Economics 312 Spring 2014

Project #4 Assignment Due: Midnight, Monday, March 3

Partner assignments

Kai Addae	Chris Weber
Prakher Bajpai	Timothy Tyree
Emmanuel Enemchukwu	Cole Sprague
Julian Haft	Will Schmid
Paapa hMensa	Alec Recinos
Julia Hofmann	Stephanie Radoslovich
Daniel Hope	Mat Olson
Mark Jarrett	John Mills
James LaBelle	Dylan McKenna
Theo Landsman	Dean Young
Colleen Werkheiser	Austin Weisgrau

This project uses a subset of a data set that accompanies another popular econometrics textbook. A link to the data set is on the assignment page or you can download it from academic.reed.edu/economics/parker/s13/312/asgns/datasets/Teachers93.dta. The data are for high schools in Michigan in 1993. There are 10 variables in the dataset. In addition to some basic descriptive characteristics of the schools, the dataset includes average teacher salary and average value of teacher benefits, which are the focus of our analysis.

Economic theory suggests that what matters to workers and employers is the total value of compensation, not the form that it takes. So, for example, assuming that the worker would choose to buy at least $1000 worth of health insurance, it should not make much difference whether the employer provides the $1000 of insurance or raises the worker's annual salary by $1000 so that the worker can purchase the insurance herself. Thus, in this situation, a school district that increases benefits should be able to lower salaries commensurately and still attract the same number and quality of teachers. Of course, if the worker values the insurance at less than $1000, then she might value the increased salary more than the benefits. In this case the reduction in salary might need to be smaller than the increase in benefits in order to convince the worker to remain in the job.

The fundamental question that we explore in this project is whether an increase in benefits is associated with a decrease in salary, either partial or complete. The variables in the dataset are in shown in the table below:

Variable name	**Description**
lnchprg	Share of students enrolled in subsidized school lunch program
enroll	School enrollment
staff	Number of staff members per 1000 students
expend	Total expenditures per student in dollars
salary	Average teacher salary in dollars
benefits	Average teacher benefits in dollars
droprate	School dropout rate
gradrate	School graduation rate
math10	Share of students passing state 10th grade math test
sci11	Share of students passing state 11th grade science test

Our regressions will have some variant of salary and benefits as the dependent variable, with the other variables in the table as potential explanatory variables. In order to be comfortable with this procedure, we must be confident that causality runs from the regressors to the dependent variable and not the other way (or both ways). One might question the exogeneity of some of these prospective regressors: would any of them be affected by a change in teacher compensation? In private firms an increase in salaries or benefits would naturally lead to a rise in total expenditures, so expenditures would clearly not be exogenous. However, public school districts' expenditures are usually determined by government budget processes and respond mostly to changes in state and local tax revenues. It is plausible that changes in teacher compensation would not affect tax revenues, so it may be reasonable to treat expenditures as exogenous. Are better-compensated teachers more successful? If so, then changes in compensation might affect the dropout, graduation, and test score variables. We will assume that the student-outcome variables and total expenditures by the district are *not* affected by teacher compensation. Our results are only as reliable as that assumption.

1. Determinants of total compensation

a. Define a variable *totcomp* for total teacher compensation to be the sum of salary and benefits. As with wage equations that we have examined, the conventional specification is to use the log of compensation on the left-hand side of the regression. Thus, our compensation regression would look something like

$$\log\left(totcomp_i\right) = \mathbf{x}_i\boldsymbol{\beta} + e_i,$$

where $\mathbf{x}_i$ is a (row) vector of characteristics of school *i* that might affect compensation. Which of the variables in the dataset *should* affect overall compensation? Why?

b. What other variables would you want? Why? Based on your beliefs about their likely correlation with the included variables, how do you think their omission would affect the estimates of the included coefficients?

