Report Section

Johnny Rasnic

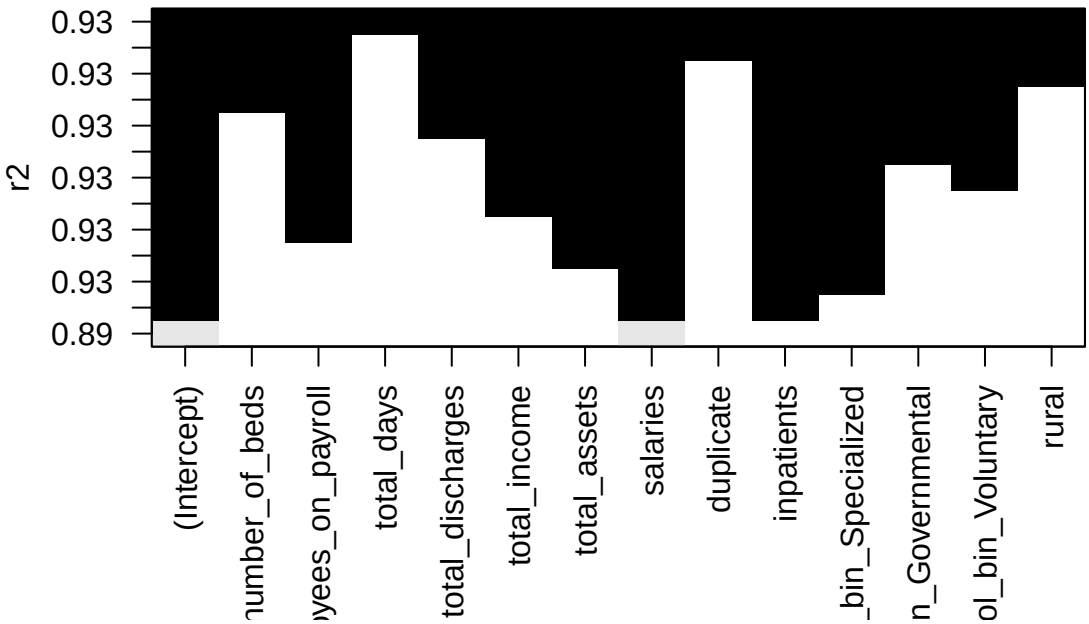
2024-05-12

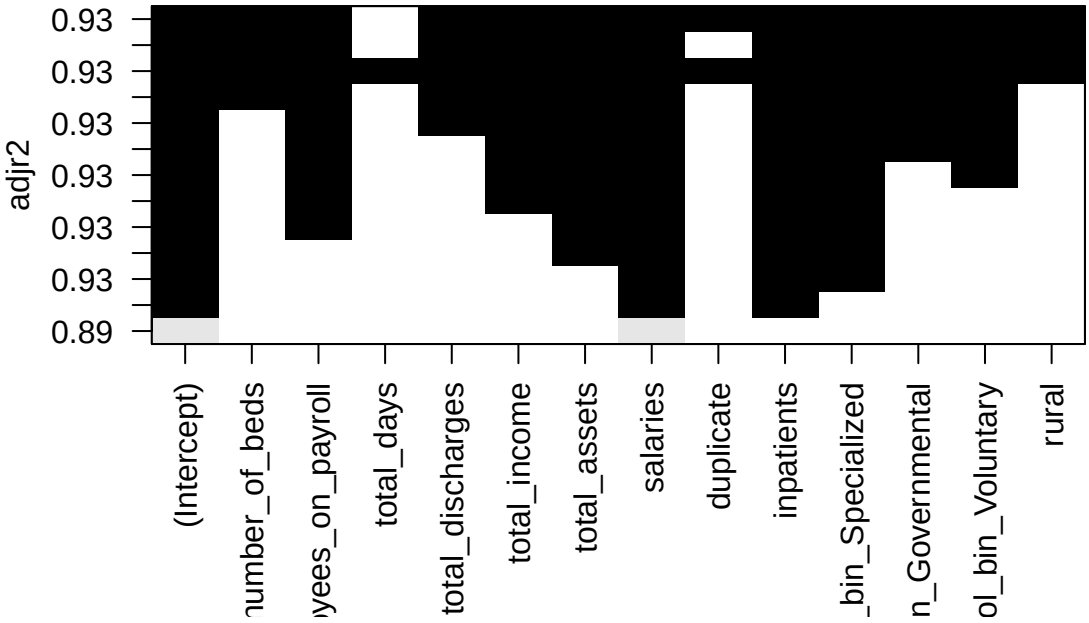
Variable Selection Analyses

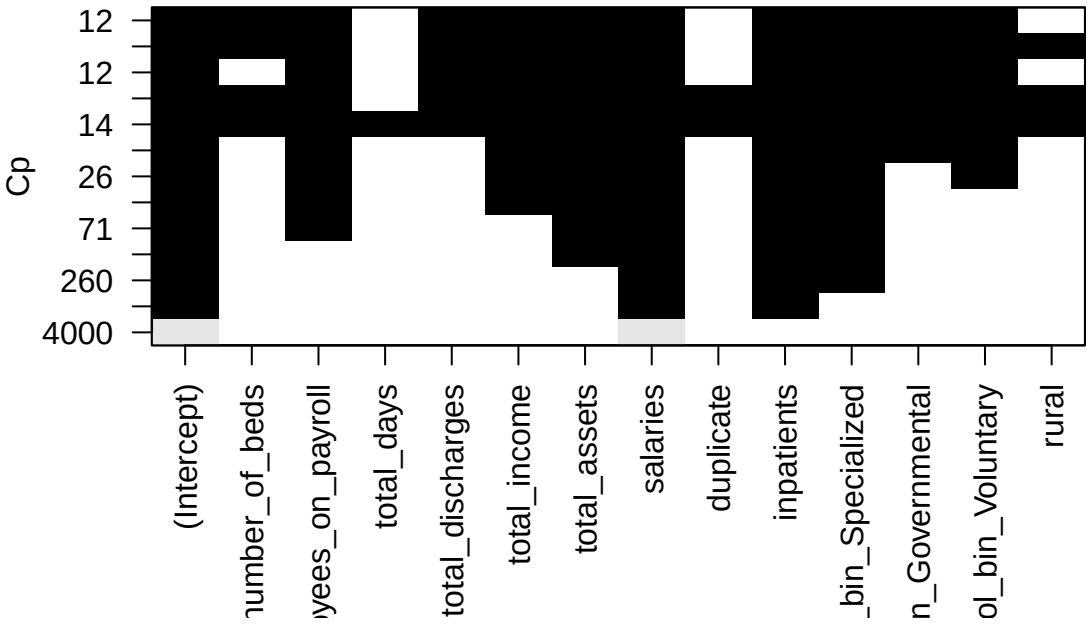
Best Subset

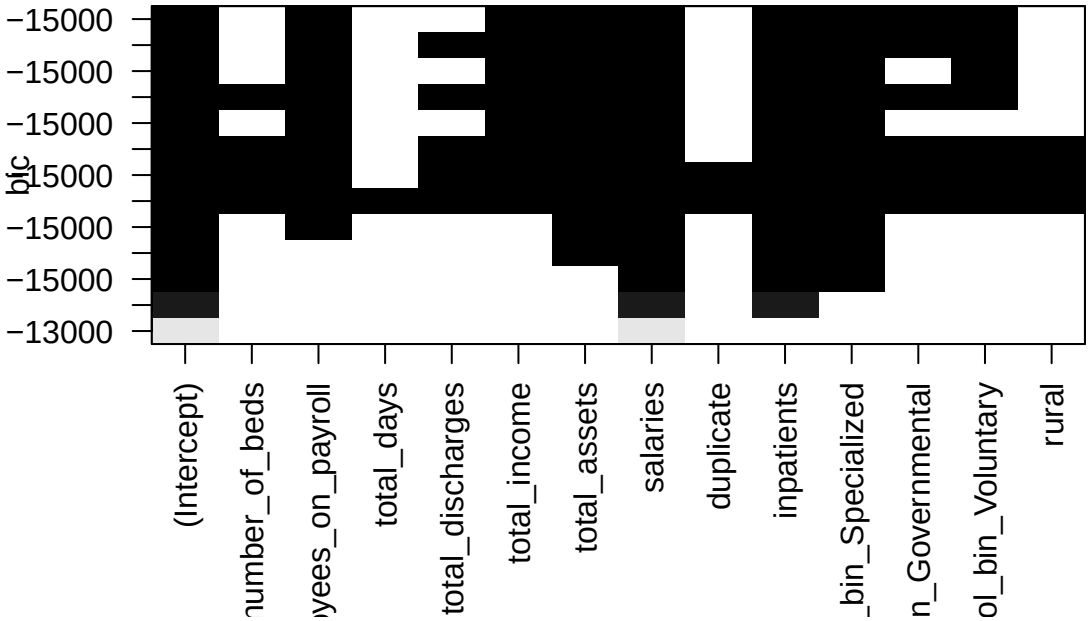
Results

In the plots below, we can see what variables are selected depending on what criterion we wish to use (R^2, adjusted R^2,etc.) for the full dataset. For the last plot, we run 10-fold CV to find the best model size to minimize the test MSE averaged across the folds. As we can see, we gain a significant reduction in the MSE from adding a non-intercept term, however, after that, the gains from additional variables in the model are comparatively quite small. Based on the 10-fold CV, we choose a model with 8 variables.

