Problem Set 1

Due date: September 30, 2022

In this problem set you will estimate conditional distributions of wages give a set of individual characteristics and then use these estimates to generate counterfactual distribution functions. The conditional distribution of outcome Y given a set of covariates X is given by

$$F(y|x) = P(Y \leq y|X = x) = E\big[1\{Y \leq y\}|X = x\big]$$

Using the observation that the conditional CDF is a conditional expectation function, we can estimate it using standard regression techniques. Since the dependent variable $1\{Y \leq y\}$ is binary, we can model the conditional CDF as

$$F(y|x) = G\big(\beta_y' B(x)\big)$$

where G is some chosen link function, $B(x)$ is a dictionary of transformations of the covariates (e.g. interactions, powers and so on) that includes a constant, and β_y is a vector of coefficients that varies by choice of threshold y.

The model is quite flexible, even for a given choice of link function since we can include a flexible set of transformations in $B(x)$, and since we allow the coefficients β_y to vary by threshold. As an example, consider a classical normal regression model in which

$$Y_i = \beta' B(X_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

which implies the conditional CDF $F(y|x) = \Phi((y - \beta' B(x))/\sigma)$. In this simple model the parameters β are the same at every y. However, by allowing the slopes to vary we allow the conditional distribution of Y given X to be non-normal, and for the conditional variance to vary by y also.

Given estimates $G\big(\widehat{\beta}_y' B(x)\big)$, for a grid of values $(y_1, \ldots, y_K)$ we can estimate the curve $F(y|x)$

for any choice of fixed X. We could also consider estimating the conditional quantile function

$$Q(u|x) = \inf\{t \in T : F(t|x) \geq u\} \wedge \sup\{t \in T\}$$

where T is the support of Y. This is as straightforward as flipping the axis and mirroring the curve.

In the below set of explanations, you will estimate counterfactual CDFs for log wages, conditional on a set of individual characteristics. The data are from the Current Population Survey (CPS) in 2012 and contain observations on workers aged 25-64 years, working more than 35 hours per week for at least 50 weeks of the year (the sample contains only white workers and excludes some groups such as the self employed). There are 29 217 observations in total. The outcome of interest is log hourly wage (the ratio of yearly earnings to hours worked), and control variables include

- 5 marital status indicators: widowed, separated, divorced, never married, married
- 6 educational indicators: 0-8 years schooling (hsd08), high-school dropout (hd911), high school grad (hsg), some college (sc), college grad (cg), and advanced degree (ad)
- 4 regional dummies: midwest (mw), south (so), west (we), north-east (ne)
- experience (included as a quartic, exp1 - exp4)
- gender (a dummy equal to one for female)

Part 1

Provide summary statistics for the variables broken down by gender. Are there any significant differences in the distribution of variables between women and men?

Part 2

We will estimate the conditional CDF using a probit model, separately for men and women, i.e.

$$\widehat{F}_j(y|x) = \Phi\big(B(x)'\widehat{\beta}_{j,y}\big)$$

where Φ is the CDF of a standard normal variable and $j = \{w, m\}$ indicates either the subsample of women or of men.

Write down a set of moment conditions that identify the parameters β_y. Is your model exactly identified or overidentified? Is the set of moment conditions you have constructed efficient?

Part 3

Pick some grid of values for y, $(y_1, \ldots, y_K)$ - one sensible way to do this would be to choose evenly spaced quantiles of the distribution of log-wage, e.g. the 2.5, 5, 7.5, ... per cent quantiles.

Estimate the parameters of the probit model separately for men and women at each value of y using GMM and the moment conditions you have constructed above. This gives two sets of estimated coefficients, $\widehat{\beta}_{w,y}$ and $\widehat{\beta}_{m,y}$, for each value of y.

Which parameters are most important in determining wages? Which coefficients differ most between men and women?

Using these estimates construct the counterfactual distribution of women's log wages, given the estimated wage structure of men. Specifically, for each value of y compute

$$\widehat{F}_{m|w}(y) = \int \widehat{F}_m(y|x) dF_w(x) = \int \Phi\big(B(x)'\widehat{\beta}_{m,y}\big) dF_w(x)$$

where $F_w(x)$ is the observed distribution of covariates for women. This is the estimated proportion of women that would earn less than or equal to y, if women's wages were determined by the estimated returns to covariates of men, $\widehat{\beta}_{m,y}$.

