Spatial Econometrics

Dani Arribas-Bel

2016-04-18

This session[1] is based on the following references, which are good follow-up's on the topic:

- Session III of Arribas-Bel (2014). Check the "Related readings" section on the session page for more in-depth discussions.

This tutorial is part of Spatial Analysis Notes, a compilation hosted as a GitHub repository that you can access it in a few ways:

- As a download of a `.zip` file that contains all the materials.
- As an html website.
- As a pdf document
- As a GitHub repository.

Dependencies

The illustration below relies on the following libraries that you will need to have installed on your machine to be able to interactively follow along[2]. Once installed, load them up with the following commands:

```
# Layout
library(tufte)
# For pretty table
library(knitr)
# Spatial Data management
library(rgdal)
# Pretty graphics
library(ggplot2)
# Pretty maps
library(ggmap)
# Various GIS utilities
library(GISTools)
# For all your interpolation needs
library(gstat)
# For data manipulation
library(plyr)
# Spatial regression
library(spdep)
```


[2] You can install package `mypackage` by running the command `install.packages("mypackage")` on the R prompt or through the `Tools --> Install Packages...` menu in RStudio.

Before we start any analysis, let us set the path to the directory where we are working. We can easily do that with `setwd()`. Please replace in the following line the path to the folder where you have placed this file -and where the `house_transactions` folder with the data lives.

```
setwd('/media/dani/baul/AAA/Documents/teaching/u-lvl/2016/envs453/code/GIT/kde_idw_r/')
#setwd('.')
```

Data

To explore ideas in spatial regression, we will be using house price data for the municipality of Liverpool. Our main dataset is provided by the Land Registry (as part of their Price Paid Data) but has been cleaned and re-packaged into a shapefile by Dani Arribas-Bel.

Let us load it up first of all:

```
hst <- readOGR(dsn = 'house_transactions', layer = 'liv_house_trans')
```

```
## OGR data source with driver: ESRI Shapefile
## Source: "house_transactions", layer: "liv_house_trans"
## with 6324 features
## It has 18 fields
```

```
## NOTE: rgdal::checkCRSArgs: no proj_defs.dat in PROJ.4 shared files
```

The tabular component of the spatial frame contains the followig variables:

```
names(hst)
```

```
##  [1] "pcds"       "id"         "price"
##  [4] "trans_date" "type"       "new"
##  [7] "duration"   "paon"       "saon"
## [10] "street"     "locality"   "town"
## [13] "district"   "county"     "ppd_cat"
## [16] "status"     "lsoa11"     "LSOA11CD"
```

The meaning for most of the variables can be found in the original Land Registry documentation. The dataset contains transactions that took place during 2,014:

```
# Format dates
dts <- as.Date(hst@data$trans_date)
# Set up summary table
tab <- summary(dts)[c('Min.', 'Max.')]
tab
```

```
##         Min.         Max.
## "2014-01-02" "2014-12-30"
```

Although the original Land Registry data contain some characteristics of the house, all of them are categorical: *is the house newly built? What type of property is it?* To bring in a richer picture and illustrate how continuous variables can also be included in a spatial setting, we will augment the original transaction data with Deprivation indices from the CDRC at the Lower Layer Super Output Area (LSOA) level.

Let us read the `csv` in:

```
imd <- read.csv('house_transactions/E08000012.csv')
```

The table contains not only the overall IMD score and rank, but some of the component scores, as well as the `LSOA` code:

```
names(imd)
```

```
##  [1] "LSOA11CD"   "imd_rank"   "imd_score"
##  [4] "income"     "employment" "education"
##  [7] "health"     "crime"      "housing"
## [10] "living_env" "idaci"      "idaopi"
```

That bit of information, `LSOA11CD`, is crucial to be able to connect it to each house transaction. To "join" both tables, we can use the base command `merge`, which will assign values from `imd` into `hst` making sure that each house transaction get the IMD data for the LSOA where it is located:

```
db <- merge(hst, imd)
```

The resulting table, `db`, contains variables from both original tables:

```
names(db)
```

```
##  [1] "LSOA11CD"   "pcds"       "id"
##  [4] "price"      "trans_date" "type"
##  [7] "new"        "duration"   "paon"
## [10] "saon"       "street"     "locality"
## [13] "town"       "district"   "county"
## [16] "ppd_cat"    "status"     "lsoa11"
## [19] "imd_rank"   "imd_score"  "income"
## [22] "employment" "education"  "health"
## [25] "crime"      "housing"    "living_env"
## [28] "idaci"      "idaopi"
```