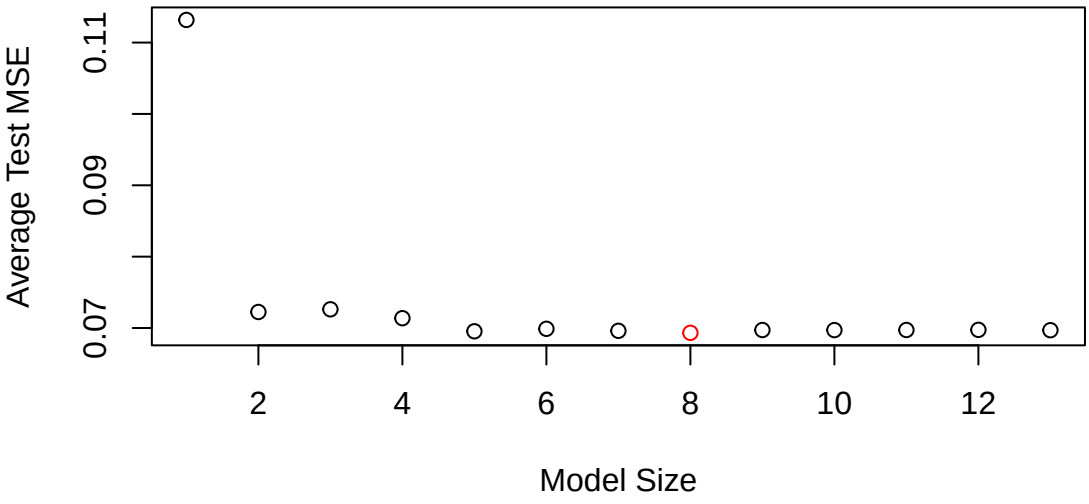



adjr2
0.93
0.93
0.93
0.93
0.93
0.93
0.89
(Intercept)
number_of_beds
yees_on_payroll
total_days
total_discharges
total_income
total_assets
salaries
duplicate
inpatients
_bin_Specialized
n_Governmental
ol_bin_Voluntary
rural


Cp
12
12
14
26
71
260
4000
(Intercept)
number_of_beds
yees_on_payroll
total_days
total_discharges
total_income
total_assets
salaries
duplicate
inpatients
_bin_Specialized
n_Governmental
ol_bin_Voluntary
rural



