Predictive Analysis of Hospital Costs: A Comparative Study of Statistical Learning Techniques

•••

Group 3
Lindsay Knupp, Rose Porta, Johnny Rasnic

Motivation

What influences a hospital’s total operating costs?

- Amount of employees? Location? Number of inpatients?
- Collected data from Centers for Medicare and Medicaid Services with over 5000 hospitals during 2020 fiscal year
- Performed analyses to assess predictors’ relative importance

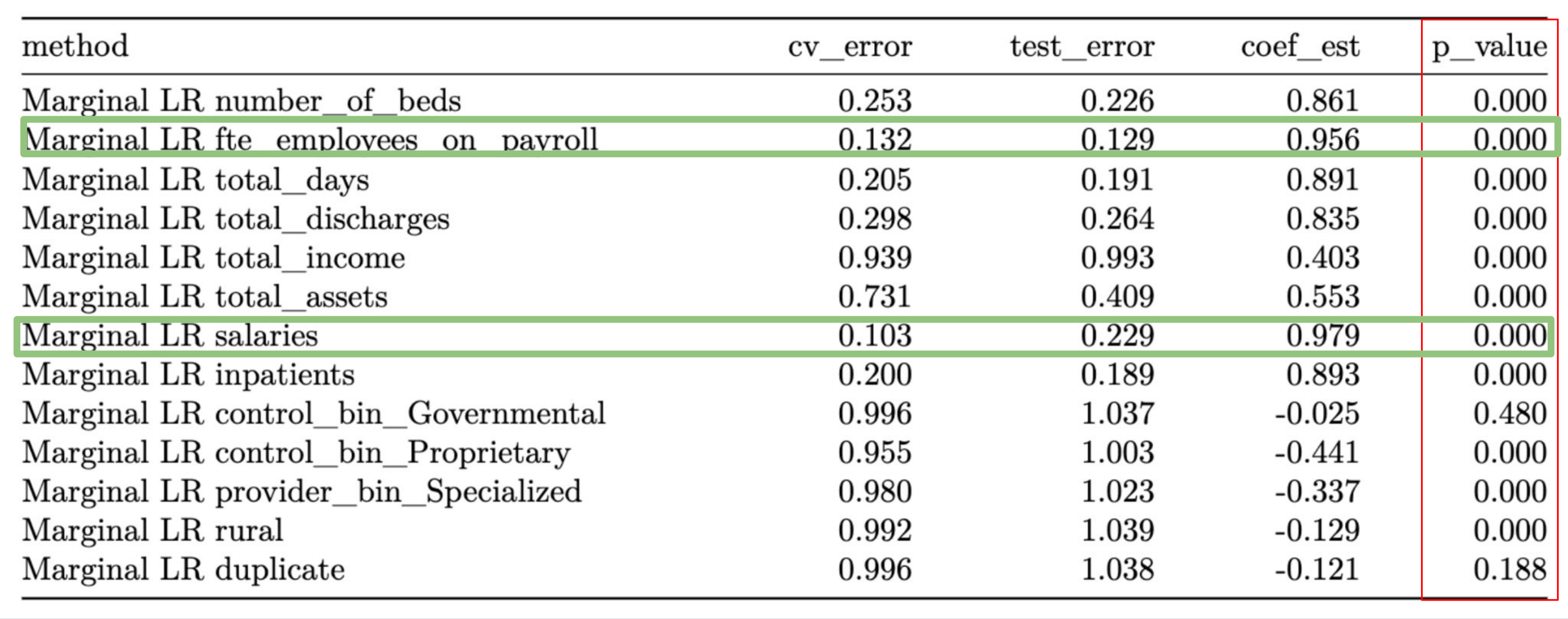
Exploratory Data Analysis

- Response total costs is very right-skewed
- Most relationships between predictors and response are linear
- Notable collinearity between several of the predictors
  - Example: salaries and number of employees

