دانشگاه الزهرا - دانشکده علوم اجتماعی و اقتصاد

اقتصادسنجی کارشناسی ارشد

(استنتاج آماری در معادلات k متغیره)

استاد: دکتر صفرزاده

اسفند ۹۸

استنتاج آماری در معادلات k متغیره

استنتاج آماری به مفروضاتی که تصریح روابط رگرسیون بر آن استوار شده است، بستگی دارد.

مفروضات

- فرض بر این است که ماتریس X غیر تصادفی و دارای رتبه کامل ستونی k است. استنتاج آماری مشروط به ارزش‌های نمونه‌ای متغیرهای ماتریس X انجام می‌شود. بنابراین مولفه‌های این ماتریس در نمونه‌های تکراری تغییر نمی‌کند. برای اینکه بردار ضرایب b به صورت یگانه بدست بیاید؛ لازم است ستون ماتریس X دارای استقلال خطی باشند.
- جملات اختلال دارای ویژگی‌های زیر است:

$$E(\varepsilon)=0 \ \& \ \mathrm{var}(\varepsilon)=E(\varepsilon\varepsilon')=\sigma^2 I$$

استنتاج آماری در معادلات k متغیره (ادامه)

به عبارت دیگر:

$$E(\varepsilon) = E\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ . \\ . \\ . \\ E(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ . \\ 0 \end{bmatrix} = 0$$

$$E(\varepsilon\varepsilon') = E\left(\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_n \end{bmatrix}\right) = \begin{pmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \dots & E(\varepsilon_1\varepsilon_n) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & \dots & E(\varepsilon_2\varepsilon_n) \\ . \\ . \\ . \\ E(\varepsilon_n\varepsilon_1) & E(\varepsilon_n\varepsilon_2) & \dots & E(\varepsilon_n^2) \end{pmatrix}$$

$$= \begin{pmatrix} \mathrm{var}(\varepsilon_1) & \mathrm{cov}(\varepsilon_1,\varepsilon_2) & \dots & \mathrm{cov}(\varepsilon_1,\varepsilon_n) \\ \mathrm{cov}(\varepsilon_2,\varepsilon_1) & \mathrm{var}(\varepsilon_2) & \dots & \mathrm{cov}(\varepsilon_2,\varepsilon_n) \\ . \\ . \\ . \\ \mathrm{cov}(\varepsilon_n,\varepsilon_1) & \mathrm{cov}(\varepsilon_n,\varepsilon_2) & \dots & \mathrm{var}(\varepsilon_n) \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ . \\ . \\ . \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 I$$

استنتاج آماری در معادلات k متغیره (ادامه)

ماتریس فوق، ماتریس واریانس- کوواریانس جملات اخلال است؛ که واریانس‌ها روی قطر اصلی و کوواریانس‌ها خارج آن قطر قرار دارند.

این ماتریس دو فرض قوی را در خود جای داده است.

فرض اول این است که واریانس جملات خطا ثابت است؛ این شرط تحت عنوان همسانی واریانس (Homoscedasticity) شناخته می‌شود. نقطه مقابل این فرض عدم تساوی واریانس‌ها است که تحت عنوان ناهمسانی واریانس (Heteroscedasticity) معرفی می‌شود.

فرض دوم عدم وجود همبستگی به صورت دودویی بین جملات اخلال است. در داده‌های مقطعی این فرض به معنی کوواریانس صفر بین جملات اخلال مقاطع مختلف و در داده‌های سری زمانی به معنی کوواریانس صفر بین جملات اخلال در دوره‌های زمانی مختلف است. وقتی این فرض نقض می‌شود، گفته می‌شود که جملات اخلال دچار خودهمبستگی یا همبستگی سریالی شده است.

میانگین b

برآورد کننده حداقل مربعات معمولی برای ضرایب b به صورت زیر بود:

$$b = (X'X)^{-1} X'Y$$

اگر به جای Y از تابع رگرسیون جامعه جایگذاری شود:

$$b = (X'X)^{-1} X'(X\beta + \varepsilon) = \beta + (X'X)^{-1} X'\varepsilon$$

از این رابطه می‌توان به رابطه زیر رسید.

$$b - \beta = (X'X)^{-1} X'\varepsilon$$

اگر از رابطه فوق ارزش انتظاری گرفته شود؛ عملگر وقفه از متغیرهای غیرتصادی رد شده و پشت جزء تصادفی قرار خواهد گرفت. از فرض اول کلاسیک رابطه زیر برای میانگین شرطی ضرایب بدست خواهد آمد.

$$E(b - \beta) = (X'X)^{-1} X'E(\varepsilon)$$

$$E(b) = \beta$$

واریانس b

واریانس ضرایب b هم به صورت زیر بدست خواهد آمد:

$$\operatorname{var}(b) = E\left[(b-\beta)(b-\beta)'\right]$$

$$E\left[(b-\beta)(b-\beta)'\right] = E[(X'X)^{-1}X'\varepsilon\varepsilon'X'(X'X)^{-1}]$$

$$= [(X'X)^{-1}X'E(\varepsilon\varepsilon')X'(X'X)^{-1}]$$

$$= \sigma^{2}(X'X)^{-1}$$

$$\operatorname{var}(b) = \sigma^{2}(X'X)^{-1}$$

برآورد واریانس جملات اخلال (σ^2)

از جلسات قبل به یاد داریم که جمله خطای رگرسیون به صورت زیر بدست می‌آید.

