پاسخ تمرین سری نهم

۱- یک مدل رگرسیون ساده را درنظر بگیرید که اثر داشتن کامپیوتر شخصی را بر نمره را برای دانشجویان در یک دانشگاه بزرگ تخمین می‌زند:

$$GPA = \beta_0 + \beta_1 PC + u$$

PC متغیر دودویی برای داشتن PC است.

i. چرا PC میتواند با جزءخطا همبستگی داشته باشد.

مشخص است که وضعیت اقتصادی بر عملکرد دانش آموزان تاثیر می‌گذارد. جزءخطا شامل عوامل مختلفی از جمله درآمد خانواده است که تاثیر مثبتی بر GPA دارد و همچنین با احتمال زیادی با داشتن PC همبسته است.

ii. توضیح دهید چرا PC ممکن است با درآمد سالیانه والدین مرتبط باشد. آیا به این معنی است که درآمد والدین یک متغیر ابزاری خوب برای PC است؟ چرا؟

خانوار با درآمد بالا، می‌توانند برای فرزندان خود کامپیوتر بخرند. بنابراین درآمد خانواده قطعا دومین شرط برای متغیر ابزاری بودن را برآورده می‌کند یعنی به متغیر توضیحی به‌صورت درونزا ارتباط دارد. اما همانطور که در قسمت قبل دیدیم faminc اثری مثبت بر GPA دارد. بنابراین شرط اول برای یک متغیر ابزاری خوب برای faminc برقرار نیست. اگر faminc را داشتیم آن را در معادله کنترل می‌کردیم. اگر این تنها متغیر حذف‌شده بود که با PC در ارتباط است می‌توانستیم معادله را با OLS تخمین بزنیم.

iii. **فرض کنید ٤سال پیش دانشگاه به نیمی از دانشجویان که بطور تصادفی انتخاب شدند کمک هزینه خرید کامپیوتر داده است. توضیح دهید چگونه می‌توان از این اطلاعات استفاده کرد برای ساخت یک متغیر ابزاری برای PC.**

این یک آزمایش است برای اینکه داشتن کامپیوتر دارای تاثیر است یا نه. بعضی از دانش‌آموزان که با کمک هزینه، کامپیوتر می‌خرند بدون این کمک هزینه نمی‌توانند یعنی کسانی که کمک هزینه دریافت نکردند هنوزهم ممکن است کامپیوتر داشته باشند. متغیر مجازی grant را به اینصورت تعریف می‌کنیم که اگر دانش‌آموز کمک هزینه دریافت کرده باشد برابر یک است و در غیر اینصورت صفر. سپس اگر grant بطور تصادفی اختصاص داده شود با جزءخطا ارتباط ندارد یعنی با درآمد خانواده و سایر عوامل اقتصادی و اجتماعی در جزءخطا ارتباطی ندارد. بعلاوه

grant باید با PC ناهمبسته باشد یعنی احتمال داشتن یک کامپیوتر بایدبطور معنادار برای دانش آموزانی که کمک هزینه را دریافت می‌کنند بالاتر باشد. اگر دانشجویان کم درآمد اولویت دانشگاه برای اعطای کمک هزینه باشند آنگاه grant با جزءخطا همبسته است و IV ناسازگار خواهد بود.

۲- فرض کنید می‌خواهید اثر حضور در کلاس را بر عملکرد دانش‌آموزان ارزیابی کنید. مدل به‌صورت زیر است:

$$\boldsymbol{stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + u}$$

$stndfnl$ نمره امتحان است $atndrte$ متغیر نرخ حضور در کلاس،

i. فرض کنید $dist$ فاصله‌ی محل زندگی دانشجویان تا کلاس درس باشد. آیا این متغیر با جزءخطا همبستگی دارد یا نه؟

به نظر میرسد همبستگی وجود ندارد چون کلاس‌های درس معمولا برای دانشجویان خاص تخصیص نمی‌یابد.

ii. فرض کنید $dist$ و u همبسته نباشند. چه فروض دیگری لازم است تا $dist$ یک متغیر ابزاری برای $tndrte$ باشد؟