Quantitative Response Results

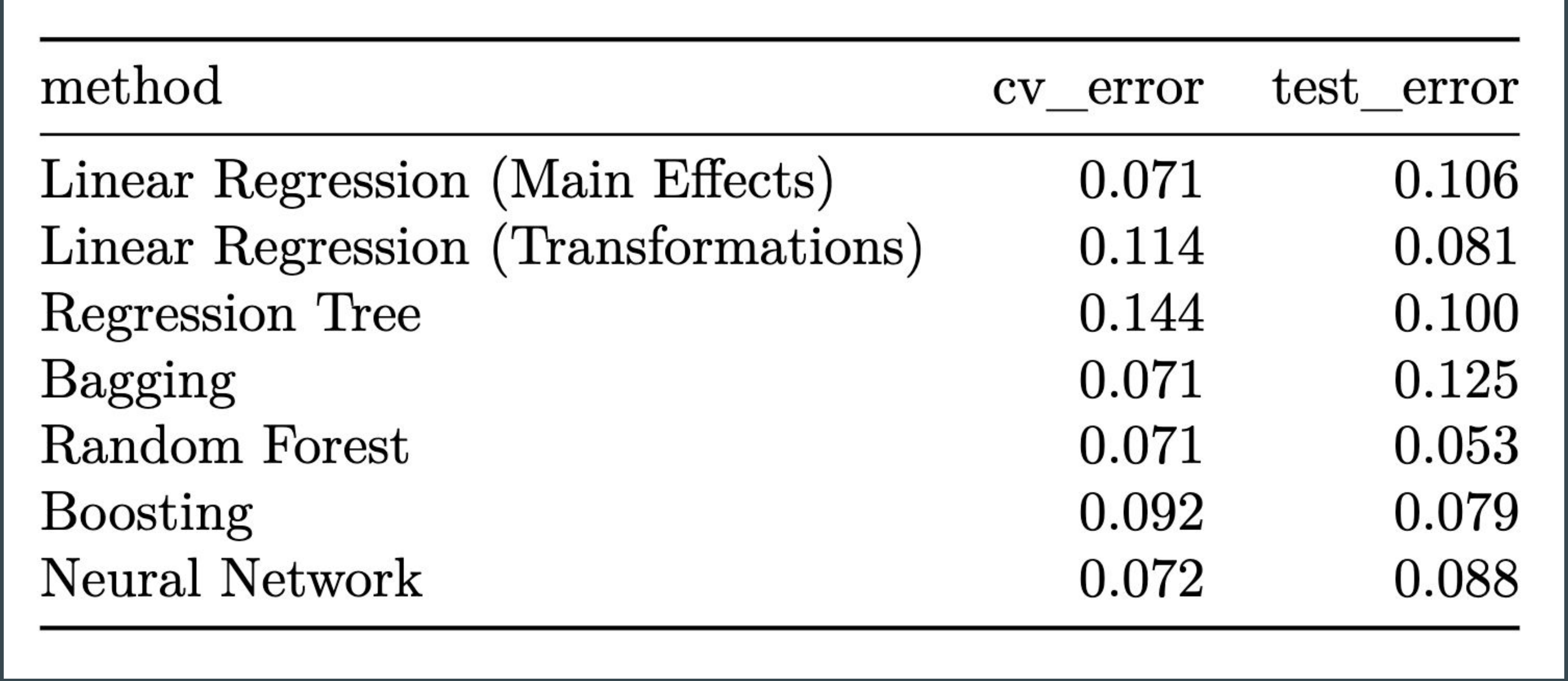
- Most Important Predictors: Salaries and Number of Employees
- However, each predictor is associated with the response
- Random Forest has lowest test MSEP, but test error is similar across all methods

Simple Linear Regression Results:



method	cv_error	test_error	coef_est	p_value
Marginal LR number_of_beds	0.253	0.226	0.861	0.000
Marginal LR fte_employees_on_payroll	0.132	0.129	0.956	0.000
Marginal LR total_days	0.205	0.191	0.891	0.000
Marginal LR total_discharges	0.298	0.264	0.835	0.000
Marginal LR total_income	0.939	0.993	0.403	0.000
Marginal LR total_assets	0.731	0.409	0.553	0.000
Marginal LR salaries	0.103	0.229	0.979	0.000
Marginal LR inpatients	0.200	0.189	0.893	0.000
Marginal LR control_bin_Governmental	0.996	1.037	-0.025	0.480
Marginal LR control_bin_Proprietary	0.955	1.003	-0.441	0.000
Marginal LR provider_bin_Specialized	0.980	1.023	-0.337	0.000
Marginal LR rural	0.992	1.039	-0.129	0.000
Marginal LR duplicate	0.996	1.038	-0.121	0.188

