Erasmus School of Economics

MOOC Econometrics

Case Project – House Prices

Notes:

- See website for how to submit your answers and how feedback is organized.
- This exercise uses the datafile CaseProject-HousePrices and requires a computer.
- The dataset CaseProject-HousePrices is available on the website.
- Perform all tests at a 5% significance level.

Goals and skills being used:

- Experience the processes of variable transformation and model selection.
- Apply tests to evaluate models, including effects of endogeneity.
- Study the predictive ability of a model.

Background

This project is of an applied nature and uses data that are available in the data file Capstone-HousePrices. The source of these data is Anglin and Gencay, "Semiparametric Estimation of a Hedonic Price Function"(Journal of Applied Econometrics 11, 1996, pages 633-648). We consider the modeling and prediction of house prices. Data are available for 546 observations of the following variables:

- sell: Sale price of the house
- lot: Lot size of the property in square feet
- bdms: Number of bedrooms
- fb: Number of full bathrooms
- sty: Number of stories excluding basement
- drv: Dummy that is 1 if the house has a driveway and 0 otherwise
- rec: Dummy that is 1 if the house has a recreational room and 0 otherwise
- ffin: Dummy that is 1 if the house has a full finished basement and 0 otherwise
- ghw: Dummy that is 1 if the house uses gas for hot water heating and 0 otherwise
- ca: Dummy that is 1 if there is central air conditioning and 0 otherwise
- gar: Number of covered garage places
- reg: Dummy that is 1 if the house is located in a preferred neighborhood of the city and 0 otherwise
- obs: Observation number, needed in part (h)

Erasmus

Questions

(a) Consider a linear model where the sale price of a house is the dependent variable and the explanatory variables are the other variables given above. Perform a test for linearity. What do you conclude based on the test result?

(b) Now consider a linear model where the log of the sale price of the house is the dependent variable and the explanatory variables are as before. Perform again the test for linearity. What do you conclude now?

(c) Continue with the linear model from question (b). We now consider possible transformation of the lot size variable. We can consider either the variable itself, or a log transformation of this variable. Which of these do you prefer? (Keep all other explanatory variables included without transformation.)

(d) Consider now a model where the log of the sale price of the house is the dependent variable and the explanatory variables are the log transformation of lot size, with all other explanatory variables as before. We now consider interaction effects of the log lot size with the other variables. Construct these interaction variables. How many are individually significant?

(e) Perform an F-test for the joint significance of the interaction effects from question (d).

(f) Now perform model specification on the interaction variables using the general-to-specific approach. (Only eliminate the interaction effects.)

(g) One may argue that some of the explanatory variables are endogenous and that there may be omitted variables. For example, the 'condition' of the house in terms of how it is maintained is not a variable (and difficult to measure) but will affect the house price. It will also affect, or be reflected in, some of the other variables, such as whether the house has an air conditioning (which is mostly in newer houses). If the condition of the house is missing, will the effect of air conditioning on the (log of the) sale price be over- or underestimated? (For this question no computer calculations are required.)

(h) Finally we analyze the predictive ability of the model. Consider again the model where the log of the sale price of the house is the dependent variable and the explanatory variables are the log transformation of lot size, with all other explanatory variables in their original form (and no interaction effects). Estimate the parameters of the model using the first 400 observations. Make predictions on the log of the price and calculate the MAE for the other 146 observations. How good is the predictive power of the model (relative to the variability in the log of the price)?

Erasmus