متغیر $dist$ باید با $atndrte$ همبسته باشد.

iii. فرض کنید متغیر حاصلضرب $priGPA.antdrate$ را به مدل اضافه کنیم:

$$\boldsymbol{stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA.antdrate + u}$$

اگر $atndrte$ با u همبسته باشد آنگاه $priGPA.antdrate$ نیز اینچنین است. چه چیزی می‌تواند متغیر ابزاری خوبی برای $priGPA \cdot atndrte$ باشد؟ (راهنمایی: اگر $E\{u|priGPA, ACT, dist\} = 0$ باشد، $priGPA, ACT, dist$ برونزا هستند آنگاه هر تابعی از $priGPA$، ACT و $dist$ با u ناهمبسته است.)

به متغیر ابزاری برای priGPA·atndrte نیاز داریم (حتی اگر priGPA برونزا باشد atndrte برنزا نیست و بنابراین ضرب آن‌ها با جزءخطا همبسته است). با فرض برونزایی که $E\{u|priGPA, ACT, dist\} = 0$ هر تابعی از ACT، $priGPA$ و $dist$ نیز با u ناهمبسته است بطور خاص priGPA·dist نیز با جزءخطا ناهمبسته است. اگر dist با atndrte همبسته باشد آنگاه priGPA·dist با priGPA·atndrte همبسته است بنابراین می‌توانیم معادله زیر را تخمین بزنیم:

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA.antdrate + u$$

۳- مدل رگرسیون زیر را در نظر بگیرید

$$\boldsymbol{y = \beta_0 + \beta_1 x + u}$$

و z یک متغیر ابزاری برای x باشد. نشان دهید تخمین‌زن ابزاری $\widehat{\beta_1}$ به‌صورت زیر است:

$$\widehat{\boldsymbol{\beta_1}} = \frac{\overline{y_1} - \overline{y_0}}{\overline{x_1} - \overline{x_0}}$$

که $\overline{x_0}$ و $\overline{y_0}$ میانگین x_i و y_i است با $z_i = 0$.

$$\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x}) = \sum_{i=1}^{n} z_i\,(x_i - \bar{x})$$

$$and\ we\ have \sum_{i=1}^{n} z_i(y_i - \bar{y}) = \sum_{i=1}^{n} z_i y_i - \left(\sum_{i=1}^{n} z_i\right)\bar{y} = n_1\bar{y}_1 - n_1\bar{y}$$

$$where\ n_1 = \sum_{i=1}^{n} z_i\ \ and\ use\ the\ fact\ that \frac{(\sum_{i=1}^{n} z_i y_i)}{n_1} = \bar{y}_1$$

$$\bar{y} = \left(\frac{n_0}{n}\right)\bar{y}_0 + \left(\frac{n_1}{n}\right)\bar{y}_1 \quad where\ n_0 = n - n_1$$

$$\rightarrow \bar{y}_1 - \bar{y} = \left[\frac{n - n_1}{n}\right]\bar{y}_1 - \left(\frac{n_0}{n}\right)\bar{y}_0 = \left(\frac{n_0}{n}\right)(\bar{y}_1 - \bar{y}_0)$$

$$so\ we\ have \sum_{i=1}^{n} z_i(y_i - \bar{y}) = (\frac{n_0 n_1}{n})(\bar{y}_1 - \bar{y}_0)$$

$$and\ similarly\ we\ have \sum_{i=1}^{n} z_i(x_i - \bar{x}) = (\frac{n_0 n_1}{n})(\bar{x}_1 - \bar{x}_0)$$

$$\widehat{\beta_1} = \frac{(\bar{y}_1 - \bar{y}_0)}{(\bar{x}_1 - \bar{x}_0)}$$

۵- با توجه به رگرسیون‌های زیر فرض کنید $\sigma_u = \sigma_x$ بطوریکه تغییر جمعیت در جزآخطا مثل x است. فرض کنید متغیر ابزاری z کمی با u همبستگی دارد: $corr(z,u) = 0.1$ همچنین فرض کنید zوx با هم همبستگی داشته باشند $corr(z,x) = 0.2$.

