Predicting Health Care Costs by Two-part Model with Sparse Regularization

Atsuyuki Kogure
Keio University, Japan

July, 2015

Abstract

We consider the problem of predicting health care costs using the two-part model with the sparse regularization. The data we use are health care insurance claims data and health screening data of 10,000 individuals over the period from 2010 through 2012, randomly taken from several private health care plans in Japan. We construct a predictive model using the year 2011 health care cost variable as the response variable and all the year 2010 variables as explanatory variables. Then we measure the accuracy of the model for predicting the year 2012 health care costs by comparing the actual year 2012 health care cost data with the prediction from the model with the explanatory variables replaced by those of the year 2011. The results show that the sparse regularization techniques improves the prediction accuracy.

1 Introduction

We consider the problem of predicting health care costs using sparse regularized regression techniques. Our aim is to explore the possibility of the sparse techniques for improving the accuracy to predict an individual's cost in the next year when applied to the two-part model. The data we use are health care insurance claims data and health screening data of 10,000 individuals over the period from 2010 through 2012, which were randomly taken from several private health care plans in Japan[1]. For our purpose we estimate two-part models using the total medical expenditures in the year 2011 as a response with all the variables in the year 2010 as explanatory variables. There are 27 variables available in each year. To account for possible nonlinearities in the regression equations, we estimate the model with all the two-way interactions between the covariates having 378 explanatory variables in total.

We construct a predictive model using the year 2011 health care cost variable as the response variable and the year 2011 variables as explanatory variables. Then we measure the accuracy of the model for predicting the 2012 health care cost by comparing the actual 2012 health care cost data with the prediction based on the model with the explanatory variables replaced by those of the year 2011. The results show that the sparse techniques improve the prediction accuracy, similar to the findings reported in Loginov et. al (2013). However, we have applied only the basic form of the lasso regularization at present. Since the original lasso was proposed, a lot of advancements have been made including the elastic net (Zhou and Hastie, 2005), the adaptive lasso (Zhou, 2006) and the Bayesian lasso (Park and Cassela, 2008).

In spite of the improvement, we note that the prediction error is still large with the sparse regularization. This problem mainly arises from the difficulty in predicting the inpatient medical expenditure as the prediction error is considerably made smaller by excluding the inpatient medical expenditure from the total medical expenditure.

[1]The data were kindly supplied by Japan Medical Data Center (JMDC).

2 Two-part models

Health care insurance claims data take nonnegative values but have a substantial proportion of values at zero. Modeling such "zero-inflated" data is challenging and many methods have been proposed including the Tobit model (Tobin, 1958), the two-part model (Duan et al, 1983), the sample selection model (Heckman, 1979) and the Tweedie model (Jorgensen, 1987), among which we here choose the two-part model which seems thus far the most popular method for the modeling of health care expenditures.

2.1 Regression models

The two-part model uses two regression equations to separate the modeling of the insurance claim data into two part. The first part refers to whether the claim outcome is positive. Conditional on its being positive, the second part refers to its level. To be more specific, let Y_i denote the claim amount for individual i $(i = 1, 2, \ldots, n)$ and let $\boldsymbol{x}_i$ denote the vector of explanatory variables associated with it. Then the conditional distribution of Y_i given $\boldsymbol{x}_i$ is

$$f_Y(y_i; \boldsymbol{\theta}|\boldsymbol{x}_i) = \begin{cases} \Pr(Y_i = 0; \boldsymbol{\theta}_1|\boldsymbol{x}_i) & \text{if } y_i = 0 \\ f_Y(y_i; \boldsymbol{\theta}_2|\boldsymbol{x}_i, y_i > 0) \Pr(Y_i > 0; \boldsymbol{\theta}_1|\boldsymbol{x}_i) & \text{if } y_i > 0 \end{cases} \tag{1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ denotes a pair of the parameter vectors of the parameters in the first and the second parts.

