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

Lindsay Knupp, Rose Porta, Johnny Rasnic

2024-05-14

## Dataset

Our data was collected from Centers for Medicare & Medicaid Services on April 03, 2024. It features information about hospitals through annual cost reports in 2020. It had 5712 observations and we used 90% of them as our training set and 10% as our test set. We were interested in understanding how certain characteristics of hospitals like location, number of full time equivalent employees, or number of beds, for example, affected the hospitals' total operating costs. To perform qualitative analyses, we categorized each hospital as above or below the median.

While the data is roughly annual, certain hospitals reported for different fiscal year lengths. To normalize, we divided some of our variables by the length of their cost reporting period to obtain daily estimates like average number of inpatients per day or average salary expense per day. Variables that are reported as "averages per day" are denoted with the word "average" in the predictor table below. Some hospitals were listed multiple times with distinct reporting periods. We learned that this could correspond to a change in control of the hospital. For example, the hospital could have been sold and transitioned from a voluntary to a governmental hospital. Duplicate hospitals were left in the dataset and a dummy variable, `duplicate` was added to indicate its status.

There were 13 different categories of control ranging from "Voluntary Non-Profit-Church" to "Governmental-Federal". To reduce our number of categories, we re-binned this variable to only include the broad categories: "Voluntary", "Proprietary", and "Governmental". We followed a similar procedure for the 12 different categories of provider type ranging from "Children" to "Cancer" to "General Long Term". In this case, we re-binned provider type to only distinguish between "General" and "Specialized" care. We classified "General Short Term", "General Long Term", and "Religious Non-Medical Health Care Institution" as "General" care and classified "Cancer", "Psychiatric", "Rehabilitation", "Children", "Reserved for Future Use", "Other", "Extended Neoplastic Disease Care", "Indian Health Services", and "Rural Emergency Hospital" as "Specialized" care.

To improve accuracy on methods like Lasso regression, we scaled our numerical variables using the `scale()` function which centered and scaled our data appropriately. The following table includes the ranges of the predictors and response before they were re-scaled.

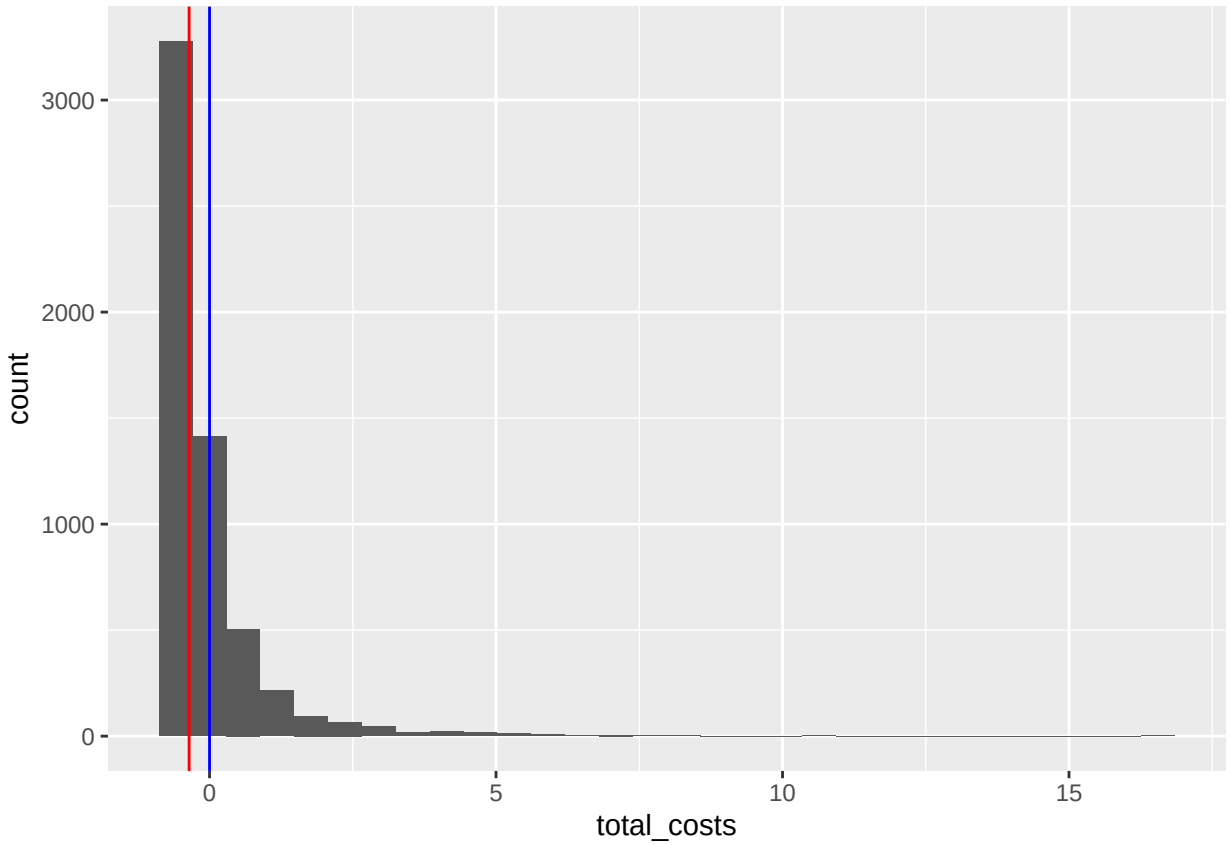
| Variables | Pre-scaled Range | Descriptions |
|---|---|---|
| **Predictors** | | |
| Number of Beds | [1-2,791] | Total number of available beds including adult beds, pediatric beds, birthing room, and newborn ICU beds |

| | | |
|---|---|---|
| FTE employees on payroll | [0.05-26,941.09] | Average number of full time-equivalent employees |
| Total hospital days | [1-772,819] | Total number of inpatient days (i.e. days all patients spent in the hospital) |
| Total discharges | [0.0027-462.63] | Average number of discharges including deaths |
| Total income | [-$6,129,919, $11,516,626] | Average income including net revenue from services given to patients |
| Total assets | [-$636,856,458, $29,465,487,958] | Total current assets |
| Salaries | [$128.51, $9,032,294.85] | Average salary expenses |
| Inpatients | [0.0033-2123.13] | Average number of inpatients |
| Rural versus Urban | [2487 rural, 3225 urban] | Location of hospital defined as rural or urban |
| Type of control | [2927 voluntary, 1728 proprietary, 1057 governmental] | Type of control under which hospital is conducted |
| Type of provider | [4779 general, 993 specialized] | Type of services provided |
| Duplicate hospital | [132 duplicates, 5580 non duplicates ] | Whether or not hospital was listed multiple times |
| **Response** | | |
| Total costs | [$2,718.28, $16,000,980.58] | Total hospital costs |
| Costs bin | [2856 above median, 2856 below median] | Whether or not total hospital costs was above/below median |

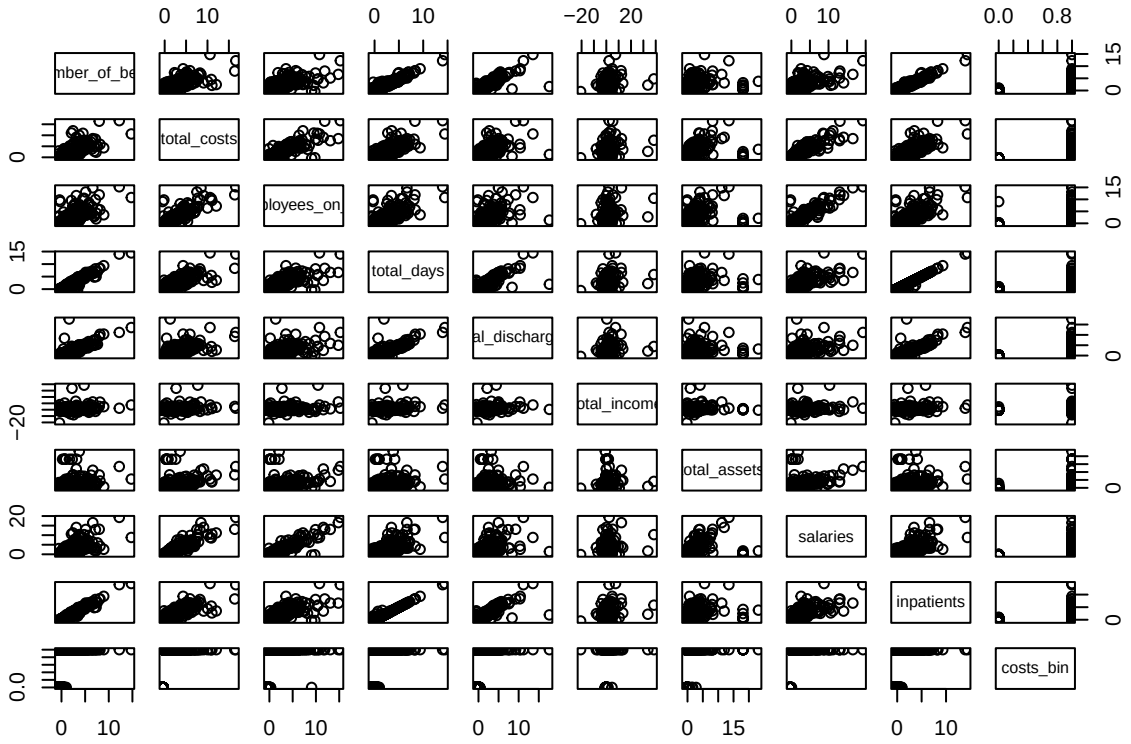
# Exploratory Data Analysis

In order to get a visual sense of the relationships in the data, we created a few exploratory plots.

The plot below displays the distribution of (scaled) total costs, where the red line represents the median and the blue line represents the mean. We can see that the distribution of total costs is very right-skewed. For this reason, we chose to define our binary response based on whether or not total costs was above the median, not the mean.



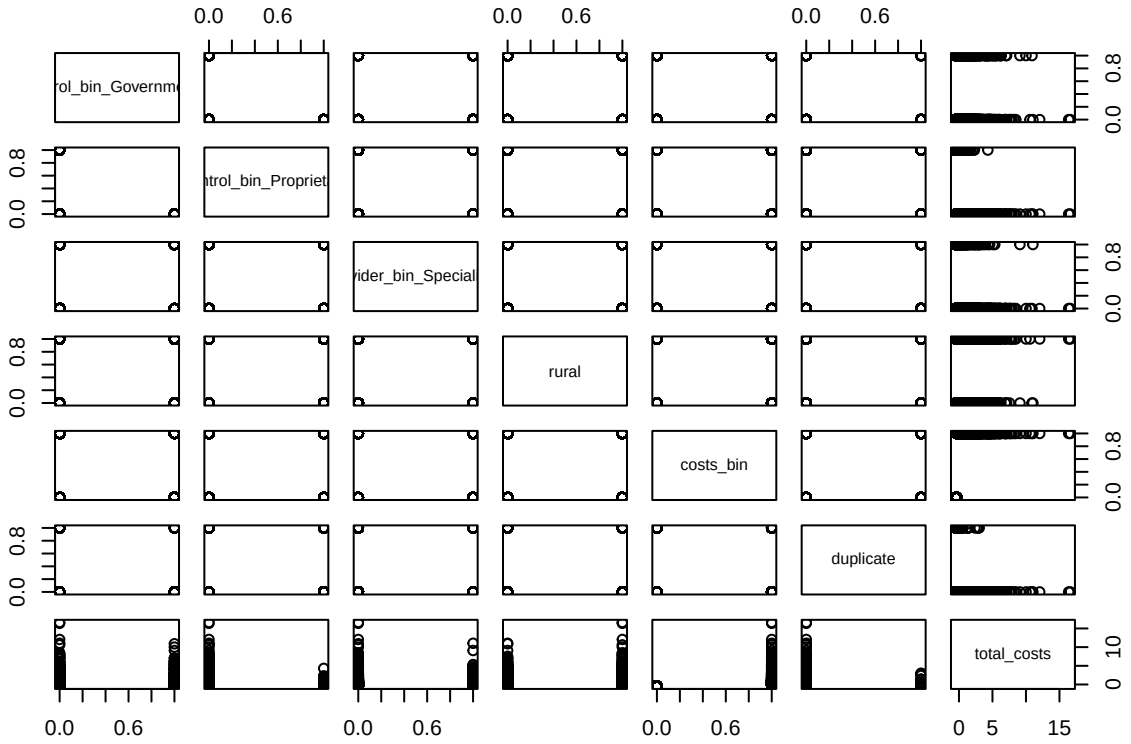

The pairs plots below visualize the relationships between all variables in the data (using the scaled data). We separated the data into two pairs plots, one including only numeric predictors and one including only categorical predictors, in order to see the relationships more clearly.