$$plim\ \widehat{\beta_{1,IV}} = \beta_1 + \frac{corr(z,u)}{corr(z,x)} . \frac{\sigma_u}{\sigma_x}$$

$$\boldsymbol{plim\widehat{\beta_{1,OLS}} = \beta_1 + corr(x,u).\frac{\sigma_u}{\sigma_x}}$$

i. بایاس تقریبی در متغیر ابزاری چیست؟

$$plim\ \widehat{\beta_{1,IV}} = \beta_1 + \frac{corr(z,u)}{corr(z,x)} = \beta_1 + \left(\frac{0.1}{0.2}\right) = \beta_1 + 0.5$$

بنابراین اریب آن ۰.۵ است.

ii. چقدر همبستگی باید بین x و u وجود داشته باشد تا بایاس تقربی OLS بیشتر از 2sls باشد؟

$$plim\widehat{\beta_{1,OLS}} = \beta_1 + corr(x,u) > \beta_1 + 0.5 \rightarrow corr(x,u) > 0.5$$

بنابراین برای اینکه اریب **OLS** بیشتر از **2sls** باشد باید شرط بالا برقرار باشد.

۷- در زیر یک مدل ساده برای اندازه‌گیری تاثیر برنامه انتخاب مدرسه بر عملکرد آ زمون را می‌بینید:

$$\boldsymbol{score = \beta_0 + \beta_1 choice + \beta_2 faminc + u_1}$$

$score$، نمره آزمون، $choice$ متغیر دودویی است که نشان می‌دهد دانش آموز در انتخاب مدرسه مشارکت کرده و $faminc$ درآمد خانوار است. متغیر ابزاری برای $choice$، grant است که مقدار پولی است که به دانش آموزان برای شهرینه مدارس داده می‌شود. مبلغ کمک هزینه با سطح درآمد خانوار متفاوت است، برای همین $faminc$ را کنترل کردیم.

i. حتی با وجود $faminc$ در معادله، چرا $choice$ با جزءخطا همبسته است؟

در یک سطح مشخص درآمد، بعضی از دانش‌آموزان توانایی و انگیزه بیشتری دارند و خانواده آنها بیشتر از فرزندانشان حمایت می‌کنند. بنابراین مشکل self-selection وجود دارد یعنی دانش‌آموزانی که بهتر هستند با احتمال بیشتری در انتخاب مدرسه شرکت می‌کنند.

ii. اگر در هرطبقه درآمد مبلغ کمک هزینه بطور تصادفی اختصاص داده شود. آیا grant با جزآ خطا ناهمبسته است؟

از آنجایی که جزءخطا، درآمد را در بر ندارد، تصادفی اختصاص دادن کمک هزینه در طبقات درآمدی به این معناست که تعیین کمک هزینه با عوامل غیرقابل مشاهده مانند توانایی دانش‌آموزان، انگیزه و حمایت خانواده همبسته نیست.

iii. معادله فرم کاهش یافته را برای $choice$ بنویسید. چه چیز برای grant نیاز است که تاحدی با $choice$ همبسته باشد؟

فرم کاهش‌یافته برابراست با:

$$choice = \pi_0 + \pi_1 faminc + \pi_2 grant + \nu_2$$

برای اینکه $grant$ با $choice$ ناهمبسته باشد باید $\pi_2 \neq 0$ باشد. به عبارت دیگر باید $grant$ بر $choice$ اثر داشته باشد. این امر منطقی به نظر می‌رسد مشروط بر اینکه مبلغ کمک هزینه در هرطبقه درآمد متفاوت باشد.

iv. معادله فرم کاهش یافته را برای $score$ بنویسید. توضیح دهید چرا مفید است. (راهنمایی: چگونه ضریب grant را تفسیر می‌کنید؟)

فرم کاهش یافته برای $score$ به صورت زیر است:

$$score = \alpha_0 + \alpha_1 faminc + \alpha_2 grant + \nu_1$$

این معادله به ما اجازه می‌دهد تا مستقیما تاثیر افزایش کمک هزینه بر نمره آزمون را با ثابت نگه‌داشتن درآمد خانواده تخمین بزنیم.

