دانشگاه الزهرا – دانشکده علوم اجتماعی و اقتصاد

اقتصاد سنجی کارشناسی ارشد

استاد: دکتر صفرزاده

رگرسیون‌های چند متغیره

در جلسات گذشته دیدیم که معادلات نرمال برای رگرسیون‌های چند متغیره در شکل ماتریسی به صورت زیر بود:

$$(X'X)b = X'Y \qquad (1)$$

این معادلات نشان می‌دهد که برآورد کننده OLS برای بردار b با داده‌ها مرتبط است.

اگر بخواهیم معادلات نرمال را برای رگرسیون دو متغیره $(k = 2)$ به دست آوریم خواهیم داشت:

$$X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}$$

$$X'X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} n & \sum X \\ \sum X & \sum X^2 \end{pmatrix}$$

$$\& \quad X'Y = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum Y \\ \sum XY \end{pmatrix}$$

$$\begin{pmatrix} n & \sum X \\ \sum X & \sum X^2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} \sum Y \\ \sum XY \end{pmatrix} \Rightarrow \begin{cases} nb_1 + b_2 \sum X = \sum Y \\ b_1 \sum X + b_2 \sum X^2 = \sum XY \end{cases}$$

که این همان معادلات نرمال به دست آمده برای رگرسیون‌های دو متغیره در اقتصاد سنجی دوره کارشناسی است. طبیعتا برای رگرسیون‌های سه متغیره هم به همین روال معادلات نرمال به صورت زیر به دست خواهد آمد:

$$\begin{cases} nb_1 + b_2\sum X_2 + b_3\sum X_3 = \sum y \\ b_1\sum X_2 + b_2\sum X_2^2 + b_3\sum X_2X_3 = \sum X_2y \\ b_1\sum X_3 + b_2\sum X_2X_3 + b_3\sum X_3^2 = \sum X_3y \end{cases}$$

نتایج حاصل از معادلات نرمال

از رابطه (۱) داریم:

$$(X'X)b = X'y \;\; \& \;\; y = xb + e$$

$$(X'X)b = X'(Xb + e) = (X'X)b + X'e \Rightarrow X'e = 0 \quad (2)$$

این رابطه یکی از نتایج بنیادی حداقل مربعات است.

اولین مولفه در رابطه فوق به صورت $\sum e = 0$ خواهد بود بنابراین:

$$\bar{e} = \bar{y} - b_1 - b_2\bar{X}_2 - \cdots - b_k\bar{X}_k = 0$$

یعنی میانگین جملات خطا صفر بوده و صفحه رگرسیون از نقطه میانگین فضای k بعدی رد می‌شود.

مولفه‌های دیگر آن رابطه به صورت $i = 2, \ldots, k \quad \sum_t X_{it} e_t = 0$ خواهد بود. مفهوم این شرط آن است که هر کدام از جملات (متغیرهای) توضیحی مستقل از جمله خطا هستند. به عبارت دیگر همبستگی نمونه‌ای آنها با جمله خطا صفر است؛ بنابراین بردار رگرسیون $\hat{Y} = Xb$ نیز مستقل از e خواهد بود.

$$\hat{Y}'e = (Xb)'e = b'X'e = 0$$

تجزیه مجموع مجذورات

چون رگرسورها مستقل از جمله خطا هستند؛ امکان تجزیه مجموع مجذورات را ممکن می‌سازد. یعنی می‌توان بردار Y را به دو جزء توضیح داده‌شده و توضیح داده‌نشده (خطا) تجزیه کرد.

$$Y = \hat{Y} + e = Xb + e$$

$$Y'Y = \left(\hat{Y} + e\right)'\left(\hat{Y} + e\right) = \hat{Y}'\hat{Y} + e'e = b'X'Xb + e'e \quad (3)$$

در واقع $Y'Y$ همان مجموع مجذورات ارزش‌های واقعی Y است.

اما در رگرسیون ما بیشتر به دنبال تحلیل تغییرات در Y هستیم که در آمار و اقتصادسنجی به صورت مجموع مجذورات انحراف از میانگین نمونه‌ای ارائه می‌شود یعنی:

$$\sum (Y_t - \bar{Y})^2 = \sum Y_t^2 - n\bar{Y}^2$$

بنابراین برای تحلیل مجموع مجذورات تغییرات کل بهتر است از طرفین رابطه (3) عبارت $n\bar{Y}^2$ کم شود.

$$(Y'Y - n\bar{Y}^2) = (b'X'Xb - n\bar{Y}^2) + e'e \quad (4)$$

$$TSS = ESS + RSS$$

معادلات در شکل انحراف از میانگین

یک رویکرد جایگزین برای بیان رگرسیون‌های خطی بیان آن به صورت انحراف از میانگین‌های نمونه‌ای است.

