(a) Consider linear regression of $Y$ on $(D, X)$. Show that the coefficient of $D$ can not consistently estimate the average wage gap between sectors: $\mathbb{E}[Y_1 - Y_0]$.

Notice that $Y = DY_1 + (1-D)Y_0$

$$= Y_0 + (Y_1 - Y_0)D$$

$$= X'\beta_0 + U_0 + (X'\beta_1 + U_1 - X'\beta_0 - U_0)D$$

$$= X'\beta_0 + U_0 + (X'\beta_1 - X'\beta_0)D + (U_1 - U_0)D$$

$$= X'\beta_0 + (X\beta_1 - X'\beta_0)D + \underbrace{(U_1 - U_0)D + U_0}_{\varepsilon}$$

where $cov(\varepsilon, D) \neq 0$, So the OLS does not work.

There exists endogenity problem. Thus, we cannot use D to consistently estimate the $E[Y_1 - Y_0]$.

(b) Discuss the economic interpretation of $\mathbb{E}[Y_d|D = d, X = x]$ and $\mathbb{E}[Y_d|X = x]$, respectively.

The $E[Y_d | D=d, X=x]$ indicate the expected earning of the group of individuals with characteristics $x$ work in sector D is $E[Y_d | D=d, X=x]$ on average.

In sector 1, the expected earning is.

$E[Y_1 | D=1, X=x]$

In sector 0, the expected earning is

$E[Y_0 | D=0, X=x]$

The $E(Y_d | X=x)$ indicate that the average earnings for group of individuals with characteristics X. It includes all workers, no matter which sector they are work with.

(c) One key parameter measures the consequence caused by self-selection. It is summarize best by $\Delta_d(x) = \mathbb{E}[Y_d|D = d, X = x] - \mathbb{E}[Y_d|X = x]$. Express $\Delta_d(x)$ explicitly as a function of $X$ and model parameters.

$$\Delta_d(x) = E[Y_d | D = d, X = x] - E[Y_d | X = x]$$

$$= E[x'\beta_d + U_d | D = d, X = x] - E[x'\beta_d + U_d | X = x]$$

Notice that $(U_1, U_0) | X \sim N((0,0), \sigma_1^2, \sigma_0^2, \sigma_{01})$

$$\Delta_d(x) = E[x'\beta_d | D = d, x = x] - E[x'\beta_d | X = x]$$

Notice that $D = 1[Y_1 \geq Y_0]$

So $P(D=1) = P(Y_1 \geq Y_0)$

Suppose $E[X'\beta_d | D=1, X=x] = x'\beta_d$

$$\Delta_d(x) = E[X'\beta_d | D = d, X = x] -$$

$$(E[x'\beta_d | D=1, X=x] P(D=1 | X=x) + E[x'\beta_d | D=0, X=x] P(D=0 | X=x))$$

$$= E[X'\beta_d | D = d, X = x] -$$

$$(E[x'\beta_d | D=1, X=x] P(Y_1 \geq Y_0) + E[x'\beta_d | D=0, X=x] P(Y_1 < Y_0))$$

$$= X'\beta_d - x'\beta_1 P(Y_1 \geq Y_0) - x'\beta_0 P(Y_1 < Y_0)$$

$$= x'\beta_d - x'\beta_1 P(Y_1 \geq Y_0) - x'\beta_0 (1 - P(Y_1 \geq Y_0)))$$

$$= x'\beta_d - x'\beta_1 P(Y_1 \geq Y_0) - x'\beta_0 + x'\beta_0 P(Y_1 \geq Y_0)$$

$$= x'\beta_d - x'\beta_0 - (x'\beta_1 - x'\beta_0) P(Y_1 \geq Y_0)$$

For $d = 1$

$$\Delta d(x) = x'\beta_1 - x'\beta_0 - (x'\beta_1 - x'\beta_0) P(Y_1 \geq Y_0)$$

$$= (x'\beta_1 - x'\beta_0)(1 - P(Y_1 \geq Y_0))$$

$$= (x'\beta_1 - x'\beta_0) P(Y_1 < Y_0)$$

For $d = 0$

$$\Delta d(x) = x'\beta_0 - x'\beta_0 - (x'\beta_1 - x'\beta_0) P(Y_1 \geq Y_0)$$

$$= -(x'\beta_1 - x'\beta_0) P(Y_1 \geq Y_0)$$

(d) Propose a way to estimate model parameters. Write down the sample objective function that you will maximize/minimize to calculate the estimates.