To estimate $\Pr(Y_i > 0; \boldsymbol{\theta}_1|\boldsymbol{x}_i)$ for the first part in (1), we choose a logistic regression

$$\Pr(Y_i > 0; \boldsymbol{\theta}_1|\boldsymbol{x}_i) = \frac{1}{1 + \exp\{-\boldsymbol{\theta}_1'\boldsymbol{x_i}\}}.$$

Several specifications have been applied for estimating $f_Y(y_i; \boldsymbol{\theta}_2|\boldsymbol{x}_i, y_i > 0)$ for the second part in (1). Given $Y_i > 0$, one might simply assume the standard linear normal model

$$Y_i = \boldsymbol{\theta}_2'\boldsymbol{x}_i + \varepsilon_i,$$

where ε_i's are errors and assumed to be independent and identically distributed as a normal distribution with zero mean and a constant variance σ^2. As a way to alleviate a severe non-normality of the response variable, many authors turn to the log normal specification:

$$\log Y_i = \boldsymbol{\theta}_2'\boldsymbol{x_i} + \varepsilon_i$$

Some authors have pointed out a difficulty due to the transformation of Y_i and suggested to use a generalized linear model (GLM); see, for example, Blough, et al.(1999). A natural GLM specification may be to use the Gamma regression

$$f_Y(y_i; \boldsymbol{\theta}_2|\boldsymbol{x}_i, y > 0) = \frac{(s/\mu_i)^s}{\Gamma(s)} y^{s-1} \exp\left(-\frac{sy}{\mu_i}\right)$$

with the link function

$$\log(\mu_i) = \boldsymbol{\theta}_2'\boldsymbol{x}_i$$

where $\mu_i = \mathrm{E}[Y_i]$ and s is the shape parameter.

2.2 Prediction

Let $\hat{Y}$ denote a prediction of Y based on $\boldsymbol{x}$ and adopt the mean squared prediction error

$$\mathrm{E}\left[\left(Y-\hat{Y}\right)^2\right]$$

to measure the prediction accuracy. Then the optimal prediction is given by the the conditional mean of Y given $\boldsymbol{x}$. For the two-part model, it is given as

$$\begin{aligned}\mathrm{E}[Y;\boldsymbol{\theta}|\boldsymbol{x}] &= \mathrm{E}[Y;\boldsymbol{\theta}_2|Y>0,\boldsymbol{x}]\Pr(Y>0;\boldsymbol{\theta}_1|\boldsymbol{x}) + \mathrm{E}[Y;\boldsymbol{\theta}_2|Y=0,\boldsymbol{x}]\Pr(Y=0;\boldsymbol{\theta}_1|\boldsymbol{x}) \\ &= \mathrm{E}[Y;\boldsymbol{\theta}_2|Y>0,\boldsymbol{x}]\Pr(Y>0;\boldsymbol{\theta}_1|\boldsymbol{x}) \end{aligned} \tag{2}$$

For the normal case and the Gamma GLM cases, (2) becomes

$$\begin{aligned}\mathrm{E}[Y;\boldsymbol{\theta}|\boldsymbol{x}] &= \mathrm{E}[\boldsymbol{x}'\boldsymbol{\theta}_2+\varepsilon]\times\Pr(Y>0|\boldsymbol{x}) \\ &= \frac{\boldsymbol{x}'\boldsymbol{\theta}_2}{1+\exp\{-\boldsymbol{x}'\boldsymbol{\theta}_1\}}\end{aligned} \tag{3}$$

For the log normal case, (2) becomes

$$\begin{aligned}\mathrm{E}[Y;\boldsymbol{\theta}|\boldsymbol{x}] &= \mathrm{E}[\exp\{\boldsymbol{x}'\boldsymbol{\theta}_2\}\exp\{\varepsilon_i\}]\times\Pr(Y>0|\boldsymbol{x}) \\ &= \frac{\exp\{\boldsymbol{x}'\boldsymbol{\theta}_2+\sigma^2/2\}}{1+\exp\{-\boldsymbol{x}'\boldsymbol{\theta}_1\}}\end{aligned}$$

