Economics 274 *Reading Guide* *Winter 2015*

Migration Paper

Gustafsson, Björn and Sai, Ding (2014). Why is There No Income Gap between the Hui Muslim Minority and the Han Majority in Rural Ningxia, China? ***China Quarterly***, 220(December), pp. 968-987.

1. What is the structure of the paper? They provide section headings, but you ought to think about the overall purpose of each section ("motivation for paper" "previous research" "analytic method" and so on).

2. What do they claim as the main question they seek to answer? subsidiary questions?

3. What do they find? How do they find it? One approach is to look at variables that are statistically significant, that is, empirically affect the likelihood of migration. Another is to ask what variables are economically significant, that is, make a substantial difference to actual outcomes. That is, a variable can be statistically significant in empirical work, correlating very well with the likelihood of migration, but be so small in magnitude that it really doesn't matter.

4. What do they list as the defects in their approach?

5. What additional questions does their work raise?

6. What policy implications follow from their work?

Technical issues

Gini coefficient: this is the area between a 45° line and the actual distribution. If income is distributed evenly, then the bottom 1% will be 1% of the population and so on, and you get a 45° line and a Gini coefficient of 0. If you have most people with no income, and a few with very high incomes, then the actual distribution will lie near the x-axis and the Gini coefficient will be nearly 1. So the metric ranges over [0, 1] and a bigger number means a greater degree of inequality.

Regression: fitting a line to the data to estimate the statistical relationship between the variable of interest and various explanatory variables. Basic results generally are in a table with asterisks to indicate whether there is a relationship, and how precisely the coefficient is estimated.

The underlying idea is that the relationship the authors want to estimate is:

> probability migrate = f(household income & wealth variables, ethnicity variables, network variables, family variables, personal characteristics)

which in practice, given available data (the authors had a hand in crafting the survey) means:

> household income, land per household member;
> ethnic composition of village, own ethnicity;
> proportion migrants in village;
> number of children, number of elderly;
> age, education, gender.

Now once we have data, we can then fit a curve to the data, with parameters that will help us interpret how they affect the likelihood to migrate. However, here the data are either "yes" or "no" for whether someone migrated. That causes major problems.

To see why this is a problem, graph (say) income versus migration. We have a lot of points at y=0 on the x-axis and a lot of points at y=1. If we fit a line, it will run through the midpoint of the y=0 dots and the midpoint of the y=1 dots. So if we plug values into the fitted equation migration probability = $\beta_0 + \beta_1$ income and plug in observed income levels, we'd end up with probabilities less than 0 and greater than 1. Having more variables doesn't eliminate this problem, and there's no natural way to figure out what the value of the coefficients (the β) represent.

To get around this, we need to use a function that translates a line onto the interval [0, 1]. One such is to use a function such as:

$Y = e^x/(1+e^x)$ [a logit regression] or Y = cumulative inverse normal distribution $\Phi(x)$ [a probit regression].

The computer statistics package Stata has built-in commands to do such regressions, and to generate appropriate measures of goodness of fit. The estimated coefficients then represent the change in probability from an increase in the magnitude of a variable, and can be read in the standard manner.