# Real Estate Practice Assignment

June 15, 2020

This assignment deals with two datasets based on a subsample of publicly available records of property transactions in Taiwan from 2000Q1 to 2012Q3:

- **local_building_raw.dta**: details on sales of buildings, such as the date and location, price, and characteristics of the building or apartment unit.
- **local_land_raw.dta**: details on sales of land parcels, including the date and location, price, and the size and shape of the plot.

The files have several features in common with other datasets we will be using extensively in this project. The label attached to each variable gives the original Chinese name of the variable included in the raw data.

The first dataset contains 17 variables. Here are some details on the ones which are not self-explanatory:

- roadwidth: length in meters of the front-facing road adjacent to the land where the property is located.
- street_section: description of the property's location on the street indicated in the address string.
- main_use: a string variable containing a note on the main use of the property (e.g. commercial storefront, industrial plant, apartment building, etc.).
- build_mat: a string variable describing the building materials.
- construct_date: the original construction date, usually formatted as yyymm, where "yyy" is the ROC year and "mm" is the month. The variables "rocyear" and "quarter" describe the transaction quarter-year in the ROC calendar.
- property_floor: the number of floors in the unit being sold. 全 ("all floors") refers to cases where the entire building is sold.
- num_floor: the number of floors in the building involved in the transaction – may include areas not bought and sold.
- land_use: a string variable containing a note on the main use of the land. This variable is generally more detailed than main_use, as it may describe uses for land shared with the building being sold, in addition to the main use of the building itself.

The second dataset for land-only transactions contains only 14 variables. There are three variables which are specific to land-only transactions: shape, which gives a rough classification of the parcel as square, regular, or irregular-shaped, and the width and depth (both in meters) of the land parcel. Note that the "unit_price" variable is defined slightly differently across the two datasets. For building transactions, the unit price is the price per unit of floor space (build_area). For land transactions, the unit price is the price per unit of land area (land_area).

We would like you to perform basic analysis of the data to get an idea of your programming and organization skills. In what follows, assume that you are producing output that is going to be submitted to a top journal. Keep in mind that since these files contain the raw data, there may be typos you will need to account for in order to proceed.

Also, we would like your code to be as clear and organized as possible – one do file that runs everything is preferable. Ideally the code should contain comments outlining the purpose of non-obvious steps. For instance, renaming a variable is a simple enough step that it does not require a comment, but changing values of a variable should trigger a comment describing why you made such a choice.

Please complete the following tasks and put all outputs into a single zip folder. Use the "plotplainblind" scheme to produce graphs.

1. Using local_land_raw.dta, compute the median unit price for each quarter-year and city combination.

2. Create a time series graph which plots the median unit price over time for all of Taiwan (including all 22 regions) alongside separate series for each of Taiwan's six major cities (Taipei, New Taipei, Taoyuan, Taichung, Tainan, and Kaohsiung). Do this for the full sample time period of 2000Q1 to 2012Q3. A common procedure when making this type of plot is to "normalize" each series by dividing prices by the value in a base period. Use 2000Q1 as the base period. The resulting series are sometimes called **median price indices**. Save the above time series graph in .gph and .eps formats.

3. Redo the first two steps, restricting to transactions of land parcels within the top 20% by transaction value.

4. Now combine the two raw data files into one file. Create a variable(s) that allows you to keep track of which transactions are (i) land-only, (ii) building-only, and (iii) a building-land bundle. You may need to combine several variables to identify these transaction types.

5. In addition, create the following variables which can be used to help determine why one property may be more/less valuable:

   - The age of the building being sold in years (if applicable).
   - Convert num_floor and property_floor to numeric variables.
   - Sort the properties into five categories based on use: single-family housing, apartments, commercial (offices/stores), industrial, and agricultural. Call this variable "broad_use_code." You may find it helpful to combine the existing variables "land_use" and "main_use."
   - A single numeric variable which identifies distinct neighborhoods within the same region.

   Save the resulting file as **local_alltrans_clean.dta**. All numerical variables should be saved as numeric type, not as strings.

6. Pool all of the transactions together to estimate the following regression model:

$$\log P_{i,t} = \sum_{\tau=1}^{T} \beta_\tau \cdot \mathbb{1}\{\tau = t\} + \gamma' \cdot \mathbf{X_{i,t}} + \epsilon_{i,t}$$

where $i$ refers to individual transactions, and $t$ refers to a quarter-year. Use the log transaction value as the dependent variable. The vector $\mathbf{X_{i,t}}$ in the above model should include: road width, dummies for "broad_use_code," dummies for the number of floors in the unit, dummies for neighborhood, a quadratic in age, a quadratic in floor space, and a quadratic in land area (see step 5 for details).

Transform the coefficient estimates $\exp(\widehat{\beta}_t)$ and plot them as a time series with the 95% confidence intervals. As in step 2, normalize the coefficients so that $\widehat{\beta}_t$ is equal to 1 in 2000Q1. This step creates an **hedonic price index**.[1]

7. Finally, use the variable you created in step 4 to do the following:

   (i) Create six time series graphs, one for each of Taiwan's six major cities. Each graph should plot the median unit price index over time for the three transaction types.

   (ii) Combine the six time series graphs into a single figure displayed as two rows of three graphs (a $2 \times 3$ grid).

   (iii) Based on the series in these graphs, generate a simple table which computes growth rates for each region-transaction type combination between three periods: 2000Q1-2007Q4, 2008Q1-2010Q4, and 2011Q1-2012Q3.

[1] Alternatively, you can create a table that can easily be integrated into LaTeX which reports all of the estimated regression coefficients and standard errors.