NCG612 - Housing Project

Martin Charlton and Chris Brunsdon

Semester 2 2018-2019

Introduction

Housing Valuation is an area in which statistical models can play a role. The models which are frequently used can also be used to model other price structures, for example, hotel room charges.

The project is concerned with finding the most reliable determinants on property prices. The dataset is a subset of anonymised mortgage records for the area that is known as Greater London. The purchase price (which is different from the asking price) is available, as a a series of the characteristics of the property.

The goal is to find the best group of predictors of property price.

Data

The table below lists the names of the variables. The data are anonymised - first the data were coded to unit postcode level (in this UK this represents about 20 addresses). The grid references have been further spatially jittered.

Variable	Description
Easting	Easting in m
Northing	Northing in m
Purprice	Purchase Price in GBP
BldIntWr	Built between 1918 and 1939
BldPostW	Built between 1945 and 1959
Bld60s	Built between 1960 and 1969
Bld70s	Built between 1970 and 1979
Bld80s	Built between 1980 and 1989
TypDetch	Detached property
TypSemiD	Semi-detached property
TypFlat	Flat or apartment
GarSingl	Single Garage
GarDoubl	Double Garage
Tenfree	Leasehold/Freehold indicator
CenHeat	Central heating
BathTwo	Two or more bathrooms
BedTwo	Two bedrooms
BedThree	Three bedrooms
BedFour	Four bedrooms
BedFive	Fie bedrooms
NewPropD	New property
FlorArea	Floor area in square metres
NoCarHh	Proportion of households without a car
CarspP	Cars per person in neighborhood
ProfPct	Proportion of Households with Professional Head
UnskPct	Proportion of Households with Unskilled head
RetiPct	Proportion of residents retired
Saleunem	Not known

Variable	Description
Unemploy	Unemployed workers
PopnDnsy	Local population density

The variables are organised into several groups, and for the most part represent the levels in a categorical variable expanded into dummy (0/1) variables.

Dates: these represent the time period in which the property was constructed. The omitted category is *built pre-1914.* If the property was built in the 1980s, the Bld80s variable will be 1, and the others in this group will be zero; if all are zero, the property was built before 1914.

Tenure: these represent the type of building. The omitted category is *bungalow.*

Garage: The omitted category is *hardstanding* - that is, no garage.

Bedrooms: the omitted category is *one bedroom*

There are also a series of *neighbourhood* variables taken from the Census of Population. There are no measures of income in the UK census, so some socio-economic variables are available as proxies.

A few of the variables may seem rather unusual. The data are over 25 years old, and there were many older properties without central heating. The sales came at the end of a decade when property prices had started to rise quickly following the deregulation of financial sector under Margaret Thatcher's government and the passing of legislation allowing council tenants to buy their own council houses.

Method

A commonly used methodology in valuation is known a hedonic modelling. It owes its origins to Kelvin Lancaster in the 1960s. The price of house is modelled as a function of its attributes, and some of the characteristics of neighbourhood. Importance is placed in the interpretation of the parameter estimates.

For example if we fitted a model of the form:

$$Price = \beta_0 + \beta_1 FloorArea + \epsilon$$

The variation in price is modelled as a function of the variation in floorsize. The β_1 parameter would represent the price per unit floor area (price per square metre), *ceteris paribus.* Anything not in the model is represented by the variation in the error term.

If we added another variable, say, *CenHeat*, so that the model becomes:

$$Price = \beta_0 + \beta_1 FloorArea + \beta_2 CenHeat + \epsilon$$

we can interpret the β_2 parameter as the increase in price, *ceteris paribus*, if Central Heating is present in the property. Recall that this variable has the value 0 for *no central heating* and 1 for *central heating.* It means that we effectively has two parallel regression lines with the gradient β_1: for for properties without central heating and one for properties *with* central heating.

The modelling methodology allows for the incorporation of *interaction terms.* If we added a term for FloorArea x CenHeat then effectively we obtain a separate relationship for properties with and without central heating: the sum of the β_0 and β_2 coefficients yields the intercept for the central heated houses, and the sum of the β_1 and interaction term coefficient would give the gradient value (interpreted as price per unit floorspace for properties with central heating).

Where the variables represent a characteristic such as age, then you will have to add them as a group. You may wish to recode some of the variables. For example the 3,4 and 5+ bedroom dummies could be recoded as a single 3+ bedroom dummy.

One of the aims of undertaking price variation modelling is model building - which variables represent the trade off between adequacy of predictive power versus parsimony. Another goal is whether the parameter estimates seem sensible - are there likely to be any problems with collinearity? Do the residuals deviate from the desirable properties of zero mean, independent, and homoscedastic? Are there any unusually large residuals - why might these be unusual? Might location play a role? The Regency terraces in Central London might have some rather different characteristics than the terraces once occupied by the working class in Tower Hamlets.

Spatial Variation

One of the issues is whether there might be a spatial component to the models that can be built. With some boundaries of the London Boroughs, it would be possible to use a spatial join to add the Borough codes as dummy variables. The resulting parameter estimates would indicate whether there was spatial varation in the intercept term.

In our book we fitted separate models for each borough to start with… another alternative would be to use an interaction term for the floorspace term (with the boroughs). You then obtain an estimate of the spatial variation in the influence of floorspace change on price by borough.

If you intend to use GWR, then be aware that there some 12500 records in the data. This means you would have to think carefully about the scalability of GWR functions in the GWmodel library. Perhaps a single computation of the distance matrix would be one route, or you might think about dividing the data into training and validation sets.

Output

As before the output will be a single authored report of no more than 5000 words. To reach that goal you can work either singly or in groups. You will need to think carefully about data exploration. There are many ways of fitting linear models to these data of which OLS is but one option (it's often used in practice).

You will need to

1. Describe your project, its aim(s) and objectives
2. Describe the methods for property price prediction
3. Describe the data - and how it relates to the prediction approach
4. Discuss the results
5. Conclusions

It goes without saying that you should present the code you use, suitably commented. You will also need an executive summary.

Martin Charlton and Chris Brunsdon