Comparison of MSE For All Methods

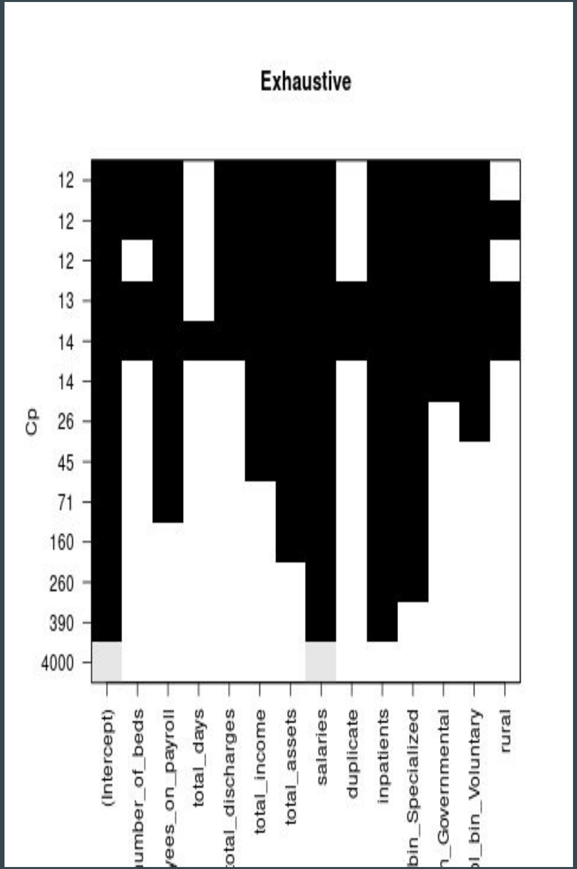


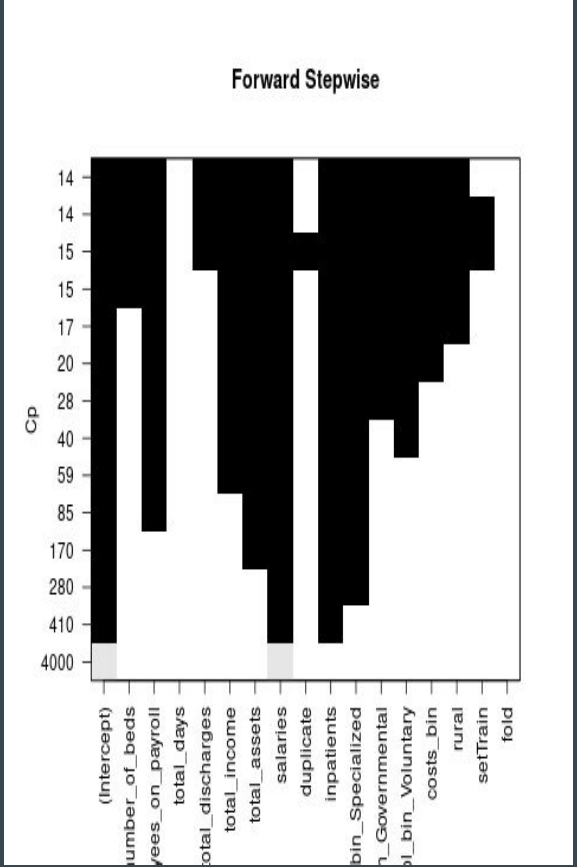
method	cv_error	test_error
Linear Regression (Main Effects)	0.071	0.106
Linear Regression (Transformations)	0.114	0.081
Regression Tree	0.144	0.100
Bagging	0.071	0.125
Random Forest	0.071	0.053
Boosting	0.092	0.079
Neural Network	0.072	0.088