2.3 Estimation

In practice, the parameters must be estimated from the data. The standard practice for the estimation is to maximize the likelihood

$$\prod_{i=1}^{n} f_Y(y_i;\boldsymbol{\theta}|\boldsymbol{x_i}) = \prod_{i=1}^{n}\left[\Pr(Y_i>0;\boldsymbol{\theta}_1|\boldsymbol{x}_i)^{b_i}\Pr(Y_i=0;\boldsymbol{\theta}_1\boldsymbol{x}_i)^{1-b_i}\right]\prod_{i\in N_1} f_Y(y_i;\boldsymbol{\theta}_2|\boldsymbol{x}_i,y_i>0), \tag{4}$$

where b_i is defined as

$$b_i \equiv \begin{cases}1 & \text{if } y_i=0\\ 0 & \text{if } y_i>0\end{cases}$$

and N_1 denotes the set of observations with $y_i>0$. Thus the likelihood factors into two parts and estimation can be done separately for each likelihood.

3 Sparse regularized regression

Let p denote the number of the parameters in the model. Given a fixed parameter size p it is well known that the maximum likelihood estimation method in 2.3 provides an asymptotically optimal estimation as the sample size n becomes large. In practice, however, n is fixed, and the issue in over-fitting arises. Given a fixed sample size n, the fitting of the larger model always exceeds that of the smaller model.

One approach to dealing with this issue is to select a subset of explanatory variables that we believe to be related to the response variable. In this approach we fit a separate regression model for each possible combinations of the p explanatory variables and select a single best model using a criterion such as AIC or BIC from among the 2^p possibilities. However, this quickly becomes infeasible as p gets large (NP-hard problem).

Lasso (Tibshirani, 1996) is a relatively new approach to prevent such over-fitting by penalizing models with larger parameter values. It is a shrinkage regression method, but unlike the traditional shrinkage method such as the ridge regression, it makes some of the parameters become exactly zero, and thus is called a sparse regularization. With the property of sparsity, lasso is used as a selection method for explanatory variables in high dimensional data environment where the traditional methods such as AIC and BIC criterion suffer computational difficulty. See Vidaurre et. al. (2013) for a review on the lasso regularization.

We use the health care insurance claims data and health screening data with 10,000 individuals from 2010 to 2012. To estimate regression models for the insurance claim as the response variable, 27 covariates are available in each year. To account for possible nonlinearity, we incorporate all the two-way interactions between the variables into the regression equations, which inflates the number of explanatory variables to 378. To deal with such a high dimensionality, we utilize the lasso method for each model in the first and second parts separately.

The lasso regularization is applied by maximizing the objective function

$$\begin{aligned} l(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} \log f_Y(y_i;\boldsymbol{\theta}|\boldsymbol{x_i}) - \pi_1(\boldsymbol{\theta}_1) - \pi_2(\boldsymbol{\theta}_2) \\ &= \frac{1}{n}\sum_{i=1}^{n} \log\left[\Pr(Y_i>0;\boldsymbol{\theta}_1|\boldsymbol{x}_i)^{d_1}\Pr(Y_i=0;\boldsymbol{\theta}_1|\boldsymbol{x}_i)^{1-d_i}\right] - \pi_1(\boldsymbol{\theta}_1) \\ &\quad + \frac{1}{n}\sum_{i\in N_1} \log f_Y(y_i;\boldsymbol{\theta}_2|\boldsymbol{x}_i, y_i>0) - \pi_2(\boldsymbol{\theta}_2) \end{aligned} \tag{5}$$

where $\pi_i(\boldsymbol{\theta}_i)$ is a penalty term for $\boldsymbol{\theta}_i = \{\theta_j^{(i)}, j = 1,\ldots,p_i\}$ defined as

$$\pi_i(\boldsymbol{\theta}_i) = \lambda_i \sum_{j=1}^{p_i} |\theta_j^{(i)}|, \quad i = 1, 2.$$

The λ_i is a positive quantity to tune the regularization, and thus called a tuning parameter. Several methods have been proposed to select the tuning parameter. We use the cross validation techniques for the selection.

