Problem Set #1 (2^{nd} Half)

(due Friday, December 2)

Economics 715 **Fall 2022**

Please turn in:
1) Matlab programs and output;
2) An "in words" description of what each program does;
3) Written answers to appropriate problems

Data:
Upload the CPS March 2009 data set from Bruce's website:
`https://www.ssc.wisc.edu/~bhansen/econometrics/`
This is one of the data sets contained in the zip folder available from the "Econometrics Data" link.

You will need to create one new variable: $exp = age - 6 - ed$.

Then, let's "clean" the data. Delete all observations with $exp < 9$ or $exp > 14$, $ed < 11$ or $ed > 17$.

1. (a) Run a least squares regression of $log(earnings)$ on $educ$, exp, and exp^2 (including an intercept). Take note of the education coefficient.

   (b) Consider the quantile regression specification:

   $$log(wage) = \alpha^\tau + ed \cdot \beta_1^\tau + exp \cdot \beta_2^\tau + exp^2 \cdot \beta_3^\tau + \varepsilon^\tau, \quad \text{where } \Pr(\varepsilon^\tau \leq 0 | ed, exp) = \tau \tag{1}$$

   Let $\beta^\tau = (\beta_1^\tau, \beta_2^\tau, \beta_3^\tau)$ and $x_i = (ed_i, exp_i, exp_i^2)'$.

   $$\begin{aligned} S(\alpha, \beta) \quad = \quad & \sum_i [(1-\tau) \cdot |log(wage_i) - \alpha - \beta x_i| \cdot \mathbf{1}\{log(wage_i) \leq \alpha + \beta x_i\} \\ & + \tau \cdot |log(wage_i) - \alpha - \beta x_i| \cdot \mathbf{1}\{log(wage_i) > \alpha + \beta x_i\}] \end{aligned}$$

   Minimizing the criterion function S yields estimates of the quantile coefficients. The *quantreg* function in matlab will estimate using the criterion function S (google search and download the script). Using *quantreg*, run quantile regression of equation (1) for quantiles $\tau = .5$ and $.75$. Let $\hat{\beta}_1^{.75}$ denote the education coefficient in the .75 quantile regression. We will use this value below.

   (c) The *quantreg* function will also produce standard errors using the bootstrap. What is the standard error for $\hat{\beta}_1^{.75}$?

2. (a) Instead of minimizing the criterion S, you can use GMM to estimate equation (1). What's the moment function for GMM? What's the GMM criterion that gets minimized?

   (b) Using the identity matrix as your weight matrix, estimate equation (1) by GMM for quantiles $\tau = .5$ and $.75$.

(c) Derive the asymptotic variance-covariance of the GMM quantile regression coefficient estimators for $\tau = .75$.

(d) Estimate the asymptotic variance-covariance using the simplest form where independence of errors and regressors is assumed. To estimate the density of residuals, fit a normal distribution to the estimated residuals and then use the density estimate from the fitted normal.

(e) Use your estimated asymptotic variance-covariance matrix to obtain a standard error for your estimate of $\beta_1^{.75}$. What is this standard error?

3. Now let's try a different method to estimate the quantile regression coefficients. We will use Classical Minimum Distance (see Newey-McFadden). Note that the covariates (ed, exp) are actually discrete valued. Call the observations with a fixed value of (ed, exp) a "cell", e.g. the $(ed = 12,\ exp = 10)$ cell. You may need to set a minimum number of observations per cell.

   (a) Obtain the $\tau = .75$ quantile of $\log(wage)$ for each ed, exp cell.

   (b) Derive the asymptotic variance-covariance of your quantile estimators across all the cells.

   (c) Estimate the asymptotic variance-covariance. Any density value should be estimated from the corresponding fitted normal distribution.

   (d) For each discrete (ed, exp), you now have an estimated $\log(wage)$ quantile and you have the estimated variance-covariance for these quantiles. Now you can fit a regression line through these $\log(wage)$ quantile values to estimate equation (1) again. But instead of using OLS to estimate the regression line, use FGLS to take advantage of your estimated variance-covariance matrix. (This is Classical Minimum Distance!)

   (e) Derive the asymptotic variance-covariance for your FGLS coefficient estimates.

   (f) Estimate your asymptotic variance-covariance and use the result to obtain the standard error for your estimate of $\beta_1^{.75}$. What is this standard error?

4. You now have three different ways to estimate $\beta_1^{.75}$.

   For each of the three estimation methods, there is a corresponding "bagged" estimation method. You obtain the bagged estimator for each method as follows:
   (i) take B (for B large) nonparametric bootstrap samples;
   (ii) for each bootstrap sample, compute the estimator based on the bootstrap sample;
   (iii) compute the sample average of the B bootstrap estimates - this is the bagged estimator.

   (a) Take a random sample of size 400 from the cleaned CPS data. Obtain your three estimates of $\beta_1^{.75}$ and the three corresponding bagged estimates. Compare your estimates to your estimate $\hat{\beta}_1^{.75}$ from the CPS sample in problem 1(b).

   (b) Now repeat part (a) J times (for J large) and keep track of the performance of your three estimation methods and their corresponding bagged estimates. Treating the estimate $\hat{\beta}_1^{.75}$ from the CPS sample in problem 1(b) as the true value, compute the bias and variance of each estimation method over the J samples. For each of the three estimation methods, which performs better - the estimation method or its corresponding bagged estimation method?