Kanyao Han
MACSS 30100

Problem Set #1

First Part

Journal article: Yao Lu & Ran Tao. 2017. "Political Organizational Structure and Collective Action: Lineage Networks, Semiautonomous Civic Associations, and Collective Resistance in Rural China." *American Journal of Sociology* 122 (6): 1726-74

Model: This article uses a fixed-effects linear probability model to exame the relationship between organization structure and collective petitions. It is worth noting that μ_t is the intercept in year t, α_i is the fixed effect specific to i village but constant over time.

$$P(y_{it} = 1|x) = \mu_t + \alpha_i + \beta * L_{it} * O_i\Gamma + X_{it}\Delta + \epsilon_{it}$$

Endogenous Variables: L_{it}, O_i, X_{it}

1. L_{it} indicates land requisition in village i in year t.
2. O_i is a vector of variables related to village organizations, including senior association and lineage structures (Oligopoly, Monopoly, Temple and Church).
3. X_{it} is a vector of control variables including per capita annual income(log), village population(log), number of enterprise, proportion of several population features (working only in agricultural sector, away as migrant workers, working-age population and ethnic minority), per capita household farmland(log), proportion of arable land that is irrigated, per capita fiscal transfer from upper-level government(log) and villagers working in upper-level government.

Exogenous Variable: The occurrence of collective petition for village i in year t (y_{it}) is the exogenous variable and its conditional probability $P(y_{it} = 1|x)$ is on the left side of the equation.

Model Type: First, although it contains time t, this model is a typical **static** panel data model because the changes at time t have an immediate effect on the dependent variables. Second, as a linear probability model, it is **linear**. Third, it is also **stochastic** since it contains error term ϵ_{it}.

Missing variable: The variable "gender" might be valuable because villagers with different gender may have different collective potential. Therefore, it should be put into the vector of control variables.

Second Part

Model

I adopt the binary logistic model to explore whether someone decides to get married. The variables in my model contains Y_i (decision to get married, P(Y_i=1|X) in equation represents the probability of decision to get married), X_H (homosexual identity, yes or no), X_L (legal same-sex marriage, yes or no), X_S (sex, female or male), X_R (Republican, yes or no), X_E (education level, year of education) and X_I (income). The last two variables are continuous and the others are binary.

$$log(\frac{P(Y_i = 1|X)}{1 - P(Y_i = 1|X)}) = \beta_0+\beta_1 X_H+\beta_2 X_L+\beta_3 X_H X_L+\beta_4 X_S+\beta_5 X_I+\beta_6 X_S X_I+\beta_7 X_R+\beta_8 X_E+\epsilon_i$$

Key factors and factor selection

I assume that homosexual identity, same-sex marriage legality and their interaction terms will be key factors because illegality of same-sex marriage means homosexual people nearly have no opportunity of deciding to marriage. Besides, whether a people is Republican probably have a big influence on marriage because the Republican people typically value family.

Besides, income is very likely to have some influence due to the fact that developed countries usually have a low marriage rate than developing country. I also interact sex with income since I assume that male people with low income has low ability to marry but low income may have a smaller influence on female group. Finally, given that well-educated people usually reflect the value of marrage more often and are more likely to doubt it, I believe it would also have some influence on the decision to marrage.

Preliminary test

I can adopt four methods to do preliminary tests. First, I can compute the standard errors of the coefficients and then the p-values of all variables. Those which have sufficiently small p-values are statistically significant. Second, I will drop one or two variables that seem relatively unimportant, compute the maximum log-likelihood of different models and the significance of their difference and hence compare the fitness of different models. Third, I will compute different error rates through cross-validation method. Although this method cannot directly indicate significance of individual variables, it can still assess the predictive power of them if dropping some variables would drastically increase or decrease the error rate. Finally, looking for a new set of data as test dataset and then testing the real predictive power of the model would be a better method.