University of Illinois		Department of Economics
Fall 2005	Econ 508	Roger Koenker

Economics 508: Applied Econometrics
Problem Set 5

This problem concerns predicting productivity of new workers in a large American manufacturing firm. There are five variables: y_i – an observed standardized physical productivity measure for the i^{th} worker after the initial training period, sex_i – a dummy variable for the workers' sex (males are 1) dex_i – a score on a physical dexterity exam administered before the worker was hired, lex_i – the number of years of education of the worker, and *quit* – whether the person quit within the first six months (quitters are 1). The last two columns of the data provide actual duration of employment and a censoring indicator, respectively. If the censoring indicator is 0 then the corresponding duration is censored. These last variables are used only in Question 5.

1. Estimate the model

$$y = \alpha_0 + \alpha_1 sex + \alpha_2 dex + \alpha_3 lex + \alpha_4 lex^2 + u.$$

   (a) Test the hypotheses: $H_0 : \alpha_3 = \alpha_4 = 0$ and $H_0 : \alpha_4 = 0$. Interpret the results of the tests in economic terms.

   (b) Given the results of part a) draw a diagram illustrating the dependence of "mean productivity" on education. Set dexterity at its mean and $sex = 0$. Interpret the picture. How does it change for men? Suppose you thought the whole *shape* of the education effect was different for men and women; reestimate your respecified model. Does this improve things?

   (c) Use the δ-method and/or the bootstrap to construct a confidence interval for $lex^* =$ level of education maximizing expected productivity.

2. Now consider the possibility that the *dispersion and perhaps even the shape* of the conditional density of productivity depends on the $sex - dex - lex$ variables.

   (a) Propose a quantile regression model of this type, estimate and interpret it. For this purpose, redoing the prior plots of mean productivity for several quantiles would be helpful.

(b) Admitting that the whole distribution of productivity changes with the observable covariates leads to a much more complex, and richer, view of the employers decision problem. Suppose that the firm chooses a cutoff of 14 for productivity so that workers who do not acheive this level after one year on the job are dismissed. What proportion of the workers at various education levels (assume mean dexerity scores) would be retained? How would this be likely to affect hiring decisions?

(c) Now suppose that it is very difficult to fire less productive workers, and that the employer want to hire workers to maximize the probability that they would be able to acheive productivity 13. Suggest a hiring strategy.

3. Now consider a similar model for quits

$$logit(P(quit = 1)) = (\beta_0 + \beta_1 sex + \beta_2 dex + \beta_3 lex + \beta_4 lex^2)$$

where $quit = 1$ if the worker quit within the first 6 months after employment, and is 0 otherwise.

(a) Estimate this model by logit, interpret the estimated parameters, in particular the estimated education effect. Draw a picture as in part (1b.) above of the probability of quitting as a function of years of education. Explain the connection between the parameter estimates and the picture.

(b) Explore the effect of gender along the lines of question 1b.

(c) Evaluate the *logit* specification by computing the Pregibon diagnostic suggested in class and interpret.

(d) Presumably, there is a fixed cost of hiring and training so there is an incentive on the part of the firm to avoid hiring workers who are likely to quit after only a few months. How would your findings be expected to influence the firms willingness to hire workers of various education levels?

4. Now we wish to reconsider the *sexdexlex* productivity model of Part 1 exploring the consequences of "sample selectivity". Suppose instead of observing the entire sample of 683 individuals, we instead observed productivity only for those who didn't quit.

(a) Use the Heckman two-step procedure to estimate the productivity equation of Part 1, using only the non-quitters.

(b) Compare and contrast the results from (1.) with your previous results using the full sample, and the results from (naively) applying OLS to the restricted sample. In particular, discuss how the inferences drawn above are altered by the sample selection of non-quitters.

5. Finally we would like to consider a more detailed analysis of quit behavior based on a sample of (censored) survival times rather than the binary dependent variable used in question 2.

   (a) Exploratory data analysis of these survival times is usefully done via the Kaplan Meier estimator. Investigate the efect of gender on quit behavior by estimating seperate survival curves for men and women. Then stratify the sample into three education levels: less than 12 years, exactly 12, and more than 12, and plot the corresponding KM curves for the three groups and interpret.

   (b) Now estimate a Cox proportional hazard model like the one used in question 3.) based on the survival time data and interpret the model, comparing with the results in that question.