Article from:

ARCH 2013.1 Proceedings

August 1- 4, 2012

Michael V. Loginov, Emily Marlow, Victoria Potruch

PREDICTIVE MODELING IN HEALTHCARE COSTS USING REGRESSION TECHNIQUES

Michael Loginov, Emily Marlow, Victoria Potruch

University of California, Santa Barbara

Introduction

- Building a model that predicts an individual's cost to an insurer

Introduction

- Building a model that predicts an individual's cost to an insurer
- Goal: Determine future healthcare costs using prior costs, demographics, and diagnoses

Introduction

- Goal: Determine future healthcare costs using prior costs, demographics, and diagnoses

- Accurate health insurance rate-setting

Introduction

- Goal: Determine future healthcare costs using prior costs, demographics, and diagnoses

- Accurate health insurance rate-setting

- Identify individuals for medical management

Introduction

- Goal: Determine future healthcare costs using prior costs, demographics, and diagnoses

- Accurate health insurance rate-setting

- Identify individuals for medical management

- Measure risk for fund transfer between insurers in new health insurance exchange after 2014

Data

- Data set of health insurance claims from 2008 to 2009
- 30,000 individuals
- 133 variables

Data

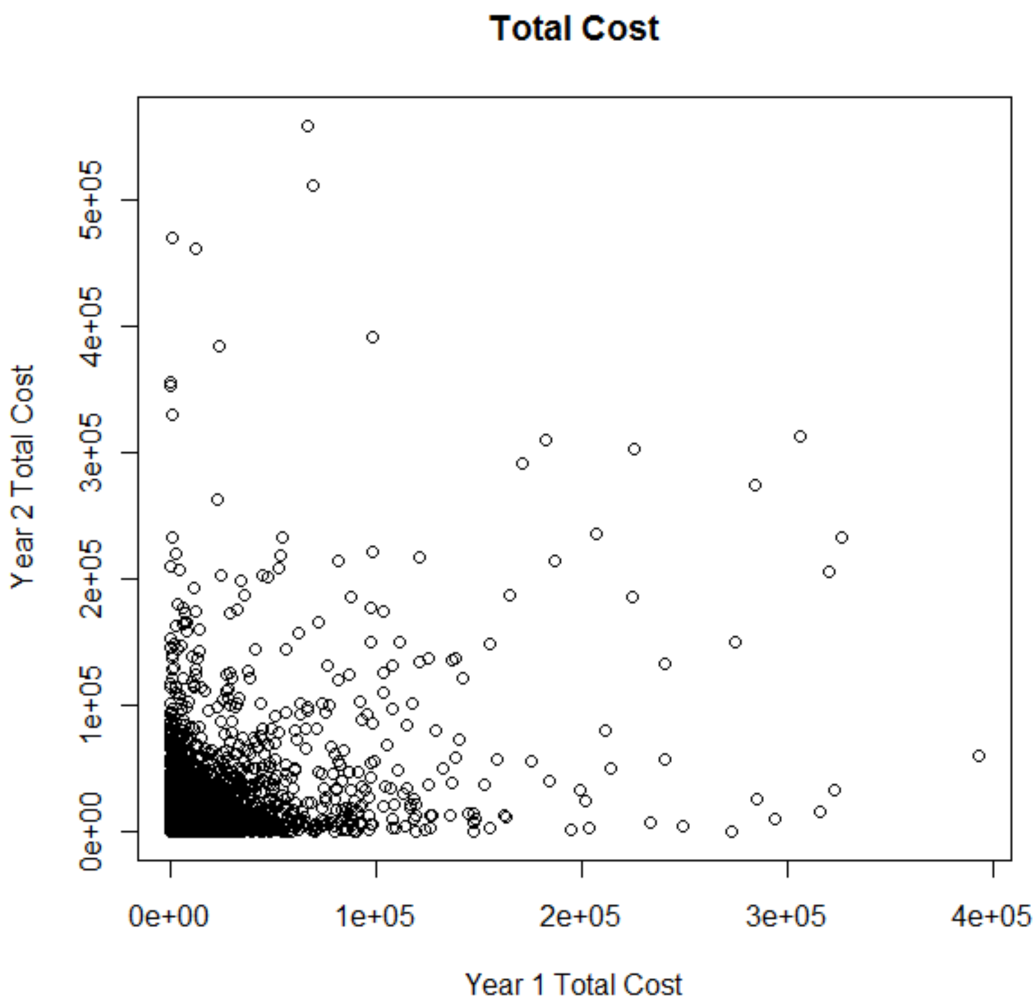
A1 | f_x | member_id

	U	V	W	X	Y	Z
1	gender	allow_current_rx	allow_current_ip	allow_current_op	allow_current_prof	allow_current_total
2	F	0	0	0	0	0
3	F	418.66	1945	177.73	5944.2	8485.59
4	M	0	0	1462.74	595.29	2058.03
5	M	0	0	592.79	447.23	1040.02
6	M	0	0	0	401.02	401.02
7	F	50.75	0	2754.42	823.55	3628.72
8	F	0	0	5108.5	2567.65	7676.15
9	M	0	0	0	250.43	250.43
10	M	0	0	2737.87	2984.43	5722.3
11	F	0	0	245.5	2524.75	2770.25
12	F	1053.74	0	0	701.12	1754.86
13	M	1537.39	65608	1696.6	19134.24	87976.23
14	M	0	0	1596.32	2623.7	4220.02
15	F	0	0	8894.3	6605.56	15499.86
16	F	0	0	258	199.42	457.42
17	F	0	0	0	257.82	257.82
18	F	520.12	0	0	181.7	701.82
19	F	306.93	0	0	101.6	408.53
20	F	1753.23	0	119.29	672.57	2545.09

Data

- Numeric variables: age, total cost, categorical costs
- Binary variables: flags for hospital and PCP visits, flags for HCCs
- String variables: gender, self funded or fully insured