On a single plot show: (1) $F_m(y)$, the CDF of men's wages, (2) $F_w(y)$, the CDF of women's wages, and (3) $\widehat{F}_{m|w}(y)$ the estimated counterfactual CDF.

Part 4

We can use the counterfactual CDF to decompose the difference in the CDF of log wages for women and men.

$$F_{m|m} - F_{w|w} = (F_{m|m} - F_{m|w}) + (F_{m|w} - F_{w|w})$$

where $F_{m|m}$ is the CDF given men's covariates and wage structure, which is simply the unconditional distribution of men's wages F_m, and similarly $F_{w|w} = F_w$. The first component $(F_{m|m} - F_{m|w})$ is a 'composition effect' , which represents the difference in wage distributions that is explained by differences in the covariate distributions of men and women. The second term $(F_{m|w} - F_{w|w})$ is a 'price effect' or 'discrimination effect' that represents differences in the distributions that are due to differences is the coefficients $\beta_{m,y}$ and $\beta_{w,y}$, e.g. if the return to education differs across gender.

This decomposition is essentially a distributional version of the Kitagawa-Blinder-Oaxaca decomposition. Note that we could alternatively have computed the decomposition by using $F_{w|m}$ rather than $F_{m|w}$ as the counterfactual. The two choice can in principle give different results.

One computes the value of the composition effect using the men's wage structure, while the other computes it at the women's wage structure and so how we value different covariates is affected. Similarly, the price effect is averaged either over the men's covariate distribution or the women's, which again can affect the result.

Plot the estimated composition and price effects. Which plays the larger role in explaining differences in wages in this case?

Part 5

We will now compute confidence bands for the estimated distributions. Since we have set the problem up using moment conditions and estimated with GMM, we could compute standard errors using the estimated covariance matrix for the coefficients along with the delta-method (since our object of interest is a smooth function of the covariates). However, we will instead proceed by the bootstrap.

For each estimated CDF $\widehat{F}(y)$ (the empirical CDF of wages for men, for women, and the estimated counterfactual CDF), obtain bootstrap confidence sets as follows:

1. Draw B bootstrap samples from the data and compute corresponding estimates $\widehat{F}_b^*(y)$ for each value of y.

2. Compute a bootstrap estimate of the variance for each value of y as

$$\widehat{s}^2(y) = \frac{1}{B}\sum_b \left(\widehat{F}_b^*(y) - \widehat{F}(y)\right)^2$$

3. For each bootstrap draw, define

$$m_b^* = \max_y \left|\frac{\widehat{F}_b^*(y) - \widehat{F}(y)}{\widehat{s}(y)}\right|$$

   Compute the critical value $c_{1-\alpha}$ as the $(1-\alpha)$ quantile of $\{m_b^*\}_{b=1}^B$.

4. Construct a (jointly valid) confidence set as

$$\widehat{F}(y) \pm c_{1-\alpha}\widehat{s}(y)$$

The confidence set is jointly valid in the sense that it will contain the true CDF at each y simultaneously with probability $1-\alpha$.

Plot the three estimated CDFs along with their confidence bands.

Comment 1: Confidence bands for the quantile functions can be computed flipping and mirroring the distribution confidence bands in the exact same way as for the point estimates. (Note that the upper confidence band will become the lower)

Comment 2: Neither the estimated counterfactual CDF nor the estimated confidence bands are guaranteed to be non-decreasing. We can fix this by simply enforcing the known shape restrictions onto our estimated curves (and confidence bands). There are a number of ways to impose monotonicity on functions, but one simple way is to just resort the $\widehat{F}(y)$ into ascending order. Chernozhukov, Fernandez-Val and Galichon (2008) show that rearrangement to satisfy shape constraints weakly improves both accuracy of estimated curves and coverage of confidence bands.

Part 6

Finally, let's construct confidence bands for the estimated composition and price effects. Let $[u_1, u_2]$ be a confidence interval for an estimate U, and $[v_1, v_2]$ and confidence interval for V. Then we can construct a confidence interval for $V - U$ as

$$[v_1 - u_2, v_2 - u_1]$$

This is simply set $\{v - u : v \in [v_1, v_2], u \in [u_1, u_2]\}$, sometimes known as the Minkowski difference.

Using your estimated confidence sets for $F_m(y)$, $F_w(y)$, and $\widehat{F}_{m|w}(y)$, construct confidence sets for the price and composition effects. Plot these along with the point estimates. Comment on the results.