In the plot above visualizing the relationships between all numeric predictors and the response, we find that total costs appears to have fairly strong positive linear relationships with the following predictors: employees on payroll, total days, bed days, total discharges, and salaries. Total costs appears to have weak or no relationship with total income and total assets. For the binary response, the predictors which have positive relationships with total cost have a pattern such that binary costs = 0 corresponds to a higher concentration of points with low values of the predictor, and binary costs = 1 corresponds to a wider range of values for the predictor. This same pattern shows up also for total assets and total income, which did not look associated with continuous total cost.

Further, we find very high pairwise positive collinearity between (1) total days, bed days, and total discharges, and (2) salaries and employees on payroll.

Interestingly, despite total costs being so skewed, the relationships between each predictor and the response look very linear, and none of them appear non-linear.




Looking at the categorical predictors above, it is less obvious to see relationships between the predictors and response, but there does appear to be some relationship between duplicate and total costs as well as between proprietary control and total costs. For each of these relationships, it seems that when the predictor value equals zero, the total costs are low, while when the predictor value equals one, there is more. of a range of total costs values. However, it is hard to tell if this is a true relationship or if it simply results from unequal numbers of observations in each predictor category. Most total costs values are small, so if there are a smaller number of data points in one category, it would be likely that most of those have small total costs values, while if there are more points in a category, it would look like a wider range.

Focusing on the plot of total costs against binary cost (the two response variables), we notice that all total costs values below the median are very small, while those above the median have a much wider range of values. This reflects the skewness of total costs visualized above.

# Quantitative Outcome Analyses

## Marginal simple linear regressions

### Assumptions

Linear Regression analysis has four key assumptions:

1. Linear Relationship between predictor and response
2. Independence of Errors
3. Constant Variance

4. Errors are normally distributed

In order to assess these assumptions, we created pairwise scatterplots between the response (total costs) and each predictor. We see from the scatterplots that the relationship between the response and each numeric predictor appears approximately linear, and there are no obvious violations of non-constant variance. A histogram of total costs shows that the response variable is notably right-skewed, indicating possible concern that the errors may be non-normal. However, the linear relationships between each predictor and the response indicate that linear regression is a suitable model. We tried log-transforming total costs to better meet the normality of errors assumption, but we found that this transformation made the relationships between each predictor and the response notably non-linear, indicating that this transformation is not useful to improve the fit of the data to the model assumptions.

For independence of errors, it is relatively reasonable to assume that the total costs of one hospital would not impact those of another hospital, so this assumption is reasonably met. There could be small violations if, for example, two hospitals are in close proximity and one has more capacity than the other. In this scenario the one with lower capacity may redirect patients to the higer capacity one, making the costs decrease for the smaller one and increase for the larger one.

**Results**

The results of the simple linear regression models are summarized in the table below. The first column of the table represents the estimated Mean Squared Error of Prediction (MSEP) by cross validation. The second column represents the MSEP for the held-out test set. The third column represents the coefficient estimate for the predictor. The fourth column represents the p-value associated with the t-test for whether or not each coefficient estimate is equal to zero.

| 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 |

| method | cv_error | test_error | coef_est | p_value |
|---|---|---|---|---|
| Marginal LR duplicate | 0.996 | 1.038 | -0.121 | 0.188 |

We notice that almost all of the p-values equal zero, indicating strong evidence that each predictor is truly associated with the response, total costs. The indicator variable for Government control (`control_bin_Governmental`) has a large p-value, but the indicator variable for Proprietary control (`control_bin_Proprietary`) has a small p-value, indicating evidence that in general, the type of control is truly associated with total costs. The only other predictor with a large p-value is duplicate. This indicates that we do not have evidence that change of ownership of a hospital during the fiscal year is related to the total costs of the hospital.

Looking at the cross validation (CV) and test errors, we see that they are smallest for the models including number of beds, number of employees, total days, salaries, and inpatients. This indicates that these predictors are more strongly associated with the response than the others.

## Multiple linear regression

### Assumptions

The assumptions for multiple linear regression are the same as those for simple linear regression outlined in the previous section.

### Results

The results for the main effects multiple linear regression model are summarized below.

| | |
|---|---|
| Observations | 5141 |
| Dependent variable | total_costs |
| Type | OLS linear regression |

| | |
|---|---|
| F(13,5127) | 5871.91 |
| $R^2$ | 0.94 |
| Adj. $R^2$ | 0.94 |

The adjusted $R^2$ value of 0.94 indicates that 94 percent of the variability in total costs can be explained by the predictors. This $R^2$ value is quite high. The estimate of the test MSEP using CV is 0.071, and the test MSEP based on the held-out set is 0.106. These error rates are lower than any of the individual marginal linear regression error rates, indicating that when we include all predictors in one model, we can predict total costs more accurately than if we only include one predictor.

Looking at the coefficient estimates and p-values in the table below, we see that most predictors still have very small p-values, but some which were significant in the marginal linear regressions become insignificant in the multiple linear regression. Specifically, total days, number of beds, and total discharges have large

| | Est. | S.E. | t val. | p |
|---|---|---|---|---|
| (Intercept) | 0.02 | 0.01 | 3.00 | 0.00 |
| number_of_beds | -0.01 | 0.02 | -0.61 | 0.54 |
| fte_employees_on_payroll | 0.11 | 0.01 | 8.64 | 0.00 |
| total_days | 0.06 | 0.05 | 1.16 | 0.25 |
| total_discharges | 0.02 | 0.01 | 1.46 | 0.15 |
| total_income | 0.03 | 0.00 | 7.20 | 0.00 |
| total_assets | 0.04 | 0.00 | 9.99 | 0.00 |
| salaries | 0.58 | 0.01 | 51.17 | 0.00 |
| inpatients | 0.23 | 0.06 | 4.25 | 0.00 |
| control_bin_Governmental | -0.01 | 0.01 | -1.54 | 0.12 |
| control_bin_Proprietary | -0.03 | 0.01 | -3.84 | 0.00 |
| provider_bin_Specialized | -0.07 | 0.01 | -6.65 | 0.00 |
| rural | 0.01 | 0.01 | 1.72 | 0.08 |
| duplicate | 0.02 | 0.02 | 0.77 | 0.44 |

Standard errors: OLS

p-values in the multiple regression when all three had very small p-values in the marginal models. This indicates that there is likely correlations between the predictors such that once we have already accounted for some of them, others do not add much further information. This is consistent with the pairs plot which shows strong linear relationships between several pairs of predictors.

## Multiple linear regression with interactions and transformations:

Based on our exploratory data analysis, there were no obvious transformations or interactions which were needed in order to satisfy the linear regression assumptions. However, we added a few transformations and interactions for experimentation purposes. For transformations, we added a quadratic term for salaries and total days. For interactions, we added one interaction between number of employees and total income as well as between number of employees and provider type.

### Assumptions

The assumptions for multiple linear regression are the same as those for simple linear regression outlined in the previous section.

### Results

The results for the multiple linear regression model with interactions and transformations are summarized below.

| | |
|---|---|
| Observations | 5141 |
| Dependent variable | total_costs |
| Type | OLS linear regression |

| | |
|---|---|
| F(17,5123) | 4556.17 |
| $R^2$ | 0.94 |
| Adj. $R^2$ | 0.94 |

The adjusted $R^2$ value of 0.94 indicates that this model with transformations and interactions explains about the same proportion of variablity in total costs as the main effects model. One interesting finding, however, is that the p-value for the quadratic term for total days has a significant p-value despite the main-effect for total days not being significant.

The estimate of the test MSEP using CV is 0.114, and the test MSEP based on the held-out set is 0.081. These error values are very similar to those for the main effects model.

## Regression Tree (with pruning)

### Assumptions

For the regression tree, the only assumption we are making is that the response is continuous. There are no parametric assumptions.

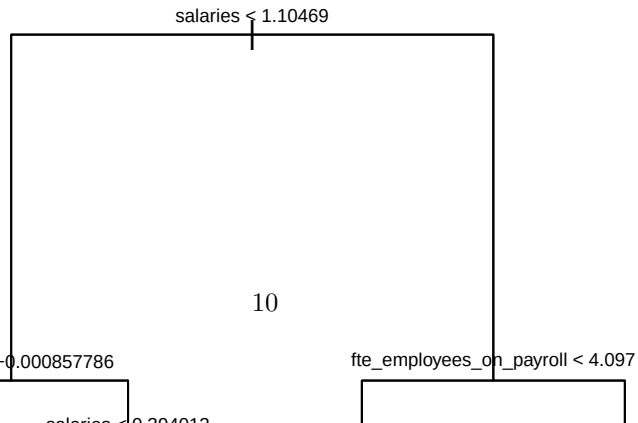
### Results

A plot of the pruned regression tree is displayed below. The optimal pruning size was found to be 9 via cross validation.

We see that almost all splits in the tree are made based on the salaries predictor, indicating that salaries is the most important predictor. There are only two other predictors included in the tree, which are number of employees and total income.

| | Est. | S.E. | t val. | p |
|---|---|---|---|---|
| (Intercept) | 0.02 | 0.01 | 2.88 | 0.00 |
| number_of_beds | -0.02 | 0.02 | -1.35 | 0.18 |
| fte_employees_on_payroll | 0.10 | 0.01 | 7.57 | 0.00 |
| total_discharges | 0.01 | 0.01 | 0.52 | 0.60 |
| total_income | 0.05 | 0.01 | 7.82 | 0.00 |
| total_assets | 0.04 | 0.00 | 10.08 | 0.00 |
| inpatients | 0.21 | 0.06 | 3.89 | 0.00 |
| control_bin_Governmental | -0.02 | 0.01 | -1.79 | 0.07 |
| control_bin_Proprietary | -0.02 | 0.01 | -2.61 | 0.01 |
| provider_bin_Specialized | -0.07 | 0.01 | -6.38 | 0.00 |
| rural | 0.01 | 0.01 | 1.49 | 0.14 |
| duplicate | 0.02 | 0.02 | 0.64 | 0.53 |
| poly(total_days, 2, raw = TRUE)1 | 0.05 | 0.05 | 0.94 | 0.35 |
| poly(total_days, 2, raw = TRUE)2 | 0.01 | 0.00 | 5.09 | 0.00 |
| poly(salaries, 2, raw = TRUE)1 | 0.66 | 0.02 | 42.19 | 0.00 |
| poly(salaries, 2, raw = TRUE)2 | -0.01 | 0.00 | -6.87 | 0.00 |
| fte_employees_on_payroll:total_income | -0.00 | 0.00 | -4.44 | 0.00 |
| fte_employees_on_payroll:provider_bin_Specialized | -0.01 | 0.01 | -0.62 | 0.54 |

Standard errors: OLS



The estimate of the test MSEP using CV is 0.144, and the test MSEP based on the held-out set is 0.100. These error values are very similar to those for the multiple linear regression models.

## Bagging

### Assumptions

Bagging involves taking averages across multiple regression trees, so there are no additional assumptions.

