Homework I

OLS - SE - Non-parametric models

Fernando Rios-Avila

Part I: OLS and Standard errors

Consider the dataset `hhprice.dta`, available from `frause` repository. The dataset contains information on 10k house values with data on house characteristics and location.

1. Using data for all houses with information on latitude and longitude, estimate a log linear model on the determinants of TownHouse prices (`type_h==1`). Interpret the Results and discuss significance of the coefficients.

Be mindful of using data with sufficient variation, specially if using categorical variables. (for example, if using categorical variables, avoid using groups with fewer than 10 observations).

2. Consider the code here, adapt the code to estimate robust Standard errors and robust standard errors clustered by `postcode` for your model.
    - Is there any reason to believe errors are related to `postcode`?
    - Does any of your conclusions (regarding significance) change when using clustered standard errors?

Part II: Model selection and Cross validation

As explained in class, some times models are selected based on their capability to make out-sample predictions. This is what methods like Lasso and Ridge use to select among a myriad of possible models.

One way to evaluate the predictive power of a model is to use cross validation.

Using the same dataset as in Part I, but now consider two models: - The same one you propose in Part I - A model where all variables are interacted with each other (over parametrized model)

1. Report the number of parameters used, the goodness of fit (R2 and adjusted R2) of the models, and comment on these results.

2. Implement a simple one round 5-fold Crossvalidation procedure to evaluate the predictive power of the two models.

   Process:

   - Split the data into 5 random groups of equal size
   - Using the first 4 groups, estimate the two models, and predict the values for the 5th group
   - Repeat the process for all possible combinations of 4 groups (there are 5 possible combinations)
   - For each model, calculate the MSE (mean squared error) of the predictions.

The MSE is estimated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the actual value of the dependent variable, and $\hat{y}_i$ is the predicted value of the dependent variable using the different models.

- What do you conclude?
- Repeat the process again, and see if your conclusions change.

Part III: Non-parametric models

- In real-estate, Location of a house is a very important determinant of house prices. One way to measure this location effect is to use the distance to the city center, as the center may offer the most amenities and services.

- Consider your model from Part I. If you have not done so, add the distance (log distance) to the city center as a regressor.

- Following the code here, adapt it to estimate Robinson (1988) Root-N-Consistent Semiparametric regression

  The model:

$$lprice = x\beta + \theta(distance) + e$$

- Compare the results of other coefficients with the OLS model. What do you conclude? Does modeling distance non-parametrically matter?

- Based on your results, Is the effect of distance on house prices linear?

- A second way to explore the role of distance on housing price-determination is by looking at the model coefficients across houses at different distances from the city center.

- Classify houses into 5 groups based on their distance to the city center, and estimate the model for each group. (this is a rough approach to the Smooth varying coefficient model)

  - What do you conclude? Does distance affect how other amenities impact house prices?