Data

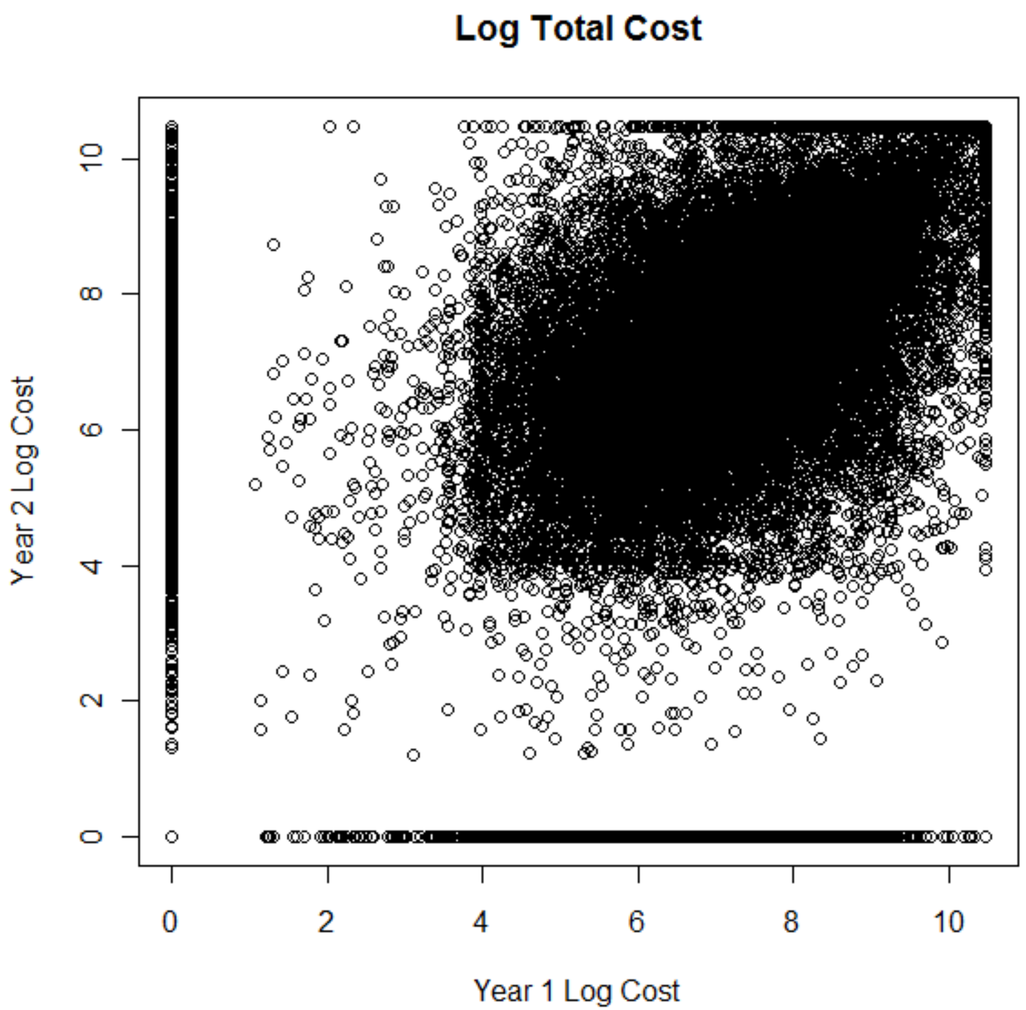


Total Cost
Year 2 Total Cost
Year 1 Total Cost
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05

Data

- Log transformation

Data



Data

- Log transformation
- Truncation

Data

- Log transformation
- Truncation
- Creation of “interaction” variables

Data

- Set of n=10,000 individuals is used to create the model
- Another sample of m=10,000 is used to test predictive power

Methods

- Linear regression: assume the data follows

$$y=\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + N(0,\sigma^2)$$

- y is an individual's log year 2cost
- x_k is the value of a parameter, such as age

Methods

- Linear regression: assume the data follows

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + N(0, \sigma^2)$$

- y is an individual's log year 2cost
- x_k is the value of a parameter, such as age
- Build a model by estimating the coefficients $\beta_1, \ldots, \beta_n$ and σ^2 with least squares estimates

Methods

- To reduce the number of predictors needed for the model we implement **Lars**, the use of **least angle regression** with the **least absolute shrinkage and selection operator**

Methods

- Least angle regression: creating a linear regression model one variable at a time
  - Standardize all variables
  - Choose the parameter that is most highly correlated with y, and perform simple linear regression with that one parameter

Methods

- Least angle regression: creating a linear regression model one variable at a time
  - Standardize all variables
  - Choose the parameter that is most highly correlated with y, and perform simple linear regression with that one parameter
  - Find the parameter most correlated with the residuals and repeat

