I_j^{(c)}}{r+n^{(c)}})\frac{n^{(c)}+\alpha_c}{n+\alpha_0+\alpha_1}P(\Phi_d\mid X_{1,:},\cdots,X_{n,:})\,d\Phi_d\end{aligned}$$

where $\Phi_d=\{r,\phi_1,\cdots,\phi_d,C_1,\cdots,C_n\}$
We use Markov chain samples to evaluate above integration. It is similar to calculate the denominator.

Calculating the Correction Factor

We need only evaluate

$$P(|\mathrm{Cor}(Z,x)| \leq \gamma | x, r, \phi, C_1, \cdots, C_n) \qquad (*)$$

which is the integrand in the correction factor, replacing the notation $X_{:,d+1}$ with Z, $X_{:,1}$ with x for simplicity.

Note that

$$\mathrm{Cor}(Z,x) = \frac{(0-\bar{x})N_{01} + (1-\bar{x})N_{11}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{N_{,1} - n^{-1}(N_{,1})^2}} \tag{1}$$

where $N_{01} = \sum_{i=1}^{n} I(x_i = 0, Z_i = 1)$, $N_{11} = \sum_{i=1}^{n} I(x_i = 1, Z_i = 1)$, $N_{,1} = N_{01} + N_{11} = \sum_{i=1}^{n} I(Z_i = 1)$.

Let $B_{\gamma,x} = \{(N_{01}, N_{11}) \mid |\mathrm{Cor}(Z;x)| \leq \gamma\}$

$$(*) = \sum_{(N_{01},N_{11}) \in B_{\gamma,x}} P(N_{01}, N_{11} \mid x, r, \phi, C_i, i = 1, \cdots, n)$$

Calculating the Correction Factor(Cont.)

Decomposing N_{01}, N_{11} into two parts by the group,

$$
\begin{aligned}
& P(N_{01}, N_{11} \mid x, r, \phi, C_i's) \\
= & \sum_{N_{01}^{(0)}+N_{01}^{(1)}=N_{01}, N_{11}^{(0)}+N_{11}^{(1)}=N_{11}} P(N_{01}^{(0)}, N_{11}^{(0)} \mid r, \phi, C_i = 0) \cdot P(N_{01}^{(1)}, N_{11}^{(1)} \mid r, \phi, C_i = 1)
\end{aligned}
$$

where $N_{01}^{(c)} = \sum_{i=1}^{n} I(x_i = 0, Z_i = 1, C_i = c), N_{11}^{(c)} = \sum_{i=1}^{n} I(x_i = 1, Z_i = 1, C_i = c)$

$$
\begin{aligned}
& P(N_{01}^{(c)}, N_{11}^{(c)} | C_i = c, r, \phi) \\
= & \left(\begin{array}{c} N_{0,}^{(c)} \\ N_{01}^{(c)} \end{array} \right) \left(\begin{array}{c} N_{1,}^{(c)} \\ N_{11}^{(c)} \end{array} \right) \frac{\prod_{k=0}^{N_{01}^{(c)}+N_{11}^{(c)}-1} (r\phi + k) \prod_{k=0}^{N^{(c)}-(N_{01}^{(c)}+N_{11}^{(c)})-1} (r(1-\phi)+k)}{\prod_{k=0}^{N^{(c)}-1} (r+k)}
\end{aligned}
$$

where $N_{0,}^{(c)} = \sum_{i=1}^{n} I(x_i = 0, C_i = c), N_{1,}^{(c)} = \sum_{i=1}^{n} I(x_i = 1, C_i = c)$

Simulation Studies

Two data sets were generated using the model described here. For each dataset, $\theta_1^{(0)} = 0.7, \theta_1^{(1)} = 0.3$, the number of variables is 10000, the training set has 50 cases from each group, the testing set has 300 cases from each group. The difference for two data sets is that $r = 400$ for one dataset, representing weak relationship between $X_{:,1}$ and $X_{:,2}, \cdots, X_{:,d}$ and $r = 10$ for the other dataset, representing strong replationship. 1, 9, 99, 999, 1999 variables are selected respectively to form 5 sub data sets.

We compare the corrected method and the uncorrected method by looking at how the predictive probabilities are close to the true proportions of “1” and by looking at posterior correlation between the labels and the response.

