Untitled

Linh Nguyen Madeline Lin

4/7/2019

Exercise 3

Exercise 3.1

Build the best predictive model possible for price.

I preprocess the data as follow:

1. Clear out all the observations with missing values in the data.
2. Scale down the size by $\frac{1}{10000}$ to abide the lasso computation's requirement so it can converge.
3. Omit building with leasing rate = 0 because it makes no sense to charge a rent for unoccupied buildings (they are either under construction or abandoned or other reasons)

To find best predictive model, I will deploy stepwise regression and lasso regression.

a. Stepwise regression using LEED and EnergyStar explicitly and using green rating (implying LEED and Energy Star is green certified). In both models, I started with the null model by regressing rent on one, followed by adding new variables in the forward selection method. This is the baseline model to run stepwise regression

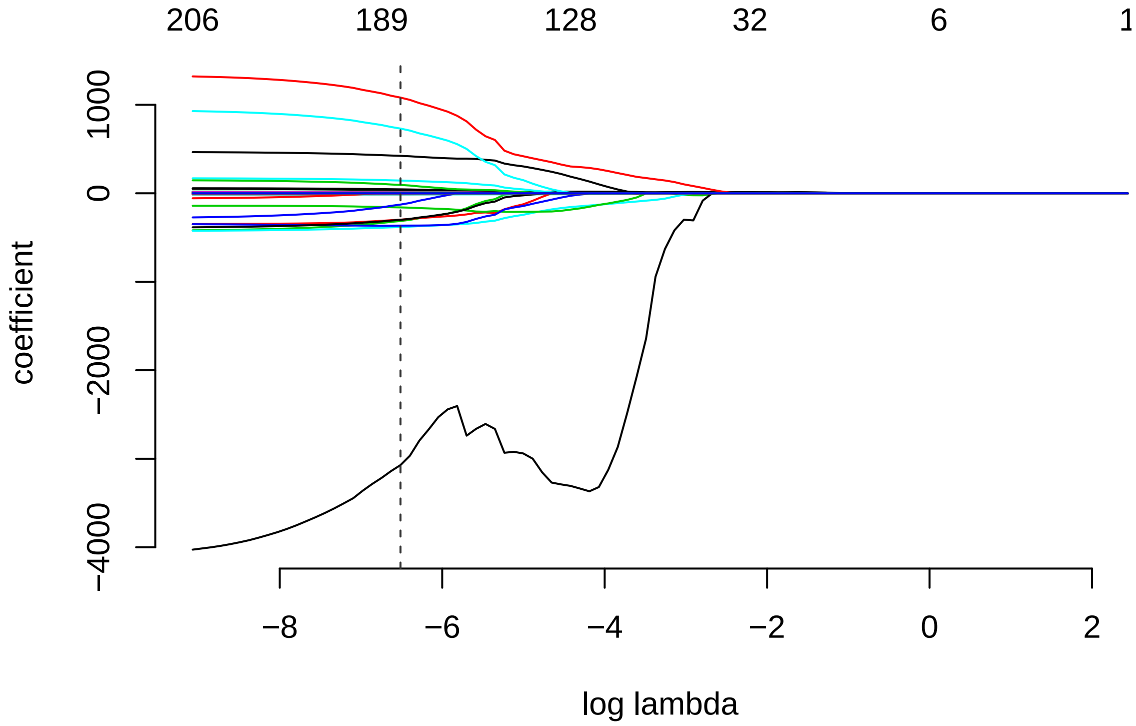
Model using LEED and EnergyStar explicitly:

```
## Rent ~ cluster_rent + size + class_a + class_b + cd_total_07 +
##     age + cluster + net + Electricity_Costs + hd_total07 + leasing_rate +
##     LEED + amenities + cluster_rent:size + size:cluster + cluster_rent:cluster +
##     class_b:age + class_a:age + cd_total_07:net + cd_total_07:hd_total07 +
##     cluster_rent:age + size:leasing_rate + size:Electricity_Costs +
##     size:class_a + age:Electricity_Costs + cluster_rent:leasing_rate +
##     cluster_rent:net + cluster_rent:LEED + Electricity_Costs:hd_total07 +
##     size:cd_total_07 + cluster:Electricity_Costs + class_a:cd_total_07 +
##     cluster:hd_total07 + cluster_rent:hd_total07 + size:age +
##     size:class_b + class_b:amenities + size:amenities + Electricity_Costs:amenities +
##     cluster_rent:amenities + cluster:leasing_rate + age:cluster +
##     size:hd_total07 + age:LEED
```