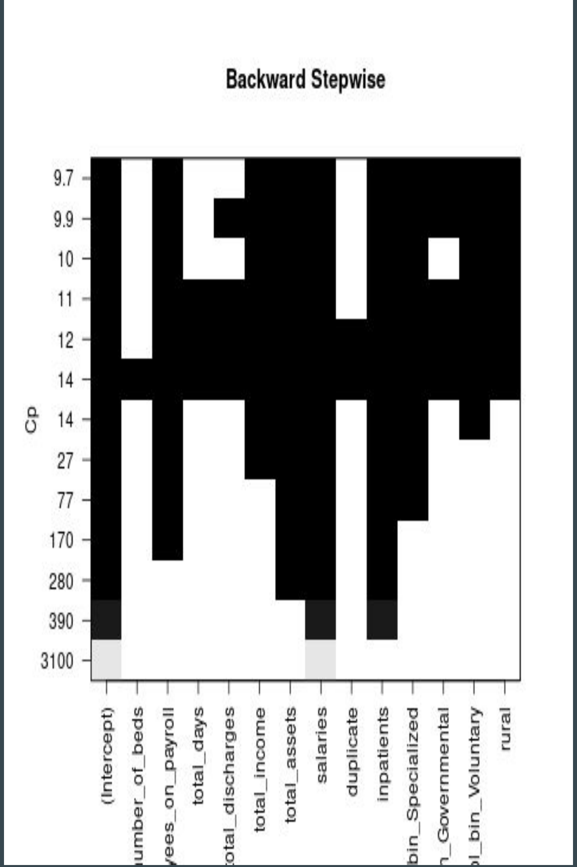
Variable Selection Results

We used the regsubsets library

- Exhaustive
- Forward stepwise
- Backward stepwise
- In all cases, we got 8 variables as the optimal model size according to cross-validated test MSE.

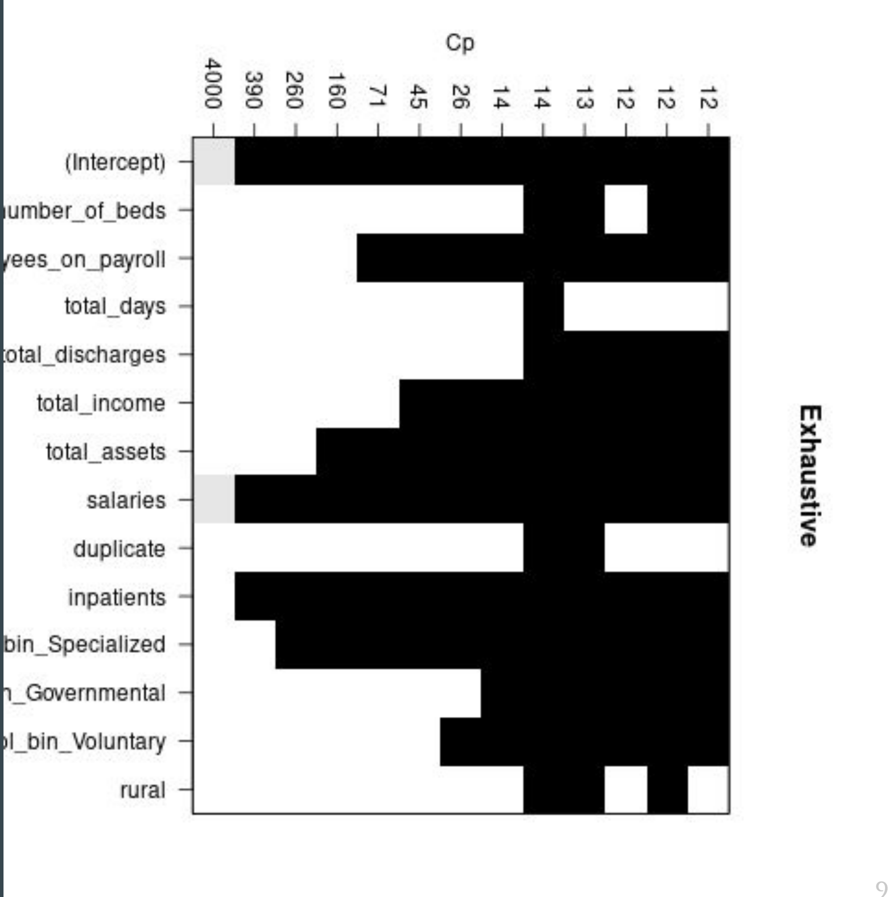

Exhaustive
Cp
12
12
12
13
14
14
26
45
71
160
260
390
4000
(Intercept)
umber_of_beds
yees_on_payroll
total_days
otal_discharges
total_income
total_assets
salaries
duplicate
inpatients
bin_Specialized
n_Governmental
ol_bin_Voluntary
rural


