Case 2 prediction of house prices February 27

Notes:

1) Only one upload (in Canvas/discussions) by team now. Upload it in each group if your teammate belongs to a different tutorial.

2) Provide comments on the student who is at position + 1 after you on the list. Take the one after him/her if that student did not submit a file.

Start with the file `housing_sample.wf1` that contains data for 1460 US housing prices denoted SALEPRICE. The other series are the potential regressors that I want you to start to include in a multiple regression framework. Regressors are TOTALBSMTSF, BEDROOMABVGR. See see appendix below for a description.

There are many additional characteristics of those houses in the full file `housing_price.xls` (see again appendix for a description). To add variables to the small `housing_sample.wf1` EViews file, go to the Excel sheet and copy the array with the data (including headers) that you paste in EViews. Check (show) whether data have been correctly imported. Save the file.

A. Your first task:

- In groups of 2 people, find the determinants of the **log sale prices** with series or transformation of series TOTALBSMTSF, BEDROOMABVGR. By transformation I mean you can take the log of TOTALBSMTSF, to include squares, etc. For BEDROOMABVGR you can create one dummy variable by number of bedrooms by adding as many regressors (minus one for the benchmark) as the number of bedrooms with @expand(BEDROOMABVGR,@drop(1)).
- You are totally free, there is not a good model that I know, nor a final answer.
- Explain briefly what you have tried and the steps you made before you have reached the final equation that you will keep for prediction.
- Explain your final results accurately: values of coefficients, goodness of fit, significance of the parameters, misspecification, etc.
- Compare your small model with a big data model selection framework in which you include other variables that are in `housing_price.xls.`
- I describe below steps in more details.

B. The estimation steps small OLS

- Take the natural **log of the house price** series as your dependent variable in order to probably decrease the presence of heteroskedasticity.

- Find the best model for

  `log(sale)=f(explanatory variable)`

  using the explanatory variables. Note that you might take for the regressors, the series, dummy variable indicators, but also some transformations such as logs of series, nonlinear functions like squared, interactions. When a series have several input values like 0, 1, 2, 3, 1,...a way to split it in several 0, 1 dummies is by using the command @expand(.). Be careful that you have to choose a benchmark which becomes the intercept. For instance if you take bedrooms=1 you write @expand(bedrooms, @drop(1)).

- There might be outliers, missing variables or weird observations. For instance I noticed that there are houses with a basement of 0 square feet. You might decide to delete those variables. In that case you can use in the sample box a statement like

  `if TotalBsmtSF` > 0

  or to use the surface of the first floor and to replace the 0 by that values. Again these are practical issues that you will face in most datasets, hence decide what to do. There isn't a good answer, these are real data.

- Try to find the best, still manageable (that makes sense) model (using $\bar{R}^2$, information criteria), use t-tests and F-tests to delete non significant explanatory variables.

- Look at mispecification tests to guide you:

  - non normality (might detect outliers),
  - heteroskedasticity (might tell you to control for a group that you forgot),
  - non linearity (might tell you to look at interaction variables, power functions), etc.

- It will probably be the case that some mispecification will remain. Use HCSE for instance if heteroskedasticity does not disappear after you have improved the model.

C. The estimation steps for big data and model selection

- Add as many variables as you wish from the whole database. EViews lite won't probably work!

- Use selection methods (unidirectional t-tests, lasso, etc). You can play with threshold options to change the critical value of the t-test or p-values. Try several options and look at the final model. Don't forget to standardize variables in lasso! I would use the sample standard deviation option.

- There are categorial variables: say variables with entries "good, excellent, normal "...instead of 0 1 dummy. To use (expand) those variables in a regression and e.g. assuming the first category is the benchmark for the variable bsmtqual

  `@expand(bsmtqual,@dropfirst)`

  Note that you may choose another benchmark with the EXACT name of the category (character specific). This means in my case that

  `log(saleprice)` c `@expand(bsmtqual, @drop("Gd"))`

  works but not

  `log(saleprice)` c `@expand(bsmtqual, @drop("gd"))`

- Questions are:
  - Is the larger model better. Look at $\bar{R}^2$, information criteria but you will probably have issues to compare the small model with a larger one as you won't have the same number of observations. This is because some of the regressors you add are not available for the entire sample.
  - Does the large model make sense to you (from values of coefficients)?

D. Prediction: reestimate both B and C models (the small one and your favorite large one):

- I assume that you have obtained two good models for $\log(Saleprice)$. I would like to predict the level SALEPRICE on the period 1401 to 1460 and compare the out of sample model performances. Indeed what you want to give to clients is an estimate of a house price, not its log. I also want to measure whether the performances of your models are adequate for a part of the sample for which we know the realisations (e.g. here the last 60 observations).