c. Explore possible regression models with log(*totcomp$_i$*) as the dependent variable and decide which variables (other than salary or benefits) should be included and excluded. (In terms of linearity vs. nonlinearity, convention [as interpreted by Jeff] suggests taking logs of the variables that are not already percentages: *enroll*, *staff*, and *expend*, but leaving the variables that are percentages already in linear form.) Show in a single outreg table the candidate regressions you think are most promising, and discuss which one you prefer.

d. If *all* school-district expenditures were on staff compensation, then

$$totcomp = A \times \frac{expend}{staff},$$

with the constant *A* rescaling to account for *expend* being measured in dollars per student, *staff* in number per 1000 students, and *totcomp* in dollars per teacher. This suggests that, to the extent that school budgets are dominated by teacher compensation, the elasticity of compensation with respect to expenditures should be near one and the elasticity with respect to staff should be near minus one. Any increase in expenditures either goes to increasing compensation or hiring more staff. Using a log-log specification, test these hypotheses individually and jointly, and interpret your results.

e. Is multicollinearity a problem in your regression? Are there any significant outliers in your regression or any "high-leverage" observations? (The lvr2plot command is very useful here to identify observations that have large squared residuals [outliers] or values of **x** that are far from the mean [high leverage].) If you find a few extreme residuals, create the residual series are look at the observations with large residuals to see what makes them different. If you find that there are a few highly influential observations, use the predict *varname* , leverage command to create a new variable that contains a "leverage estimator" for each observation. Use the Stata manuals (pdf) and other resources (as needed) to determine what this leverage estimator is and what it means (and explain the intuition in your report). Find the observations that have large leverage and examine them to see what makes them different. If you have an observation or two that have extreme leverage, what happens if you eliminate them from the sample? (Note that unless there are identifiable characteristics of $\mathbf{x}_i$ for the large-squared-residual observations that make them obviously different and inappropriate to the model, you cannot delete them based on their *y* values. That would be selecting observations non-randomly based on *e*, which would violate the random-sample assumption.)

2. Testing the tradeoff between benefits and salary

a. We are interested in the effect of an increase in benefits on salary. Does salary fall to keep total compensation at the predicted level? Since $totcomp_i = salary_i + benefits_i = \mathbf{x}_i\boldsymbol{\beta} + e_i$, the most obvious way of testing how increases in benefits affect salary would be to run a regression such as $salary_i = \mathbf{x}_i\boldsymbol{\beta} + \gamma\left(benefits_i\right) + e_i$ and testing the estimated value of γ to see whether it is zero (no effect) or –1 (complete offset). Run this regression (using the appropriate controls **x** based on your analysis in question one) and perform these tests.

b. The procedure in part a is problematic for two reasons: (1) unobserved shocks that change salary may also affect benefits, which means that the regressor *benefits* is correlated with the error, and (2) we generally think that it is better to model salary equations with a log dependent variable. We can get around these problems by expressing total compensation as $totcomp = salary\left(1 + \frac{benefits}{salary}\right)$. Take the log to get $\log\left(totcomp\right) = \log\left(salary\right) + \log\left(1 + bs\right)$, where *bs* is the benefits/salary ratio. We can further approximate $\log\left(1 + bs\right) \approx bs$ as long as the value of *bs* is not too large, giving us $\log\left(totcomp\right) = \log\left(salary\right) + bs$. If benefits are valued equally with salary (so that only total compensation matters to teachers), then $\log\left(salary_i\right) = -bs_i + \log\left(totcomp_i\right) = -bs_i + \mathbf{x}_i\boldsymbol{\beta} + e_i$. Why is it more plausible that *bs* would be unrelated to the salary disturbance term than that the level of benefits would be unrelated to that disturbance? Use a regression of this form (using the results of the previous problem to determine the appropriate controls) to test whether (1) benefits have no value to teachers, or (2) benefits are valued equally with salary by teachers. How do your results compare to those of the previous test?

c. What conclusions do you draw about the effects of changes in benefits on salary and total compensation? In thinking about external validity, would you expect these conclusions to hold for teachers in other states? Would you expect them to hold for workers in other professions? How would you use your results to analyze the likely effects of rising health-insurance premiums on salaries?