PS3

April 27, 2024

1. (a)

$$
\begin{aligned}
H &= \sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 d_i - \widehat{\beta}_2 z_i - \widehat{\beta}_3 d_i z_i)^2 \\
\frac{\partial H}{\partial \widehat{\beta}_0} &= 2\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 d_i - \widehat{\beta}_2 z_i - \widehat{\beta}_3 d_i z_i)(-1) = 0 \\
\frac{\partial H}{\partial \widehat{\beta}_1} &= 2\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 d_i - \widehat{\beta}_2 z_i - \widehat{\beta}_3 d_i z_i)(-d_i) = 0 \\
\frac{\partial H}{\partial \widehat{\beta}_2} &= 2\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 d_i - \widehat{\beta}_2 z_i - \widehat{\beta}_3 d_i z_i)(-z_i) = 0 \\
\frac{\partial H}{\partial \widehat{\beta}_3} &= 2\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 d_i - \widehat{\beta}_2 z_i - \widehat{\beta}_3 d_i z_i)(-d_i z_i) = 0
\end{aligned}
$$

   (b) $\bar{y}_{11}$ represents the average income of rural female, $\bar{y}_{10}$ represents the average income of urban female, $\bar{y}_{01}$ represents the average income of rural male and $\bar{y}_{00}$ represents the average income of urban male.

   (c)

$$
\begin{aligned}
&\sum_{i=1}^{n} d_i z_i (y_i - b_0 - b_1 d_i - b_2 z_i - b_3 d_i z_i) \\
&= \sum_{i=1}^{n} d_i z_i \left[y_i + \bar{y}_{00}(-1 + d_i + z_i - d_i z_i) + \bar{y}_{01}(-z_i + d_i z_i) + \bar{y}_{10}(-d_i + d_i z_i) + \bar{y}_{11}(-d_i z_i)\right] \\
&= \sum_{i=1}^{n} \left[y_i d_i z_i + \bar{y}_{00}(-d_i z_i + d_i z_i + d_i z_i - d_i z_i) + \bar{y}_{01}(-d_i z_i + d_i z_i) + \bar{y}_{10}(-d_i z_i + d_i z_i) + \bar{y}_{11}(-d_i z_i)\right] \\
&= \sum_{i=1}^{n} \left[y_i d_i z_i + \bar{y}_{00} \cdot 0 + \bar{y}_{01} \cdot 0 + \bar{y}_{10} \cdot 0 + \bar{y}_{11}(-d_i z_i)\right] \\
&= \sum_{i=1}^{n} [y_i d_i z_i + \bar{y}_{11}(-d_i z_i)] \\
&= \sum_{i=1}^{n} y_i d_i z_i + \bar{y}_{11} \sum_{i=1}^{n}(-d_i z_i) \\
&= \sum_{i=1}^{n} y_i d_i z_i - \frac{\sum_{i=1}^{n}(y_i d_i z_i)}{\sum_{i=1}^{n}(d_i z_i)} \sum_{i=1}^{n}(d_i z_i) \\
&= 0
\end{aligned}
$$

   (d)

$$
\begin{aligned}
\mathbb{E}(\bar{y}_{11}) &= \mathbb{E}\left(\frac{\sum_{i=1}^{n} y_i d_i z_i}{\sum_{i=1}^{n} d_i z_i}\right) \\
&= \frac{\sum_{i=1}^{n} d_i z_i \mathbb{E}(y_i)}{\sum_{i=1}^{n} d_i z_i} \\
&= \mathbb{E}(y | d = 1, z = 1)
\end{aligned}
$$

Similarly, we can get $\mathbb{E}(\bar{y}_{10}) = \mathbb{E}(y|d=1,z=0)$, $\mathbb{E}(\bar{y}_{01}) = \mathbb{E}(y|d=0,z=1)$ and $\mathbb{E}(\bar{y}_{00}) = \mathbb{E}(y|d=0,z=0)$. Then consider the following equations:

$$
\begin{aligned}
\beta_0 &= \mathbb{E}(\widehat{\beta}_0) = \mathbb{E}(b_0) = \mathbb{E}(\bar{y}_{00}) = \mathbb{E}(y|d=0,z=0) \\
\beta_1 &= \mathbb{E}(\widehat{\beta}_1) = \mathbb{E}(b_1) = \mathbb{E}(\bar{y}_{01}) - \mathbb{E}(\bar{y}_{00}) = \mathbb{E}(y|d=0,z=1) - \mathbb{E}(y|d=0,z=0) \\
\beta_2 &= \mathbb{E}(\widehat{\beta}_2) = \mathbb{E}(b_2) = \mathbb{E}(\bar{y}_{10}) - \mathbb{E}(\bar{y}_{00}) = \mathbb{E}(y|d=1,z=0) - \mathbb{E}(y|d=0,z=0) \\
\beta_3 &= \mathbb{E}(\widehat{\beta}_3) = \mathbb{E}(b_3) = \mathbb{E}(\bar{y}_{11} - \bar{y}_{10} - \bar{y}_{01} + \bar{y}_{00}) \\
&= \mathbb{E}(y|d=1,z=1) - \mathbb{E}(y|d=1,z=0) - \mathbb{E}(y|d=0,z=1) + \mathbb{E}(y|d=0,z=0)
\end{aligned}
$$

(e) The disparity in income improvement effects between men and women transitioning from urban to rural areas.