Best Subset Selection

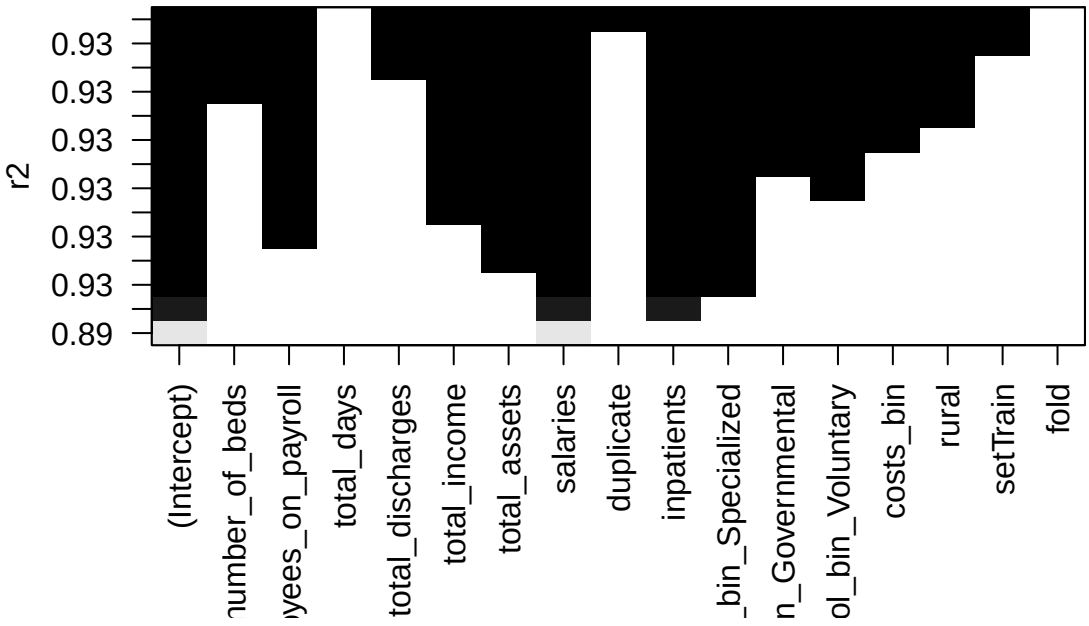


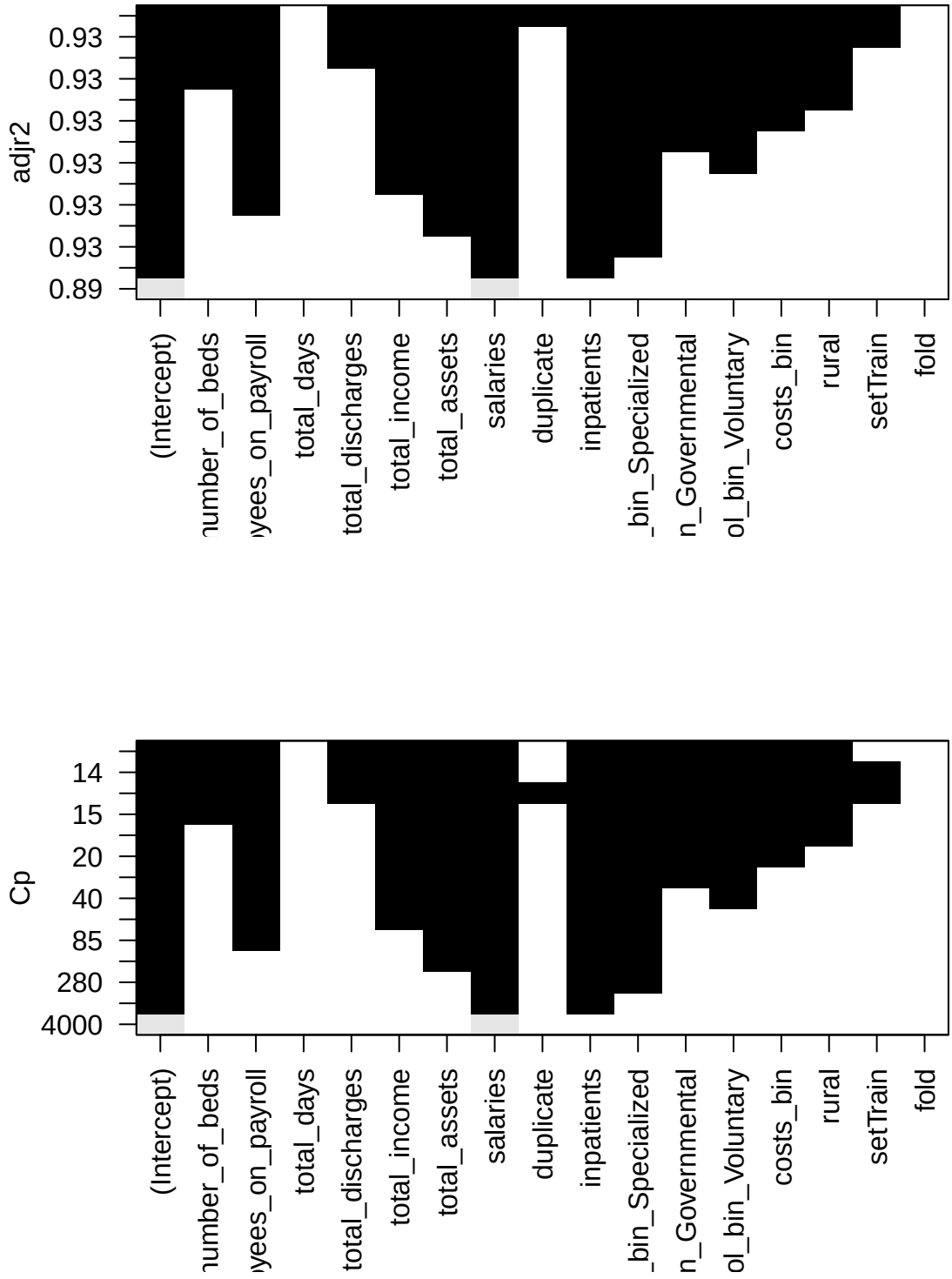
	Coefficient.Estimate
(Intercept)	-0.0195336
fte_employees_on_payroll	0.1033635
total_income	0.0202764
total_assets	0.0399147
salaries	0.5312923
inpatients	0.3318892
provider_bin_Specialized	-0.0828036
control_bin_Governmental	0.0397232
control_bin_Voluntary	0.0501687

