

# quantoutcome

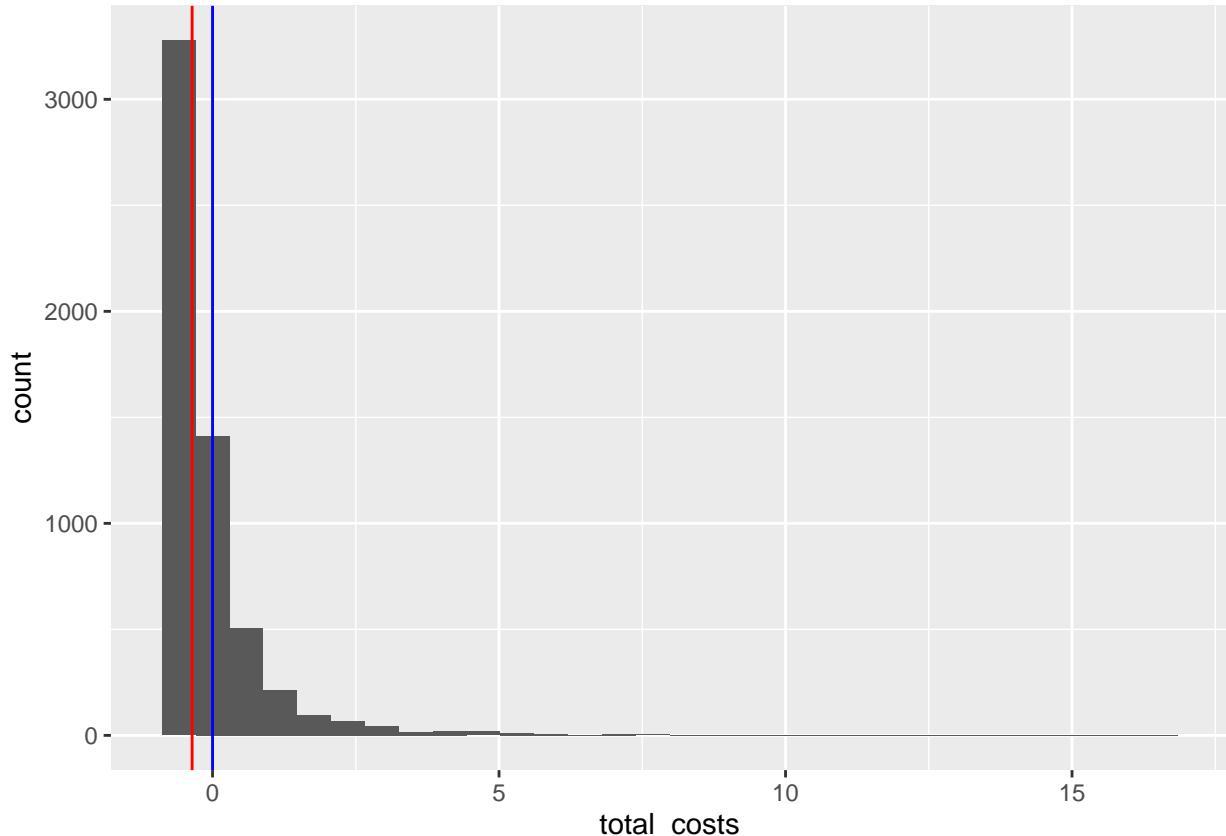
Lindsay Knupp

2024-05-13

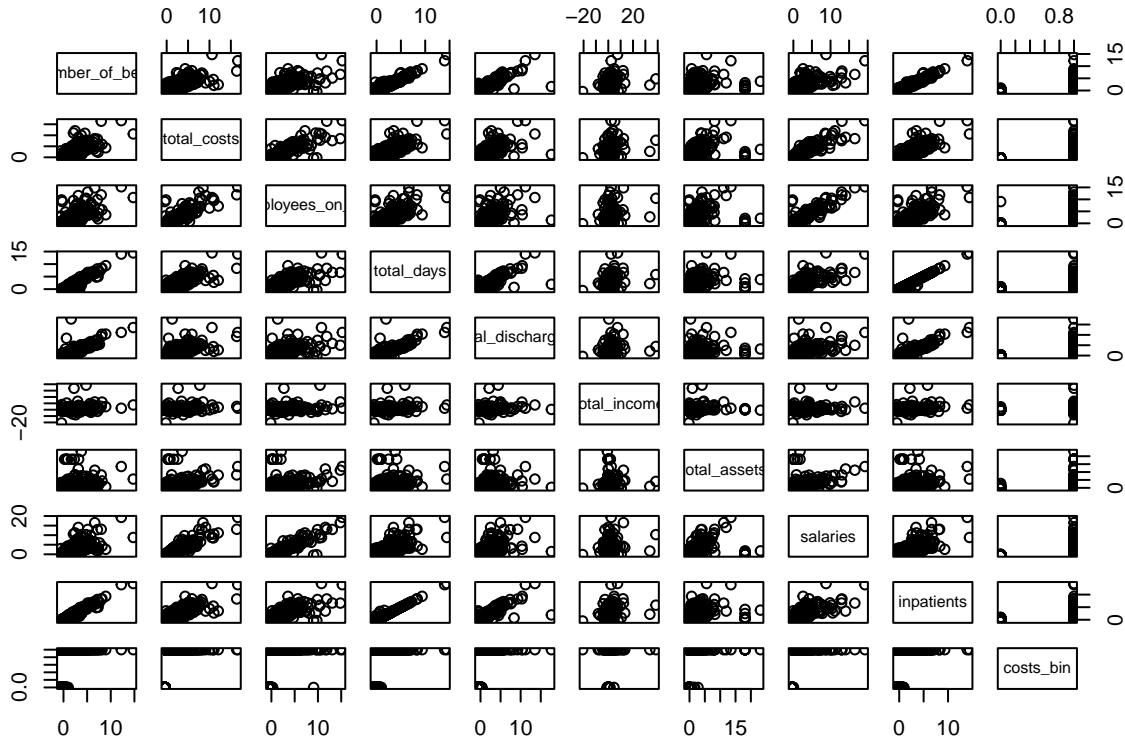
## Exploratory Data Analysis

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.