$ATE = E[\Delta] = E[Y_1 - Y_0]$

Suppose we do observe both $Y_1$ & $Y_0$ for everyone

$$\widetilde{\hat{ATE}} = \frac{1}{n}\sum_{i=1}^{n}(Y_{1i} - Y_{0i})$$

For people with $D_i = 1$, the $Y_{0i}$ is missing

$D_i = 0$, the $Y_{1i}$ is missing

For individual i with $D_i = 1$

Let $X_i$ be the characteristics, now we find j, with same charatertistics $X_i = X_j$ and $D_j = 0$. Thus $Y_j = Y_{0j}$ is observed

We can set $M^0(i)$ be the set of matched observation from the group with $D=0$ (control Group) and $M^1(i)$ be the set of matched observations from treatment group $(D=1)$

# $M^0(i)$ is the set of matched observation from the control group

# $M^1(i)$ is the set of matched observation from the treatment group

If $D_i=1$, then $Y_{1i} = Y_i$ is observed and

$$\hat{Y}_{0i} = \frac{1}{\# M^0(i)} \sum_{i \in M(i)} Y_i$$

If $D_i=0$, then $Y_{0i} = Y_i$ is observed and

$$\hat{Y}_{1i} = \frac{1}{\# M^1(i)} \sum_{i \in M(i)} Y_i$$

$$\widetilde{\hat{A}TE} = \frac{1}{n} \sum_i \{ D_i (Y_i - \hat{Y}_{0i}) + (1-D_i)(\hat{Y}_{1i} - \hat{Y}_i) \}$$

We need to select the $X_j$ with "small distance" to $X_i$

$$d(x_i, x_j) = w_1 (X_{i1} - X_{j1})^2 + W_2 (X_{i2} - X_{j2})^2$$

$$= \sum_{k=1}^{K} W_K (X_{iK} - X_{jK})^2$$ for $x$ is $K$ dimension

and $W_K$ are the weights assigned to $K$

$$M^0(i) = \{ j : d(x_i, x_j) < c, D_j = 0 \}$$

$$M^1(i) = \{ j = d(x_i, x_j) < c, D_j = 1 \}$$

When the control group is large, we can find many $X_j$ such that the match is very precise. (i.e. small $d(x_i, x_j)$)

and bias is eliminate)

Thus, in this process, we will select $d(x_i, x_j)$ as the objective function and try to minimize the $d(x_i, x_j)$ for gaining a precise estimate.

Q2

The paper “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records” written by Joshua D. Angrist research on the long-term labor market consequences of participating military service during the Vietnam era. This result can help the government to implement appropriate military manpower policy for compensating the veterans’ service.

This paper's outcome variable is veterans' civilian earnings, and the explanatory variable is the dummy variable veteran status (whether join the military service). The main difficulty in the econometric modeling is that the veteran status could be influenced by other potential possibilities, which causes endogeneity. For instance, men with few civilian opportunities are more likely to serve in the armed forces. The failure of controlling for differences in other individual characteristics related to earnings makes the measurements of the effects of civilian earnings by veteran status biased. The traditional OLS model cannot work since some unobservable confounders make the explanatory variables correlated with the regression model's error terms.