Model using green certified category:

```
## Rent ~ cluster_rent + size + class_a + class_b + cd_total_07 +
##     age + cluster + net + Electricity_Costs + hd_total07 + leasing_rate +
##     green_rating + amenities + cluster_rent:size + size:cluster +
##     cluster_rent:cluster + class_b:age + class_a:age + cd_total_07:net +
##     cd_total_07:hd_total07 + cluster_rent:age + size:leasing_rate +
##     size:Electricity_Costs + size:class_a + age:Electricity_Costs +
##     cluster_rent:leasing_rate + cluster_rent:net + Electricity_Costs:hd_total07 +
##     size:cd_total_07 + cluster:Electricity_Costs + cluster:hd_total07 +
##     class_a:cd_total_07 + size:age + size:class_b + size:hd_total07 +
##     cluster_rent:hd_total07 + green_rating:amenities + size:amenities +
##     class_b:amenities + cluster_rent:amenities + Electricity_Costs:amenities +
##     cluster:leasing_rate + age:cluster
```

b. Lasso regression using LEED and EnergyStar separately, then second lasso regression combining them into a single “green certified” category. I created sparse matrix including all interactions terms then run lasso regression on it

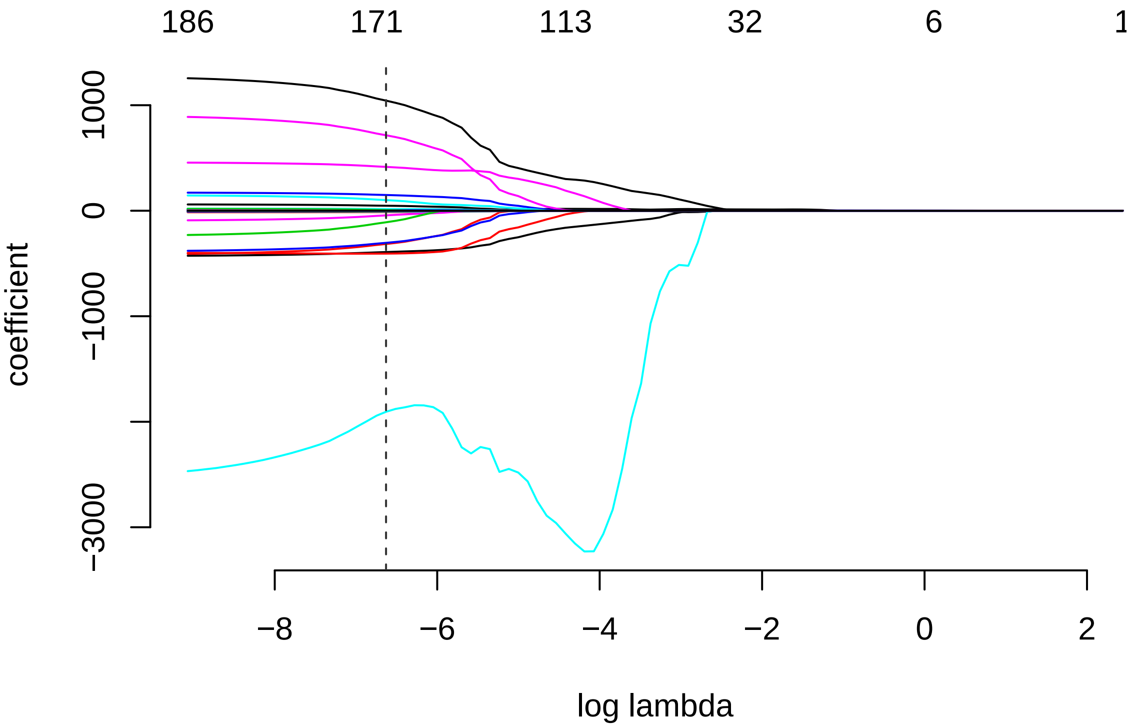


Minimum AIC occurs at segment 78.

