Data Mining and Statistical Learning: Exercise 3

Tejaswi Pukkalla

April 8, 2019

Question 1:: Model Selection and Regulaization

We go back to the dataset GreenBuildings from one of our earlier exercises in an attempt to build the best predictive model of price possible for us. Before delving further into various models of pricing, we first clean the data by removing any incomplete observations with blank values. We plan to explore several different models of pricing and compare them together to see which would be a better fit for the question in hand. For the buildings with green rating, we also add additional classification on the basis of their classes (a,b or c).

As a side note, we consider all the buildings with green rating as green certified and do not distinguish between LEED or EnergyStar rating as we do not believe there would be much of a statiscally significant difference between LEED OR EnergyStar rating. Our expectation is that both entities would maintain the same high standards of certification.

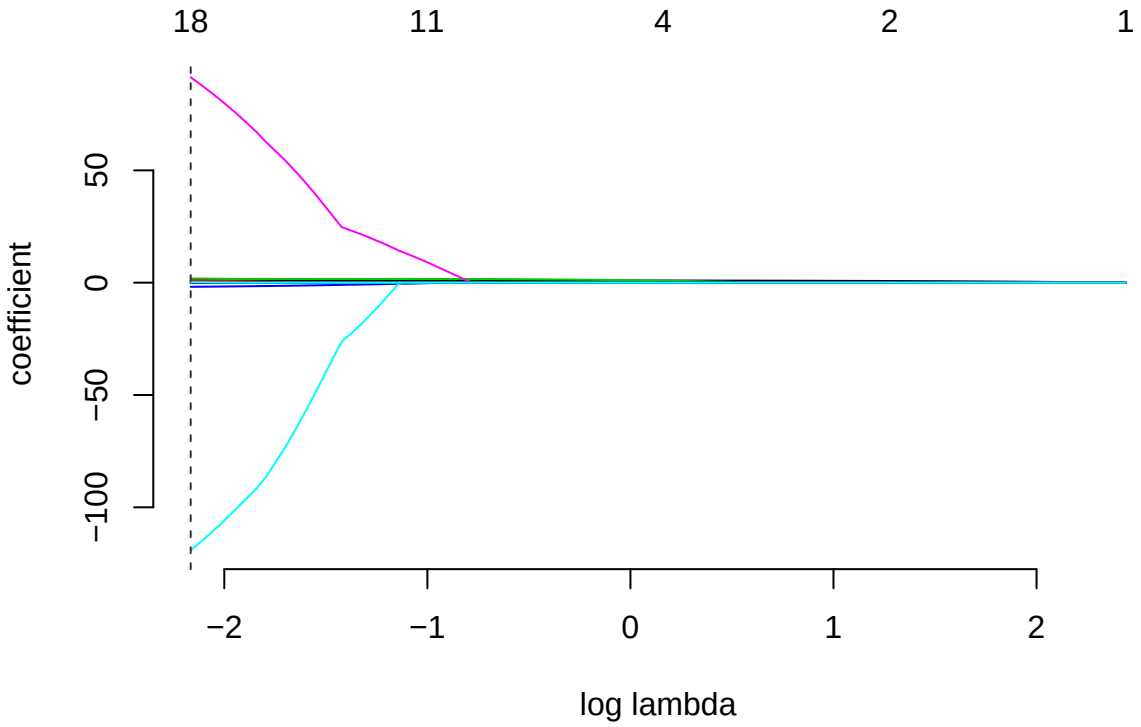
We start the process with a simple linear model, where we believe the important factors that determine the price of the building are the size of the building, number of stories, its class in case it is green rated, amenities represented with the help of indicator variables, fuel costs and clustering them on the basis of proximity.We follow it up with other models such as forward selection, backward selection, stepwise selection, for all the buildings together. We get the following number of factore being used by each model.

```
##              model_type Number_of_variables
## 1   lm_medium_all Model                  14
## 2  lm_forward_all Model                  36
## 3 lm_backward_all Model                  68
## 4     lm_step_all Model                  45
```

We see that models have different number of factors determining the prices. As our goal is to determine if there is and quantify the average change in rental income per square foot for buildings with green certification, holding all other variables constant, we must see if our target variable green rating is even used in these models.So now, let's see if the rental income is affected by the green rating.

```
##              model_type green_rating_value
## 1   lm_medium_all Model  0.541569057763953
## 2  lm_forward_all Model   1.24640605983934
## 3 lm_backward_all Model                 NA
## 4     lm_step_all Model    1.1492771990006
```