### Results

The variable importance metrics for bagging are summarized in the table below. The Mean MSE Increase refers to the average percent increase in MSE when the predictor is excluded. This is computed by permuting the out-of-bag portion of the data. The Node Purity refers to the increase in node purity accounted for by the predictor. For both of these importance measures, a larger value indicates higher importance.

| | Mean MSE Increase | Node Purity Increase |
|---|---|---|
| number_of_beds | 9.694294 | 38.3591981 |
| fte_employees_on_payroll | 19.483684 | 760.0131614 |
| total_days | 15.647406 | 47.2789876 |
| total_discharges | 16.351773 | 51.0071867 |
| total_income | 3.465100 | 39.1531516 |
| total_assets | 10.155649 | 92.1079080 |
| salaries | 64.634045 | 4018.2544946 |
| inpatients | 12.650244 | 48.7096693 |
| control_bin_Governmental | -1.867176 | 3.9657913 |
| control_bin_Proprietary | 3.895454 | 0.2718931 |
| provider_bin_Specialized | 7.739045 | 4.3490918 |
| rural | -1.529207 | 2.8266191 |
| duplicate | 2.242368 | 0.4227698 |

We see that salaries by far has the highest node purity increase and also has the highest mean increase in MSE when the variable is removed. The second most important variable seems to be number of employees. These results are consistent with the plot of the single regression tree, and they indicate that much of the variability in hospital costs can be accounted for by the money spent on paying employees.

The estimate of the test MSEP using CV is 0.071, and the test MSEP based on the held-out set is 0.125. These error values are very similar to those for the multiple linear regression models and the single regression tree.

## Random Forest

### Assumptions

Random Forest involves taking averages across multiple regression trees, so there are no additional assumptions.

### Results

The variable importance metrics for random forest are summarized in the table below. Their interpretation is the same as for bagging.

| | Mean MSE Increase | Node Purity Increase |
|---|---|---|
| number_of_beds | 7.0392835 | 268.9865987 |
| fte_employees_on_payroll | 17.6791577 | 1237.7064785 |
| total_days | 9.2230340 | 483.6220110 |
| total_discharges | 10.4088151 | 360.9172543 |
| total_income | 0.1925477 | 72.4556145 |
| total_assets | 9.9302091 | 300.5235321 |
| salaries | 20.5029953 | 1624.8458791 |
| inpatients | 11.6779113 | 692.4830620 |
| control_bin_Governmental | -0.5305487 | 5.6536003 |
| control_bin_Proprietary | 2.4712200 | 4.3315256 |
| provider_bin_Specialized | 4.6305656 | 8.5510975 |
| rural | 1.9012842 | 4.0501618 |
| duplicate | 2.5463912 | 0.9453899 |

We see that the ordering of variable importance is similar to that for bagging, with salaries and number of employees being the top two most important.

The estimate of the test MSEP using CV is 0.071, and the test MSEP based on the held-out set is 0.053. These error values are a little lower than those for the previous methods.

## Boosting

### Assumptions

Boosting involves manipulating and combining multiple regression trees, so there are no additional assumptions.

**Results**

The tuning parameter was chosen to be 0.1 by cross validation.

The relative influence of each predictor is summarized in the table below. The relative influence for boosting is similar to variable importance for bagging and random forest, but the method for computing it is slightly different. For boosting, the relative influence is computed as follows: for each split in the tree, compute the decrease in MSE. Then, average the improvement for each variable across all trees where that variable is included. A higher relative influence corresponds to a larger average decrease in MSE. A main difference in the computation compared to the Mean MSE Decrease from bagging and random forest is that for boosting, we are computing the mean decrease based on the entire training set, not only the out of bag portions. There is another method for computing importance which uses out of bag samples only, but the method described above is more widely used, so we chose that one.

| | Relative Influence |
|---|---|
| salaries | 60.7001581 |
| fte_employees_on_payroll | 16.2764509 |
| inpatients | 7.8161505 |
| total_days | 5.1148605 |
| total_discharges | 4.5885032 |
| total_assets | 3.3395015 |
| total_income | 1.2132591 |
| number_of_beds | 0.9511163 |
| control_bin_Governmental | 0.0000000 |
| control_bin_Proprietary | 0.0000000 |
| provider_bin_Specialized | 0.0000000 |
| rural | 0.0000000 |
| duplicate | 0.0000000 |

We see that salaries and number of employees have the highest influence by far compared to the others, consistent with the importance metrics from the previous methods.

The estimate of the test MSEP using CV is 0.092, and the test MSEP based on the held-out set is 0.079. These error values are on the lower side compared to the previous methods, but not as low as for random forest.

## Neural Network

### Assumptions

The only assumption for the neural network is that the response is continuous. There are no parametric assumptions.

### Results

We created a neural network with four hidden layers and one output layer. Out of the five total layers, three are dense layers and two are dropout layers. For the two hidden dense layers, we used relu activation functions. For the two hidden dropout layers, we used dropout rate 0.4 and 0.3 respectively. The output layer uses a linear activation function because our response is continuous. We structured the network as alternating dense layers and dropout layers to help prevent over-fitting.

The estimate of the test MSEP using CV is 0.072, and the test MSEP based on the held-out set is 0.088. These error values are on the lower side compared to the previous methods, but not as low as for random forest.

# Summary Table Quantitative Outcome

The below table summarises the error rates for all methods applied to the quantitative response, including all predictors.

| method | cv_error | test_error |
|---|---|---|
| Marginal LR number_of_beds | 0.253 | 0.226 |
| Marginal LR fte_employees_on_payroll | 0.132 | 0.129 |
| Marginal LR total_days | 0.205 | 0.191 |
| Marginal LR total_discharges | 0.298 | 0.264 |
| Marginal LR total_income | 0.939 | 0.993 |
| Marginal LR total_assets | 0.731 | 0.409 |
| Marginal LR salaries | 0.103 | 0.229 |
| Marginal LR inpatients | 0.200 | 0.189 |
| Marginal LR control_bin_Governmental | 0.996 | 1.037 |
| Marginal LR control_bin_Proprietary | 0.955 | 1.003 |
| Marginal LR provider_bin_Specialized | 0.980 | 1.023 |
| Marginal LR rural | 0.992 | 1.039 |
| Marginal LR duplicate | 0.996 | 1.038 |

| 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 | 0.115 |
| Random Forest | 0.071 | 0.053 |
| Boosting | 0.092 | 0.077 |
| Neural Network | 0.072 | 0.088 |

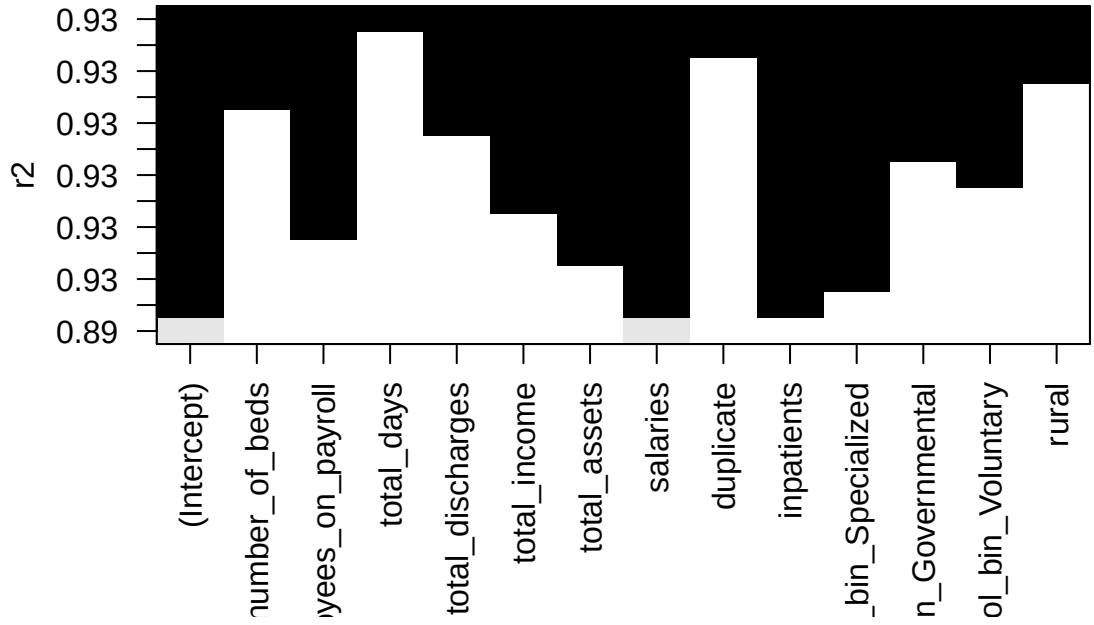
We note that the random forest method has the lowest MSEP.