Forward Stepwise
Cp
14
14
15
15
17
20
28
40
59
85
170
280
410
4000
(Intercept)
umber_of_beds
yees_on_payroll
total_days
otal_discharges
total_income
total_assets
salaries
duplicate
inpatients
bin_Specialized
n_Governmental
ol_bin_Voluntary
costs_bin
rural
setTrain
fold

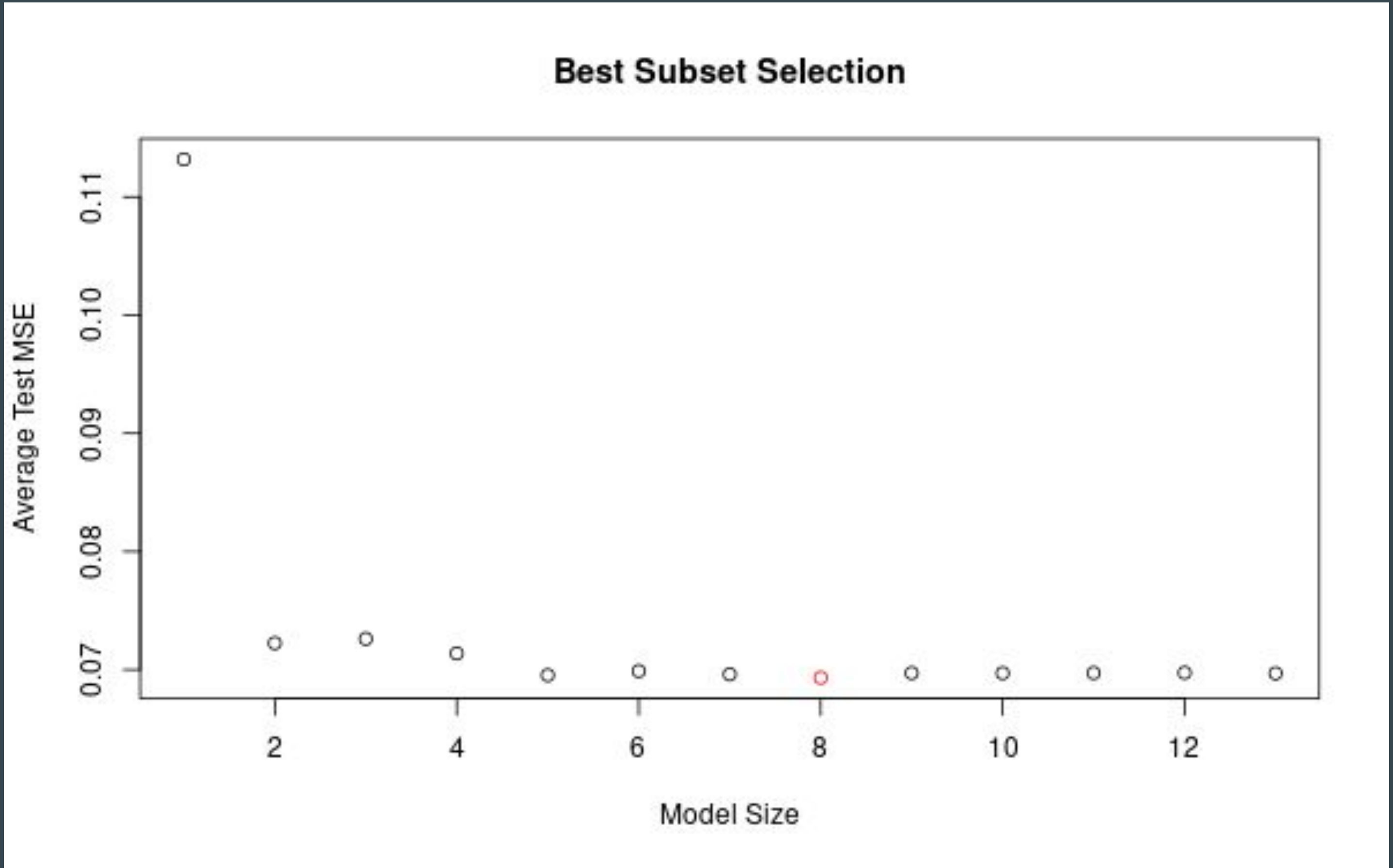

Backward Stepwise
Cp
9.7
9.9
10
11
12
14
14
27
77
170
280
390
3100
(Intercept)
umber_of_beds
yees_on_payroll
total_days
otal_discharges
total_income
total_assets
salaries
duplicate
inpatients
bin_Specialized
n_Governmental
ol_bin_Voluntary
rural