Simulation Result I: r =400

	1 Variables Selected						9 Variables Selected						99 Variables Selected					
	Uncorrected			corrected			Uncorrected			corrected			Uncorrected			corrected		
Category	NO.	Pred	Freq1	NO.	Pred	Freq1	NO.	Pred	Freq1	NO.	Pred	Freq1	NO.	Pred	Freq1	NO.	Pred	Freq1
0-0.1	0	-	-	0	-	-	132	0.06	0.48	0	-	-	236	0.02	0.47	0	-	-
0.1-0.2	0	-	-	0	-	-	10	0.12	0.30	0	-	-	40	0.16	0.50	0	-	-
0.2-0.3	0	-	-	0	-	-	60	0.25	0.45	0	-	-	10	0.25	0.20	0	-	-
0.3-0.4	340	0.35	0.47	0	-	-	106	0.33	0.50	39	0.38	0.51	28	0.35	0.57	10	0.37	0.50
0.4-0.5	0	-	-	340	0.48	0.47	2	0.40	0.50	293	0.45	0.47	14	0.46	0.57	169	0.47	0.44
0.5-0.6	0	-	-	260	0.50	0.50	12	0.56	0.50	239	0.53	0.50	23	0.55	0.17	421	0.51	0.50
0.6-0.7	260	0.66	0.50	0	-	-	110	0.67	0.46	29	0.62	0.52	15	0.65	0.53	0	-	-
0.7-0.8	0	-	-	0	-	-	38	0.72	0.50	0	-	-	19	0.75	0.53	0	-	-
0.8-0.9	0	-	-	0	-	-	23	0.89	0.43	0	-	-	33	0.86	0.52	0	-	-
0.9-1.0	0	-	-	0	-	-	107	0.94	0.54	0	-	-	182	0.97	0.53	0	-	-

999 Variables Selected						1999 Variables Selected						Complete Data Set		
Uncorrected			corrected			Uncorrected			corrected					
NO.	Pred	Freq1	NO.	Pred	Freq1	NO.	Pred	Freq1	NO.	Pred	Freq1	NO.	Pred	Freq1
295	0.01	0.40	0	-	-	306	0.01	0.39	0	-	-	0	-	-
21	0.15	0.33	0	-	-	22	0.14	0.41	0	-	-	0	-	-
12	0.26	0.25	0	-	-	11	0.24	0.64	0	-	-	0	-	-
8	0.34	0.38	0	-	-	17	0.35	0.29	0	-	-	0	-	-
9	0.45	0.33	600	0.49	0.49	17	0.46	0.65	600	0.49	0.49	600	0.49	0.49
9	0.56	0.44	0	-	-	8	0.53	0.25	0	-	-	0	-	-
14	0.65	0.50	0	-	-	15	0.65	0.60	0	-	-	0	-	-
10	0.75	0.50	0	-	-	5	0.76	0.40	0	-	-	0	-	-
19	0.84	0.58	0	-	-	21	0.86	0.67	0	-	-	0	-	-
203	0.99	0.64	0	-	-	178	0.98	0.64	0	-	-	0	-	-