Since the original lasso was proposed, a lot of advancements have been made. The elastic net (Zou and Hastie, 2005) uses the penalty term

$$\pi(\boldsymbol{\theta}) = \lambda\left[\alpha\sum_{j=1}^{p}|\theta_j| + (1-\alpha)\sum_{j=1}^{p}\frac{\theta_j^2}{2}\right]$$

to overcome the situation where $p > n$. Here α is a value between 0 and 1. If it is 1, then the elastic net becomes the standard lasso. If it is 0, then the elastic net reduces to the ridge regression. The adaptive lasso (Zou, 2006) adopts the penalty term

$$\pi(\boldsymbol{\theta}) = \lambda\sum_{j=1}^{p} w_j|\theta_j|$$

with

$$w_j = 1/|\hat{\theta}_j|^{\gamma}, \gamma > 0,$$

where $\hat{\theta}_j$ is a consistent estimator for θ_j. It intends to ease the penalty on the parameters with large values of $\hat{\theta}_j$. Park and Casella (2008) noted that the penalty term in the lasso corresponds to the prior distribution

$$f(\boldsymbol{\theta}|\sigma,\lambda) = \prod_{j=1}^{p}\frac{\lambda}{2}\exp\left\{-\lambda\frac{|\theta_j|}{\sigma}\right\}$$

under the standard Bayesian framework and proposed a Bayesian procedure called the Bayesian lasso.

4 Analysis of Japanese Medical Data

In this section we present a simple data analysis. We estimated the two-part model of the logistic regression as the first part and the standard linear model as the second part with and without the lasso regularization. We considered two cases. For Case 1 we used 27 variables described in the next section as explanatory variables. For case 2 we used the 27 variables plus all the possible two-way interactions between them, which results in 378 explanatory variables in total.

4.1 Data

For each year from 2010 to 2012 we have the following observations on each individual:

- Demography variables
  sex (SEX), age (AGE)
- Health screening variables
  body mass index (BMI), systolic blood pressure (SBP), diastotic blood pressure (DBP), neutral (NF), HDL cholesterol (HDLC), LDL cholesterol (LDLC), glutamate oxaloacetic transaminase (GOT), glutamate pyruvate transaminase (GPT), Gamma-Glutamyl Transpeptidase (GGT), fasting blood sugar (FBS), hemoglobin A1c (HbA1c), urinal sugar (US), fasting blood sugar (FBS)
- Frequency and severity variables
  total Medical expenses (TME), inpatient medical expenses (IME), outpatient medical expenses (OME), pharmaceutical expenses (PE) hospital inpatient days (HID), outpatient visits (OV);
- Medical practice variables
  emergency medical care (EMC), cholesterol medication (CM), diabetes medication (DM), blood pressure medication (BPM)
- Disease Type variables
  hyperlipidemia (HL), diabetes (DB), high blood pressure (HBP), liver disease (LD), Highblood Pressure (HBP), Liver Desease (LD).

We omit the outpatient medical expenses (OME) variable from the estimation because there is a exact linear relationship among total medical expenses (TME), inpatient medical expenses (IME), outpatient medical expenses (OME), and pharmaceutical expenses (PE) such that

$$\text{TME} = \text{IME} + \text{OME} + \text{PE}.$$

The target for the prediction is the total medical expenses (TME). The summary statistics of the target variable in year 2010 to 2012 are given in Table 1.

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max	SD
TME.2010	0	5,590	29,820	91,920	92,950	12,570,000	278,740
TME.2011	0	8,868	38,350	108,600	115,000	9,389,000	291,642
TME.2012	0	9,178	39,640	117,100	120,800	6,767,000	308,818

Table 1: Summary statistics of total medical expenditures

4.2 Case 1

We use 27 variables described in the above section as explanatory variables.