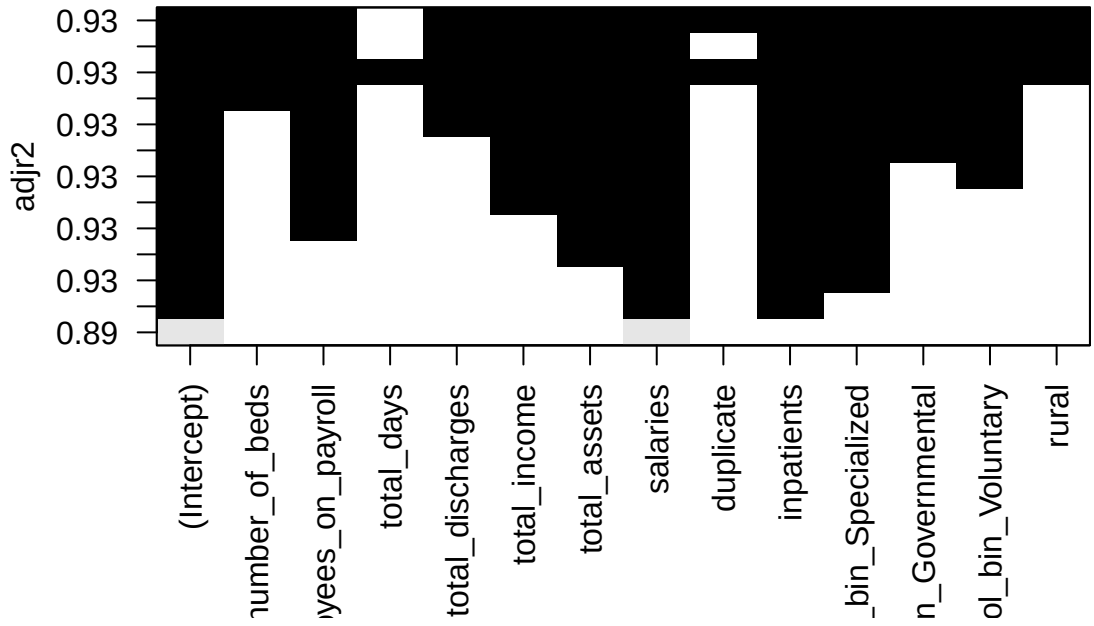
# Variable Selection Analyses

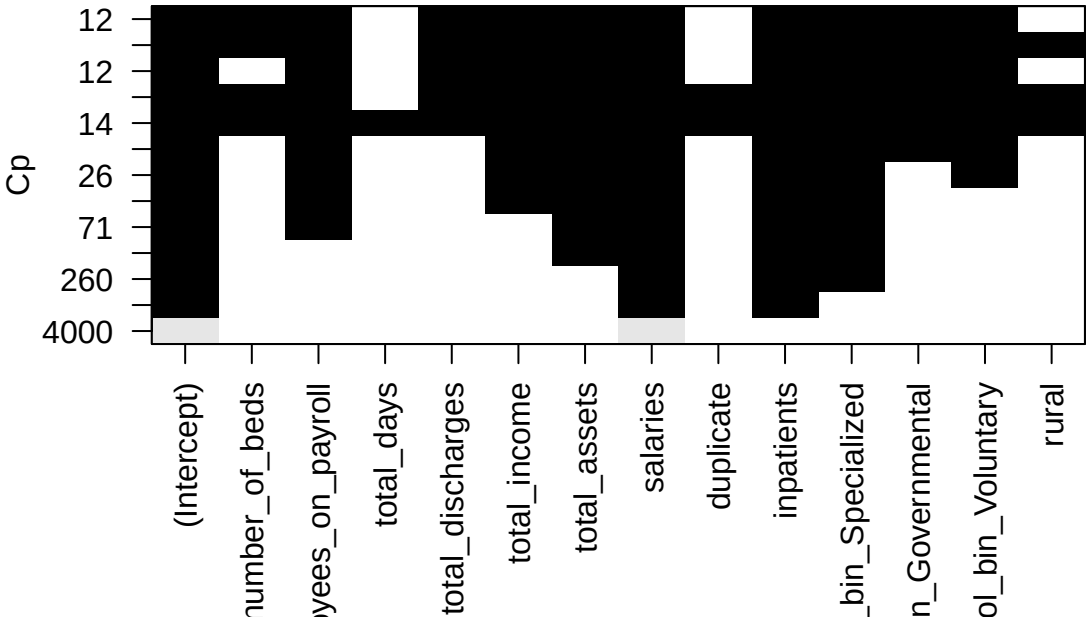
## Best Subset

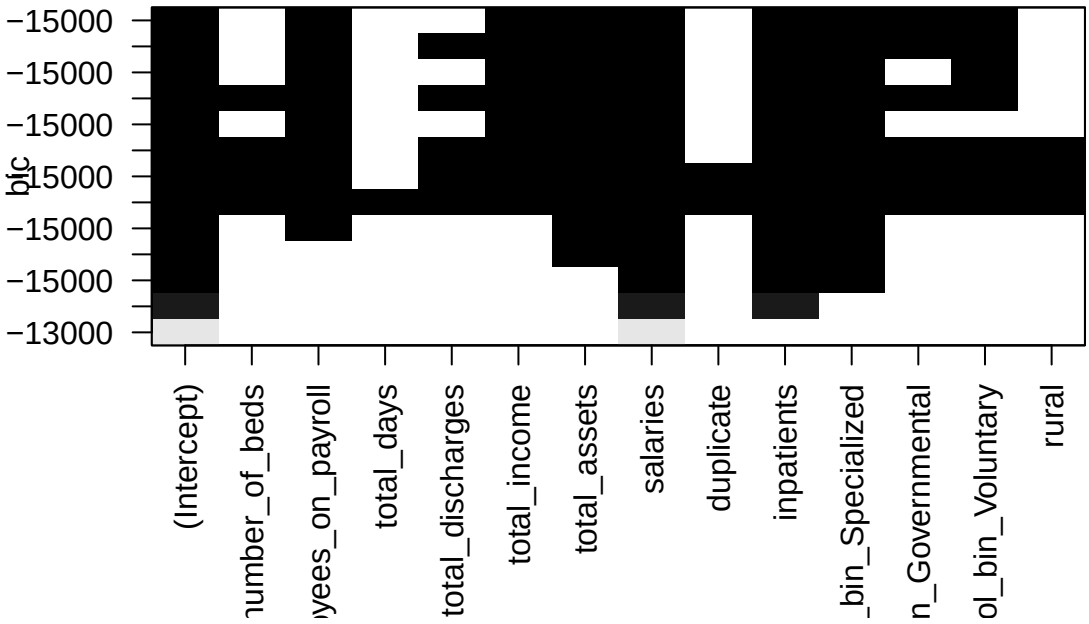
### Results

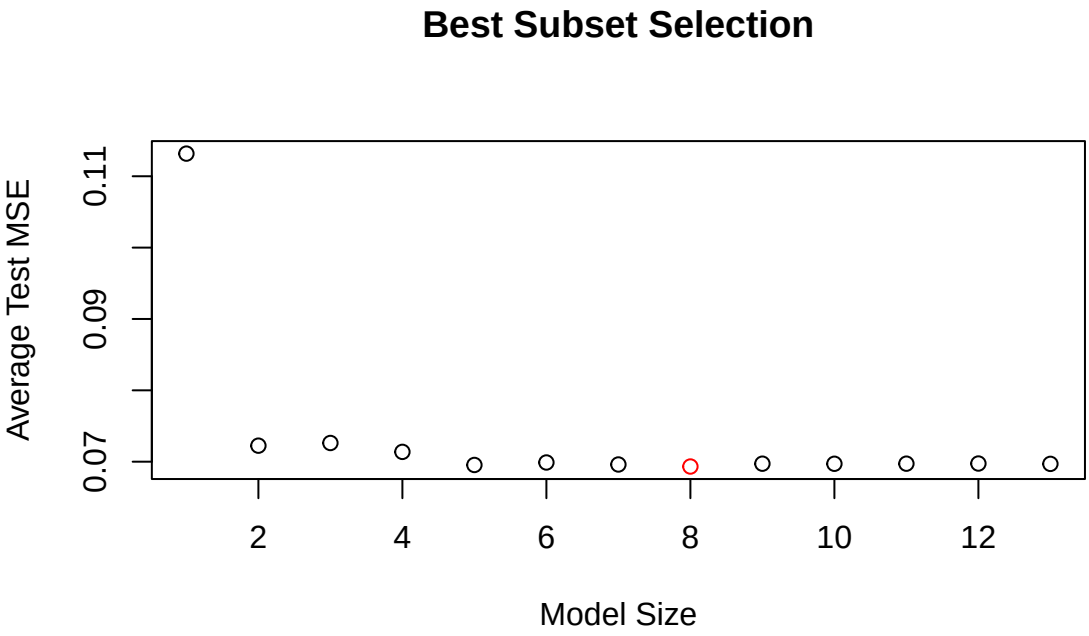
In the plots below, we can see what variables are selected depending on what criterion we wish to use ($R^2$, adjusted $R^2$,etc.) for the full dataset. Each black box signifies that that variable was chosen. For the last plot, we run 10-fold CV to find the best model size to minimize the test MSE averaged 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. We obtained a cross validation MSEP of 0.0613 and a test MSEP of 0.105.


r2
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


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


bic
−15000
−15000
−15000
−15000
−15000
−15000
−13000
(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

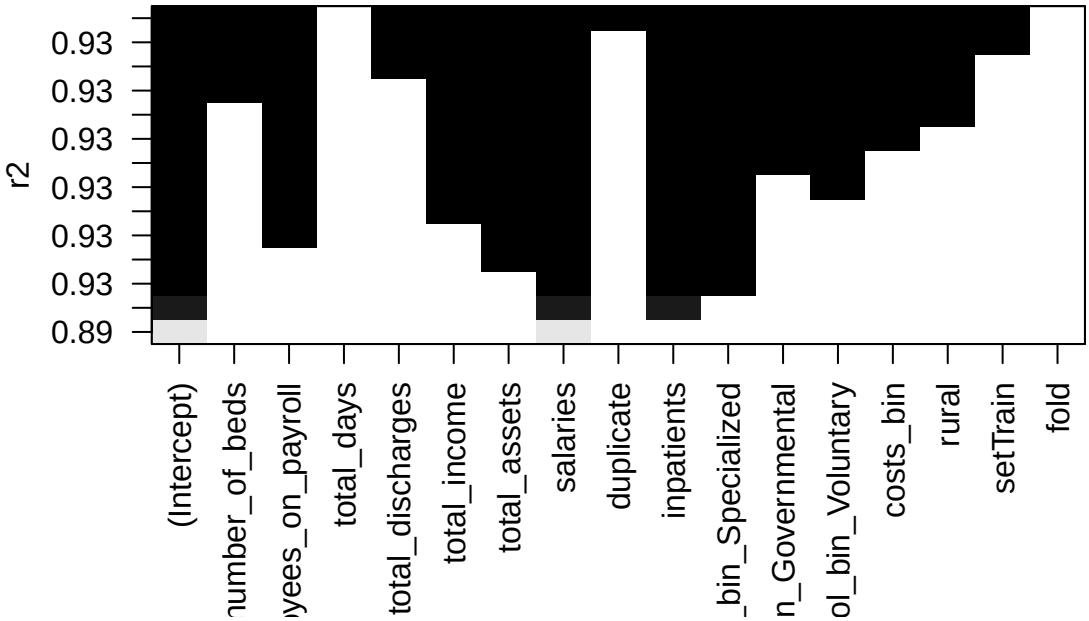


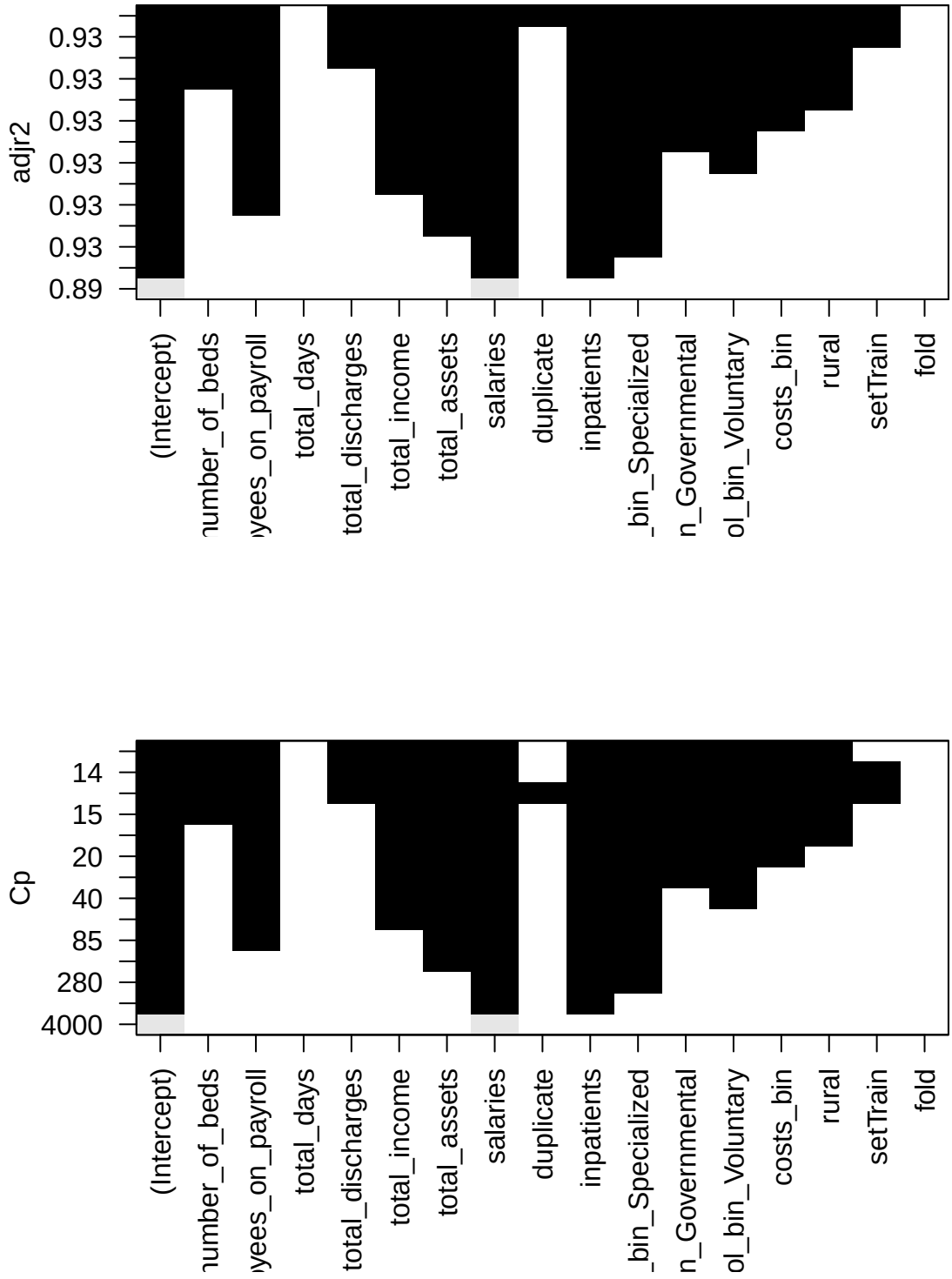

| | Coefficient.Estimate |
|---|---|
| (Intercept) | -0.0195336 |
| fte_employees_on_payroll | 0.1033635 |
| total_income | 0.0202764 |
| total_assets | 0.0399147 |
| salaries | 0.5312923 |
| inpatients | 0.3318892 |
| provider_bin_Specialized | -0.0828036 |
| control_bin_Governmental | 0.0397232 |
|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