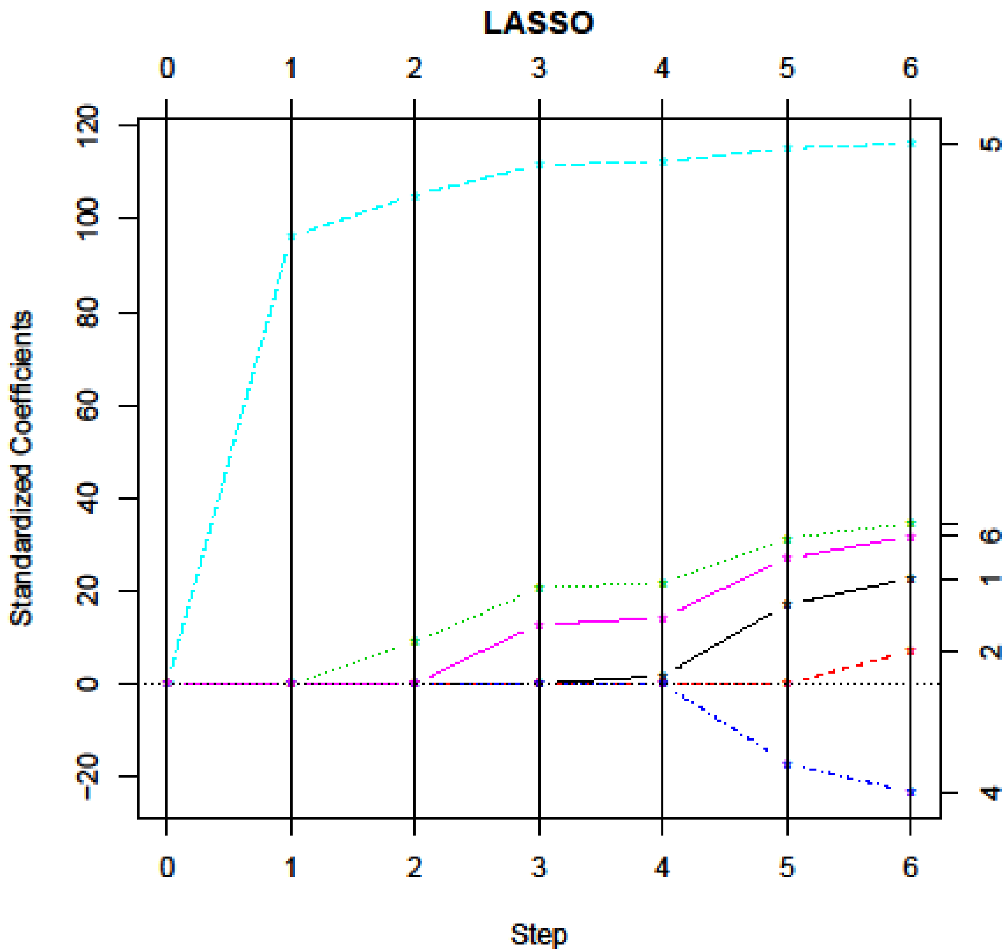
Methods

- Lasso uses a constraint λ on the sum of the standardized regression coefficients:

$$\text{Maximize } \sum(y-\hat{y})^2 \text{ subject to } \sum|\beta\sim| \leq \lambda$$

- $\hat{y}$ is the predicted value of y using the estimates of $\beta_1,\ldots,\beta_n$
- $\beta\sim$ coefficients are standardized
- λ is arbitrary

Methods



LASSO
0
1
2
3
4
5
6
120
100
80
60
40
20
0
-20
Standardized Coefficients
5
6
1
2
4
Step