For some of our analysis, we will need the coarse postcode of each house, rather than the finely specified one in the original data. This means using the available one

```
head(db@data['pcds'])
```

```
##      pcds
## 62 L1 0AB
## 63 L1 0AB
## 64 L1 0AB
## 65 L1 0AB
## 66 L1 0AB
## 67 L1 0AB
```

to create a new column that only contains the first bit of the postcode (`L1` in the examples above). The following lines of code will do that for us:

```
db$pc <- as.character(lapply(strsplit(as.character(db$pcds), split=" "),
```

Given there are 6,324 transactions in the dataset, a simple plot of the point coordinates implicitly draws the shape of the Liverpool municipality:

```
plot(db)
```

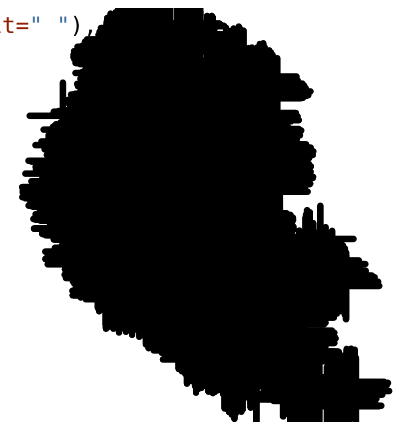

Figure 1: Spatial distribution of house transactions in Liverpool

Non-spatial regression, a refresh

Before we discuss how to explicitly include space into the linear regression framework, let us show how basic regression can be carried out in R, and how you can begin to interpret the results. By no means is this a formal and complete introduction to regression so, if that is what you are looking for, I suggest the first part of Gelman and Hill (2006), in particular chapters 3 and 4.

The core idea of linear regression is to explain the variation in a given (*dependent*) variable as a linear function of a series of other (*explanatory*) variables. For example, in our case, we may want to express/explain the price of a house as a function of whether it is new and the degree of deprivation of the area where it is located. At the individual level, we can express this as:

$$P_i = \alpha + \beta_1 NEW_i + \beta_2 IMD_i + \epsilon_i$$

where P_i is the price of house i, NEW_i is a binary variable that takes one if the house is newly built or zero otherwise and IMD_i is the IMD score of the LSOA where i is located. The parameters β_1, β_2, and β_3 give us information about in which way and to what extent each variable is related to the price, and α, the constant term, is the average house price when all the other variables are zero. The term ϵ_i is usually referred to as "error" and captures elements that influence the price of a house but are not whether the house is new or the IMD score of its area. We can also express this relation in matrix form, excluding subindices for i[3].

Essentially, a regression can be seen as a multivariate extension of simple bivariate correlations. Indeed, one way to interpret the β_k coefficients in the equation above is as the degree of correlation between the explanatory variable k and the dependent variable, *keeping all the other explanatory variables constant*. When you calculate simple bivariate correlations, the coefficient of a variable is picking up the correlation between the variables, but it is also subsuming into it variation associated with other correlated variables –also called confounding factors[4]. Regression allows you to isolate the distinct effect that a single variable has on the dependent one, once we *control* for those other variables.

Practically speaking, running linear regressions in `R` is straightforward. For example, to fit the model specified in the equation above, we only need one line of code:

```
m1 <- lm('price ~ new + imd_score', db)
```

We use the command `lm`, for linear model, and specify the equation we want to fit using a string that relates the dependent variable (`price`) with a set of explanatory ones (`new` and `price`) by using a tilde ~ that is akin the = symbol in the mathematical equation. Since we are using names of variables that are stored in a table, we need to pass the table object (`db`) as well.

In order to inspect the results of the model, the quickest way is to call `summary`:

```
summary(m1)
```

```
##
## Call:
## lm(formula = "price ~ new + imd_score", data = db)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##  -184254   -59948   -29032    11430 26434741
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)   235596      13326  17.679
## newY            4926      19104   0.258
## imd_score      -2416        308  -7.843
##             Pr(>|t|)
## (Intercept)  < 2e-16 ***
```

[3] In this case, the equation would look like

$$P = \alpha + \beta_1 NEW + \beta_2 IMD + \epsilon$$

and would be interpreted in terms of vectors and matrices instead of scalar values.

[4] **EXAMPLE** Assume that new houses tend to be built more often in areas with low deprivation. If that is the case, then NEW and IMD will be correlated with each other (as well as with the price of a house, as we are hypothesizing in this case). If we calculate a simple correlation between P and IMD, the coefficient will represent the degree of association between both variables, but it will also include some of the association between IMD and NEW. That is, part of the obtained correlation coefficient will be due not to the fact that higher prices tend to be found in areas with low IMD, but to the fact that new houses tend to be more expensive. This is because (in this example) new houses tend to be built in areas with low deprivation and simple bivariate correlation cannot account for that.

```
## newY           0.797
## imd_score   5.12e-15 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## Residual standard error: 509000 on 6321 degrees of freedom
## Multiple R-squared:  0.009712,   Adjusted R-squared:  0.009398
## F-statistic: 30.99 on 2 and 6321 DF,  p-value: 4.027e-14
```

A full detailed explanation of the output is beyond the scope of this note, so we will focus on the relevant bits for our main purpose. This is concentrated on the `Coefficients` section, which gives us the estimates for the β_k coefficients in our model. Or, in other words, the coefficients are the raw equivalent of the correlation coefficient between each explanatory variable and the dependent one, once the polluting effect of confounding factors has been accounted for[5]. Results are as expected for the most part: houses tend to be significantly more expensive in areas with lower deprivation (an average of GBP2,416 for every additional score); and a newly built house is on average GBP4,926 more expensive, although this association cannot be ruled out to be random (probably due to the small relative number of new houses).