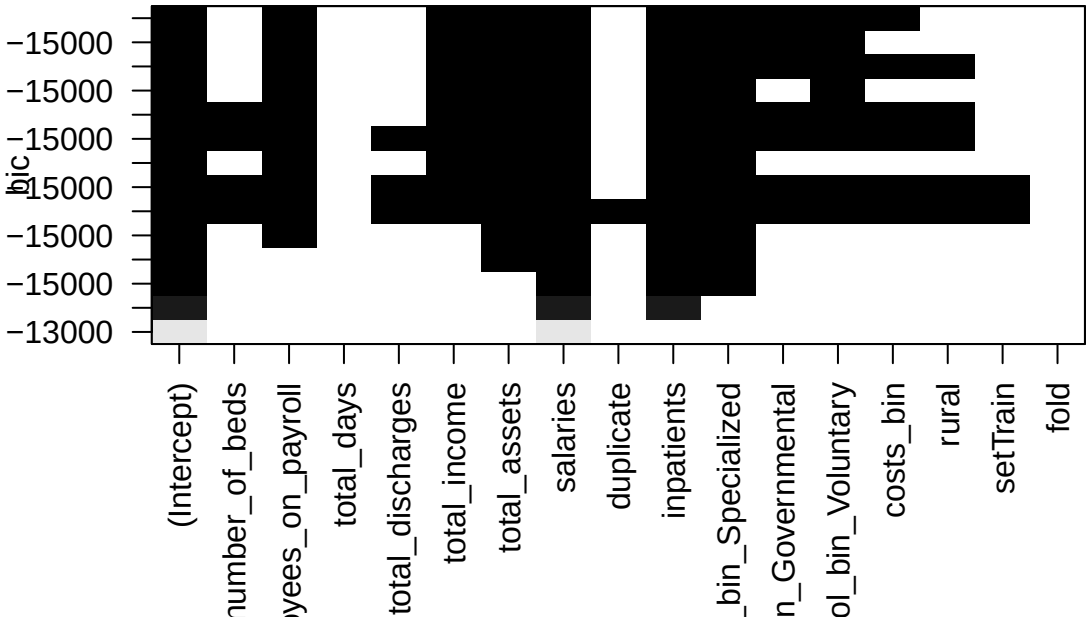
## Forward Stepwise

### Results

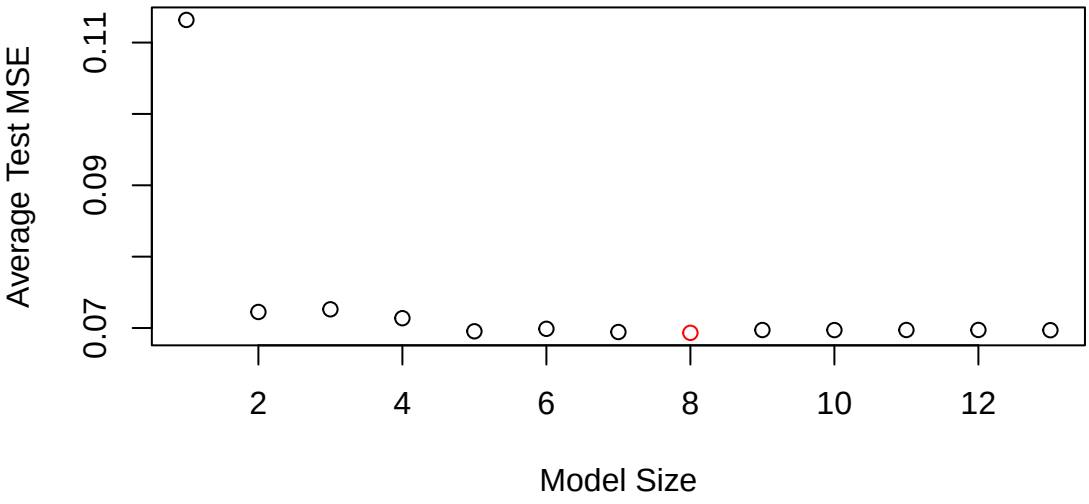
We see near identical results between the exhaustive best subsets selection and the forward/backward stepwise methods. We obtained a cross validation MSEP of 0.0613 and a test MSEP of 0.105.




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**

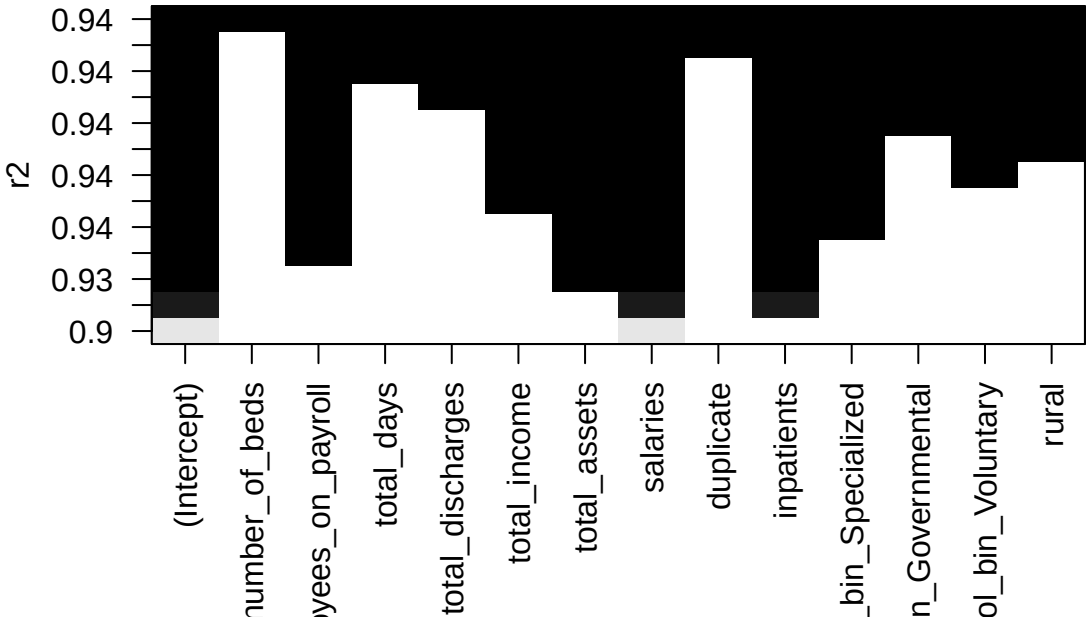


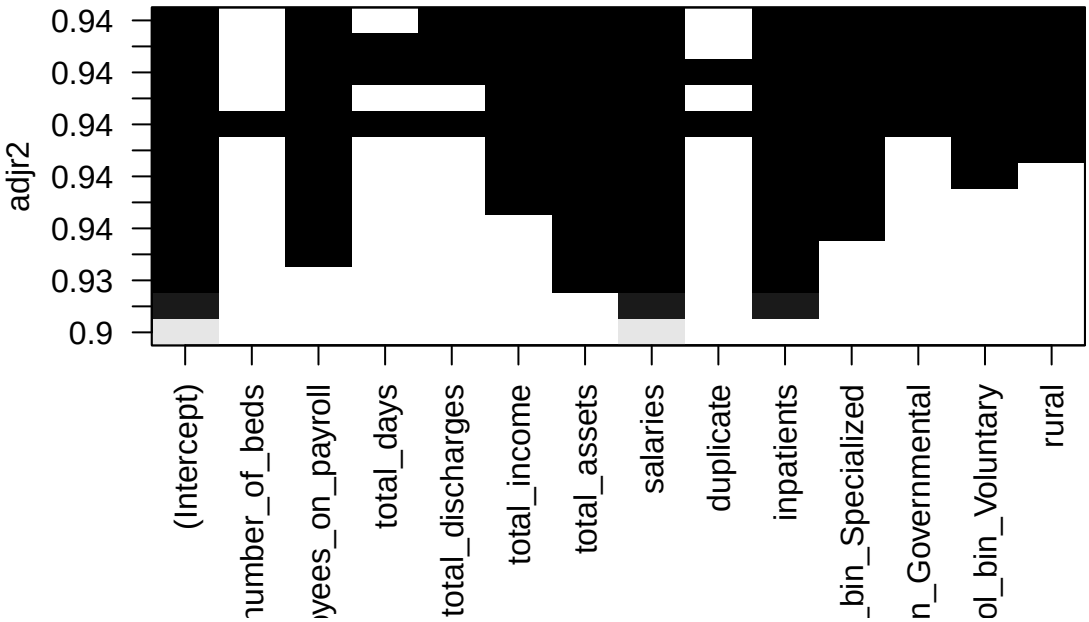
| | Coefficient.Estimate |
|---|---|
| (Intercept) | -0.0195336 |
| fte_employees_on_payroll | 0.1033635 |
| total_income | 0.0202764 |
| total_assets | 0.0399147 |
| salaries | 0.5312923 |
| inpatients | 0.3318892 |
| provider_bin_Specialized | -0.0828036 |
| control_bin_Governmental | 0.0397232 |
|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