Exhaustive

- Salaries, inpatients, and the dummy variable for whether a hospital is specialized or not seem to play a major role.
- Conversely, total days and the duplicate dummy variable in our dataset did not seem to play a major role in our models.



10-fold cross validated best model by smallest test MSE



Lasso selection results

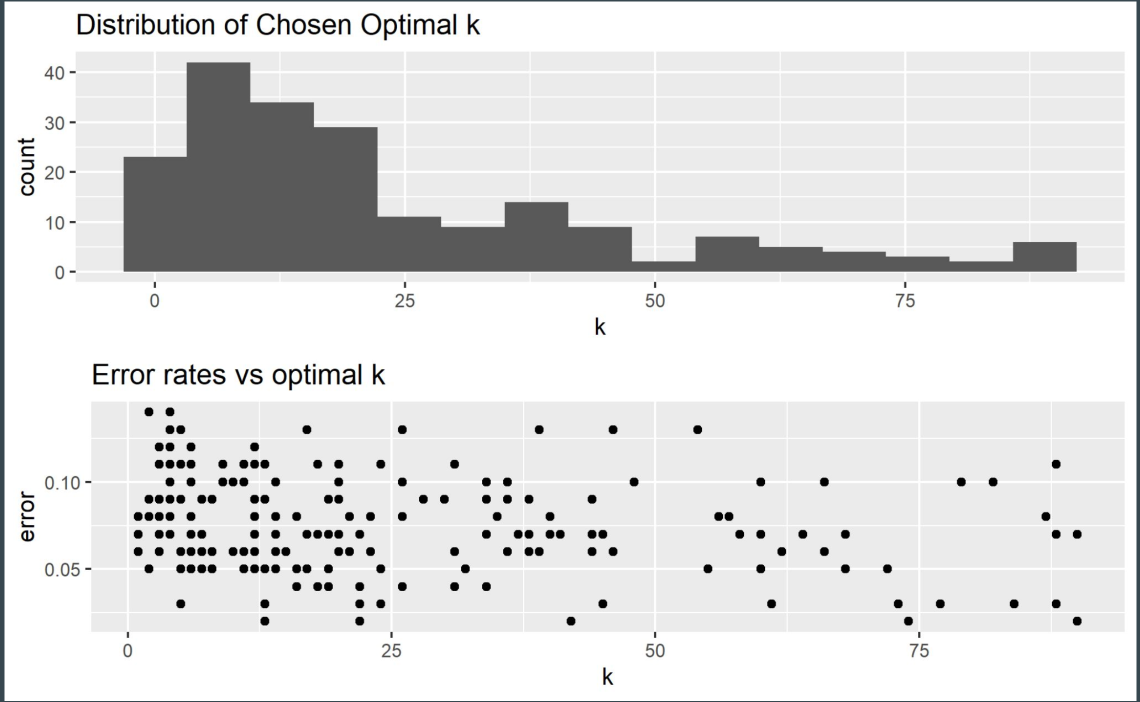
- Lasso selects 10 variables to be nonzero, dropping three in the process.
- This model size is somewhat close to the model size selected by our exhaustive and stepwise methods.
- The variables dropped are the duplicate dummy variable, the governmental dummy variable, and the number of beds.