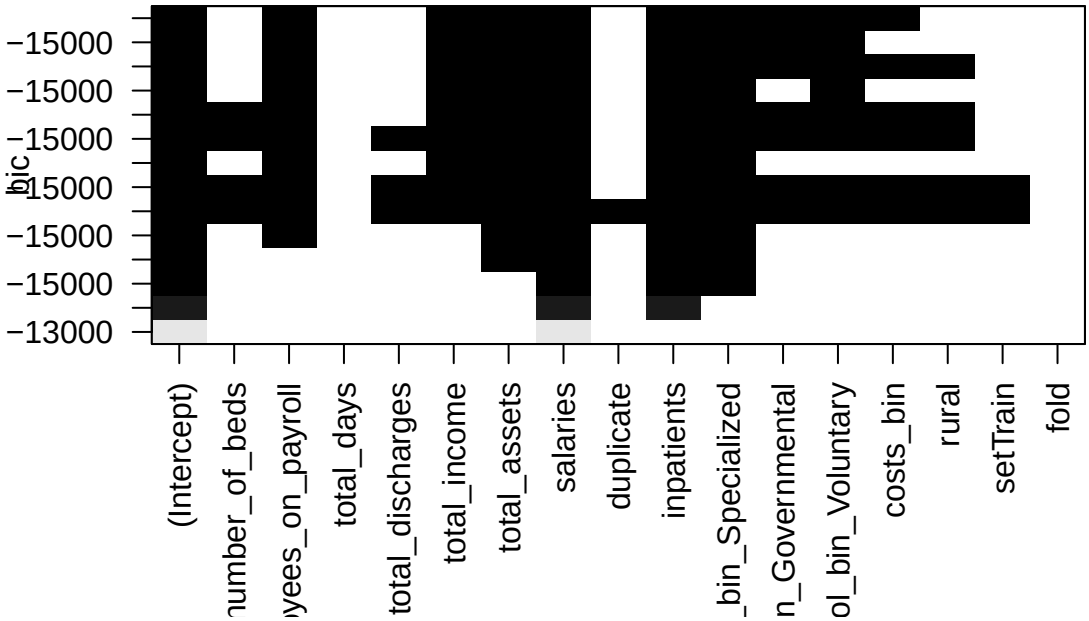
Forward Stepwise

Results

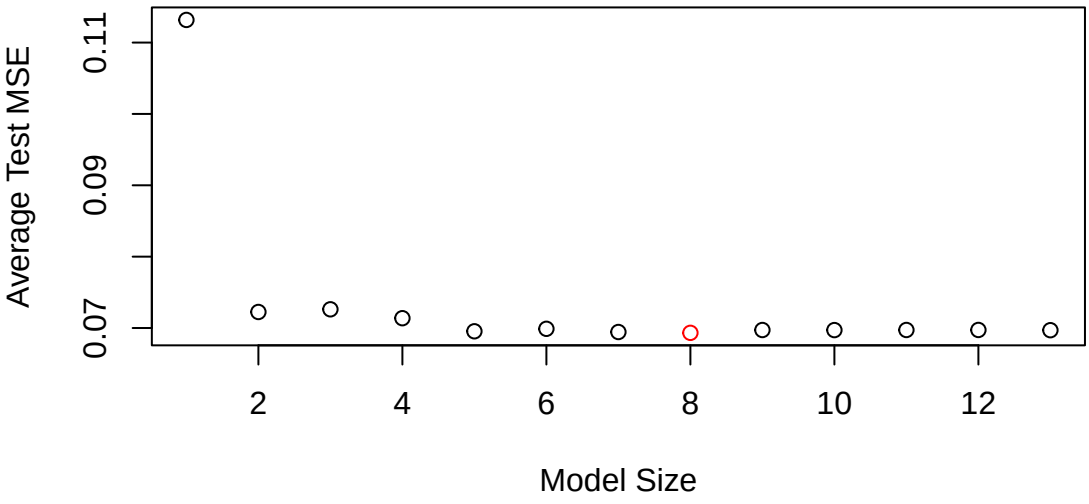
We see near identical results between the exhaustive best subsets and the forward/backward stepwise methods.




adjr2
0.93
0.93
0.93
0.93
0.93
0.93
0.89
(Intercept)
number_of_beds
yees_on_payroll
total_days
total_discharges
total_income
total_assets
salaries
duplicate
inpatients
_bin_Specialized
n_Governmental
ol_bin_Voluntary
costs_bin
rural
setTrain
fold
Cp
14
15
20
40
85
280
4000
(Intercept)
number_of_beds
yees_on_payroll
total_days
total_discharges
total_income
total_assets
salaries
duplicate
inpatients
_bin_Specialized
n_Governmental
ol_bin_Voluntary
costs_bin
rural
setTrain
fold