Thus, coefficients chose in the model are:

```
## Rent ~ cluster + size + leasing_rate + stories + age + renovated +
##     class_a + class_b + LEED + Energystar + net + amenities +
##     hd_total07 + total_dd_07 + Precipitation + Electricity_Costs +
##     cluster_rent + cluster:size + cluster:empl_gr + cluster:leasing_rate +
##     cluster:stories + cluster:age + cluster:renovated + cluster:class_a +
##     cluster:class_b + cluster:LEED + cluster:Energystar + cluster:net +
##     cluster:hd_total07 + cluster:total_dd_07 + cluster:Precipitation +
##     cluster:Gas_Costs + cluster:Electricity_Costs + cluster:cluster_rent +
##     size:empl_gr + size:leasing_rate + size:stories + size:age +
##     size:renovated + size:class_a + size:class_b + size:LEED +
##     size:Energystar + size:net + size:amenities + size:cd_total_07 +
##     size:hd_total07 + size:Precipitation + size:Gas_Costs + size:Electricity_Costs +
##     size:cluster_rent + empl_gr:leasing_rate + empl_gr:stories +
##     empl_gr:age + empl_gr:renovated + empl_gr:class_b + empl_gr:LEED +
##     empl_gr:net + empl_gr:amenities + empl_gr:hd_total07 + empl_gr:Precipitation +
##     empl_gr:Electricity_Costs + empl_gr:cluster_rent + leasing_rate:stories +
##     leasing_rate:age + leasing_rate:class_a + leasing_rate:class_b +
##     leasing_rate:LEED + leasing_rate:net + leasing_rate:amenities +
##     leasing_rate:cd_total_07 + leasing_rate:hd_total07 + leasing_rate:total_dd_07 +
##     leasing_rate:Precipitation + leasing_rate:Gas_Costs + leasing_rate:Electricity_Costs +
##     leasing_rate:cluster_rent + stories:age + stories:renovated +
##     stories:class_a + stories:class_b + stories:LEED + stories:Energystar +
##     stories:net + stories:amenities + stories:cd_total_07 + stories:hd_total07 +
##     stories:Precipitation + stories:Gas_Costs + stories:Electricity_Costs +
##     stories:cluster_rent + age:renovated + age:class_a + age:class_b +
##     age:LEED + age:Energystar + age:net + age:amenities + age:cd_total_07 +
##     age:hd_total07 + age:total_dd_07 + age:Precipitation + age:Gas_Costs +
##     age:Electricity_Costs + age:cluster_rent + renovated:class_a +
##     renovated:class_b + renovated:LEED + renovated:Energystar +
##     renovated:net + renovated:amenities + renovated:cd_total_07 +
##     renovated:hd_total07 + renovated:total_dd_07 + renovated:Precipitation +
##     renovated:Gas_Costs + renovated:Electricity_Costs + renovated:cluster_rent +
##     class_a:LEED + class_a:Energystar + class_a:net + class_a:amenities +
##     class_a:cd_total_07 + class_a:hd_total07 + class_a:total_dd_07 +
##     class_a:Precipitation + class_a:Gas_Costs + class_a:Electricity_Costs +
##     class_a:cluster_rent + class_b:LEED + class_b:Energystar +
##     class_b:net + class_b:amenities + class_b:cd_total_07 + class_b:hd_total07 +
##     class_b:total_dd_07 + class_b:Precipitation + class_b:Gas_Costs +
##     class_b:Electricity_Costs + class_b:cluster_rent + LEED:Energystar +
##     LEED:net + LEED:amenities + LEED:cd_total_07 + LEED:total_dd_07 +
##     LEED:Precipitation + LEED:Gas_Costs + LEED:Electricity_Costs +
##     LEED:cluster_rent + Energystar:net + Energystar:amenities +
##     Energystar:cd_total_07 + Energystar:total_dd_07 + Energystar:Precipitation +
##     Energystar:Electricity_Costs + Energystar:cluster_rent +
##     net:amenities + net:cd_total_07 + net:total_dd_07 + net:Precipitation +
##     net:Gas_Costs + net:Electricity_Costs + net:cluster_rent +
##     amenities:hd_total07 + amenities:total_dd_07 + amenities:Precipitation +
##     amenities:Gas_Costs + amenities:Electricity_Costs + amenities:cluster_rent +
##     cd_total_07:hd_total07 + cd_total_07:total_dd_07 + cd_total_07:Precipitation +
##     cd_total_07:Electricity_Costs + cd_total_07:cluster_rent +
##     hd_total07:total_dd_07 + hd_total07:Precipitation + hd_total07:cluster_rent +
##     Precipitation:Gas_Costs + Precipitation:Electricity_Costs +
##     Precipitation:cluster_rent + Gas_Costs:Electricity_Costs +
##     Gas_Costs:cluster_rent + Electricity_Costs:cluster_rent
```

For Lasso regression with green-certified variable:



Minimum AIC occurs at segment 79.