فرض کنید:

$$Y_t = b_1 + b_2X_{2t} + b_3X_{3t} + \cdots + b_kX_{kt} + e_t \qquad t = 1, \dots, n \qquad (5)$$

میانگین نمونه‌ای برای این رگرسیون به صورت زیر است:

$$\bar{Y} = b_1 + b_2\bar{X}_2 + b_3\bar{X}_3 + \cdots + b_k\bar{X}_k \qquad \& \qquad \bar{e} = 0 \qquad (6)$$

اگر رابطه (6) را از (5) کم کنیم:

$$y_t = b_2x_{2t} + b_3x_{3t} + \cdots + b_kx_{kt} + e_t \qquad t = 1, \dots, n \qquad (7)$$

در واقع حروف کوچک در رابطه (7) بیانگر انحراف از میانگین نمونه‌ای هستند. در شکل انحراف از میانگین b_1 (عرض از مبدا) در ظاهر دیده نمی‌شود ولی می‌توان از رابطه زیر آن را بازیابی کرد:

$$b_1 = \bar{Y} - b_2\bar{X}_2 - \cdots - b_k\bar{X}_k$$

ضرایب به دست آمده برای شیب‌های رگرسیون‌های (5) و (7) از طریق روش حداقل مربعات معمولی یکسان خواهد بود همچنین جمله‌ی خطای دو رگرسیون هم یکسان خواهد بود.

رگرسیون در شکل انحراف از میانگین را به طور ساده‌تر می‌توان با استفاده از ماتریس زیر به دست آورد:

$$A = I_n - \left(\frac{1}{n}\right) ii' \qquad (8)$$

i یک بردار ستونی است که از n تا (1) تشکیل شده است. می‌توان نشان داد ماتریس A متقارن (symmetric) و هم‌قوه یا پوچ‌توان (Idempotent) است. (به عنوان تمرین این دو ویژگی را نشان دهید)

این ماتریس به هر بردار پیش ضرب شود مشاهدات آن بردار را به صورت انحراف از میانگین در می‌آورد.

بنابراین خواهیم داشت:

$$\begin{cases} Ae = e \\ Ai = 0 \end{cases} \qquad (9)$$

اگر معادلات حداقل مربعات را به صورت زیر بنویسیم:

$$Y = Xb + e = [i \quad X_2]\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + e$$

که در آن X_2 یک ماتریس $n * (k-1)$ از رگرسورهاست و b_2 نیز بردار ضرایب است که $(k-1)$ ضریب را در بر می‌گیرد. $b_1, b_2, \dots, b_k$

اگر رابطه رگرسیون بالا را در A پیش ضرب کنیم:

$$AY = [0 \quad AX_2]\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + Ae = AX_2 b_2 + e$$

$$Y^* = X_* b_2 + e \qquad (10)$$

رگرسیون به صورت انحراف از میانگین در می‌آید.

آنجا که $Xe = 0$ است، $X_*'e = 0$ خواهد بود. اگر رابطه (10) را در X_*' پیش ضرب کنیم، خواهیم داشت:

$$X_*'Y^* = (X_*'X_*)b_2$$

که همان شکل آشنای معادلات نرمال در رابطه (1) است. b یک بردار ستونی است که $(k-1)$ شیب را در بر می‌گیرد و عرض از مبدا را شامل نمی‌شود.

با استفاده از رابطه (10) تجزیه مجذورات را می‌توان به شکل زیر نوشت:

$$Y_*'Y_* = b_2'X_2'X_2b_2 + e'e \quad (11)$$

$$TSS = ESS + RSS$$

از روی این رابطه می‌توان ضریب تعیین را به دست آورد:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \qquad (12)$$

R^2 بیانگر بخشی از تغییرات کل در بردار Y که به وسیله ترکیب خطی رگرسورها توضیح داده می‌شود.

$$\bar{R}^2 = 1 - \frac{\frac{RSS}{n-k}}{\frac{TSS}{n-1}} \qquad (13)$$

در واقع صورت و مخرج رابطه (13) برآوردکننده‌های نااریبی از واریانس جمله خطا و متغیر Y است. می‌توان رابطه بین R^2 و $\bar{R}^2$ را به صورت زیر نشان داد:

$$\bar{R}^2 = 1 - \frac{\mathrm{n}-1}{\mathrm{n}-\mathrm{k}}(1-R^2) = \frac{1-k}{n-k} + \frac{n-1}{n-k}R^2 \qquad (14)$$

برای مقایسه برازش تصریح‌های مختلف با تعداد رگرسورهای متفاوت معمولا معیارهای شوارتز و آکائیک هم به صورت زیر مورد استفاده قرار می‌گیرد:

$Schwartz\ Criterion$:

$$Sc = \ln\frac{e'e}{n} + \frac{k}{n}\ln n$$

$Akaike\ information\ Criterion$:

$$AIc = \ln\frac{e'e}{n} + \frac{2k}{n}$$