۸- فرض کنید می‌خواهید تست کنید که آیا دخترانی که در دبیرستان‌های دخترانه شرکت می‌کنند نسبت به دخترانی که در دبیرستان‌های مختلط شرکت می‌کنند، در درس ریاضی بهتر هستند یا نه. یک نمونه تصادفی از دختران دبیرستانی در یک ایالت از امریکا در اختیار دارید و

نمرات آن‌ها نمرهدر یک آزمون ریاضی استاندارد است. girlhs یک متغیر دامی است که نشان می‌دهد دانش‌آموز در دبیرستان دخترانه شرکت می‌کند یا خیر.

i. چه عوامل دیگری را می‌توانید در این معادله کنترل کنید؟

عواملی مثل درآمد خانواده، سابقه خانوادگی و تحصیلی، تحصیلات پدر و مادر و .... .

ii. معادله مربوط به *score* و *girlhs* و سایر عواملی که در قسمت قبل ذکر کردید را بنویسید.

$$score = \beta_0 + \beta_1 girlhs + \beta_2 faminc + \beta_3 meduc + \beta_4 feduc + u_1$$

iii. فرض کنید حمایت و انگیزه دادن والدین از عوامل غیرقابل اندازه‌گیری در جزءخطا در بخش قبل هستند. آیا این ها با girlhs ارتباط دارند؟

والدینی که حامی هستند وبه دختران خود انگیزه می‌دهند، با احتمال بیشتری دختران خود را در مدرسه دخترانه ثبت‌نام می‌کنند. بنابراین به‌نظر می‌رسد girlhs با جزءخطا ارتباط دارد.

iv. فروض مورد نیاز برای اینکه تعداد دبیرستان‌های دخترانه در شعاع بیست مایلی خانه یک دختر، متغیر ابزاری معتبر برای girlhs باشد را بیان کنید.

برای اینکه متغیر ابزاری گفته شده معتبر باشد باید دو شرط زیر را برقرار کند:

۱-باید با جزء خطا ناهمبسته باشد و ۲- با girlhs همبسته باشد. شرط دوم به‌نظر می‌رسد برقر ار است و می‌توانیم با فرم کاهش‌یافته زیر آن را تست کرد:

$$\text{girlhs} = \pi_0 + \pi_1 faminc + \pi_2 meduc + \pi_3 feduc + \pi_4 numghs + v_2$$

که $numghs$ تعداد دبیرستان‌های دخترانه موجود در فاصله ۲۰ مایلی خانه یک دختر است.

اما برقراری شرط اول مشکل تر است. دبیرستان‌های دخترانه تمایل دارند مکان‌هایی را بیابند که تقاضا در آنجا بیشتر است و این تقاضا منعکس کننده‌ی جدیت افراد جامعه باشد. بعضی از مناطق بطور متوسط دانش‌آموزان بهتری دارد و این موضوع ربطی به درآمد خانواده و تحصیلات والدین ندارد و ممکن است به $numghs$ وابسته باشد.

v. فرض کنید زمانی که فرم کاهش یافته را برای girlhs تخمین می‌زنید، متوجه می‌شوید که ضریب numghs (تعداد دبیرستان‌های دخترانه در شعاع ۲۰ مایلی) منفی و از نظر آماری معنی‌دار است. آیا با تخمین iv که numghs متغیر ابزاری مورد استفاده برای girlhs است موافقید؟

هرچه تعداد دبیرستان‌های دخترانه در فاصله ۲۰ مایلی بیشتر باشد، دانش‌آموزان با احتمال کمتری در دبیرستان دختزانه شرکت می‌کنند. با این تفسیر به‌نظر می‌رسد متغیر ابزاری مناسبی نیست. چراکه به‌نظر می‌رسد متغیرهایی در فرم کاهش یافته وجود دارند که سبب برون‌زایی شدند اگر بتوانیم بگوییم این متغیرها

همان‌هایی هستند که در جزءخطا در معادله اصلی حضور دارند آنگاه شرط دوم نقض شده و متغیر ابزاری با جزءخطا همبسته است پس متغیر ابزاری مناسبی نیست.