We can cast the research question into the Potential Outcome Model (POM). The potential outcome $Y_0$ is the earnings for male who did not join the army and $Y_1$ is the earnings for male who joined the army. The binary treatment $D_i$ is the veteran status, $D_0 = 0$ represents man is not enlist and $D_1 = 1$ represents man does enlist. We use $Y_i = D_i Y_1 + (1 - D_i) Y_0$ to indicate the individual earnings and use the change of earnings for the same person $\Delta = Y_1 - Y_0$ to measure the individual treatment effect. For anyone, either $Y_1 \text{ } or \text{ } Y_0$ is observed, depending on D. Thus, we need to compare between different individuals and estimate average treatment effect. We can write the linear model in detail as:

$$Y_{cit} = \beta_c + \delta_t + \alpha_i D_i + u_{it}$$

where $Y_{cit}$ is the earnings of man i in cohort c at time t. The $\beta_c$ and $\delta_t$ is the cohort effect and period effects for all cohorts, respectively. $\alpha$ is the effect of enlist on earnings and $u_{it}$ is a residual.

Due to the endogeneity problem, we need to introduce the instrumental variable Z for using the POM. The binary instrument $Z$ represents the draft-eligible status decided by a lottery. The Random Sequence Number (RSN) from 1-365 is assigned to the birthdate and it represent the lottery number for each male. There is a ceiling determined by the Defense Department, and the man with assigned birthdate number (lottery number) that smaller than the ceiling will be considered as ‘draft-eligible’. In order to use the POM, we need to test the assumptions for the instrumental variable. First, we need to check the eligibility of draft on earnings. Based on the rule, the man will be called to enlist if he has a lottery number that lower than ceiling, which implies correlation. Second, the independence/endogeneity must hold. The draft-eligible status $Z$ is randomly decided by lottery numbers. By this way, $Z$ must independent with the male earning $Y_i$. Draft-eligible status will not change any unobservable terms that may influence the earnings like the education background. The endogeneity is hold. The only thing the lottery influence is the veteran status, so that the Z only impact the earnings through the explanatory variables: veteran status. Therefore, all assumptions for instrumental variable are hold.

With the usage of instrumental variable, we will have following estimator for

$\widehat{\alpha_\iota} = \frac{\overline{y^e}-\overline{y^n}}{\widehat{p^e}-\widehat{p^n}}$ $where\ \hat{p}\ is\ the\ prportion\ of\ the\ cohort\ that\ entering\ the\ miliary\ and\ \bar{y}$ is the mean earnings. e and n denote draft-eligible and draft-not-eligible. This estimator helps to adjusts the earning differences by draft-eligibility status. Based on POM, the 2SLS procedure will estimate the local average treatment effect of veteran status on civilian earning with instrumental variable draft-eligible status. The estimates are unbiased from the influence that certain types of men are more likely to enlist and other unobservable factors. Thus, this result will provide accurate and helpful information for policymakers. For instance, it helps the government implement an appropriate compensation policy for veterans who lose two years of civilian labor market experience on average due to enlist.

This paper (Angrist,1990, AER) and the POM discussed in Imbens and Angrist (1994, Econometrica), the methods both mention that how the 2SLS is used to measure the local average treatment effect. In 1994, they restrict the condition 2: Monotonicity to ensures that "the instrument affects the participation or selection decision in a monotone way" (Imbens and Angrist, 1994, Econometrica). In draft-lottery cases, it requires the lottery participants who will enlist with lottery number k would also enlist with lottery number l and l<k. This assumption works only for the complier who follow the assignment decision, always taker and never taker, but not defier. In 1990, one key defect is that there exists a group of males who enlist but not because of lottery. So, they might become the defier which influence the results. Moreover, both paper indicates comparing the average outcome Y and treatment D at two different values of instrument Z can help to estimate the local average treatment effect.

Reference

Angrist, J. D. (1990). The draft lottery and voluntary enlistment in the Vietnam era: Evidence form Social Security Administrative Records. *American Economic Association,* 313-336. doi:10.3386/w3514

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica, 62*(2), 467. doi:10.2307/2951620