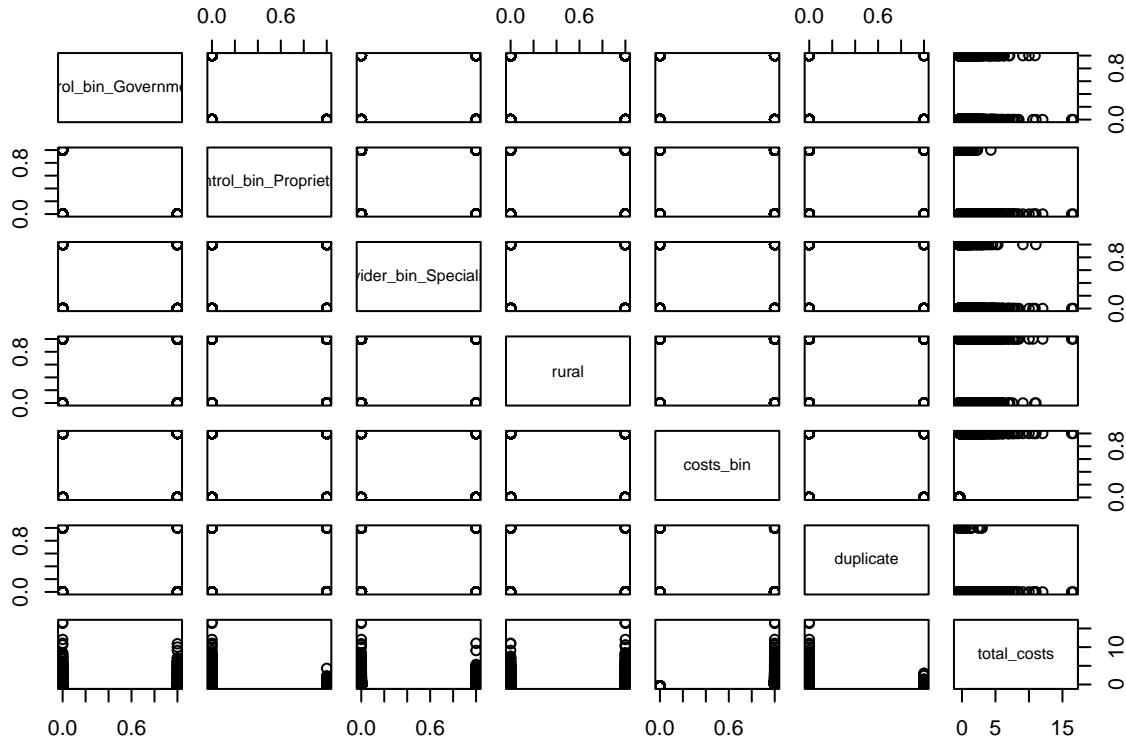
The pairs plots below visualize the relationships between all variables in the data (using the scaled data). We separated the data into two pairs plots, one including only numeric predictors and one including only categorical predictors, in order to see the relationships more clearly.



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. 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 higher 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
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 <sup>2</sup>	0.94
Adj. R <sup>2</sup>	0.94

	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

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 above, 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 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 <sup>2</sup>	0.94
Adj. R <sup>2</sup>	0.94

	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 adjusted  $R^2$  value of 0.94 indicates that this model with transformations and interactions explains about the same proportion of variability 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)

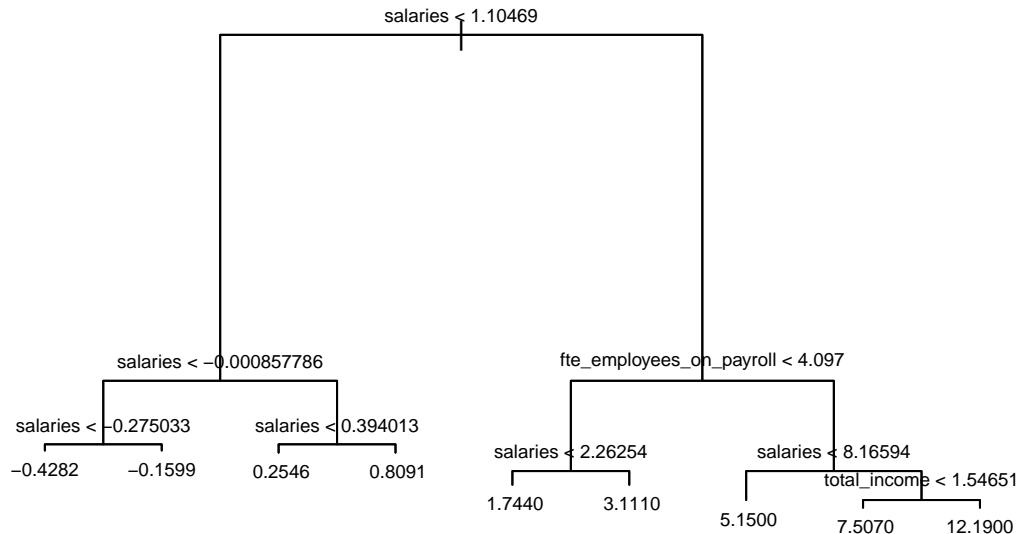
### Assumptions

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.



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.