۱۰- در یک مقاله اخیر، اثر حضور در یک مدرسه کاتولیک بر احتمال حضور در کالج بررسی شده‌است. college یک متغیر باینری است که برابر یک است اگر دانش آموز در کالج شرکت کند و در غیر اینصورت صفر است. CathHS یک متغیر باینری است که یک است اگر در مدرسه کاتولیک شرکت کند. مدل به صورت زیر است:

$$college = \beta_0 + \beta_1 CathHS + other\ factors + u$$

که $other\ factors$ شامل جنسیت، نژاد، درآمد خانواده و تحصیلات والدین است.

i. چرا $CathHS$ ممکن است با u همبسته باشد؟

دانش‌آموزان بهتر و جدی‌تر تمایل به رفتن به کالج دارند و همین نوع دانش‌آموزان ممکن است به دبیرستان‌های خصوصی و به‌طور دقیق‌تر کاتولیک جذب شوند. بنابراین در اینجا نیز مشکل self-selection داریم زیرا دانش‌آموزان خودشان مدارس کاتولیک را انتخاب می‌کنند به جای اینکه تصادفی به آنها اختصاص داده شوند.

ii. نویسندگان دیتای مربوط به نمره کسب‌شده را در هنگام تحصیل دارند. با استفاده از این متغیرها برای بهبود برآورد ceteris paribus از حضور در دبیرستان کاتولیک، چه می‌توان کرد؟

نمره معیاری از توانایی دانش‌آموزان است بنابراین می‌تواند به عنوان یک متغیر پراکسی در رگرسیون OLS استفاده شود. داشتن این معیار در رگرسیون OLS موجب بهبود می‌شود نسبت به وقتی که هیچ پراکسی برای توانایی دانش‌آموزان وجود نداشته باشد.

iii. CathRel یک متغیر دودویی است که برابر یک است اگر دانش‌آموز کاتولیک باشد. دو مورد از الزامات مورد نیاز برای اینکه در مدل قبلی، IV معتبری برای CathHS باشد را بحث کنید. کدام یک از این‌ها قابل آزمایش است؟

شرط اول اینست که CathRel با انگیزه و توانایی دانش‌آموز و سایر عوامل موجود در جزءخطا ناهمبسته باشد. این شرط زمانی برقرار است که بزرگ شدن به‌عنوان یک کاتولیک موجب بهتر شدن دانش‌آموز نشود (برعکس حضور در یک دبیرستان کاتولیک). به نظر می‌رسد این شرط برقرار است و می‌توان فرض کرد دانش‌آموزان کاتولیک توانایی‌های ذاتی بهتری نسبت به غیرکاتولیک‌ها ندارند. اینکه آیا کاتولیک بودن ارتباطی با انگیزه دانش‌آموزان یا آمادگی برای دبیرستان ندارد مساله جدی‌تر است. شرط دوم اینست که کاتولیک بودن (با کنترل

کردن سایر عوامل برونزا) در حضور در یک دبیرستان کاتولیک تاثیر داشته باشد. این شرط را می‌توانیم با فرم کاهش‌یافته زیر تست کنیم:

$$CathHS = \pi_0 + \pi_1 CathRel + other\ factor + v_1$$

iv. عجیب نیست که کاتولیک بودن تاثیر مثبتی در حضور در دبیرستان کاتولیک دارد. به نظر شما CathRel یک متغیر ابزاری قانع‌کننده برای CathHS است؟

به مقاله مراجعه شود. به‌نظر نمی‌رسد کاتولیک بودن از مسیر دیگری غیر از شرکت در مدرسه کاتولیک بر نمره افراد تاثیر بگذارد و همین که نمی‌توانیم برای نقض آن مثالی بزنیم کافیست که متغیر ابزاری مناسب است.