Problem Set 3

MGMT 737

1. **Quantile Regression**. This analysis will use the dataset from Problem Set 1, `lalonde_nsw.csv` (which I will refer to as NSW), as well as the dataset from Problem Set 2, `lalonde_psid.csv` (which I will call PSID).

   (a) We will begin by defining an estimation approach for doing quantile regression that doesn't require linear programming. This approach comes from Gary Chamberlain (in Chamberlain (1994), and discussed in Angrist et al. (2006)).

   Let X be a (discrete) right hand side variable with J discrete values. For each j value of $X = x_j$, calculate $\hat{\pi}_\tau(x) = Q_\tau(Y|X_j)$, which is the τ percentile of the outcome variable, conditional on the value of X, and $\hat{p}_j$, which is the empirical probability of $X = x_j$. Do so using the PSID dataset for $X =$ education, for $\tau = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$, using `re78` as the outcome variable.

   (b) Given these inputs, the quantile regression slope estimates is just

   $$\hat{\beta}_\tau = \arg\min_b \sum_j (\hat{\pi}_\tau(x_j) - x_j b)^2 \hat{p}_j.$$

   This is a simple (weighted) linear regression (or minimum distance problem), with the diagonal weight matrix $\hat{W} = diag(\hat{p}_1, \ldots, \hat{p}_J)$. Estimate $\hat{\beta}_\tau$ for the education example above.

   (c) Our variance estimator is the sum of two terms (coming from uncertainty in the QCF, and the estimation of the slope conditional on those terms), V and D:

   $$\begin{aligned}
   V &= \left(\boldsymbol{x}'\hat{W}\boldsymbol{x}\right)^{-1} \boldsymbol{x}'\hat{W}\Sigma\hat{W}\boldsymbol{x} \left(\boldsymbol{x}'\hat{W}\boldsymbol{x}\right)^{-1}, \\
   \Sigma &= diag(\sigma^2_{\tau,1}/p_1, \ldots, \sigma^2_{\tau,J}/p_J) \\
   D &= \left(\boldsymbol{x}'\hat{W}\boldsymbol{x}\right)^{-1} \boldsymbol{x}'\hat{W}\Delta\hat{W}\boldsymbol{x} \left(\boldsymbol{x}'\hat{W}\boldsymbol{x}\right)^{-1}, \\
   \Delta &= diag((\pi_{\tau,1} - x_1\beta_\tau)^2/p_1, \ldots, (\pi_{\tau,J} - x_J\beta_\tau)^2/p_J).
   \end{aligned}$$

   Everything here should be straight forward to estimate, except for $\sigma^2_{\tau,j}$. To do this, define the following order statistics:

   $$\begin{aligned}
   b_j &= \max\left\{1, \text{round}\left(\tau N_j - z_{1-\alpha/2}\sqrt{\tau(1-\tau)N_j}\right)\right\} \\
   t_j &= \min\left\{N_j, \text{round}\left(\tau N_j + z_{1-\alpha/2}\sqrt{\tau(1-\tau)N_j}\right)\right\},
   \end{aligned}$$

   where round(a) rounds to the closest integer, and $z_{1-\alpha/2} = 1.96$, typically, and N_j is the number of observations in the jth bin of X. Then,

   $$\hat{\sigma}^2_{\tau,j} = N_j \left(\frac{y_{j(t_j)} - y_{j(b_j)}}{2z_{1-\alpha}}\right)^2. \tag{1}$$

   Report the standard error on your estimates, which is calculated as $\sqrt{(V+D)/N}$

   (d) Finally, using the NSW dataset, calculate the $\tau = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$ treatment effects, and their standard errors.

2. **Poisson Regression**. This analysis will use the dataset `Detroit.csv`. You should use built-in pacakges for this – in `R`, use the `fixest` library. In `Stata`, use the `reghdfe` and `ppmlhdfe` packages.

(a) Run an OLS regression of `flows` (the number of workers who work in `home_ID` and work in `work_ID`) on `distance_Google_miles`, and include `home_ID` and `work_ID` fixed effects (absorb them), and cluster on `home_ID`. Report the coefficient and standard error on `distance_Google_miles`.

(b) Run an OLS regression of `log(flows)` on `log(distance_Google_miles)` and include `home_ID` and `work_ID` fixed effects , omitting the cells with zero flows. Cluster on `home_ID`. Report the coefficient and standard error on `log(distance_Google_miles)`.

(c) Repeat part 1b, but instead of omitting the zero cells, run the OLS regression of `log(c + flows)` for c = 0.1, 1 and 10. Compare how your coefficients change.

(d) Finally, repeat part 1a using Poisson regression, and contrast the estimates to Part b and c.

3. **Duration Modeling**. This analysis will use the dataset `acs_duration.csv`. The `acs_duration.csv` dataset is from the American Community Survey in 2019, and has heads of households' responses to the question "How long have you lived in this home?" (`moving_approx` – in reality, this value is given as a range in the public data – I have imputed using the midpoint. A fun exercise left to the reader is to think about how to generalize this problem using ranges.) and homeownership (`homeowner` vs. `renter`).

   (a) Using the ACS data,write down how to estimate the unconditional probability that a household stays in a home for T or more years, using the available data. Estimate this for $T = 7$ and report the value.

   (b) Calculate the hazard rate for each observed value of `moving_approx`. Report this value for $T = 7$.

   (c) Recalculate these hazard values for $T = 7$ for homeowners and renters. Contrast the difference in hazard rates over time.