4.2.1 Analysis without the lasso regularization

Table 2 shows the results of the logistic regression of a binary variable I_i

$$I_i = \begin{cases} 1 & \text{if TME.2011}_i > 0 \\ 0 & \text{if TME.2011}_i = 0 \end{cases}$$

on the 27 variables in year 2010. The marks on the far right of the table indicate levels of significance probabilities. Table 3 shows the results of the linear regression of $Y_i = \text{TME.2011}$ for TME.2011 > 0 on the 27 variables in year 2010.

Coefficients:	Estimate	Std. Error	z value	p values	
(Intercept)	8.203e-01	4.712e-01	1.741	0.08173	.
SEX	3.192e-01	7.803e-02	4.091	4.30e-05	***
AGE.2010	-1.660e-03	3.639e-03	-0.456	0.64828	
BMI.2010	1.069e-02	1.128e-02	0.948	0.34335	
SBP.2010	-2.718e-03	3.325e-03	-0.818	0.41364	
DBP.2010	-1.160e-03	4.548e-03	-0.255	0.79871	
NF.2010	6.820e-04	4.642e-04	1.469	0.14177	
HDLC.2010	4.299e-04	2.304e-03	0.187	0.85196	
LDLC.2010	-4.030e-04	1.057e-03	-0.381	0.70290	
GOT.2010	1.048e-02	6.380e-03	1.643	0.10043	
GPT.2010	-3.434e-03	3.438e-03	-0.999	0.31788	
GGT.2010	-2.794e-04	9.099e-04	-0.307	0.75877	
FBS.2010	-3.641e-03	2.822e-03	-1.290	0.19701	
HbA1c.2010	-9.139e-03	8.351e-02	-0.109	0.91286	
US.2010	-2.382e-01	1.308e-01	-1.820	0.06869	.
TME.2010	1.669e-06	2.274e-06	0.734	0.46297	
IME.2010	-1.209e-07	2.611e-06	-0.046	0.96307	
PE.2010	2.499e-05	6.094e-06	4.101	4.12e-05	***
HID.2010	-6.387e-02	3.171e-02	-2.014	0.04400	*
OV.2010	2.676e-01	2.297e-02	11.654	$<$ 2e-16	***
EMC.2010	-1.537e+00	6.556e-01	-2.344	0.01909	*
CM.2010	3.914e-01	3.902e-01	1.003	0.31581	
DM.2010	1.236e+00	1.067e+00	1.158	0.24674	
BPM.2010	-3.584e-01	5.366e-01	-0.668	0.50416	
HL.2010	1.234e-01	2.378e-01	0.519	0.60374	
DB.2010	4.096e-01	3.245e-01	1.262	0.20686	
HBP.2010	1.662e+00	5.565e-01	2.987	0.00282	**
LD.2010	5.895e-02	2.292e-01	0.257	0.79701	

Table 2: Coefficients of the Logistic Regression

We have constructed a predictive model using the year 2011 health care cost variable as the response variable and the year 2011 variables as explanatory variables. Then we measure the accuracy of the model for predicting the 2012 health care cost by comparing the actual 2012 health care cost data and the predicted values based on the constructed model with the explanatory variables replaced by those of year 2011.