By looking at the coefficients on these various models, we can see that while backward selection doesn't take green rating into account for its final set of factors, it is predominantly showing an effect on the other models. Green rating on average, increases the rental income of a building by 0.54 dollars for linear model, 1.25 dollars for a forward selection and 1.15 dollars per square foot approximately. As an additional method to check for, we also use the gamma-lasso regression. This regression conveniently regularizes the model selection process by minimizing the deviance of the model and at the same time penalizing it for being overly complex - which is vital for improving out-of-sample predictions. The lasso approach uses the sparse matrix that we need to setup prior to give sparse solutions which automatically select variables for us. It provides a value much closer to our linear model at 0.35 dollars increase in rental income per square foot on an average. Hence, it also acts as a cross check to our previous conclusion.




```
## 23 x 1 sparse Matrix of class "dgCMatrix"
##                            seg100
## intercept          -4.983093e+00
## CS_PropertyID       .
## cluster             5.488910e-04
## size                5.798664e-06
## empl_gr             1.327552e-02
## leasing_rate        6.356737e-03
## stories             .
## age                -1.069636e-02
## renovated          -7.212772e-02
## class_a             1.875701e+00
## class_b             2.496831e-01
## LEED                9.997667e-01
## Energystar          .
## green_rating        3.294395e-01
## net                -1.835161e+00
## amenities           4.378266e-01
## cd_total_07        -9.994124e-05
## hd_total07          2.786562e-04
## total_dd_07         .
## Precipitation       .
## Gas_Costs          -1.191041e+02
## Electricity_Costs   9.140417e+01
## cluster_rent        1.029505e+00
```

Let's now try to look into the green rating effect and if it is different for different sets of buildings. We

have already divided the original datasets into subsets, classifying the green rated buildings on their class. Now we see if the green rating effect is more pronounced for one of these over the other.

```
##                   model_type green_rating_value
## 1        lm_medium_all Model  0.541569057763953
## 2       lm_forward_all Model   1.24640605983934
## 3      lm_backward_all Model                 NA
## 4          lm_step_all Model    1.1492771990006
## 5    lm_medium_class.a Model      0.12704056752
## 6   lm_forward_class.a Model               <NA>
## 7  lm_backward_class.a Model                 NA
## 8      lm_step_class.a Model   3.36886203206814
## 9    lm_medium_class.b Model   1.51102646253808
## 10  lm_forward_class.b Model   2.82293346500777
## 11 lm_backward_class.b Model                 NA
## 12     lm_step_class.b Model   9.52264948740111
## 13   lm_medium_class.c Model   4.42963480877447
## 14  lm_forward_class.c Model   6.41033692954571
## 15 lm_backward_class.c Model                 NA
## 16     lm_step_class.c Model   9.52264948740111
```

We see that the increase in rental income is way higher for Class B and C buildings as compared to Class A buildings. At first sight, it seems counter-intuitive. But this is a good reflection of how easy it is to get confounding variables messing up your interpretation. My understanding of why it would be lower increase in rental income per square foot for Class A buildings over Class B and Class C would be as follows. On an average, the Class A buildings might be way larger than Class B building because of which while they may be priced the same way, the per square foot rate might vary. There might be other confounding variables that are present in Class B and Class C buildings but not in Class A buildings, thereby inflating their effect on the rental income.

Question 2:: What Causes What?

1. If we merely run a regression on the crime rate and number of cops in a city for a few different cities, we would get completely haphazard results because there would be way too many issues with that regression. For starters, the crime data should be relative to the population of the city. For example, if there are 20 crimes a day in a city of 200, that's at least 10% of the population being affected. However, if there are 500 crimes being committed in a city of about 20,000, that's about 2.5% of the population being affected. Hence, the crime data needs to be relative and not absolute. The same goes for number of cops. It should be the proportion of cops in population that should be measured. Another issue would be omitted variable bias, such as, a city that has no tourism attractions might not have many visitors reducing the crime rate by a large amount, but if that city has high cops for other reasons, the low crime rate gets attributed to the wrong resons of causality. Each city that we choose has its own fixed characteristics that also explains crime rate, without accounting for which, regression run on number of cops would give a very biased result.

2. By observing the rise in the number of cops, change in dependent variable, on a day that showed high terror alert, which is not related to change in the daily expected crime rate, the independent variable, they were able to create a controlled environment where there was a natural shifting of the dependent variable in question, holding all else constant. By doing that, they were able to run a regression on the crime rate and how it changed when there was an influx of cops in the city. They were able to show that the rise in number of cops reduced the daily crime rate of the city by more than 7 on an average with other factors remaining the same, statistically significant at 5%. Even after controlling for any change in the number of tourists/visitors or basically any change in the amount of potential victims, the crime rate still went down by about 6 units.

3. One of the most serious and hard to evade issue in linear regression is the omitted variable bias. It is when by omitting a variable that could potentially affect the independent variable in a significant manner, we confound the effect of the present dependent variables in the regression. The researchers had to control for metro ridership to see if there was any reduction in the number of potential victims of crime due to the terror alert that might bring the crime rate down. By controlling for this variable, they could now see how much the overall crime rate would go down due to mainly the increase in number of cops.

4. In Table 4, the researchers are basically checking if the crime rate is affected in a particular district more than the others when the number of cops increases. They find that District 1 sees an almost 10 units of reduction in the crime rate overall and about 6.2 units more than the average reduction, statistically significant at 5%. The other districts saw an overall decrease in crime rate of almost 8 units, a 0.6 units more than the overall expected drop in the crime rate even after controlling the metro riders and ensuring there wasn't much of a difference in the chances of the crime rate to drop intrinsically. With this data, we can clearly conclude that an increase in the cops does have a negative effect on the crime rate of a city.

**********THE END**********