1 Minimal Theory to Drive Estimation

Denote a region by r, and within that region there are sub-units (grid cells or administrative units) that are denoted by i. Each sub-unit in a region has agricultural production of

$$Y_{ir} = A_{ir} X_{ir}^{1-\beta_r} (L_{ir}^A)^{\beta_r} \tag{1}$$

where A_{ir} is total factor productivity, X_{ir} is land area, and L_{ir}^A is the number of agricultural workers. The value of β_r is constant across the sub-units of region r, but could be unique to region r.

With free mobility of labor within the region, workers will move across cells until the marginal product of labor in each cell is equalized. In addition, let agricultural output move freely between the sub-units of a region, so the price of agricultural goods is the same in each one. These assumptions imply that

$$\beta_r A_{ir} X_{ir}^{1-\beta} (L_{ir}^A)^{\beta-1} = \beta_r A_{jr} X_{jr}^{1-\beta} (L_{jr}^A)^{\beta-1} \tag{2}$$

for any given pair of cells i and j. We also know that

$$\sum_{j \in r} L_{jr}^A = L_r^A, \tag{3}$$

which says the sum of agricultural labor used in all sub-units in region r has to add up to total agricultural workers in region r. One can use these two relationships to solve together for

$$L_{ir}^A = A_{ir}^{1/(1-\beta_r)} X_{ir} \frac{L_r^A}{\sum_{j \in r} A_{jr}^{1/(1-\beta_r)} X_{jr}}. \tag{4}$$

This says that total agricultural employment in sub-unit i depends on its own productivity level and area, relative to the weighted sum of productivity terms across the whole region. Intuitively, any sub-unit that is particularly productive should have a greater share of the agricultural labor force employed in it. The larger is L_r^A, the larger is agricultural labor in any given cell.

Note that the fraction on the right is common to every cell in the region. For notational simplicity, write this fraction as Ω_r, giving us

$$L_{ir}^A = A_{ir}^{1/(1-\beta_r)} X_{ir} \Omega_r \tag{5}$$

as the expression for the agricultural labor employed in sub-unit i. Taking logs, we have

$$\ln L_{ir}^A = \frac{1}{1-\beta_r} \ln A_{ir} + \ln X_{ir} + \ln \Omega_r. \tag{6}$$

This linear equation is something we can estimate in a regression to recover the value of $1/(1-\beta_r)$, and hence the value of β_r. The term Ω_r is the same across all sub-units, so in a regression of $\ln L_{ir}^A$ on X_{ir} and $\ln A_{ir}$, it will be picked up in the constant.

This estimation equation depends on three assumptions. First, that β_r is the same across all sub-units. Second, that labor is free to move across the sub-units. Three, that output is free to move across the sub-units. These assumptions guide us in how we define regions for the purposes of our regressions.

From this equation, we can do the estimations. What are the implications? With only one other assumption we can get some interesting outcomes concerning urbanization. Assume that every person (agricultural and non-agricultural) consumes exactly the same amount of agricultural output. Call that c_A. This need not be a subsistence amount, or any kind of long-run equilibrium amount. All we need is that it is similar for all people.

Total agricultural output must add up as follows

$$c_A L_r = \sum_{i \in r} A_i X_i^{1-\beta_r} L_{Ai}^{\beta_r}, \tag{7}$$

where L_r is total population. We know that each L_{Ai} is described above, and can plug that in to yield

$$\frac{L_{Ar}}{L_r} = \left(\frac{c_A^{1/(1-\beta_r)} L_r}{\sum_{i \in r} A_i^{1/(1-\beta_r)} X_i} \right)^{(1-\beta_r)/\beta_r}. \tag{8}$$

This shows that the ratio of agricultural to total population depends on several factors. First, the higher is c_A, the more people need to work in agriculture. Second, the higher is the population, L_r, the larger is L_{Ar}/L_r. More people crowding onto the same land requires a higher percent of those people must work to produce food, as marginal products decline. Finally, the higher is productivity (the sum in the denominator), the lower is the share of labor in agriculture.

One thing to note is that the elasticity of L_{Ar}/L_r with respect to L_r is $(1-\beta_r)/\beta_r$. As β_r gets higher, this term gets lower. This implies that population size has less of an effect on L_{Ar}/L_r the higher is β_r.

In contrast, consider the elasticity of L_{Ar}/L_r with respect to productivity, A_i. For this purpose, imagine we are talking about productivity increases that are common to every cell within the region. Then the elasticity is $1/\beta_r$. The higher is β_r, the smaller a response. So with a high β_r, the economy will not urbanize as much (e.g. have L_{Ar}/L_r go down) for a given productivity level. Or if you like, for a given high productivity level, the urbanization rate will be relatively low.

If we want to go further with theoretical implications, then we have to start adding in new assumptions about fertility and mortality behavior, preferences for food versus urban goods, etc.. etc.. But the above mechanical relationship should be sufficient to start.

The implication of the relationship is that places with high β_r will be less responsive in terms of urbanization to productivity improvements in agriculture, quite possibly limiting their ultimate development.

2 Data and Estimation

Going back to equation (6), we need three pieces of data.

- Agricultural population, L_{Ai}. This should be available from the global population data on a cell by cell basis. I believe it has total population, as well as urban population, and so we can simply find $L_{Ai} = L_i - L_{i,urban}$. It is crude, but for our purposes is probably fine.
- Agricultural productivity, A_i. For this we have several options. But the most direct is the Ramankutty dataset, which is grid-cell by grid-cell. This has the advantage of being measured in a common frame-

work across each region, so we are not introducing any problems by using wheat suitability in one place versus rice suitability in another. We can remain agnostic as to the actual crop types grown. We can also, for the time being, not worry about trying to get all the GAEZ data, which is not convenient. If we come up with some specific hypothesis we can download a specific dataset for use.

- Agricultural area, X_i. We can probably get this on a grid-cell level from the FAO, but only for modern times. We will probably have to experiment here with using total grid-cell area, modern crop area, or some other measure here.

So in the end, for this first project, we need a dataset that is organized at the grid-cell level. I'm presuming that this is (relatively straightforward) given that the input data is all in raster format.

For each region we select, we can run a regression of the form in (6). This will give us a region-specific estimate $\hat{\beta}_r$. Once we have those estimated parameters from each region, we can compare them and see how much difference there is. Perhaps do some statistical tests to see if European $\hat{\beta}_r$ values are smaller than Asian values.

The hard part empirically is going to be deciding what belongs in a "region". We need regions that match the assumptions made in the estimation: labor and food are mobile between cells of the region. The regions should be defined in part by natural geographic limits, but also by political boundaries that would create frictions. But we want to have regions that are sub-national. For example, northern China and southern China have to be distinct because they have very distinct climate differences. There is no perfect answer here, and we will probably spend lots of time trying different region definitions.

One option is to use Koppen-Geiger climate zones to define regions within political boundaries. Another is to use watersheds, on the assumption that most people along an individual river watershed will be relatively mobile with each other and share agricultural techniques. Again, no perfect answer here.