Coefficients chose in the model are:

```
## Rent ~ cluster + size + leasing_rate + stories + age + renovated +
##     class_a + class_b + green_rating + net + amenities + hd_total07 +
##     total_dd_07 + Precipitation + Electricity_Costs + cluster_rent +
##     cluster:size + cluster:empl_gr + cluster:leasing_rate + cluster:stories +
##     cluster:age + cluster:renovated + cluster:class_a + cluster:class_b +
##     cluster:green_rating + cluster:net + cluster:amenities +
##     cluster:hd_total07 + cluster:total_dd_07 + cluster:Precipitation +
##     cluster:Gas_Costs + cluster:Electricity_Costs + cluster:cluster_rent +
##     size:empl_gr + size:leasing_rate + size:stories + size:age +
##     size:renovated + size:class_a + size:class_b + size:green_rating +
##     size:amenities + size:cd_total_07 + size:hd_total07 + size:Precipitation +
##     size:Gas_Costs + size:Electricity_Costs + size:cluster_rent +
##     empl_gr:leasing_rate + empl_gr:stories + empl_gr:age + empl_gr:renovated +
##     empl_gr:class_b + empl_gr:green_rating + empl_gr:amenities +
##     empl_gr:hd_total07 + empl_gr:Precipitation + empl_gr:Electricity_Costs +
##     empl_gr:cluster_rent + leasing_rate:stories + leasing_rate:age +
##     leasing_rate:class_a + leasing_rate:class_b + leasing_rate:green_rating +
##     leasing_rate:net + leasing_rate:amenities + leasing_rate:cd_total_07 +
##     leasing_rate:hd_total07 + leasing_rate:total_dd_07 + leasing_rate:Precipitation +
##     leasing_rate:Gas_Costs + leasing_rate:Electricity_Costs +
##     leasing_rate:cluster_rent + stories:age + stories:renovated +
##     stories:class_a + stories:class_b + stories:green_rating +
##     stories:net + stories:amenities + stories:cd_total_07 + stories:hd_total07 +
##     stories:Precipitation + stories:Gas_Costs + stories:Electricity_Costs +
##     stories:cluster_rent + age:renovated + age:class_a + age:class_b +
##     age:green_rating + age:net + age:amenities + age:cd_total_07 +
##     age:hd_total07 + age:total_dd_07 + age:Precipitation + age:Gas_Costs +
##     age:Electricity_Costs + age:cluster_rent + renovated:class_a +
##     renovated:class_b + renovated:green_rating + renovated:net +
##     renovated:amenities + renovated:cd_total_07 + renovated:hd_total07 +
##     renovated:total_dd_07 + renovated:Precipitation + renovated:Gas_Costs +
##     renovated:Electricity_Costs + renovated:cluster_rent + class_a:green_rating +
##     class_a:net + class_a:amenities + class_a:cd_total_07 + class_a:hd_total07 +
##     class_a:total_dd_07 + class_a:Precipitation + class_a:Gas_Costs +
##     class_a:Electricity_Costs + class_a:cluster_rent + class_b:green_rating +
##     class_b:net + class_b:amenities + class_b:cd_total_07 + class_b:hd_total07 +
##     class_b:total_dd_07 + class_b:Precipitation + class_b:Gas_Costs +
##     class_b:Electricity_Costs + class_b:cluster_rent + green_rating:net +
##     green_rating:amenities + green_rating:cd_total_07 + green_rating:hd_total07 +
##     green_rating:Precipitation + green_rating:Gas_Costs + green_rating:Electricity_Costs +
##     green_rating:cluster_rent + net:amenities + net:cd_total_07 +
##     net:total_dd_07 + net:Precipitation + net:Gas_Costs + net:Electricity_Costs +
##     net:cluster_rent + amenities:hd_total07 + amenities:total_dd_07 +
##     amenities:Precipitation + amenities:Gas_Costs + amenities:Electricity_Costs +
##     amenities:cluster_rent + cd_total_07:hd_total07 + cd_total_07:total_dd_07 +
##     cd_total_07:Precipitation + cd_total_07:Electricity_Costs +
##     cd_total_07:cluster_rent + hd_total07:total_dd_07 + hd_total07:Precipitation +
##     hd_total07:cluster_rent + total_dd_07:Electricity_Costs +
##     Precipitation:Gas_Costs + Precipitation:Electricity_Costs +
##     Precipitation:cluster_rent + Gas_Costs:Electricity_Costs +
##     Gas_Costs:cluster_rent + Electricity_Costs:cluster_rent
```

c. To measure predictive power, I employed k-fold cross validation. I set k = 15 and calculated 4 models' CVs. Stepwise models has smaller CVs compared to Lasso models. Between 2 stepwise model, the one with a single green certified category has lower CV.