- Reestimate your equations on a smaller sample from 1 to 1400. When you did the **Quick/Estimate Equation** window, write log(saleprice) as the dependent variable, do not generate the variables before (also for explanatory variable). So doing EViews will give you the opportunity to forecast/predict log(saleprice) or the level saleprice. You can manually get the prediction $\log(\widehat{Saleprice})$, then take $exp(\log(\widehat{Saleprice}))$ to come back to the prediction $\widehat{Saleprice}$. This is a bit tedious.

- Forecast observations 1401 to 1460. You get a graph with the predictions and the 95% confidence interval. Give a name for both forecasts and standard errors. The prediction and the SE values (not the interval) are in your workfile under the name you gave.

- Compare RMSE of model B and C. Take note of the number of observations for which RMSE is computed. It would be 60 if all series are available but less in case of missing data on the 1401 to 1460 period.

E. Give me a price

- Give me a price for a house that I want to sell together with a 95% confidence interval. Of course you might need more information if your equation has more parameters. Hence you need to approximate the missing values (compared with others in the sample, average hypothesis, intuition etc, choose and defend your point of view). Do not start with my explanatory variables to set up your model.

- LotArea: 8500

  OverallQual: 7

  OverallCond: 5

  YearBuilt: 2003

  TotalBsmtSF: 1000

  GrLivArea: 1700

  FullBath: 2

  BedroomAbvGr: 3

  GarageCars: 2

- The easiest way to do it is to resize your workfile using `Proc -> structure/resize` and to enter 1 to 1461. Then go in the observations 1461 to enter my features and some values of indicators you need from your regression. It might be that you didn't have those regressors in your equation. I don 't know what to do. For instance to provide a forecast without that variable and to look at the average difference in price to shift the price. Hint: there are no methods to do it in a neat way as it might be that I give you too many information compared to your regression, or too few (your model is larger)

Appendix; Data description, here's a brief version of what you'll find in the data description file.

SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.

MSSubClass: The building class

MSZoning: The general zoning classification

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access

Alley: Type of alley access
LotShape: General shape of property
LandContour: Flatness of the property
Utilities: Type of utilities available
LotConfig: Lot configuration
LandSlope: Slope of property
Neighborhood: Physical locations within Ames city limits
Condition1: Proximity to main road or railroad
Condition2: Proximity to main road or railroad (if a second is present)
BldgType: Type of dwelling
HouseStyle: Style of dwelling
OverallQual: Overall material and finish quality
OverallCond: Overall condition rating
YearBuilt: Original construction date
YearRemodAdd: Remodel date
RoofStyle: Type of roof
RoofMatl: Roof material
Exterior1st: Exterior covering on house
Exterior2nd: Exterior covering on house (if more than one material)
MasVnrType: Masonry veneer type
MasVnrArea: Masonry veneer area in square feet
ExterQual: Exterior material quality
ExterCond: Present condition of the material on the exterior
Foundation: Type of foundation
BsmtQual: Height of the basement
BsmtCond: General condition of the basement
BsmtExposure: Walkout or garden level basement walls
BsmtFinType1: Quality of basement finished area
BsmtFinSF1: Type 1 finished square feet
BsmtFinType2: Quality of second finished area (if present)
BsmtFinSF2: Type 2 finished square feet
BsmtUnfSF: Unfinished square feet of basement area
TotalBsmtSF: Total square feet of basement area
Heating: Type of heating
HeatingQC: Heating quality and condition
CentralAir: Central air conditioning
Electrical: Electrical system
1stFlrSF: First Floor square feet
2ndFlrSF: Second floor square feet
LowQualFinSF: Low quality finished square feet (all floors)
GrLivArea: Above grade (ground) living area square feet
BsmtFullBath: Basement full bathrooms
BsmtHalfBath: Basement half bathrooms
FullBath: Full bathrooms above grade
HalfBath: Half baths above grade
Bedroom: Number of bedrooms above basement level

Kitchen: Number of kitchens
KitchenQual: Kitchen quality
TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
Functional: Home functionality rating
Fireplaces: Number of fireplaces
FireplaceQu: Fireplace quality
GarageType: Garage location
GarageYrBlt: Year garage was built
GarageFinish: Interior finish of the garage
GarageCars: Size of garage in car capacity
GarageArea: Size of garage in square feet
GarageQual: Garage quality
GarageCond: Garage condition
PavedDrive: Paved driveway
WoodDeckSF: Wood deck area in square feet
OpenPorchSF: Open porch area in square feet
EnclosedPorch: Enclosed porch area in square feet
3SsnPorch: Three season porch area in square feet
ScreenPorch: Screen porch area in square feet
PoolArea: Pool area in square feet
PoolQC: Pool quality
Fence: Fence quality
MiscFeature: Miscellaneous feature not covered in other categories
MiscVal: $Value of miscellaneous feature
MoSold: Month Sold
YrSold: Year Sold
SaleType: Type of sale
SaleCondition: Condition of sale