[5] Keep in mind that regression is no magic. We are only discounting the effect of other confounding factors that we include in the model, not of *all* potentially confounding factors.

Finally, before we jump into introducing space in our models, let us modify our equation slightly to make it more useful when it comes to interpretating it. Virtually every house price model in the literature is estimated in log-log terms:

$$\log P_i = \alpha + \beta_1 \log NEW_i + \beta_2 \log IMD_i + \epsilon_i$$

This allows to interpret the coefficients as *elasticities*, an economic term used to capture the percentual variation that a variable experiences as a result of a one percent increase in another one. This comes in very handy because it standardizes the results across variables. To fit such a model, we can specify the logarithm of a given variable directly in the formula. Note that we do not transform `new`, as it is a binary variable.

```
m2 <- lm('log(price) ~ new + log(imd_score)', db)
summary(m2)
```

```
##
## Call:
## lm(formula = "log(price) ~ new + log(imd_score)", data = db)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -4.3060 -0.3089 -0.0149  0.2936  5.3450
##
## Coefficients:
##                 Estimate Std. Error t value
## (Intercept)     13.54414    0.03504  386.56
## newY             0.23772    0.01934   12.29
## log(imd_score) -0.57471    0.01002  -57.36
##                 Pr(>|t|)
## (Intercept)       <2e-16 ***
## newY              <2e-16 ***
## log(imd_score)    <2e-16 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## Residual standard error: 0.516 on 6321 degrees of freedom
## Multiple R-squared:  0.3449, Adjusted R-squared:  0.3447
## F-statistic:  1664 on 2 and 6321 DF,  p-value: < 2.2e-16
```

Looking at the results we can see a couple of differences with respect to the original specification. First, the estimates are substantially different numbers. This is because, although they consider the same variable, the look at it from different angles, and provide different interpretations. For example, the coefficient for the IMD, instead of being interpretable in terms of GBP, the unit of the dependent variable, it represents an elasticity: a 1% increase in the degree of deprivation is associated with a 0.57% decrease in the price of a house.[6] Second, the variable `new` is significant in this case. This is probably related to the fact that, by taking logs, we are also making the dependent variable look more normal (Gaussian) and that allows the linear model to provide a better fit and, hence, more accurate estimates. In this case, a house being newly built, as compared to an old house, is overall 23% more expensive.

6 **EXERCISE** *How does the type of a house affect the price at which it is sold, given whether it is new and the level of deprivation of the area where it is located?* To answer this, fit a model as we have done but including additionally the variable `type`. In order to interpret the codes, check the reference at the Land Registry documentation.

Spatial regression: a (very) first dip

Spatial regression is about *explicitly* introducing space or geographical context into the statistical framework of a regression. Conceptually, we want to introduce space into our model whenever we think it plays an important role in the process we are interested in, or when space can act as a reasonable proxy for other factors we cannot but should include in our model. As an example of the former, we can imagine how houses at the seafront are probably more expensive than those in the second row, given their better views. To illustrate the latter, we can think of how the character of a neighborhood is important in determining the price of a house; however, it is very hard to identify and quantify "character" perse, although it might be easier to get at its spatial variation, hence a case of space as a proxy.

Spatial regression is a large field of development in the econometrics and statistics literatures. In this brief introduction, we will consider two related but very different processes that give rise to spatial effects: spatial heterogeneity and spatial dependence. For more rigorous treatments of the topics introduced here, the reader is referred to Anselin (2003), Anselin and Rey (2014), and Gibbons, Overman, and Patacchini (2014).

Spatial heterogeneity

Spatial heterogeneity (SH) arises when we cannot safely assume the process we are studying operates under the same "rules" throughout the geography of interest. In other words, we can observe SH when there are effects on the outcome variable that are intrinsically linked to specific locations. A good example of this is the case of seafront houses above: we are trying to model the price of a house and, the fact some houses are located under certain conditions (i.e. by the sea) makes their price behave differently[7].

This somewhat abstract concept of SH can be made operational in a model in several ways. We will explore the following two: spatial fixed-effects (FE), and spatial regimes.

