Download the data-set burkina731.dta from the website. This data is a subset of the ICRISAT Burkina Faso dataset used in several papers on the syllabus. It includes input and output data on all the plots cultivated by the household heads (NOT other household members) in the sample.

A classic result in neoclassical economics is the separation theorem – that is, production and consumption decisions should be separable. This implies that, independent of consumption preferences, an individual maximizes efficient outcomes in production. We will look at whether this holds in Burkina Faso, and in the process also learn something about non-parametric regressions.

Burkina Faso individuals (indexed by i) farm multiple plots simultaneously. Assume that the allocation of factors of production across the various plots controlled by an individual is efficient, so that the separation hypothesis is true across these plots (that is, conditional on plot characteristics, output per hectare is equal on all plots cultivated by a particular individual at a given point in time). Denote as $Q_{vhtci}$ the log of yield on plot $i$ planted to crop $c$ in year $t$ by the head of household $h$ in village $v$, $X_{vhtci}$ are the observed characteristics of that plot, $\lambda_{vtc}$ is a household-year-crop fixed effect.

1. To examine the correlation between yield and other household characteristics estimate

$$Q_{vhtci} = X_{vhtci} \gamma + \alpha Z_{vhtci} + \lambda_{vtc} + \varepsilon_{vhtci}$$

where $Z_{vhtci}$ is your choice of either the log of household size or the log of the total area cultivated by household head $h$ on plots other than plot $i$.

Report your results and interpret $\alpha$.

2. One of the crucial variables in $X_{vhtci}$ is the log area of plot $i$ ($A_{vhtci}$). We have no reason to think that there is a linear relationship between
this and log yield, and might be concerned that mis-specification of this relationship is biasing our results. First, let’s look at the relationship between $A_{vhtci}$ and $Q_{vhtci}$. Estimate the relationship

$$Q_{vhtci} = g(A_{vhtci}) + \mu_{vhtci}$$

using a nonparametric estimator of $g()$. You might find the kernreg.ado program (type help kernreg) helpful, or Angus Deaton’s lowrex program available at the world bank lsms web site (http://www.worldbank.org/lsms/tools/deaton/).

3. Of course, yield depends on the other covariates as well. Respecify the equation as a partial linear model:

$$Q_{vhtci} = g(A_{vhtci}) + \tilde{X}_{vhtci}\tilde{\gamma} + \tilde{\alpha}Z_{vhtci} + \tilde{\lambda}_{vtec} + \tilde{\epsilon}_{vhtci}. \quad (1)$$

Hint: don’t try to do this with fixed effects (it can be done, but it involves several extra steps. See Porter, 1996). Instead, use a full set of village-year-crop dummy variables, and treat them as simple regressors. If we let $m$ denote the matrix of $\tilde{X}$, $Z$ and the new dummy variables all together, and $\beta$ as the vector of all the coefficients. The estimator of $\beta$ is:

$$\hat{\beta} = \left( \sum_{i=1}^{N} (m_i - \bar{E}[m|A_{vhtci}]) (m_i - \bar{E}[m|A_{vhtci}])' \right)^{-1} \left( \sum_{i=1}^{N} (m_i - \bar{E}[m|A_{vhtci}]) (Q_{vhtci} - \bar{E}[Q_{vhtci}|A_{vhtci}])' \right)$$

That is, $\beta$ is estimated by estimating (separately) the nonparametric relationships between $Q$ and $A$ and $m$ and $A$, forming the residuals, and regressing the residuals of $Q$ on the residuals of $m$. How does your estimate of the effect of your choice of $Z$ on $Q$ change compared to that obtained in question 1?

Given $\hat{\beta}$, use it to estimate the function $g()$:

$$\hat{g}(A_{vhtci}) = \bar{E}[Q|A_{vhtci}] - E[m|A_{vhtci}]\hat{\beta}.$$ 

How does this estimate of $g()$ compare to that which was obtained in question 2?