Coefficients:	Estimate	Std. Error	z value	p values	
(Intercept)	-3.117e+04	3.892e+04	-0.801	0.423204	
SEX	1.838e+02	6.109e+03	0.030	0.975992	
AGE.2010	5.241e+02	3.061e+02	1.713	0.086833	.
BMI.2010	-1.017e+03	8.889e+02	-1.144	0.252831	
SBP.2010	-1.823e+00	2.694e+02	-0.007	0.994602	
DBP.2010	5.847e+02	3.708e+02	1.577	0.114886	
NF.2010	4.158e+00	3.532e+01	0.118	0.906311	
HDLC.2010	-2.138e+01	1.853e+02	-0.115	0.908161	
LDLC.2010	6.917e+01	8.535e+01	0.810	0.417711	
GOT.2010	-8.118e+01	3.668e+02	-0.221	0.824828	
GPT.2010	-3.284e-01	2.323e+02	-0.001	0.998872	
GGT.2010	4.236e+01	6.740e+01	0.629	0.529690	
FBS.2010	-1.335e+02	2.193e+02	-0.609	0.542796	
HbA1c.2010	-3.617e+03	6.413e+03	-0.564	0.572715	
US.2010	4.335e+04	8.079e+03	5.366	8.26e-08	***
TME.2010	1.211e+00	1.968e-02	61.553	<2e-16	***
IME.2010	-1.149e+00	2.912e-02	-39.447	<2e-16	***
PE.2010	-5.253e-02	4.126e-02	-1.273	0.202995	
HID.2010	7.246e+02	1.097e+03	0.660	0.509093	
OV.2010	-9.121e+02	2.647e+02	-3.446	0.000572	***
EMC.2010	-4.385e+04	3.857e+04	-1.137	0.255718	
CM.2010	1.375e+04	1.203e+04	1.143	0.252962	
DM.2010	-1.145e+04	1.740e+04	-0.658	0.510541	
BPM.2010	5.232e+04	1.635e+04	3.200	0.001382	**
HL.2010	-1.443e+04	1.074e+04	-1.344	0.179021	
DB.2010	1.810e+04	1.208e+04	1.499	0.133893	
HBP.2010	-3.129e+04	1.637e+04	-1.911	0.056025	.
LD.2010	-2.184e+04	9.741e+03	-2.242	0.024985	*

Table 3: Coefficient Estimates of linear regression

We have calculated the predicted values of TME.2012 based on the conditional mean (3) using the estimated parameter values and year 2011 explanatory variables. In case a predicted value is negative, we converted it to zero. The summary statistics are given in Table 4. To measure the prediction accuracy we adopt the mean squared prediction error

$$\text{MSPE} = \frac{1}{10000} \sum_{i=1}^{10000} (\text{TME.2012}_i - \text{PRED.2012}_i)^2,$$

which is calculated as $(235623.3)^2$.

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max
PRED.2012	0	28,920	62,950	127,300	156,600	7,005,000
TME.2012	0	9,178	39,640	117,100	120,800	6,767,000

Table 4: Summary statistics of PRED.2012 and TME.2012: case 1 without lasso

4.2.2 Analysis with the lasso regularization

We have estimated the logistic regression with the lasso regularization for the first part and the standard linear regression with the lasso regularization for the second part. The value of λ chosen by the cross validation for the logistic regression turned out very small (0.0004223471), which resulted in the same estimation results as the logistic regression without regularization. In contrast, the λ for the linear regression is considerable (2555.633), which made the estimates of many coefficients exactly zero as seen in Table 5. Figure 1 shows the mean squared error for different values of λ.

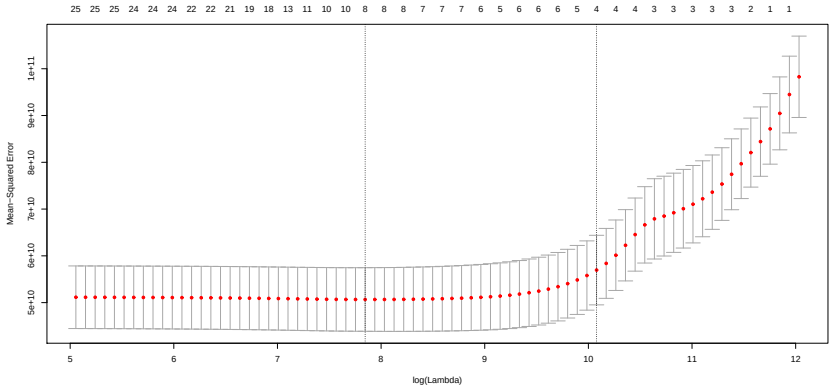

Figure 1: Mean squared error for different values of λ

We predict each individual's cost of year 2012 using the estimated model and compare the prediction errors. The summary statistics are given in Table 6. The root mean squared prediction error is calculated as 235016.2.

4.3 Case 2

We use all the two-way interactions between the variables into the regression equations, which inflates the number of covariates to 378.