Spatial FE

Let us consider the house price example from the previous section to introduce a more general illustration that relates to the second motivation for spatial effects ("space as a proxy") is what is. Given we are only including two explanatory variables in the model, it is likely we are missing some important factors that play a role at determining the price at which a house is sold. Some of them, however, are likely to vary systematically over space (e.g. different neighborhood characteristics). If that is the case, we can control for those unobserved factors by using traditional dummy variables but basing their creation on a spatial rule. For example, let us include a binary variable for every two-digit postcode in Liverpool, indicating whether a given house is located within such area (`1`) or not (`0`). Programmatically, this is straightforward:

```
# Include '-1' to eliminate the constant term and include a dummy for every area
m3 <- lm('log(price) ~ pc + new + log(imd_score) - 1', db)
summary(m3)
```

[7] **QUESTION** How would you incorporate this into a regression model that extends the log-log equation of the previous section?

```
##
## Call:
## lm(formula = "log(price) ~ pc + new + log(imd_score) - 1", data = db)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2853 -0.2744 -0.0259  0.2513  5.6489
##
## Coefficients:
##                Estimate Std. Error t value
## pcL1           12.72971    0.05821  218.70
## pcL10          12.91716    0.09260  139.49
## pcL11          12.88647    0.06906  186.61
## pcL12          13.05348    0.05467  238.77
## pcL13          12.87225    0.06367  202.18
## pcL14          12.91258    0.06755  191.17
## pcL15          13.02870    0.05659  230.22
## pcL16          13.13843    0.05023  261.54
## pcL17          13.27288    0.05493  241.63
## pcL18          13.27455    0.04544  292.14
## pcL19          13.12792    0.05460  240.44
## pcL2           13.05230    0.10959  119.10
## pcL20          12.82414    0.23068   55.59
## pcL24          12.85954    0.06995  183.83
## pcL25          13.18340    0.05240  251.59
## pcL27          13.02308    0.10661  122.15
## pcL28          12.42124    0.21096   58.88
## pcL3           12.93580    0.05477  236.16
## pcL4           12.65911    0.06721  188.34
## pcL5           12.93297    0.07759  166.69
## pcL6           12.81728    0.06921  185.20
## pcL7           12.83274    0.06977  183.93
## pcL8           12.98639    0.06628  195.93
## pcL9           12.77233    0.06106  209.16
## newY            0.29365    0.02065   14.22
## log(imd_score) -0.41098    0.01539  -26.70
##                Pr(>|t|)
## pcL1             <2e-16 ***
## pcL10            <2e-16 ***
## pcL11            <2e-16 ***
## pcL12            <2e-16 ***
## pcL13            <2e-16 ***
## pcL14            <2e-16 ***
## pcL15            <2e-16 ***
```

```
## pcL16           <2e-16 ***
## pcL17           <2e-16 ***
## pcL18           <2e-16 ***
## pcL19           <2e-16 ***
## pcL2            <2e-16 ***
## pcL20           <2e-16 ***
## pcL24           <2e-16 ***
## pcL25           <2e-16 ***
## pcL27           <2e-16 ***
## pcL28           <2e-16 ***
## pcL3            <2e-16 ***
## pcL4            <2e-16 ***
## pcL5            <2e-16 ***
## pcL6            <2e-16 ***
## pcL7            <2e-16 ***
## pcL8            <2e-16 ***
## pcL9            <2e-16 ***
## newY            <2e-16 ***
## log(imd_score)   <2e-16 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## Residual standard error: 0.4944 on 6298 degrees of freedom
## Multiple R-squared:  0.9982, Adjusted R-squared:  0.9982
## F-statistic: 1.338e+05 on 26 and 6298 DF,  p-value: < 2.2e-16
```

- Drops `L1` so it's with respect to L1
- It compares houses within `L`
- As a byproduct, it gives a measure of area desirability

Spatial regimes

Spatial dependence

- Exogenous spatial effects
- Point to further spatial regression (lag, error) –> Anselin (1988, 2003), plus Anselin & Rey (2015) for an in-depth treatment of modern approaches

Predicting house prices

- Show how to obtain a prediction for a given house
- Compare the estimate with that of interpolation?

References

Anselin, Luc. 2003. "Spatial Externalities, Spatial Multipliers, and Spatial Econometrics." *International Regional Science Review* 26 (2). Sage Publications: 153–66.

Anselin, Luc, and Sergio J. Rey. 2014. *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press LLC.

Arribas-Bel, Dani. 2014. "Spatial Data, Analysis, and Regression-a Mini Course." *REGION* 1 (1). European Regional Science Association: R1. `http://darribas.org/sdar_mini`.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/hierarchical Models*. Cambridge University Press.

Gibbons, Stephen, Henry G Overman, and Eleonora Patacchini. 2014. "Spatial Methods." CEPR Discussion Paper No. DP10135.