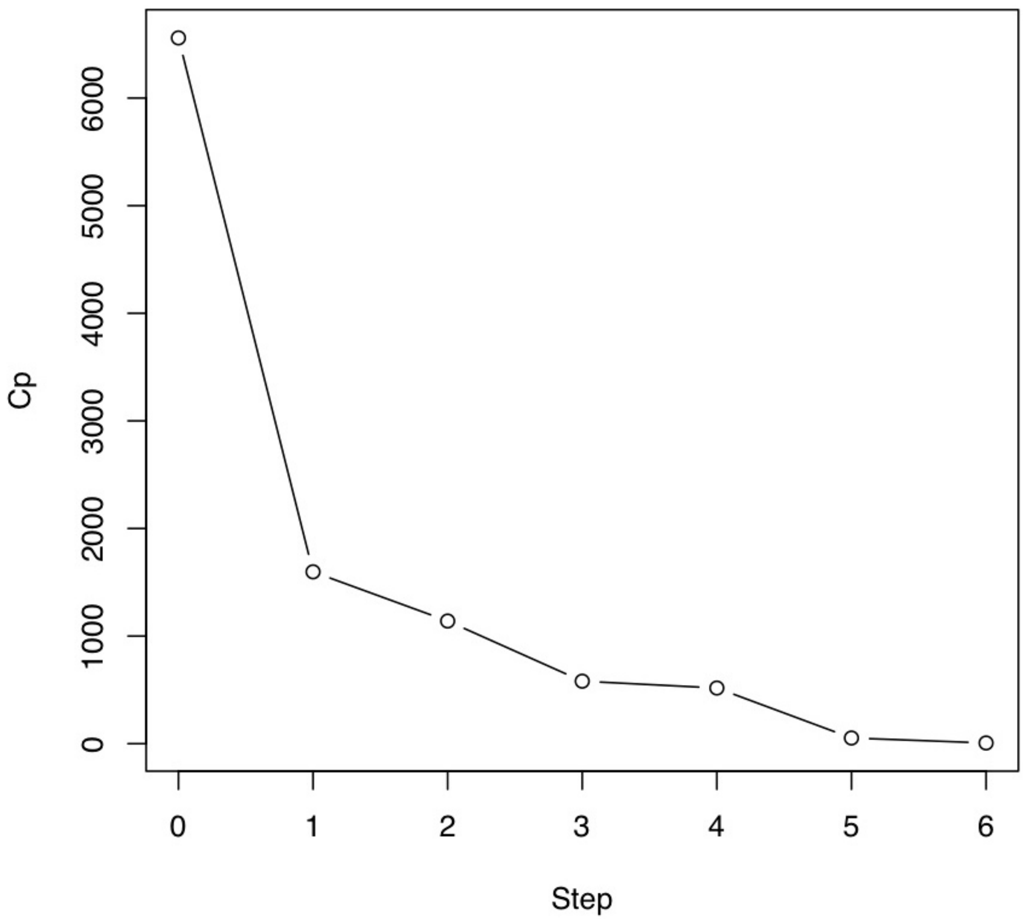
Methods

- Mallow's C_p statistic is used to choose k, the number of steps we take:

$$C_p = (1/\sigma\hat{}^2)\sum(y-\hat{y}_k)^2 - n + 2k$$

- We choose k such that C_p does not significantly decrease when k is increased

Methods



Methods

- Models are compared using **adjusted R^2** and **MSE**

$$Adjusted\ R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} * \frac{n-1}{n-k-1}$$