4.3.1 Analysis without the lasso regularization

The R squared for the linear regression increased from 0.5012 to 0.5992 with the inclusion of the interaction terms. The summary statistics are given in Table 7. The root mean squared prediction error is calculated as 257820.7.

4.3.2 Analysis with the lasso regularization

We have used the logistic regression without the lasso regularization for the first part and the linear regression with the lasso regularization for the second part. The λ chosen by the cross validation is considerable (7974.432), which made 353 out of the 378 explanatory variables become exactly zero. Figure 2 shows the mean squared error for different values of λ.

We predict each individual's cost of year 2012 using the estimated model and compare the prediction errors. The summary statistics are given in Table 8. The root mean squared prediction error is calculated as 238522.5. .

Coefficients:	Estimate	z value	lasso Estimate
(Intercept)	-3.117e+04	-0.801	-3.594384e+04
SEX	1.838e+02	0.030	·
AGE.2010	5.241e+02	1.713	2.412720e+02
BMI.2010	-1.017e+03	-1.144	·
SBP.2010	-1.823e+00	-0.007	·
DBP.2010	5.847e+02	1.577	2.919444e+02
NF.2010	4.158e+00	0.118	·
HDLC.2010	-2.138e+01	-0.115	·
LDLC.2010	6.917e+01	0.810	·
GOT.2010	-8.118e+01	-0.221	·
GPT.2010	-3.284e-01	-0.001	·
GGT.2010	4.236e+01	0.629	·
FBS.2010	-1.335e+02	-0.609	·
HbA1c.2010	-3.617e+03	-0.564	·
US.2010	4.335e+04	5.366	3.227908e+04
TME.2010	1.211e+00	61.553	1.117933e+00
IME.2010	-1.149e+00	-39.447	-1.025502e+00
PE.2010	-5.253e-02	-1.273	·
HID.2010	7.246e+02	0.660	·
OV.2010	-9.121e+02	-3.446	·
EMC.2010	-4.385e+04	-1.137	·
CM.2010	1.375e+04	1.143	·
DM.2010	-1.145e+04	-0.658	·
BPM.2010	5.232e+04	3.200	1.767420e+04
HL.2010	-1.443e+04	-1.344	·
DB.2010	1.810e+04	1.499	·
HBP.2010	-3.129e+04	-1.911	·
LD.2010	-2.184e+04	-2.242	-8.465799e+03

Table 5: Comparison of estimates of linear regressions with and without the lasso regularization.

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max
PRED.2012	10140	27,760	63,360	127,100	157,400	6,678,000
TME.2012	0	9,178	39,640	117,100	120,800	6,767,000

Table 6: Summary statistics of PRED.2012 and TME.2012: case 1 with lasso

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max
PRED.2012	0	34,040	63,900	127,600	142,600	9,848,000
TME.2012	0	9,178	39,640	117,100	120,800	6,767,000

Table 7: Summary statistics of PRED.2012 and TME.2012: case 2 without lasso

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max
PRED.2012	0	36,620	68,170	125,400	145,400	9,669,000
TME.2012	0	9,178	39,640	117,100	120,800	6,767,000

Table 8: Summary statistics of PRED.2012 and TME.2012: case 2 with lasso

4.4 Comparison of prediction errors

Table 9 and 10 compare mean squared prediction errors between the models with and without the lasso regularization for Case 1($p = 27$) and Case 2 ($p = 387$). In the tables the prediction errors

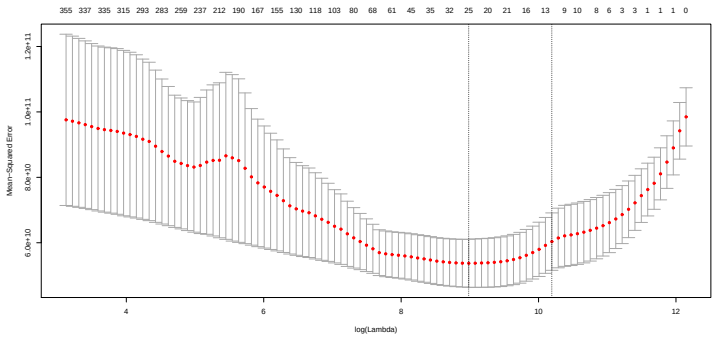

Figure 2: Mean squared error for different values of λ

for the ridge regression and the elastic net are also listed. We see that the lasso regularization improves the prediction accuracy most in either case.