Forward Stepwise Selection

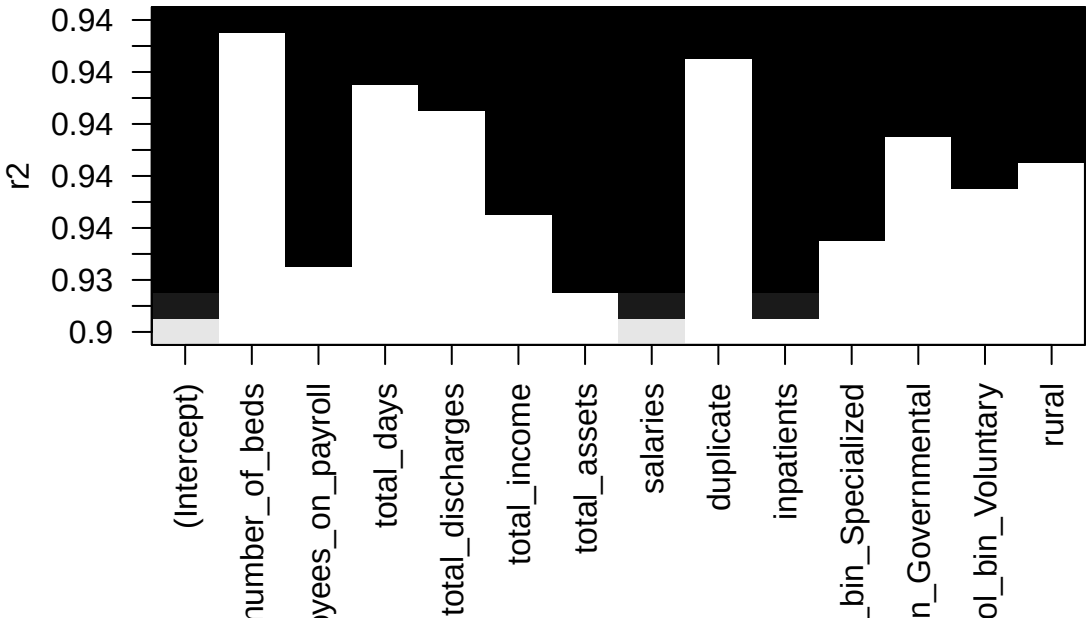


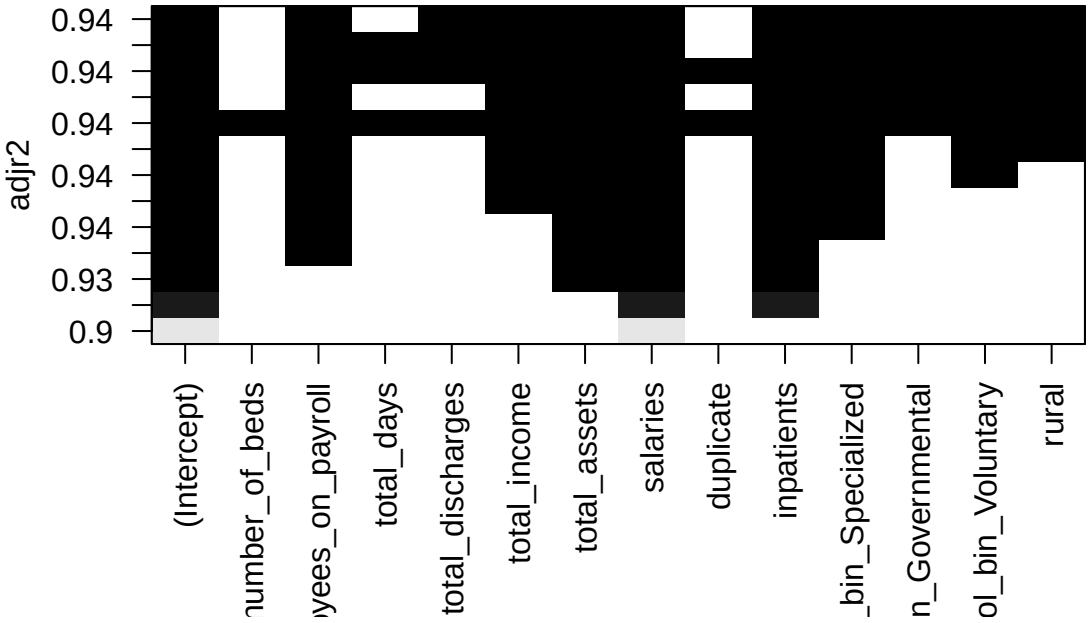
	Coefficient.Estimate
(Intercept)	-0.0195336
fte_employees_on_payroll	0.1033635
total_income	0.0202764
total_assets	0.0399147
salaries	0.5312923
inpatients	0.3318892
provider_bin_Specialized	-0.0828036
control_bin_Governmental	0.0397232
control_bin_Voluntary	0.0501687

