DATA 621: Business Analytics and Data Mining
CUNY SPS, MSc Data Analytics
Final Project Proposal

Charley Ferrari, Christina Taylor, David Stern

Introduction:

The dataset we will be using is the World Development Indicators released annually by the World Bank. The data is a time series, tracking 1345 indicators for 214 economies from 1960-2015.

Categories of these indicators include Agriculture and Rural Development, Aid Effectiveness, Climate Change, Economy and Growth, Education, Energy and Mining, Environment, External Debt, Financial Sector, Gender, Health, Infrastructure, Labor and Social Protection, Poverty, Private Sector, Public Sector, Science and Technology, Social Development, Trade, Urban Development.

Geographically, the dataset includes observations for each national economy as well as well as averages by region: World, East Asia and Pacific, Europe and Central Asia, Latin America and Caribbean, Middle East and North Africa, South Asia, and Sub Saharan Africa. There are also averages by income classification: into High income, Low or Middle income, IBRD, and IDA.

Problem statement and research question:

Our goal for this project is to construct models that can best predict specific measures of socio-economic development. Among the 20 categories, we chose three response variables representing development areas that are of most interest to us:

1. *Health*: Life expectancy at birth (both sexes)
2. *Technology*: Internet users (per 100 people)
3. *Education*: Out-of-school children of primary school age (both sexes)

The problem we intend to address is to identify which variables among the 1345 indicators best predict these outcomes. Given the number of predictors, potential permutation and combination of models is enormous. We will start by selecting categories that may directly impact (e.g. environment, poverty, labor) or are closely related (health, technology, education) to the response variables. We may also attempt to group some predictors by traditional GDP indicators (domestic expenditures, output gap, foreign trade and balance of payments, and consumer power (employment, wages, or prices). Our hope is to produce at least one powerful model, if not three, that can serve as proximate predictors for socio-economic development.

Objectives, Proposed Models, Evaluation Measures:

In preparing for model selection, we will first address the missing data challenge. Paucity of data may be a result of relatively new measurement (e.g. smartphone ownership) or disparities in the infrastructure to capture certain data. We will limit our scope to predictors that are consistently well-documented and/or aggregated.

We will then segment the countries based on income group and time frame. As stated in the assumptions, group level relationships may not apply to the individual level. Specifically, different income segments may call for different models. For our research, we intend to limit our income group to developing countries and our time frame to the last decade.

Our general approach is, in order: diagnostics, transformation, variable selection, diagnostics. Since the data is collected over time, there may be serially correlated errors. We plan to implement auto correction methods to solve this challenge. We will also experiment with Box-Cox and possibly spline/polynomial transformations to achieve normality.

Our most significant objective is effective variable selection. We expect to face serious challenges posed by the sheer number of predictors and inevitable collinearity; best subset selection may no longer be computationally feasible. Therefore, we will consider forward stepwise selection and lasso regression to limit the number of predictors and shrink highly correlated variables to zero. We are also interested in exploring Partial Least Squares as a dimension reduction technique.

We seek to find a model that balances bias and variance. To evaluate the models, we will compare the model that maximizes the adjusted R squared to those chosen stepwise by AIC and BIC. (Since our sample size is quite large, we may omit the AIC Corrected measure.) The models will then be evaluated based on test MSE using 10 fold cross-validation.

Assumptions:

1. A predictor variable’s impact on an indicator is similar among income groups. We plan to use the regional and income-level averages in the dataset to explore these relationships.
2. There may be intermediate lurking variables. Let’s say there appears to be a strong relationship between aid and social development. But development can be a result of GDP growth, which may be affected by aid effectiveness.
3. Aside from the interactions between the economy and aid, we’re assuming other classes of variables are exogenous. This includes variables like agricultural yields and climate.

Research Summary:

Evaluating the Impact of Foreign Aid on Economic Growth: A Cross-Country Study
Journal of Economic Development, Vol. 30, Number 2, December 2005, pp. 25-46

This paper used data until the mid 90’s to suggest that foreign aid has a positive impact on the economic growth of developing countries. An important assumption made here is that aid’s contribution to growth is similar for all developing countries, which may be questionable. Aside, negative effects of high aid inflows and time lags in aid-growth relationship, and correlated errors also affect the model accuracy. The limited time scope, error correction, and considerations regarding the assumption are instrumental to our model building.

A major insight we have gained from this paper is a mechanism explaining how exactly growth occurs: Business investment expands productive capacity. Economic growth starting from a depressed level of consumption can happen as consumption rises, but long term economic growth depends on investment. (Intuitively, one can think of an agricultural economy's growth due to gaining more customers, versus investing in machinery or crop rotation methods that allows farms to be more productive.) This informed our final choice of response variables. Instead of building a GDP forecasting engine as we originally planned, we will focus on predicting socio-economic development, and investigating their interaction with other development indicators: aid effectiveness (as an example of investment), for example. Part of this project will involve conceptualizing the interaction among predictors, and testing different mechanisms of causation.

Determinants of Enrollment in Primary Education: A Case Study of District Lahore
Pakistan Economic and Social Review, Vol. 46, No. 2 (Winter 2008), pp. 161-200

This paper evaluates the effect development indicators relating to households have on enrollment in primary education in the city of Lahore, Pakistan. The authors use ordinary least squares and logistic regression models to determine the impact different indicators have on gross enrollment. They find that certain indicators traditionally linked to enrollment decisions were insignificant (distance to school, education of head of household) or counterintuitive (family size, and dependency ratio). The models confirmed that dwelling ownership, family size, and literacy ration had a positive impact on enrollment. Our takeaway from this paper is twofold: not all indicators we believe to have an impact on a particular development indicator will necessarily be significant in a regression model. A good start to our model will be to hand-pick variables that we believe are linked in an entity-relationship model (e.g. life expectancy as predicted by number of hospital beds and community health workers).