Homework 3

Frank Chou

April 4, 2019

Model Selection and Regularization: Green Buildings

Revisiting an old database, **greenbuildings.csv** from exercise 1, we are going to utilize new models and model generation to better predict and evaluate the effects of green buildings on rental prices, holding all other variables constant. Starting with 7,894 rental listings, we first clean the data set by removing any listing with omitted variables, resulting in 7,820 observations. Moving forward, it is presumed that variables **Energystar** and **LEED** would have a similar approach in encouraging builders to design buildings to be more energy efficient. In other words, a building is 'green' if it received either or both of the awards. This is distinction is captured in the dummy variable **green_rating**.

Model Selection

Selecting a valid model to determine the rental effects of a green building is difficult given the number of variables and potential permutations of the regression model. To begin, a hand-crafted model will be created to estimate the effects of a green building on rental prices. Next will be three iterative models: the first would be a **foward selection** model - where a model will be created from the ground up from no variables to an optimal model; the second model would be a **backwards selection** - similar to a forward selection but in this case from the greatest number of variables then reduced to the optimal; and the third **stepwise selection** - where we start with a reasonable model (the hand crafted one) and then consider all pairwise interactions for everything in the model as well as other variables not explicitly named in the base model.

The Best model

By applying four different models to the same data set, we come to four different ways to interpret the date. In this case, the method for each of the models results in different number of coefficients for 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 the **backward model** has the most coefficient, while our **hand-built** model has the least. But the number of coefficient does not mean that a model is better. As our goal is to determine and quantify the average change in rental income per square foot in relation to green certification, holding all other variables fixed. We must see if our target variable **green_rating** is even used in these algorithmically-derived models.

Determining Green Rating Effects on Rental Income

```
##              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
```

Here we find that the **backward model** does not have the variable **green_rating** however this does not mean it doesn't have any interaction variables. But given the difficulty in interpreting interaction variables, i.e. understanding what **green_rating:amenities** means. As for our other model, we see that

a **green_rating** imparts a 0.541 increase of rental income in our hand-crafted medium model, a 1.24 in increase in the **forward model**, and 1.14 increase in our **step model**.

Determining Green Rating Effects on Rental Income by Building Class

Using the same approach in applying different models to the generalized data set, the next step is to apply the same models to subsetted data. In this case, subsetting based on the **building class** of the property in question: **class.a** and **class.b**. We find the following **green_rating** effects across both classes.

```
##                    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
```

Here we find that there is a positive effect of a **green_rating** on the three types of buildings. By comparing models that are applicable to both classes, in this case we can examine the **step model**, we can see that the effect for a **green_rating** is greater for **class.b** and **class.c** buildings than **class.a**. As for the reasoning behind this phenomena, one explanation would be that the cost of retrofitting an old building would be greater than designing a new building according to energy saving standards.

What causes what?

Question 1

The reason one cannot take crime and police data from different cities and run a regression that will have a generalized model for the entire dataset is because of exogenous factors that may be at play in one city compared to another. For example, if city A has only a population of 100, compared to city B of 100,000, doubling the number of police personnel in each city would have a different effect on crime, if we presume that police is negatively correlated with crime. At the same time, we also do not take into consideration the concept as to whether or not there are special events held in each city that would artificially increase the number of police personnel that other cities do not enact - such as Washington DC's terrorism alert that would deploy more law enforcement.

Question 2

The researchers utilized the terror alert system that was in place in Washington DC. This alert serves as a dummy variable that would artificially increase the number of police in the street that would be exogenous to the street crime norm. The researchers found that there was a small, but statistically significant effect from deploying more police in the streets from comparing days where there was no alert to days where there was an alert.

Question 3

However, the consideration here is whether or not another factor - such as tourism, would be indirectly tied to crime levels given that tourists are a prime population at risk to crime. The researchers addressed this issue by examining the relative metro ridership comparing days when the alert was in place and when it was not. If the two days are the same, this instrumental variable effectively stated that despite the alert rating, passengers and by extension tourists, still went on with their day-to-day lives.

Question 4

The model here is using a clustered regression of the effects of a high alert on specific district in Washington DC - asking what is the effect of a high alert on crime controlling for the fact that different subdivisions (districts) within Washington DC would have pre-set levels of law enforcement that is not consistent throughout the city - namely the Capitol National Mall compared to other districts. The conclusion that the researchers came to was that yes, a high alert increased police presence and decreased crime, but one you control for specific districts, most the decrease was found within one district - the Capitol Mall.