Estimation of Contract Schedules using Constrained B-Splines

Michael Chirico

February 18, 2017

Abstract

Most unionized teachers are paid according to a salary schedule (specifying wages as an increasing function of tenure and certification) explicated in contracts collectively bargained at the district level. With this in hand, teachers are able to infer their future potential wage trajectories at their own and other potential district employers. Lacking the physical contract faced by the teachers, an econometrician armed only with administrative data reporting actual wages in a given year must use some imputation techniques to deduce the wage structure. We explore the utility of natural Constrained B-Splines (COBS) to this end. COBS are an enhanced version of the traditional semiparametric splines technique enhanced by the ability to impose a monotonicity constraint on the resultant curve.

Introduction

For many years, the ubiquitous characteristic of collectively bargained teachers' contracts has been the salary table, which gives a mapping from the calendar year, a teacher's experience (their length of tenure at the current district), and their certification (typically Master's vs. Bachelor's degree) to their wage. This table gives current teachers a clear understanding of how their pay will advance as a function of their labor inputs, and thereby gives forward-looking potential teachers and potential migrant teachers a clear understanding of their potential pay arcs under a district-switching decision-making framework, especially given that this information is typically openly available.

It would behoove an econometrician seeking to understand education labor market dynamics, then, to incorporate this information on future pay into their statistical modeling framework. Unfortunately, this data is typically not available in a format lending itself to easy analysis at scale – whether locked inside idiosyncratically formatted and sporadically-available contract PDFs or hidden behind large-scale freedom of information act inquiries, the temporal and financial costs of scraping such data into a usable form can be substantial.

Much more common in empirical settings is access to teacher-year-level salary data of the form $y_{i,t} = y(\tau_{i,t}, c_{i,t}, d_{i,t}) + \varepsilon_{i,t}$, where $\tau_{i,t}$, $c_{i,t}$ and $d_{i,t}$ are the tenure, certification, and district of teacher i in year t, and $\varepsilon_{i,t}$ represents unaccount factors affecting the wage (e.g., not all teachers work full time, and many teachers supplement their income with additional duties like coaching). The goal of this paper is to give one approach and some empirical lessons for trying to estimate the underlying mapping $y(\tau, c, d)$ from such data.

Method

There are a multitude of inference/imputation techniques suitable to the inference of a latent function of unknown parametric form available in the statistician/econometrician's palette. The powerful flexibility of nonparametric approaches (local regression, splines, Random Fourier Feature expansions) is a double-edged sword; as it happens, in this particular setting, even if we know linearity is not a reasonable functional form restriction, we do know some very basic properties of the underlying tenure-wage curves that will be violated in general by uninformed estimation techniques. In particular, we know that such tenure-wage curves are non-decreasing and that they are non-negative, i.e., $y(\tau', c, d) >= y(\tau, c, d)$ whenever $\tau' >= \tau$, and $y(0, c, d) \geq 0$.

He and Ng (1999) introduce a linear programming approach to incorporating monotonicity and curvature restraints to quantile regression spline estimation techniques, and Ng and Maechler (2007) present an overview of the R package `cobs` which gives an efficient implementation of this approach (COBS standing for Constrained B-Splines). The basic idea of quantile regression spline estimation is to swap out the standard squared loss function for a quantile-dependent weighted absolute loss function to target conditional quantiles instead of conditional means. Monotonicity, point, and curvature restrictions enter as penalized terms to the objective function; `cobs` expresses this in a fashion which facilitates the application of standard linear programming techniques for efficiency, and handles internally the issues of knot selection and penalty parameter assignment through cross-validation.

Data

The state of Wisconsin's Department of Public Instruction (DPI) releases annual Salary, Position & Demographic reports through the WISEstaff data collection system, and these reports represent "a point-in-time collection of all staff members in public schools as of the 3rd Friday of September…" (DPI 2017). The crucial data elements from each report are the Highest Degree Code (certification), Total Experience in Education (tenure), Total Salary, Total Fringe/Employee Benefits, and Agency of Work Location Code (unique identifier of district). We pull data from the 1994-95 academic year (AY) through AY2015-16 for a total of 3,588,614 teacher-position-year observations (many teachers serve in multiple roles within a school/district, and each of these is filed separately in the DPI system).

Data are first fed through the matching algorithm described in further detail in a companion paper. As a consequence, 26,302 (0.7%) observations are lost on account of belonging to teachers who could not be uniquely identified in a given year of data due to sharing a first name, last name, and birth year with another teacher in the data. Specific to the exercise at hand, with data reliability and precision in mind, we make a series of further restrictions on the data.

References

DPI. 2017. "School Staff: Salary, Position & Demographic Reports." https://dpi.wi.gov/cst/data-collections/staff/published-data.

He, Xuming, and Pin Ng. 1999. "COBS: Qualitatively Constrained Smoothing via Linear Programming." *Computational Statistics* 14 (3). Heidelberg: Physica-Verlag,[1992-: 315–38.

Ng, Pin, and Martin Maechler. 2007. "A Fast and Efficient Implementation of Qualitatively Constrained Quantile Smoothing Splines." *Statistical Modelling* 7 (4). Sage Publications India Pvt. Ltd, B-42, Panchsheel Enclave, New Delhi: 315–28.