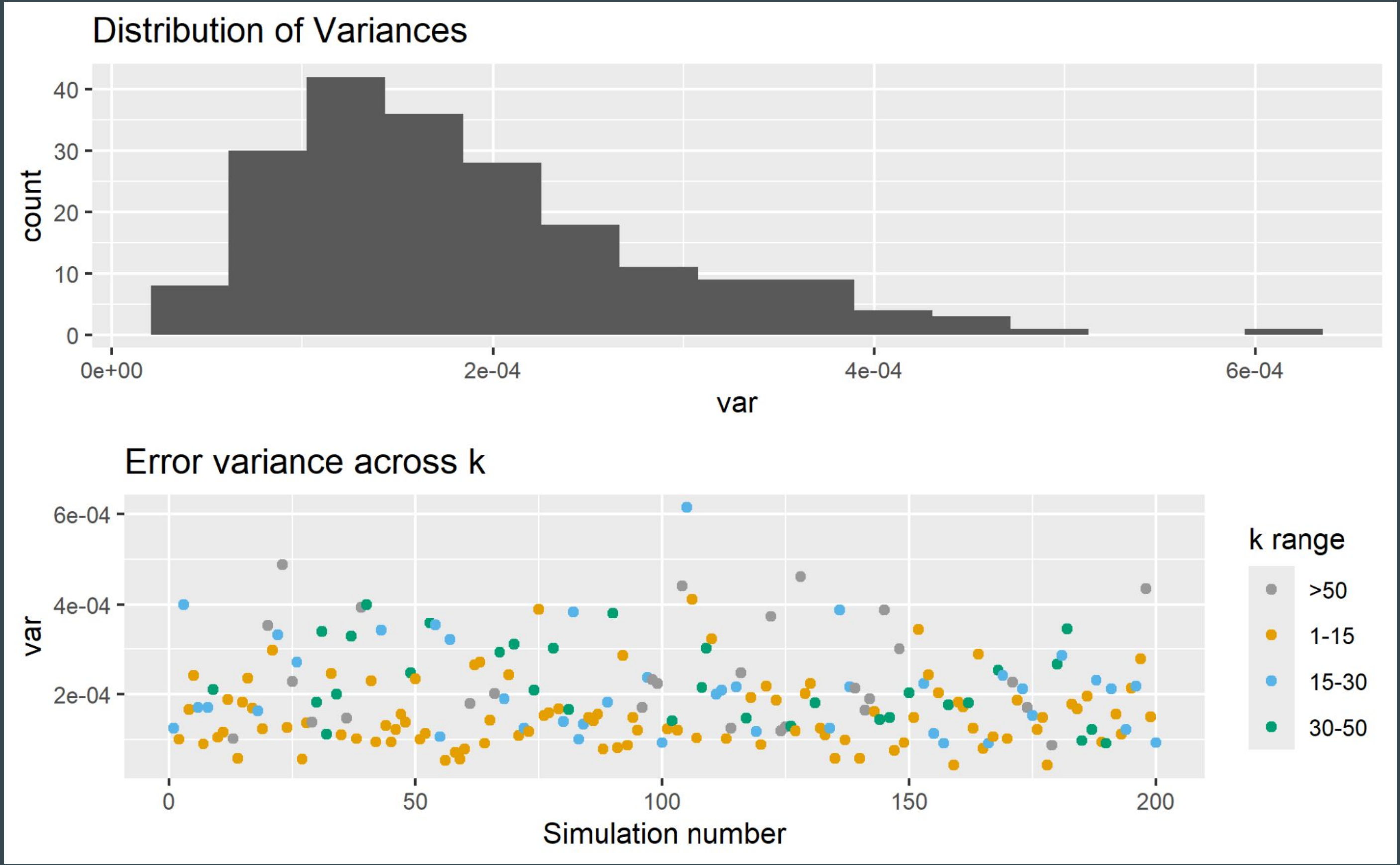
Qualitative Response Results

- Class labels: classified hospitals’ as above/below the median total costs
- Salaries consistently the most important predictor
- Multiple logistic regression, bagging, and random forest produced similar misclassification rates around 3%
- High false negative rates
  - Difficulty identifying hospitals’ with total costs above the median

Simulation Study

What is the optimal k in KNN with 200 simulated datasets?




Distribution of Variances
count
40
30
20
10
0
0e+00
2e-04
4e-04
6e-04
var
Error variance across k
var
6e-04
4e-04
2e-04
0
50
100
150
200
Simulation number
k range
>50
1-15
15-30
30-50

Conclusions

- Salaries is a very strong predictor for total costs, supported across many different analyses.
- Simple models don’t perform that much worse than the more complex models.
- For KNN, error rate does not change very much for different values of *k*; variance in optimal *k* is quite high.