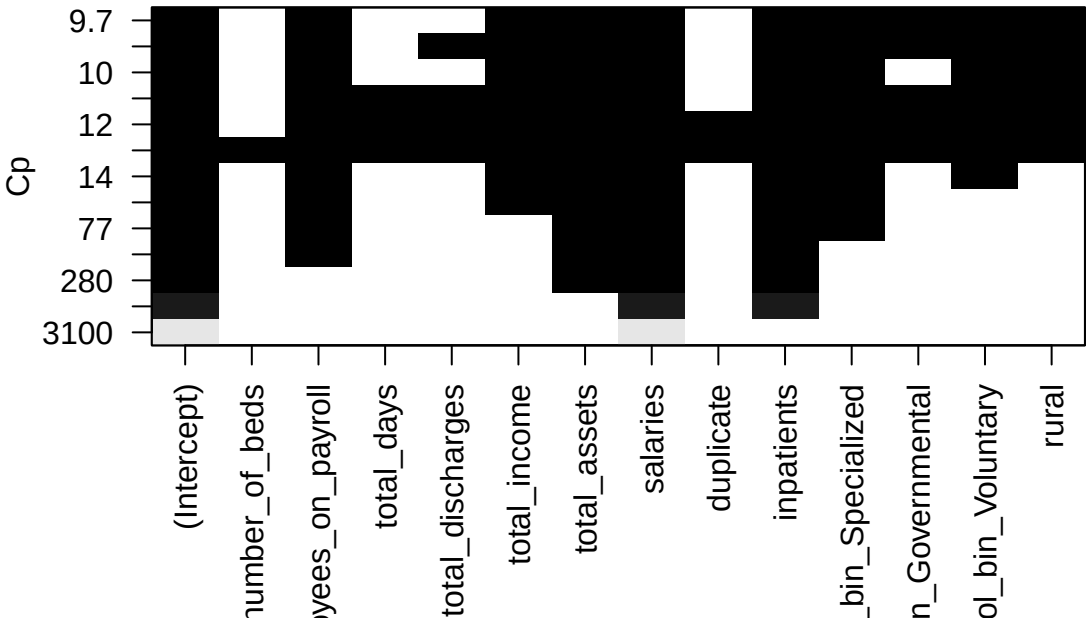
Backward Stepwise

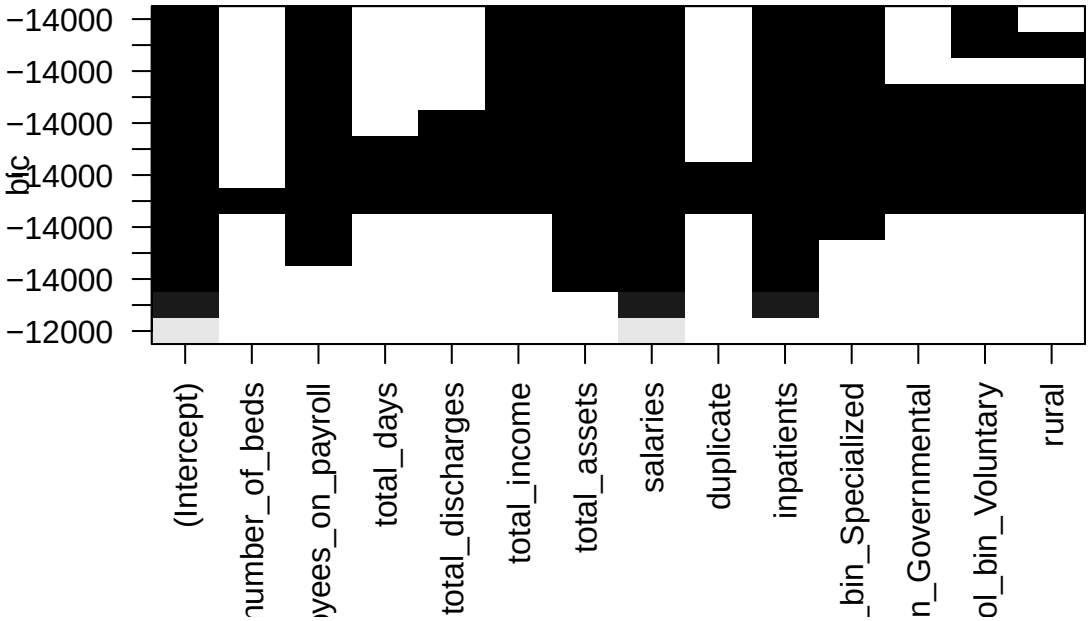
Results

[Some text here so the plots don't push to next page]