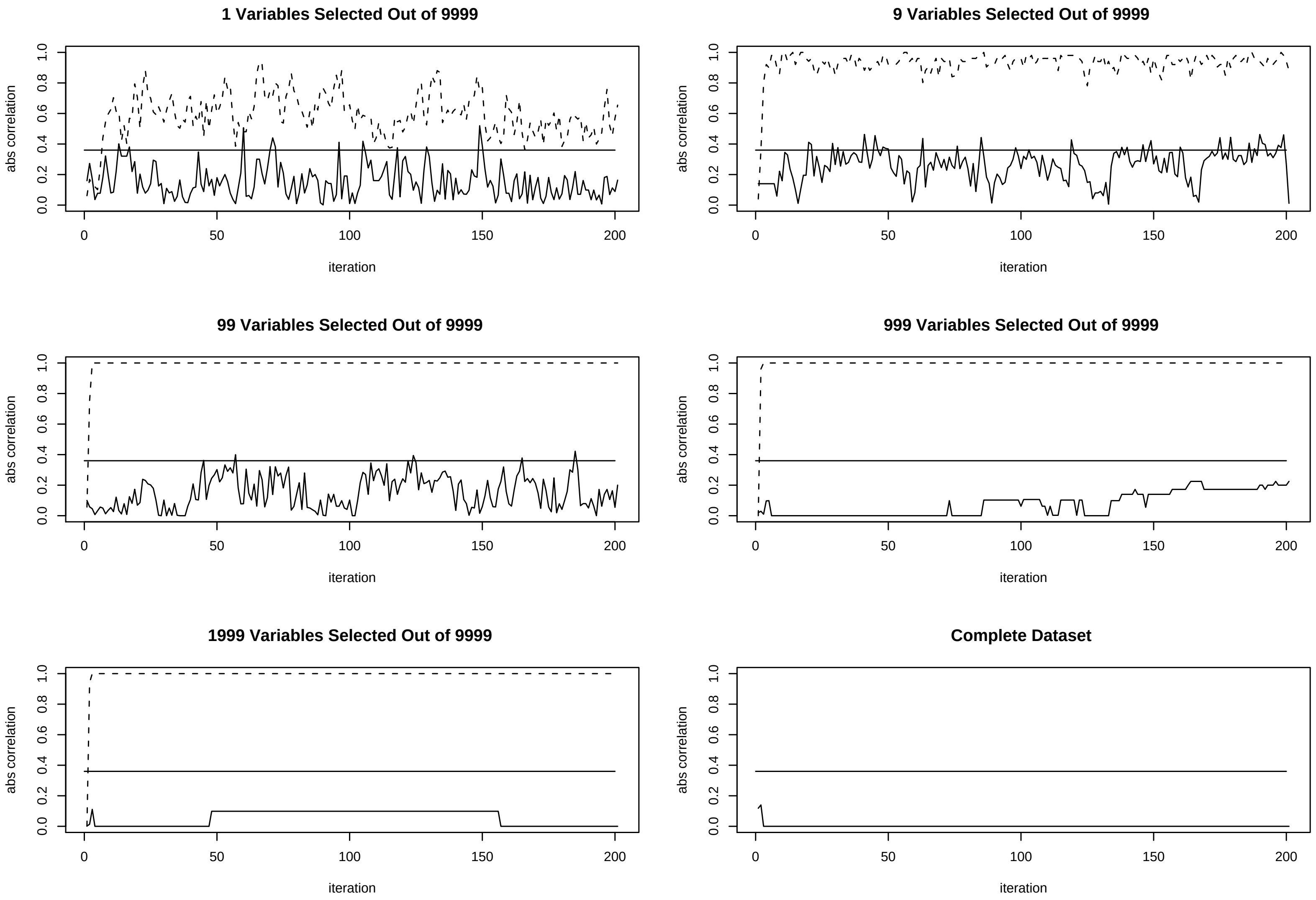


Figure 1: Plot of the abs correlations btw the MC samples of labels with the response

Simulation Result II: r =10

	1 Variables Selected						9 Variables Selected						99 Variables Selected					
	Uncorrected			corrected			Uncorrected			corrected			Uncorrected			corrected		
Category	NO.	Pred	Freq1	NO.	Pred	Freq1	NO.	Pred	Freq1	NO.	Pred	Freq1	NO.	Pred	Freq1	NO.	Pred	Freq1
0-0.1	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
0.1-0.2	0	-	-	0	-	-	240	0.14	0.35	0	-	-	0	-	-	0	-	-
0.2-0.3	262	0.27	0.40	0	-	-	38	0.23	0.53	170	0.27	0.35	300	0.27	0.32	0	-	-
0.3-0.4	0	-	-	0	-	-	33	0.34	0.45	107	0.34	0.40	0	-	-	300	0.34	0.32
0.4-0.5	0	-	-	0	-	-	30	0.45	0.47	62	0.46	0.44	0	-	-	0	-	-
0.5-0.6	0	-	-	600	0.50	0.51	23	0.55	0.52	45	0.55	0.58	0	-	-	0	-	-
0.6-0.7	338	0.70	0.58	0	-	-	18	0.64	0.61	80	0.65	0.60	300	0.70	0.69	300	0.66	0.69
0.7-0.8	0	-	-	0	-	-	36	0.75	0.58	136	0.74	0.73	0	-	-	0	-	-
0.8-0.9	0	-	-	0	-	-	62	0.86	0.61	0	-	-	0	-	-	0	-	-
0.9-1.0	0	-	-	0	-	-	120	0.93	0.74	0	-	-	0	-	-	0	-	-

999 Variables Selected						1999 Variables Selected						Complete Data Set		
Uncorrected			corrected			Uncorrected			corrected					
NO.	Pred	Freq1	NO.	Pred	Freq1	NO.	Pred	Freq1	NO.	Pred	Freq1	NO.	Pred	Freq1
0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
300	0.33	0.32	300	0.35	0.32	300	0.33	0.32	300	0.33	0.32	300	0.33	0.32
0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
300	0.67	0.69	300	0.65	0.69	300	0.68	0.69	300	0.67	0.69	300	0.67	0.69
0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
0	-	-	0	-	-	0	-	-	0	-	-	0	-	-