In conclusion, the stepwise model with "green certified" category had the minimum CV, and therefore it is the best predictive model possible for rent price.

```
## [1] 9.165846 9.163988 9.230472 9.174818
```

Use this model to quantify the average change in rental income per square foot (whether in absolute or percentage terms) associated with green certification, holding other features of the building constant.

```
##                  (Intercept)                 cluster_rent
##                        7.545                        0.851
##                         size                      class_a
##                       -0.158                        5.808
##                      class_b                  cd_total_07
##                        4.619                        0.000
##                          age                      cluster
##                        0.027                       -0.007
##                          net            Electricity_Costs
##                       -0.576                     -287.886
##                   hd_total07                 leasing_rate
##                       -0.001                       -0.029
##                 green_rating                    amenities
##                        2.295                       -1.688
##            cluster_rent:size                 size:cluster
##                        0.004                        0.000
##         cluster_rent:cluster                  class_b:age
##                        0.000                       -0.045
##                  class_a:age              cd_total_07:net
##                       -0.029                        0.001
##       cd_total_07:hd_total07             cluster_rent:age
##                        0.000                       -0.002
##            size:leasing_rate       size:Electricity_Costs
##                        0.001                        2.958
##                 size:class_a        age:Electricity_Costs
##                       -0.128                        1.654
##    cluster_rent:leasing_rate             cluster_rent:net
##                        0.002                       -0.116
## Electricity_Costs:hd_total07             size:cd_total_07
##                        0.045                        0.000
##    cluster:Electricity_Costs           cluster:hd_total07
##                        0.170                        0.000
##          class_a:cd_total_07                     size:age
##                        0.001                       -0.001
##                 size:class_b              size:hd_total07
##                       -0.090                        0.000
##      cluster_rent:hd_total07       green_rating:amenities
##                        0.000                       -2.151
##               size:amenities            class_b:amenities
##                        0.031                        1.132
##       cluster_rent:amenities  Electricity_Costs:amenities
##                       -0.066                       97.106
##         cluster:leasing_rate                  age:cluster
##                        0.000                        0.000
```

Holding all other significant features of the building fixed, green certified (LEED and EnergyStar combined) buildings are expected to be 2.295 $\$/ft^2$ per year more expensive than non-green buildings.

Assess whether the "green certification" effect is different for different buildings, or instead whether it seems to be roughly similar across all or most buildings.

As mentioned, the stepwise model with collapsed category of green certified has the best predictive power. Green certification buildings with amenities is -2.151 $\$/ft^2$ per year. It less than green certification buildings without amenities. Hence, "green certification" effect is different for buildings of with and without amenities.

One possible explanation is that the green buildings with amenities are normally considered as commercial buildings, so the buildings need to pay the energy fee according to different scheme compared to residential buildings. It can be that commercial rate is higher than residential rate. Thus, residents in the green buildings with amenities must pay more than those in the green buildings without amenities. Therefore green buildings with amenities have incentive to lower the rent fee to captivate more potential renters.

Question 2:

Probelm Summary:

The problem asks us to listen to the podcast from Planet Money. This podcast is an interview between Charles Wheelan (the guest, author of the book "Naked Statistics") and Robert Smith together with Jacob Goldstein (the hosts). The topic of this interview is about "What causes What". During the interview, they have mentioned correlation and causation of different stories, including children's education, women's health, city's crime, and so forth. Crime, which is one of the topics discussed in the interview, is the problem we need to address in this question.

So, what we need to do is to use our knowledge of statistical learniing to answer the following four questions related to econometrics based on figures and tables provided in the questions.

1) Why can't I just get data from a few different cities and run the regression of "Crime" on "Police" to understand how more cops in the streets affect crime? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city.)

Intuitively, if we want to explore the relationship between "Crime" and "Police", what we should do is to regress "Crime" on "Police" and expect that the coefficient of "Police" would be negative since normally more cops in the street will lead to a lower crime rate. However, from the dicussion of Wheelan and Smith, we can see that Smith saying "it's really hard to tease out because obviously high-crime cities have an incentive to hire a lot of cops". He meant that usually more cops will lead to low crime rate. But, it is also possible that high-crime rate cities will have more cops.