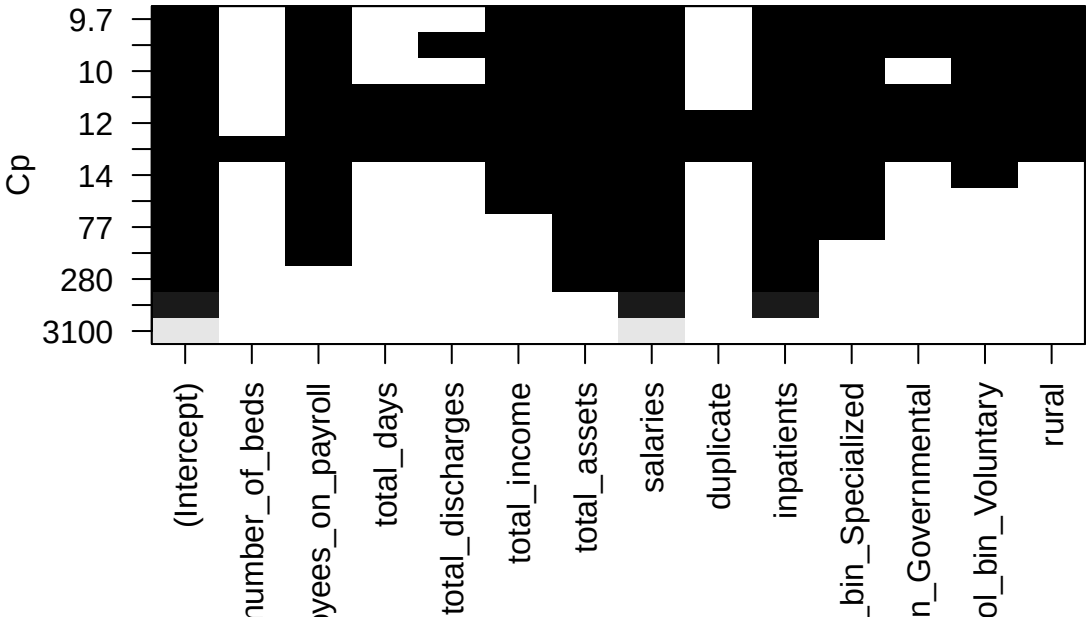
## Backward Stepwise

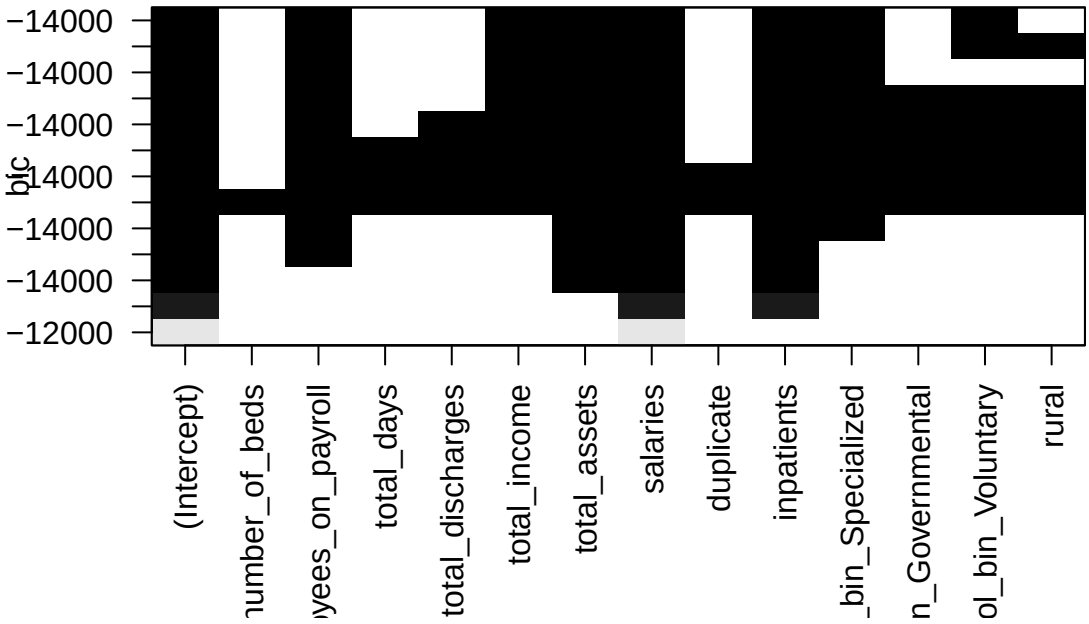
### Results

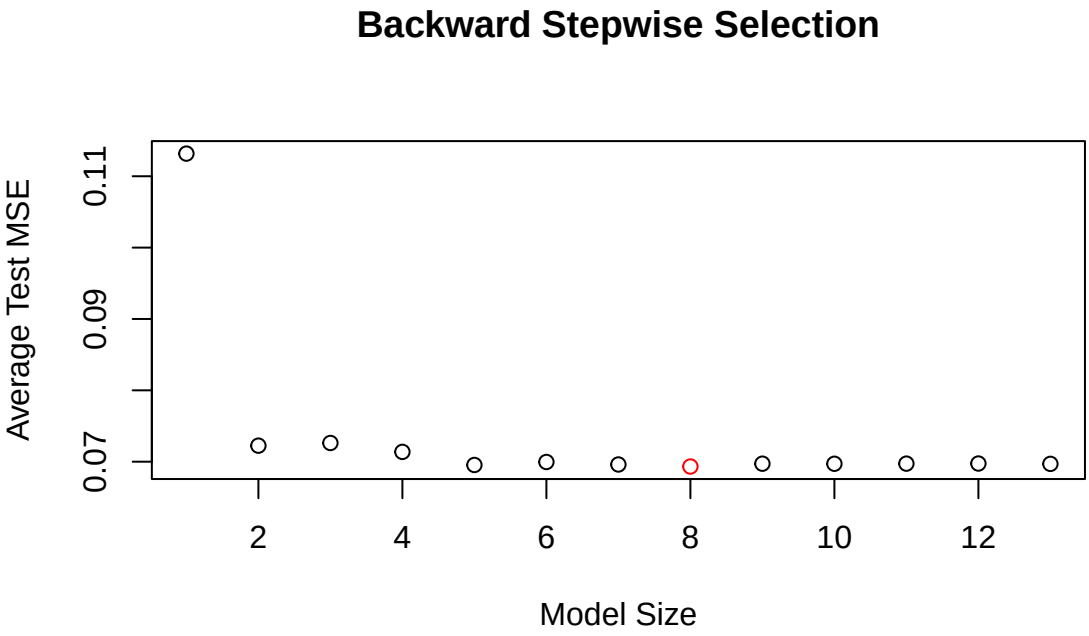
Again, we obtained a cross validation MSEP of 0.0613 and a test MSEP of 0.105.


r2
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


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


bic
−14000
−14000
−14000
−14000
−14000
−14000
−12000
(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




| | 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

### Quantitative

**Assumptions** We use the continuous `total_costs` variable as our response.

**Results** Using cross validation, we obtained $\lambda = 0.01$ as the optimal tuning parameter. Using the entire dataset, we obtained a MSEP of 0.066.

| | 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 |

### Qualitative

**Assumptions** We use the binary `costs_bin` variable as our response.

**Results** Using cross validation, we obtain a $\lambda = 0.01$ as the optimal tuning parameter. Using the entire dataset, we obtain a MSEP of 0.182.

| | 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

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`.

### Quantitative

**Assumptions** We use the continuous `total_costs` variable as our response.

**Results** Using cross validation, we obtain a $\lambda = 0.01$ as the optimal tuning parameter. Using the entire dataset, we obtain a MSEP of 0.911.

| | Coefficient.Estimate |
|---|---|
| (Intercept) | 0.0000000 |
| number_of_beds | 0.0000000 |

| | Coefficient.Estimate |
|---|---|
| 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 |

**Qualitative**

**Assumptions** We use the binary `costs_bin` variable as our response.

**Results** Using cross validation, we obtain a $\lambda = 0.01$ as the optimal tuning parameter. Using the entire dataset, we obtain a MSEP of 0.182.

| | 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 |

| | Coefficient.Estimate |
|---|---|
| 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)

#### Assumptions

We use the continuous `total_costs` variable as our response.

#### Results

We obtained a cross validation MSEP of 0.0845 and a test MSEP of 0.094.

## Summary Table Variable Selection

The following shows a comparison of the MSEPs across all the variable selection methods.

| method | cv_error | test_error |
|---|---|---|
| Best Subset | 0.0613 | 0.105 |
| Forward Stepwise | 0.0613 | 0.105 |
| Backward Stepwise | 0.0613 | 0.105 |
| Ridge (Quant.) | NA | 0.066 |
| Ridge (Qual) | NA | 0.182 |
| Lasso (Quant) | NA | 0.911 |
| Lasso (Qual) | NA | 0.182 |
| PCR | 0.0845 | 0.094 |

# Bootstrap SEs

Our bootstrap study looks at the standard errors of the coefficient estimates found through our quantitative 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) | 0.0e+00 |
| number_of_beds | 5.8e-06 |
| fte_employees_on_payroll | 7.0e-07 |
| total_days | 1.0e-06 |
| total_discharges | 3.2e-06 |
| total_income | 9.5e-06 |
| total_assets | 5.1e-06 |
| salaries | 3.2e-06 |
| duplicate | 5.7e-06 |
| inpatients | 1.9e-06 |
| provider_bin_Specialized | 6.0e-06 |
| control_bin_Governmental | 2.0e-06 |
| control_bin_Voluntary | 3.6e-06 |
| rural | 9.7e-06 |

# Qualitative Outcome Analyses

For all of our qualitative outcomes, we were trying to predict whether a hospital's total costs were above or below the median. All of the methods' error rates were comparable except for LDA which had the highest misclassification rate at about 17%. KNN had no consistent choice of an optimal $k$ across simulations and its variability inspired our simulation study.

## KNN

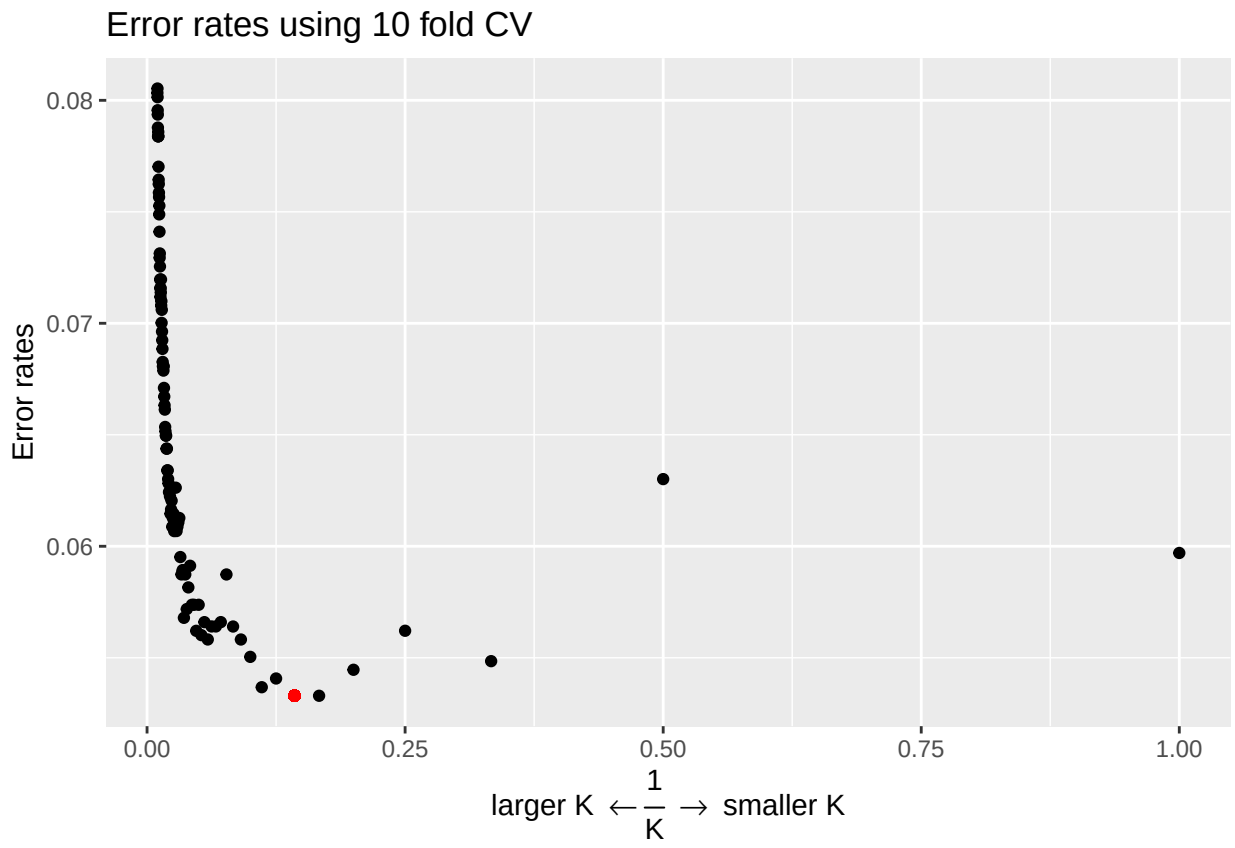
### Assumptions

We assumed that hospitals with similar predictor values have similar total costs.

### Results

We used 10 fold cross validation to first choose an optimal number of neighbors, $k$ and found $k = 7$ to be optimal with an error rate of 0.0533 using Euclidean distance. The true error rate with $k = 7$ was 0.0490 and the true/false positive and negative rates are summarized in the table below. When plotting the cross validation error rates against the chosen $k$, we see a condensed U shape. This may suggest that large $k$ suffers from high inaccuracy but too small $k$ can lead to overfitting.

| Classification Rates | Values |
|---|---|
| True positive | 0.91200 |
| True negative | 0.99300 |
| False positive | 0.00722 |
| False negative | 0.08840 |




## Multiple Logistic Regression

### Assumptions

We assume that our response is Bernoulli.