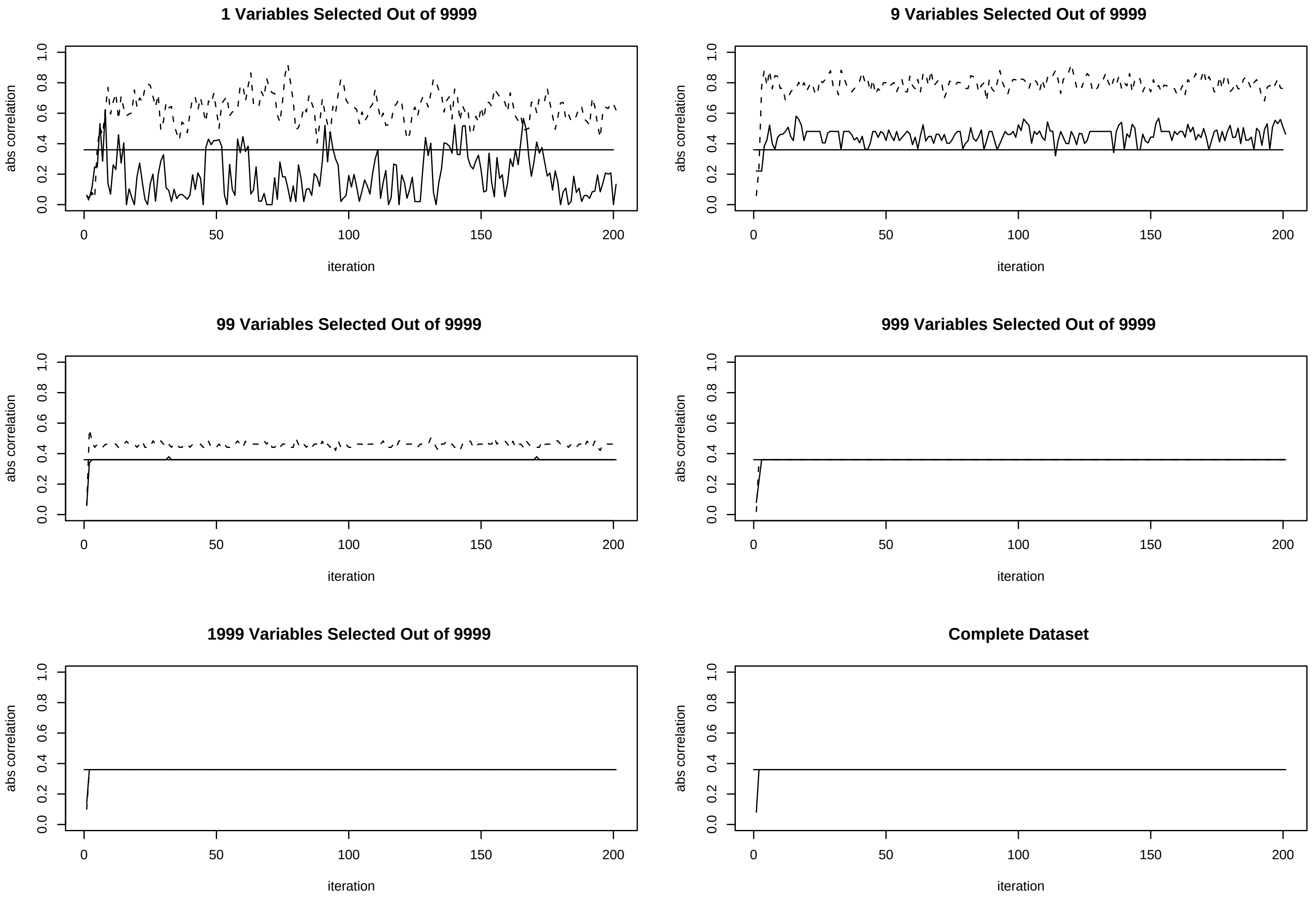


Figure 2: Plot of the abs correlations btw the MC samples of labels with the response

Comparison of Computational Time with R

NO. of Var.	1	9	99	999	1999	9999
Fraction	0.01%	0.1%	1%	10%	20%	100%
r=400	35828	34015	32996	3087	3130	11189
r=10	34532	13282	1356	2010	2959	11674

Table 1: Time for doing 200 iterations in Second: Corrected method on sub data sets and uncorrected method on complete data set

Conclusion and Discussion

- Our corrected method can correct the possible bias in all cases in this example. And it keeps the gain from selecting variables.
- When moderate fraction of variables,for example 10%,20%, are selected the computational time is much lower than working with complete dataset. Our corrected method would show more advantage when the uncorrected sampling is more dependent on the dimension of variables.
- The calculation of the correction factor is specific for this example. A more general and faster method is needed.