document for regression function

Ruofan Chen,Liquan Zhong, Sara Rahim

4/17/2021

Method

In regression_func.R there include two functions: plm_inference_state_level and plm_inference_overall. The first one is time series regression, from line 6 to line 72. The second one is panel data regression, from line 76 to line 129. This document explain the method used in these two functions.

Time series regression build a model like the following:

$$y_{it} = \beta_0 + x_{1,i,t}\beta_1 + ... + x_{5,i,t}\beta_5 + \epsilon_{i,t}$$

where i from 1 to 51, represents states (50) and District of Columbia. t from 1995 to 2018, represents the year number. x_i represents predictors found in outer data set. Here, x_1 to x_5 represents unemployment rate (unemp_rate), population density (pop_density), poverty rate (pov_rate), police officer rate (pol_officer_rate) and medium age(med_age). $\epsilon_{i,t}$ is error of the model.

For each states(when t=t_1):

$$y_{t_1} = \beta_0 + x_{1,t_1}\beta_1 + ... + x_{5,t_1}\beta_5 + \epsilon_{t_1}$$

where ϵ_{t_1} may have auto-correlation. In line 31, the ACF test has been conducted. If there is no auto-correlation,
use

$$Var(\hat{\beta}) = (X^T X)^{-1} X^T \hat{\Sigma} X (X^T X)^{-1}$$

where X represents the design matrix. $\hat{\Sigma} = MSE \cdot I$.

If auto-correlation or heteroscedasticity exist, use Newey-West method to calculate the variance of $\hat{\beta}$. We assume the auto-correlation has Lag L. The Newey-West method have the following equation:

$$Var(\hat{\beta}) = (X^T X)^{-1}(\Sigma_{t=1}^T e_t^2 x_t x_t^T + \Sigma_{l=1}^L \Sigma_{t=l+1}^T (1 - \frac{l}{L+1}) e_t e_{t-l}(x_t x_{t-l}^T + x_{t-l} x_t^T))(X^T X)^{-1}$$

Where L represents the Lag term, which is determined by the pacf plot or the corresponding test. x_i represents the transpose of i^{th} row of the design matrix.

To test the significance of the coefficients, we use p-value and choose significance level as $\alpha = 0.05$. The null hypothesis is:

$$H_0 : \hat{\beta}_j = 0$$

Then calculated t-statistics to test the significance.

$$t = \hat{\beta}_j / se(\hat{\beta}_j)$$

Panel data regression build a model like the following:

$$y_t = x_t^T \beta_t + \epsilon_t$$

where $\beta_t = (\beta_{0,t}, \beta_{1,t}, ..., \beta_{p,t})$, p is the number of predictors. Here, p=5. We assume that the β is influenced by time. Then we want to split the time effect and only want to know the predictors (fixed effect). Hence, we build the model for β:

$$\beta_{p,t} = C_p + \epsilon_t$$

where C_p is the fixed effect by predictors. If ϵ_t do not have auto-correlation: $se(C_p) = \sqrt{\frac{MSE}{T}}$. Where t is the number of year in the regression model (here is 24).

If it has auto-correlation or heteroscedasticity, use Newey-West method which is

$$(X^T X)^{-1} S (X^T X)^{-1}$$

, where

$$S = (\Sigma_{t=1}^{T} e_t^2 + \Sigma_{l=1}^{L} \Sigma_{t=l+1}^{T} 2(1 - \frac{l}{L+1}) e_t e_{t-l}/T$$

Then $se(\hat{C}_p) = \sqrt{\frac{(X^T X)^{-1} S (X^T X)^{-1}}{T}}$. T-statistics is calculated as the following, then do the significance test.

$$t = \frac{C_p}{se(C_p)}$$

Summary of the result

50 states and District of Columbia has different result. Analysis need to be carried out into states. For those states with significant coefficients, the population density is positive means that for this states, as the population density change (increase), the crime rate will change as the same direction (increase). Similarly for those whose coefficients are negative, means the crime rate will not increase or even decrease. This means that the government or the police department does do a good job to control the crime. Similarly for police officer rate, poverty rate, unemployment rate, and medium age.

For nation analysis, all the predictors are significant for violent crime and only population density is not significant for property crime. If the coefficients of one variable is positive, it depict as with the increase or decrease of the predictor, the crime rate will change in same direction. Similarly, it will opposite change if the coefficients are negative.

The violent crime calculated the Euclidean distance between two states coefficients. The North Dakota locates away from others states. For the property crime, it gathers a few states. The coefficient-violent.png and coefficient-property.png use cluster method to group these states, to show the states in map.

We are also curious about the relationship with police officer rate and crime rate. The granger causality test is carried out. granger-test-result-crime.png and granger-test-result-property.png shows the result in map.