### Results

The coefficients and standard errors associated with our model can be found below. We found that `total_discharges`, `total_assets`, `salaries`, `rural`, and `provider_bin_Specialized` were the most sta-

tistically significant. Further, most predictors increased the probability of a hospital's total costs being above the median; unsurprisingly, `salaries` stood out the most. A one unit increase in `salaries` increased the log odds of an above median classification by 50.27. It also produced a $z$ statistic of 21.001 providing strong evidence of an association between salaries and total costs.

| Coefficients | Estimate | Std. Error | z value | p-value |
|---|---|---|---|---|
| (Intercept) | 18.98 | 0.86 | 21.96 | 0.00 |
| number_of_beds | -0.84 | 0.68 | -1.23 | 0.22 |
| fte_employees_on_payroll | 0.68 | 0.33 | 2.02 | 0.04 |
| total_days | -4.44 | 5.80 | -0.77 | 0.44 |
| total_discharges | 4.94 | 0.75 | 6.55 | 0.00 |
| total_income | 0.94 | 0.36 | 2.58 | 0.01 |
| total_assets | 1.76 | 0.52 | 3.35 | 0.00 |
| salaries | 50.27 | 2.39 | 21.00 | 0.00 |
| inpatients | 3.62 | 5.91 | 0.61 | 0.54 |
| rural | -1.22 | 0.20 | -6.16 | 0.00 |
| control_bin_Governmental | -0.46 | 0.24 | -1.94 | 0.05 |
| control_bin_Proprietary | 0.44 | 0.23 | 1.93 | 0.05 |
| provider_bin_Specialized | -4.02 | 0.40 | -9.92 | 0.00 |
| duplicate | 0.56 | 0.53 | 1.06 | 0.29 |

Our estimated and true error rates were pretty close to one another with our cross validation error of 0.073 and true test error of 0.0333.

| Classification Rates | Values |
|---|---|
| True positive | 0.9460 |
| True negative | 0.9890 |
| False positive | 0.0108 |
| False negative | 0.0544 |

## Multiple Logistic Regression with Transformations

### Assumptions

We assume that our response is Bernoulli.

### Results

We decided to transform `total_income`,`fte_employees_on_payroll`, `salaries`, and `total_days` to experiment with how less significant predictors in conjunction with `salaries` affected the response. We computed polynomial models up to degree 2 for `total_days` and interaction terms between `total_income` & `fte_employees_on_payroll` and between `fte_employees_on_payroll` & `salaries`.

With our smaller model, all of our new predictors were statistically significant with extreme $z$ statistics. However, it is important to note that the standard errors associated with each coefficient were extremely high suggesting a poor fit.

| Coefficients | Estimate | Std. Error | z value | p-value |
|---|---|---|---|---|
| (Intercept) | 6.583056e+14 | 971394.8 | 677691118 | 0 |

| | | | | |
|---|---|---|---|---|
| total_income | 1.743355e+14 | 1667476.0 | 104550530 | 0 |
| fte_employees_on_payroll | 7.442677e+14 | 3289614.4 | 226247710 | 0 |
| salaries | 2.737959e+15 | 3725566.4 | 734910894 | 0 |
| total_days | 2.542708e+14 | 2645218.8 | 96124663 | 0 |
| total_days_sq | -8.915542e+13 | 360959.8 | -246995412 | 0 |
| income_emp | -2.830918e+13 | 259889.2 | -108927893 | 0 |
| sal_emp | -2.512173e+14 | 333727.6 | -752761545 | 0 |

Compared to our original multiple logistic regression, our true error rate shot up from 0.033 to 0.158 and our cross validation error rate shot up from 0.073 to 0.130. Interestingly enough though, the model perfectly predicted hospitals whose total costs were below the median with a true negative rate of 100%. But, it did misclassify hospitals whose total costs were above the median with a false negative rate of 30.6%. Overall, the transformations performed worse than our original multiple logistic model.

| Classification Rates | Values |
|---|---|
| True positive | 0.694 |
| True negative | 1.000 |
| False positive | 0.000 |
| False negative | 0.306 |

## LDA

### Assumptions

We assume that our predictors are drawn from a multivariate normal distribution and both classes share a common covariance matrix.

### Results

As stated in the introduction, LDA performed the worst with a cross validation error rate of 0.175 and a true error rate of 0.165. This may suggest that our original assumption of our predictors being sampled from a multivariate normal was incorrect or that our classes do not share a common covariance matrix. Further, it is clear that a linear decision boundary is not sufficient to classify hospitals' total costs.

| Classification Rates | Values |
|---|---|
| True positive | 0.7620 |
| True negative | 0.9130 |
| False positive | 0.0866 |
| False negative | 0.2380 |

## QDA

### Assumptions

We still assume that our predictors are drawn from a multivariate normal distribution but drop the assumption that both classes share a common covariance matrix.

#### Results

QDA did not perform much better than LDA with a cross validation error of 0.115 and a true error rate of 0.119. However, compared to LDA, QDA did a much better job of accurately classifying hospitals whose total costs were below the median reducing the false positive rate from 8.7% to 1.4%. The false negative rates stayed fairly consistent hovering around $\sim 20\%$ in both methods. Overall, QDA is adequate at predicting our class labels.

| Classification Rates | Values |
|---|---|
| True positive | 0.7820 |
| True negative | 0.9860 |
| False positive | 0.0144 |
| False negative | 0.2180 |

### Naive Bayes with Gaussian kernel

#### Assumptions

We assume that our predictors are not correlated with one another and are drawn from a multivariate normal distribution given the target class.

#### Results

With a Gaussian kernel, the naive Bayes classifier was comparable to QDA. This classifier produced a cross validation error rate of 0.127 and a true error rate of 0.137. The misclassification rates were also extremely similar and can be summarized in the table below.

| Classification Rates | Values |
|---|---|
| True positive | 0.7520 |
| True negative | 0.9820 |
| False positive | 0.0181 |
| False negative | 0.2480 |

### Naive Bayes with Kernel Density Estimation

#### Assumptions

We still assume that our predictors are not correlated with one another but drop the normal distribution assumption.

#### Results

Without assuming normality, the Bayes Classifier performs much better with a cross validation error rate of 0.073 and a true error rate of 0.0806.

| Classification Rates | Values |
|---|---|
| True positive | 0.8670 |
| True negative | 0.9750 |

| | |
|---|---|
| False positive | 0.0253 |
| False negative | 0.1330 |

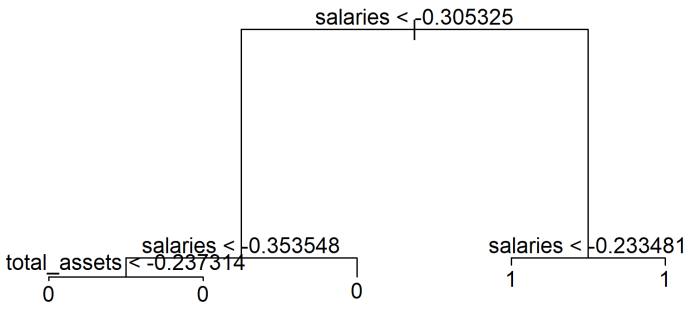
## Decision Tree with Pruning

### Assumptions

We make no assumptions on the structure of the data.

### Results

We pruned the tree using 5 terminal nodes which was found to be the optimal number of nodes using cross validation. The cross validation error rate was 0.051 compared to the true error rate of 0.0595. We also noticed that the split at the root node immediately determines how each hospital will be classified. If `salaries < -0.305`, then the hospital will be classified as having total costs below the median; otherwise, the hospital will be classified as above the median.




The false positive rate was extremely low at 1.4% while the false negative rate was slightly higher at 10%.

| Classification Rates | Values |
|---|---|
| True positive | 0.8980 |
| True negative | 0.9860 |
| False positive | 0.0144 |
| False negative | 0.1020 |

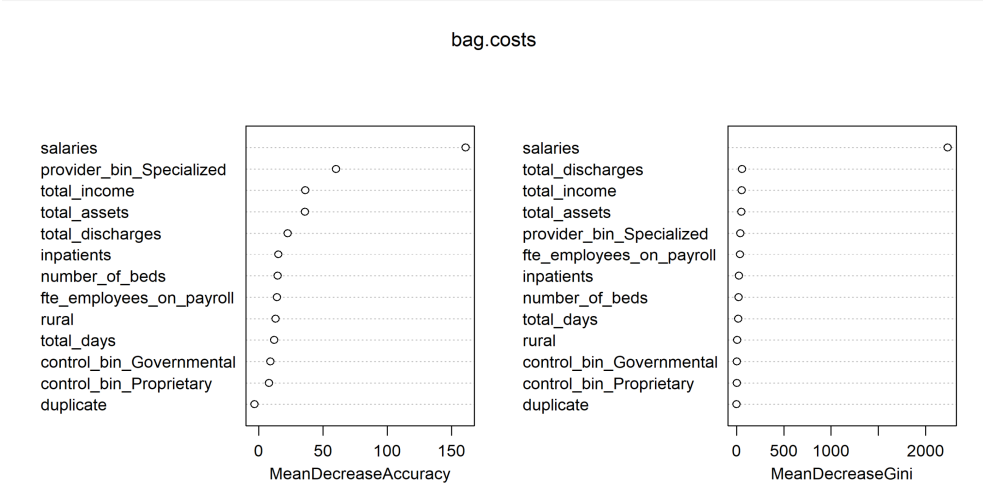
## Bagging

### Assumptions

We again make no assumptions on the structure of the data.

### Results

Bagging reduced our cross-validation error to 0.038 compared to the decision tree, a factor of about 1.3. Looking at the importance plot, `salaries` is the most importance variable. If it were to be removed from the tree, an average of 161 hospitals would be misclassified, given by the mean decrease in accuracy. Further, it's mean decrease of the Gini index is 2234.21.




Bagging also had low misclassification rates with a false positive rate of 1.81% and a false negative rate of 4.08%.

| Classification Rates | Values |
|---|---|
| True positive | 0.9590 |
| True negative | 0.9820 |
| False positive | 0.0181 |
| False negative | 0.0408 |

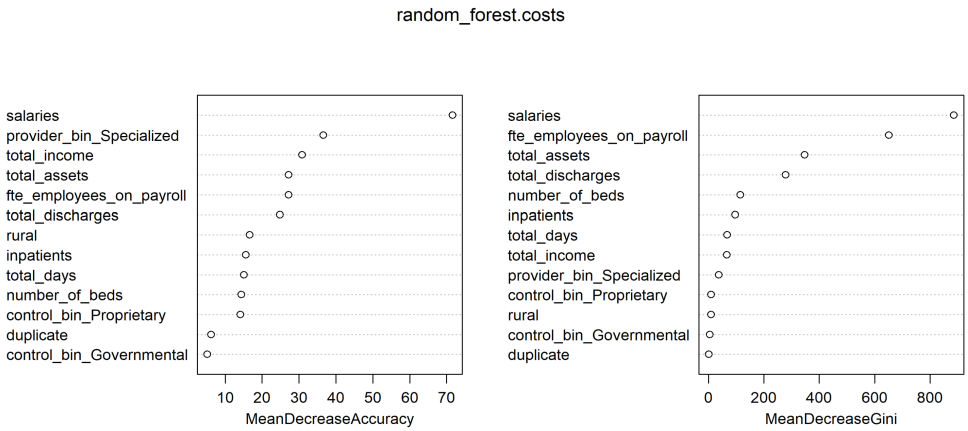
## Random Forest

### Assumptions

We again make no assumptions on the structure of the data.

### Results

Random Forest produced a cross validation error of 0.037 and a true test error 0.0298. The most important variable is `salaries` with a mean decrease in accuracy of 71.61 and a mean decrease of the Gini index is 886.61.




| Classification Rates | Values |
|---|---|
| True positive | 0.9630 |
| True negative | 0.9780 |
| False positive | 0.0217 |
| False negative | 0.0374 |

## Boosting

### Assumptions

We make no assumptions on the structure of the data.