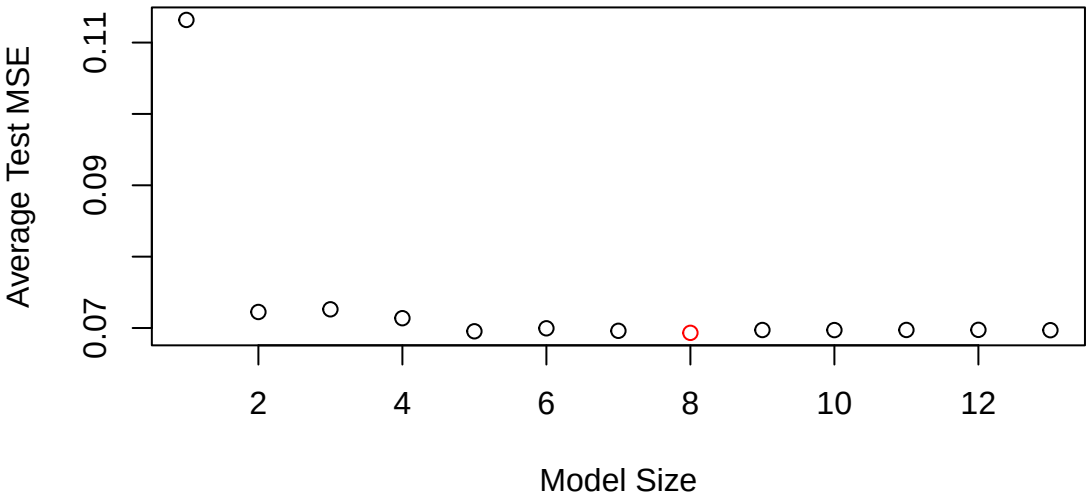



adjr2
0.94
0.94
0.94
0.94
0.94
0.93
0.9
(Intercept)
number_of_beds
yees_on_payroll
total_days
total_discharges
total_income
total_assets
salaries
duplicate
inpatients
_bin_Specialized
n_Governmental
ol_bin_Voluntary
rural


Cp
9.7
10
12
14
77
280
3100
(Intercept)
number_of_beds
yees_on_payroll
total_days
total_discharges
total_income
total_assets
salaries
duplicate
inpatients
_bin_Specialized
n_Governmental
ol_bin_Voluntary
rural



Backward Stepwise Selection



	Coefficient.Estimate
(Intercept)	-0.0195336
fte_employees_on_payroll	0.1033635
total_income	0.0202764
total_assets	0.0399147
salaries	0.5312923
inpatients	0.3318892
provider_bin_Specialized	-0.0828036
control_bin_Governmental	0.0397232
control_bin_Voluntary	0.0501687

Ridge regression

Results - Quantitative

	Coefficient.Estimate
(Intercept)	0.0000000
number_of_beds	0.0055404
fte_employees_on_payroll	0.1327093
total_days	0.1281495
total_discharges	0.0331282
total_income	0.0195163
total_assets	0.0431860
salaries	0.4976737
duplicate	0.0067522
inpatients	0.1686042
provider_bin_Specialized	-0.0261764
control_bin_Governmental	0.0145402
control_bin_Voluntary	0.0227265
rural	0.0061145

Results - Qualitative

	Coefficient.Estimate
(Intercept)	0.5000000
number_of_beds	0.2515577
fte_employees_on_payroll	0.0338598
total_days	-0.0804743
total_discharges	0.1168067
total_income	-0.0111515
total_assets	-0.0089622
salaries	-0.0053040
duplicate	-0.0011207
inpatients	-0.0965087
provider_bin_Specialized	-0.1186298
control_bin_Governmental	0.0204717
control_bin_Voluntary	0.1117096
rural	-0.1062490

Lasso

Results - Quantitative

Below we have the Lasso selection results for predicting total costs. We can notice it selects 9 variables to have non-zero coefficients, which is close to the best model found with regsubsets.

	Coefficient.Estimate
(Intercept)	0.0000000
number_of_beds	0.0000000
fte_employees_on_payroll	0.1082076
total_days	0.0779724
total_discharges	0.0266553
total_income	0.0122672
total_assets	0.0349461
salaries	0.5334924
duplicate	0.0000000
inpatients	0.2206243
provider_bin_Specialized	-0.0242279
control_bin_Governmental	0.0000000
control_bin_Voluntary	0.0086095
rural	0.0003979

Results - Qualitative

	Coefficient.Estimate
(Intercept)	0.5000000
number_of_beds	0.1508745
fte_employees_on_payroll	0.0000000
total_days	0.0000000
total_discharges	0.0547475
total_income	0.0000000
total_assets	0.0000000
salaries	0.0000000
duplicate	0.0000000
inpatients	0.0000000
provider_bin_Specialized	-0.1215522
control_bin_Governmental	0.0000000
control_bin_Voluntary	0.0961260
rural	-0.0909250

Principal Components Regression (PCR)

Results

The Test MSE for train-test split: 0.0944657820219513 and the average Test MSE across 10-folds: 0.0845142366389833.

Bootstrap SEs

Our bootstrap study looks at the standard errors of the coefficient estimates found through ridge regression, with 1000 bootstrap samples. Our data is scaled, which is why our standard errors are all approximately the same magnitude.

	Bootstrap.Standard.Error.Estimate
(Intercept)	0.00e+00
number_of_beds	6.01e-05
fte_employees_on_payroll	1.86e-05
total_days	8.00e-07
total_discharges	4.70e-05
total_income	1.22e-05
total_assets	2.83e-05
salaries	2.09e-05
duplicate	3.97e-05
inpatients	1.74e-05
provider_bin_Specialized	1.18e-05
control_bin_Governmental	2.79e-05
control_bin_Voluntary	1.13e-05
rural	5.39e-05