There exists a two-side interaction for both "Crime" and "Police". So it is kind of messey to merely run the regression of "Crime" on "Police". If we use a professional term learned from the Econometrics here, it should be called "Endogeneity". "Simultaneity" leads to a biased simple OLS regression. There should be some omitted variables which do not have an effect on crime but have an effect on police. What need to find the instrument variable in order to find a causal effect between "Police" and "Crime".

2) How were the researchers from UPenn able to isolate this effect? Briefly describe their approach and discuss their result in the "Table 2" below, from the researcher's paper.

The researchers from UPenn isolated this effect by using "Instrument Variable Method". In brief, instrument variable(Z) is corelated to the X variable, but uncorelated to u(residules) in the regression. So here, they pick "the Terrorism Alert System" as the instrument variable because terrorism is corelated to police but has nothing to do with crime. They also select the area Washington, D.C. because it is a typical area to be a terrorism target.

Therefore, on orange alert days, they can explore what happens to street crime when there are extra police on the streets because of terrorism instead of crime. If the street crime still goes down, then we can say that police have a positive deterrence on the street crime because police can make the streets safer and let things like murder, robbery, assult go down.

As it is indicated in the "Table 2" from the researcher's paper, their results are as follows:

(1) The crime rate indeed decreased during orange alert days when more cops were on the streets.

(2) Specifically, they found that high alert days were correlated with approximately an average of 7.316 unit decrease in the number of crimes committed, and, approximately an average of 6.046 unit decrease in the number of crimes committed if including a log midday Metro ridership, holding all else fixed.

(3) Both of these two above estimates were statistically significant at the 5% level of significance.

(4) With including a log midday Metro ridership, the coefficient of high-alert variable was a little bit smaller. Additionally, one unit increase in Metro ridership would lead to approximately an average of 17 crimes per day, holding all else fixed.

(5) The above estimate was statistically significant at the 1% level of significance.

3)Why did they have to control for Metro ridership? What was that trying to capture?

In the above question, there is actual one problem needs to be addressed. "the Terrorism Alert System", as an instrument variable, should not effect another variable which has an impact on "Crime". To be more specific, let's say becuase of terrorism, it is likely that robbers hide in their rooms because they're afraid of the elevated terror level, or, some tourisits from other cities don't dare to visit Washington, D.C. result from the dangerous environment as mentioned in the case. These kind of changes will lead to a lower crime rate because the number of people who commit crime and the number of people who are suffered from crime both decrease. In one word, there will be less chances for crime, resulting in a lower crime rate.

Consequently, researchers had to deal with this problem. They came up with a solution to check the number of victims by looking at ridership levels on the Metro system. Since people who visit the city usually use public transportation such as Metro, we can check whether the number of victims change or not on high-terror days. Logically, if we rerun the regression, and, the coefficient of high-alert variable is still negative after considering this factor, then we can say that police indeed has a positive impact on a lower crime rate.

As we can discover from the Table 2, by including the varibale of log midday ridership, the coefficient of the high-alert variable is still smaller than zero. This result teases out the possibility that terrorism effects the crime rate. By using the control variable Metro ridership to capture the number of victims, the researchers can verify that the accuracy of the impact that police has on crime is not confounded by the change of the number of victims.

4)Below I am showing you "Table 4" from the researchers' paper. Just focus on the first column of the table. Can you describe the model being estimated here? What is the conclusion?

The description the model being estimated here:

The model is a fixed effects specification model. It includes panel data set of different districts and add one variable of log midday metro ridership. Specifically, it runs a regression of estimating the effect police has on crime in different districts of Washington, D.C. during high alert days, holding all else fixed.

The conclusion for analyzing Column 1 are:

(1) During high-alert days, District 1 has the largest decrease, approximately an average of 2.621 units, in crime when there is one unit of police increase, holding all else fixed. The cofficient is statistically significant at 1% level.

(2) As for other districts, one unit of police increase only leads to approximately an average of 0.571 units of decrease in crime, holding all else fixed. The cofficient is not statistically significant at 1% level as the confidence interval lies on the coefficient of zero.

(3) We can suppose that District 1 is the area where most places of security importance exist. For example, the White House, Supreme Court and Capitol and so on. Therefore, masses of police are there to protect the security. Owing to the large amount of police officers, there is a strong statistically significant decrease in the crime rate during this high-alert days.