### Results

Through cross validation, the tuning parameter was selected as $\lambda = 0.08$. This produced a cross validation error of 0.036 and a true test error of 0.044, both relatively small. The variables `salaries` had the largest relative influence of 83 which was significantly larger than any of the other predictors.

| Variables | Relative Influence |
|---|---|
| salaries | 83.0005951 |
| fte_employees_on_payroll | 5.8267828 |
| total_assets | 3.6053854 |
| total_discharges | 2.6735113 |
| total_income | 1.7579816 |
| provider_bin_Specialized | 0.8658885 |
| inpatients | 0.7121837 |
| total_days | 0.5701791 |
| number_of_beds | 0.5408744 |
| rural | 0.1934688 |
| control_bin_Governmental | 0.1175751 |
| control_bin_Proprietary | 0.0771219 |
| duplicate | 0.0584523 |

| Classification Rates | Values |
|---|---|
| True positive | 0.9320 |
| True negative | 0.9820 |
| False positive | 0.0181 |
| False negative | 0.0680 |

## Neural Network

To build our neural network, we included 4 hidden layers and 1 output layer. Within our hidden layers, we had 2 dense units that used the ReLU activation function and 2 dropout units that aimed to prevent overfitting. Our output layer used a softmax activation function in order to predict the appropriate class label.

### Assumptions

There are no model assumptions.

### Results

After training our model for 10 epochs, we got a cross validation error of 0.044 and a true test error of 0.038. Both comparable with our other qualitative prediction methods.

| Classification Rates | Values |
|---|---|
| True positive | 0.915 |
| True negative | 0.993 |
| False positive | 0.007 |

| | |
|---|---|
| False negative | 0.085 |

## Summary Table Qualitative Outcome

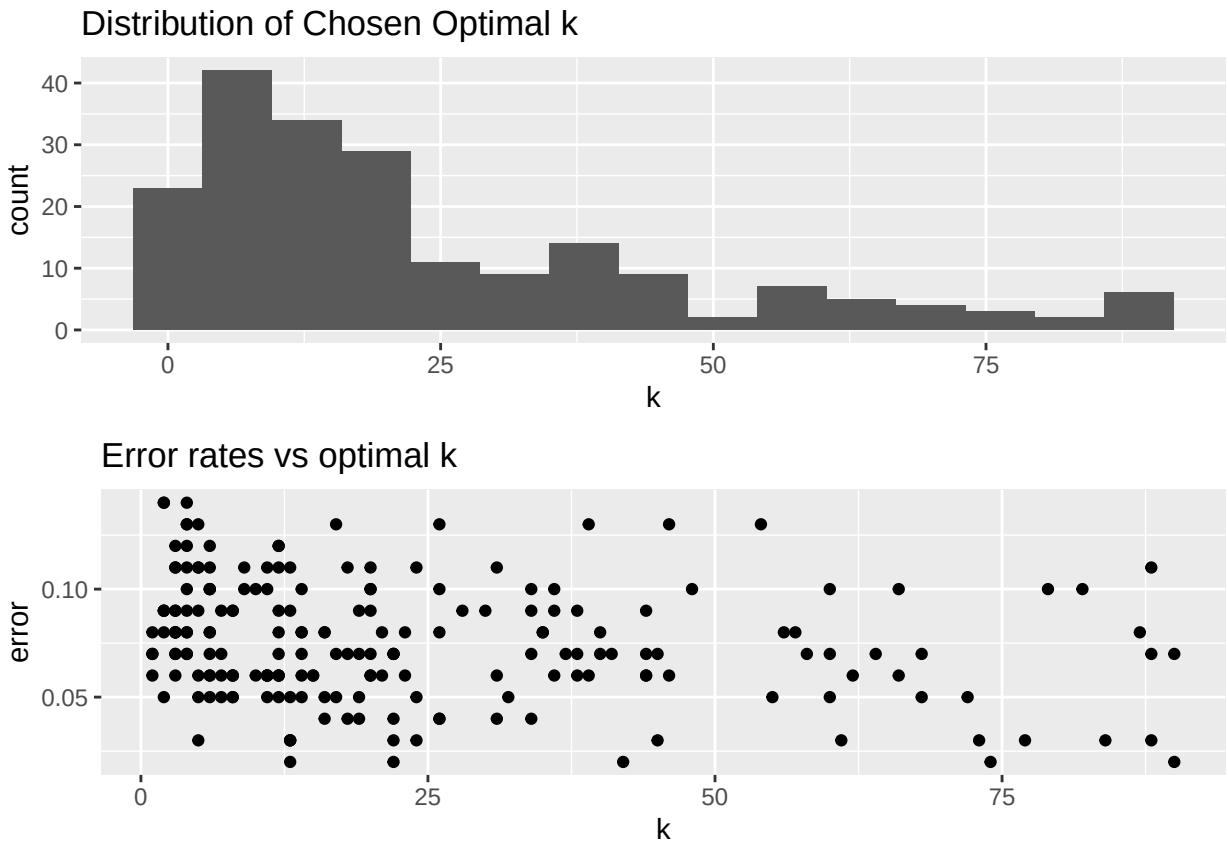
The following shows a comparison of the error rates and true/false positive and negative rates across all the qualitative outcome methods.

| method | cv_error | error_rate | true_pos | true_neg | false_pos | false_neg |
|---|---|---|---|---|---|---|
| KNN | 0.053 | 0.049 | 0.912 | 0.993 | 0.007 | 0.088 |
| Multiple Logistic Regression | 0.073 | 0.033 | 0.946 | 0.989 | 0.011 | 0.054 |
| Multiple Logistic Regression (Transformations) | 0.130 | 0.158 | 0.694 | 1.000 | 0.000 | 0.306 |
| LDA | 0.175 | 0.165 | 0.762 | 0.913 | 0.087 | 0.238 |
| QDA | 0.115 | 0.119 | 0.782 | 0.986 | 0.014 | 0.218 |
| Naive Bayes (Gaussian) | 0.127 | 0.137 | 0.752 | 0.982 | 0.018 | 0.248 |
| Naive Bayes (KDE) | 0.073 | 0.081 | 0.867 | 0.975 | 0.025 | 0.133 |
| Decision Tree | 0.051 | 0.060 | 0.898 | 0.986 | 0.014 | 0.102 |
| Bagging | 0.038 | 0.028 | 0.959 | 0.986 | 0.014 | 0.041 |
| Random Forest | 0.037 | 0.030 | 0.963 | 0.978 | 0.022 | 0.037 |
| Boosting | 0.036 | 0.044 | 0.932 | 0.982 | 0.018 | 0.068 |
| Neural Net | 0.044 | 0.039 | 0.915 | 0.996 | 0.004 | 0.085 |

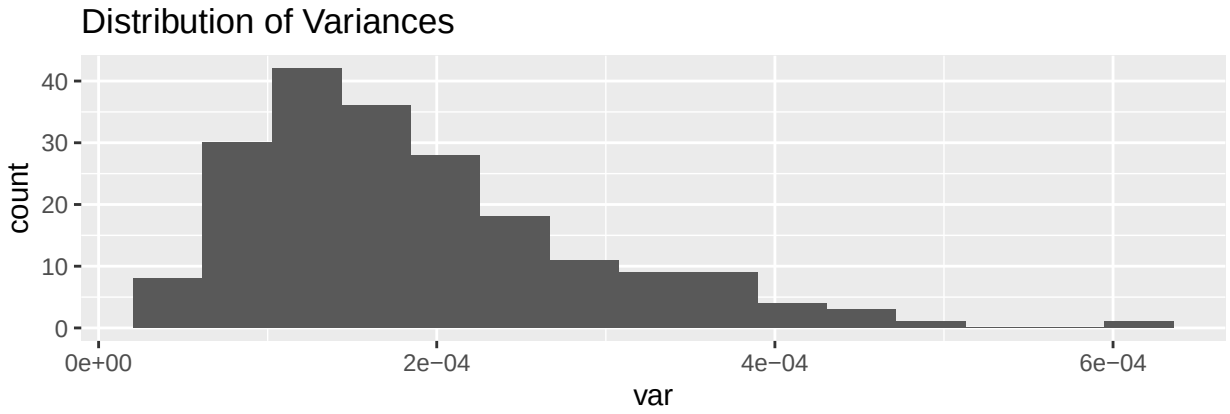
## Simulation Study

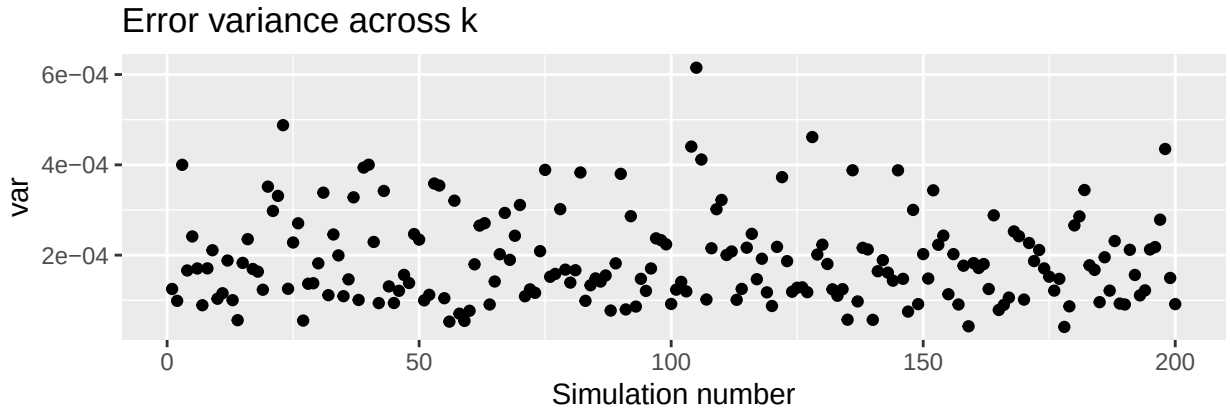
We were interested in understanding how our data affected the optimal choice of $k$ in the $k$ -nearest neighbors algorithm. We already experienced some variability when running our model through on different computers. Therefore, we wanted to see if more simulated datasets would produce the same variablility.

We generated our continuous predictors by sampling from a multivariate normal distribution. We generated our categorical predictors by sampling from a binomial distribution with our known class probabilities. We generated our "true" continuous response with our multiple linear regression model (trained on the true dataset). Normal random noise $N \sim (0, 0.25)$ was added to these predictions to generate our final continuous response. Lastly, we classified each observation as above/below the `total_costs` median to generate our class labels. We tried $k \in [1, 100]$ in each simulation and picked the best based on the error rate.




After 200, simulations, the chosen $k$ ranged from $[1, 90]$, almost the entire range of $k$s that we tried. However, most of the data is concentrated in the range from $[1, 25]$, which is consistent with our original results. We also found it interesting that in some simulations, a $k = 76$ could produce an error rate comparable to $k = 1$. This suggests that our data does not have a strong impact on the choice of $k$.




Because of the variability of our $k$, we then wondered if the best $k$ being chosen was close to other $k$ error rates, suggesting low variability of the error rates. However, there was no clear pattern emerging across the simulations. Some simulations had lower variance, all choices of $k$ produced similar error rates, and some simulations had higher variance, all choices of $k$ produced differing error rates. However, we must also note that with our scaled data, we are unfamiliar about what constitutes "small" or "large" quantities.

# Conclusions

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

We can see that salaries, inpatients, and the dummy variable for whether a hospital is specialized or not seem to play a major role overall in our prediction, since they're selected in almost all the models by size. Conversely, total days and whether a hospital had duplicate records in our dataset did not seem to play a major role in our models.

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.

We found that salaries greatly affects a hospital's total costs, which is intuitive.