$$MSE = \frac{1}{m}\sum(y-\hat{y})^2$$

- Adjusted R^2 measures goodness-of-fit

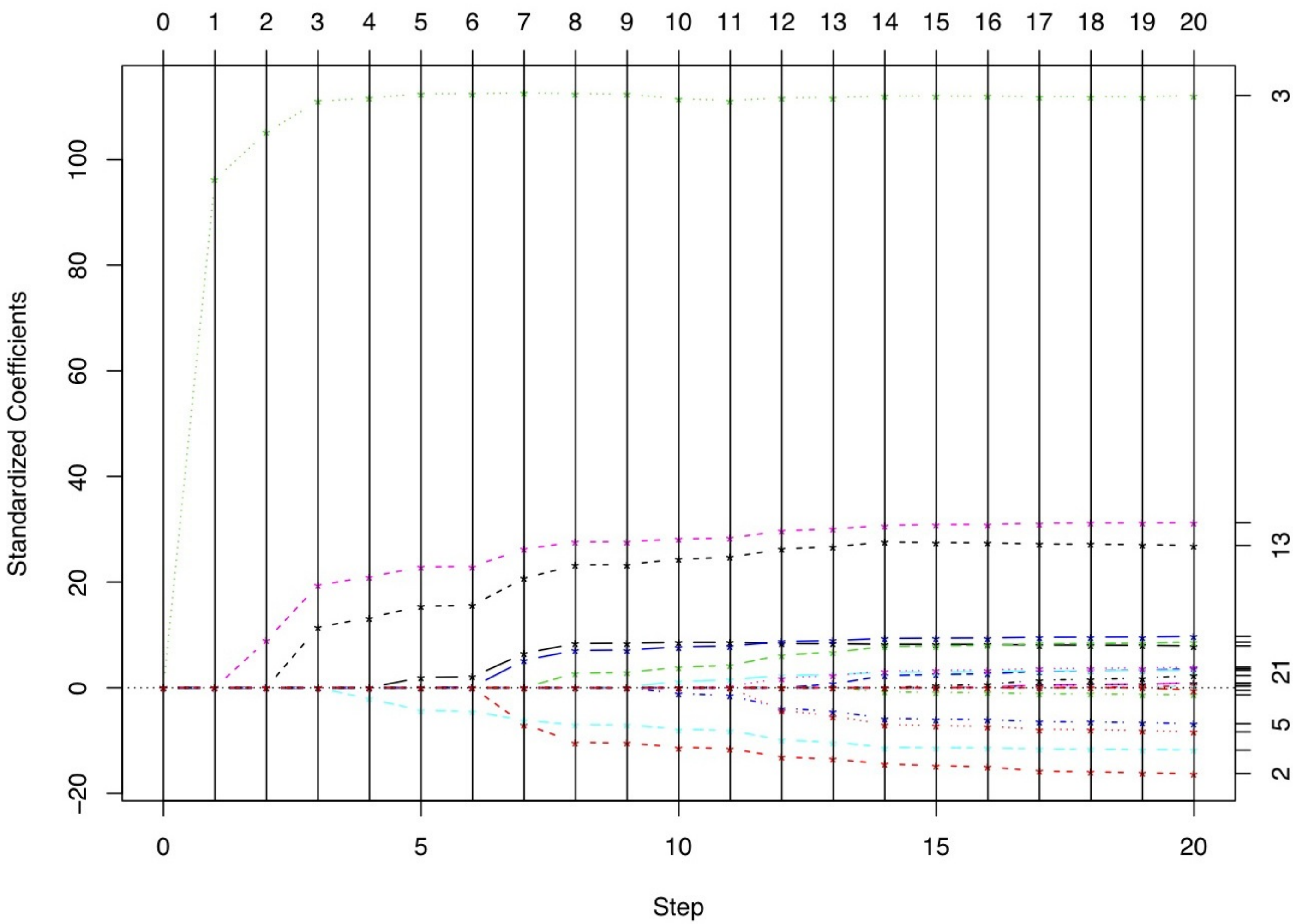
- MSE measures predictive power

Results

- Ran 4 models to compare

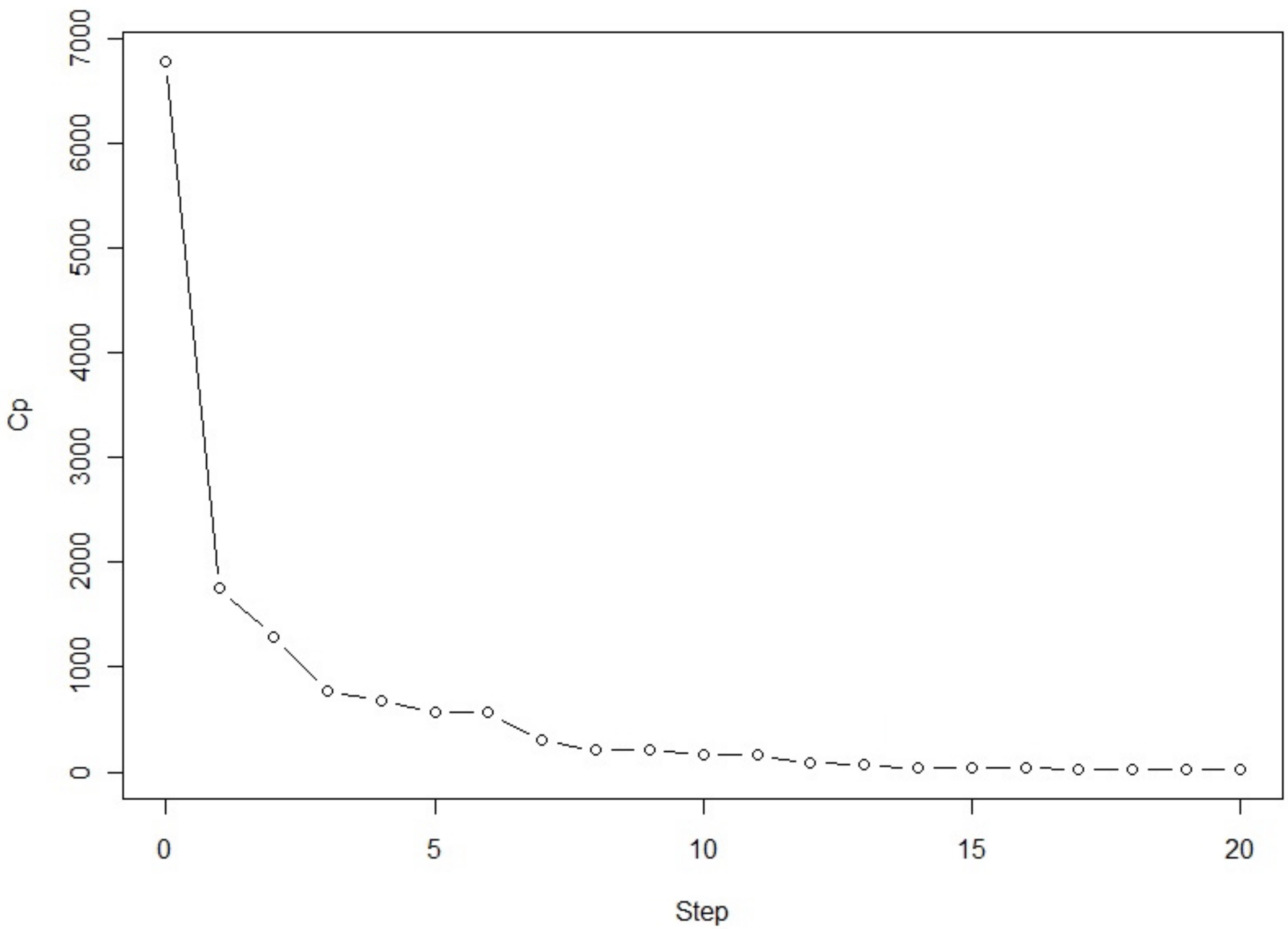
  - Model 1: Linear regression with age, gender, year 1 log cost
  - Model 2: Linear regression with all year 1 non-health data
  - Model 3: Linear regression with all data available in year 1
  - Model 4: Lars with all data available in year 1

Results



0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Standardized Coefficients
100
80
60
40
20
0
−20
0
5
10
15
20
Step
3
13
21
5
2

Results



Results

Model	Number of Variables	Adjusted R^2	MSE
Model 1	3	0.3721	6.1738
Model 2	31	0.4040	5.9146
Model 3	131	0.4069	5.8897
Model 4	13	0.4027	5.8492

- Models 3 and 4 are comparable
- Model 4 uses 118 less variables
- We use model 4 to draw conclusions

Results

Predictor	Effect on Cost
Age	+0.65% per year
Male Flag	-23.73%
Year 1 Cost	+51.24%
Male Age 15-24 Flag	-20.94%
Male Age 25-44 Flag	-23.78%
Year 1 Pharmacy Cost	+8.75%
Year 1 Inpatient Cost	-2.38%
Year 1 ER Visit Flag	+8.06%
Year 1 PCP Visit Flag	+6.66%
Year 1 PCP Visit Count	+6.47%
HCC 19: Diabetes	+28.83%
HCC 22: Metabolic/Endocrine	+22.23%
HCC 91 Hypertension	+6.36%

Acknowledgements

- In order to conduct this research we used the open source statistical software R with the package lars which includes LAR and lasso
- We used LATEX to produce our paper
- We would like to thank our faculty advisors, Ian Duncan, Raya Feldman, and Mike Ludkovski for their assistance, their guidance, and their enthusiasm for this research