(f) $H_0: \beta_1 + \beta_3 = 0$, $H_1: \beta_1 + \beta_3 \neq 0$ Then we can use F-test to test the hypothesis. The restrcited model is $y_i = \beta_0 + \beta_1 d_i + \beta_2 z_i - \beta_1 d_i z_i + u_i$. Calulate the F-statistic $F = \frac{(R_{ur}^2 - R_r^2)/1}{R_r^2/(n-4)}$ and compare it with the critical value. If $F > F_{1-0.025,1,n-4}$, we reject the null hypothesis.

(g) $H_0: \beta_1 = \beta_3 = 0$. The restrcited model is $y_i = \beta_0 + \beta_2 z_i + u_i$. The steps are similar to the previous question, but change the first parameter of F-statistic to 2, which is $F = \frac{(R_{ur}^2 - R_r^2)/2}{R_r^2/(n-4)}$. And compare it with the critical value $F_{1-0.025,2,n-4}$. If the F-statistic is larger, we reject the null hypothesis.

2. (a)

$$
\begin{aligned}
\mathbb{E}(\widehat{\beta}_1) - \tau_{ATE} &= \mathbb{E}\left[\frac{1}{n_1}\sum_{i=1}^{n} d_i y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-d_i)y_i\right] - \mathbb{E}(y(1) - y(0)) \\
&= \frac{\sum_{i=1}^{n} d_i \mathbb{E}(y_i)}{\sum_{i=1}^{n} d_i} - \frac{\sum_{i=1}^{n}(1-d_i)\mathbb{E}(y_i)}{\sum_{i=1}^{n}(1-d_i)} - \mathbb{E}(y(1)) + \mathbb{E}(y(0)) \\
&= \mathbb{E}(y(1)|d=1) - \mathbb{E}(y(0)|d=0) - \mathbb{E}(y(1)) + \mathbb{E}(y(0)) \\
&= \mathbb{E}(y(1)|d=1) - \mathbb{E}(y(0)|d=0) - [\mathbb{E}(y(1)|d=1)\cdot p_1 + \mathbb{E}(y(1)|d=0)\cdot(1-p_1)] \\
&\quad + [\mathbb{E}(y(0)|d=1)\cdot p_1 + \mathbb{E}(y(0)|d=0)\cdot(1-p_1)] \\
&= (\mathbb{E}(y(1)|d=1) - \mathbb{E}(y(1)|d=0))(1-p_1) + (\mathbb{E}(y(0)|d=1) - \mathbb{E}(y(0)|d=0))\, p_1
\end{aligned}
$$

(b) Since we have

$$y = \beta_0 + \beta_1 d + u = dy(1) + (1-d)y(0)$$

, so we can get

$$y(1) = \beta_1 + \beta_0 + u \quad y(0) = \beta_0 + u$$

, which implys that

$$\mathbb{E}(y(1)|d=1) = \mathbb{E}(y(1)|d=0) \quad \mathbb{E}(y(0)|d=1) = \mathbb{E}(y(0)|d=0)$$

, then we can get

$$\mathbb{E}(\widehat{\beta}_1) - \tau_{ATE} = 0$$

. If we further assume that $\mathbb{E}[u|d] = 0$, it won't affect the result.

(c)

$$
\begin{aligned}
y &= \beta_0' + \tau_{ATE} d + u' \\
&= \beta_0' + \mathbb{E}[y(1) - y(0)]d + u' \\
&= \beta_0' + [(\beta_1 + \beta_0 + u) - (\beta_0 + u)]d + u' \\
&= \beta_0' + \beta_1 d + u' \\
&= \beta_0 + \beta_1 d + u
\end{aligned}
$$

par So, we can get $\beta_0' = \beta_0$ and $u' = u$.

3. (a) No, since *attend* can be affected by *alcohol* (like students who drink too much will not attend the class tomorrow) . And we only want to estimate the effect of *alcohol* on *colGPA* . If we add *attend* into the model, the coefficient of *alcohol* 's meaning will be change to the effect of *alcohol* on *colGPA* when *attend* is fixed, the power of *alcohol* will be weakened.

(b) No, *gaokaoScore* and *hsGPA* can both be affected by *alcohol*, they'll weaken the power of *alcohol* on *colGPA* if we add them into the model, just like the previous question.

4. (a) See the log file.

   (b) $\widehat{\beta}_1$ means holding *restaurn*, people between 30 and 50 years old have 3.1 more cigarettes smoked per day than people below 30 years old.
   $\widehat{\beta}_1$ means holding *restaurn*, people between 50 and 70 years old have 0.92 more cigarettes smoked per day than people below 30 years old.
   $\widehat{\beta}_1$ means holding *restaurn*, people above 70 years old have 5.8 less cigarettes smoked per day than people below 30 years old.

   (c) The marginal effect of *age* on *cigs* is $\frac{\partial cigs}{\partial age} = \theta_1 + 2\theta_2 age$. Solve: $\widehat{\theta_1} + 2\widehat{\theta_2} \cdot age = 0$, we get $age = 43$.

   (d) Holding other factors fixed, the decrease in smoking per day if the city requires no smoking in restaurants.

   (e)
$$\frac{\partial E(cigs)}{\partial educ} = \begin{cases} \gamma_1 & \text{if } restaurn = 0 \\ \gamma_1 + \gamma_3 & \text{if } restaurn = 1 \end{cases} \tag{1}$$

   γ_3 means the increase of the marginal effect of *educ* on smoking per day if the city requires no smoking in restaurants.

   (f) From the Stata output, we know the p-value for $\gamma_3 = 0$ is 0.890, which is smaller than 0.95, so we can reject the null hypothesis.
   γ_3 is significant at the 5% level.