	MSPE	$\frac{\text{MSPE}}{\text{Variance of TME.2012}}$
No regularization	$(235623.3)^2$	0.5821
lasso	$(235016.2)^2$	0.5791
ridge	$(237216.8)^2$	0.5900
elastic net with $\alpha = 0.5$	$(237789.4)^2$	0.5793

Table 9: Comparison of prediction errors for case 1

	MSPE	$\frac{\text{MSPE}}{\text{Variance of TME.2012}}$
No regularization	$(257820.7)^2$	0.6970
lasso	$(238522.5)^2$	0.5966
ridge	$(239440.3)^2$	0.6012
elastic net with $\alpha = 0.5$	$(239059.9)^2$	0.5985

Table 10: Comparison of prediction errors for case 2

In spite of the improvement, we note that the prediction error is still large with the sparse regularization. This problem mainly arises from the difficulty in predicting the inpatient medical expenditure. Table 11 shows that mean squared prediction errors of the TME minus IME using the two-part model with and without the sparse regularization. We note that the mean squared prediction error relative to the variance was made smaller considerably.

	MSPE	$\frac{\text{MSPE}}{\text{Variance of (TME.2012} - \text{IME.2012)}}$
No regularization	$(118107.4)^2$	0.3244078
lasso	$(101373.9)^2$	0.2389953

Table 11: Comparison of prediction errors of TME minus IME for Case 2

For the year 2011 data we note that only 4% of the IME values are positive, they tend to be rather large. We also note that the positive IME is strongly related to the the hospital inpatient days. We are currently developing an extended two-part model to predict the IME by taking the HID into consideration.

5 Conclusions

We have considered the problem of predicting health care costs using the two-part model with the sparse regression techniques. We have constructed a predictive model using the year 2011 health care cost variable as the response variable and the year 2011 variables as explanatory variables. Then we have measured the accuracy of the model for predicting the 2012 health care cost by comparing the actual 2012 health care cost data and the predicted values based on the constructed model with the explanatory variables replaced by those of year 2011. The results show that the sparse techniques improves the prediction accuracy.

References

Blough, K., Madden, C.W., and Hornbrook, M.C. (1999), Modeling risk using generalized linear models. *Journal of Health Economics*, **18**, 153-171.

Duan, N., Manning, W. G. Jr., Morris, C. N., and Newhouse, J.P. (1983), A comparison of alternative models for the demand for medical care (Corr: V2 P413). *Journal of Business and Economic Statistics*, **1**, 115-126.

Heckman, J. (1979), Sample selection bias as a specification error. Econometrica, **47**, 153-161.

Jorgensen, B. (1987), Exponential dispersion models. *Journal of the Royal Statistical Society*, Series B, **49**, 127-145.

Loginov et. al. (2013), Predictive Modeling in Healthcare Costs Using Regression Models, *ARCH 2013.1 Proceedings*.

Park, T. and Casella, G. (2008), The Bayesian lasso. *Journal of the American Statistical Association*, **103**, 681-686.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society*, Series B, **58**, 267-288

Tobin, J. (1958), Estimation of relationships for limited dependent variables. *Econometrica*, **26**, 24-36.

Vidaurre, D., Bielza, C. and Pedro Larranaga, P. (2013), A Survey of L1 Regression. *International Statistical Review*, **81**, 361-387

Zou, H. (2006), The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418-1429

Zou, H. and Hastie, H. (2005) Regularization and variable selection via the Elastic Net, *Journal of the Royal Statistical Society*, Series B, **67**, 301-320