$$e = MY = M(X\beta + \varepsilon) \ \& \ MX = 0 \Rightarrow e = M\varepsilon$$

اگر فرم مربع جمله خطا را نوشته و امید بگیرم رابطه زیر بدست خواهد آمد:

$$E(e'e) = E(\varepsilon' M' M \varepsilon) = E(\varepsilon' M \varepsilon)$$

از آنجا که اثر (Trace) هر اسکالر، اسکالر خواهد بود؛ می‌توان نوشت:

$$\begin{aligned} E(\varepsilon' M \varepsilon) &= E(tr(\varepsilon' M \varepsilon)) = E(tr(\varepsilon'\varepsilon) M) \\ &= \sigma^2 tr(M) = \sigma^2 trI - \sigma^2 tr[X(X'X)^{-1}X'] \\ &= \sigma^2 trI - \sigma^2 tr[(X'X)^{-1}(X'X)] \\ &= \sigma^2 (n-k) \end{aligned}$$

بنابراین برآورد کننده واریانس جملات اخلال به صورت زیر خواهد بود:

$$s^2 = \frac{e'e}{n-k}$$

که برآوردکننده بدون تورشی از واریانس (σ^2) است. جذر عبارت فوق (s) انحراف معیار Y حول صفحه رگرسیون خواهد بود. این عبارت اغلب به عنوان خطای استاندارد برآورد (Standard error of estimate) یا خطای استاندارد رگرسیون (Standard error of regression) یا (SER) معرفی می‌شود.

قضیه گوس – مارکوف (Gauss-Markov Theorem)

این قضیه بنیادی‌ترین قضیه در رابطه با برآورد کننده‌های حداقل مربعات است. بر اساس این قضیه و به شرط برقراری مفروضات بیان شده، برآورد کننده‌های حداقل مربعات در میان تمامی برآورد کننده‌های خطی نااریب برای پارامتر جامعه، دارای حداقل واریانس بوده و به عبارت بهتر کاراترین هستند. (به عنوان تمرین این قضیه را ثابت کنید.)

این قضیه برای هر ترکیب خطی از این برآورد کننده‌ها هم صادق است.

نتایج قضیه گوس مارکوف را می‌توان به صورت زیر خلاصه کرد:

- ضریب برآورد شده به روش حداقل مربعات برای هر b_i، بهترین برآورد کننده خطی نااریب برای پارامتر متناظر جامعه‌ای آن یعنی β_i است.
- هر برآورد کننده خطی نااریبی از هر ترکیب خطی از پارامترهای β_i، همان ترکیب خطی b_i خواهد بود.
- بهترین برآورد کننده خطی نااریب از میانگین شرطی متغیر وابسته $(E(Y/X))$ به صورت زیر خواهد بود:

$$\hat{Y}_s = b_1 + b_2 X_{2s} + b_3 X_{3s} + \dots b_k X_{ks}$$

آزمون فرضیه‌های خطی (چند مثال)

نمونه‌های زیر را در مورد انواع فرضیه در نظر بگیرید:

- $H_0 : \beta_i = 0$. مفهوم این فرضیه این است که متغیر توضیحی X_i تاثیری روی Y ندارد. این فرضیه در مطالعات رگرسیون عمومیت دارد و معمولا تحت عنوان آزمون معنی‌داری شناخته می‌شود.
- $H_0 : \beta_i = \beta_{i0}$. که در آن β_{i0} مقداری مفروض تعریف شده است.
- $H_0 : \beta_2 + \beta_3 = 1$. این فرضیه معمولا در مورد توابع تولید استفاده می‌شود که در آن ضرایب بیانگر کشش نهاده‌های تولید هستند. در واقع فرضیه فوق دنبال آزمون کردن بازدهی ثابت نسبت به مقیاس است.
- $H_0 : \beta_3 = \beta_4 \; or \; \beta_3 - \beta_4 = 0$
- $H_0 : \begin{bmatrix} \beta_2 \\ \beta_3 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ . \\ 0 \end{bmatrix}$. این فرضیه بیان می‌کند که هیچ کدام از متغیرهای توضیحی روی متغیر وابسته اثرگذار نیستند.

به عبارت دیگر دنبال آزمون معنی‌داری کل رابطه تصریح شده است.

- $H_0 = \beta_2 = 0$. این فرضیه زمانی استفاده می‌شود که بردار پارامترها (β) به دو زیر بردار (β_1) و (β_2) افراز شده باشد. که طبیعتا هر کدام از این زیر بردارها به ترتیب k_1 و $k_2 (= k - k_1)$ ضریب را در بر می‌گیرند. این فرضیه بیان می‌کند که زیر مجموعه مشخصی از متغیرهای توضیحی هیچ نقشی در تعیین متغر وابسته ندارند.

شکل کلی آزمون فرضیه

تمام مثال‌های فوق را می‌توان در رابطه خطی کلی زیر خلاصه کرد:

$$R\beta = r$$

که در آن R یک ماتریس $q \times k$ از مقادیر ثابت معین و r نیز بردار ستونی $q \times ۱$ از مقادیر تعریف شده است.
q و k به ترتیب بیانگر تعداد پارامترهای رگرسیون و تعداد قیدهاست که همواره $q < k$ خواهد بود.

بیان ماتریسی مثال‌های ارائه شده

برای مثال‌های ارائه شده، این ماتریس‌ها به صورت زیر خواهد بود:

- $R=[0 \ldots 0\,1\,0 \ldots 0] \ \& \ r=0 \ \& \ q=1$ که عدد ۱ در جایگاه اِم قرار دارد.
- $R=[0 \ldots 0\,1\,0 \ldots 0] \ \& \ r=\beta_{i0} \ \& \ q=1$ که عدد ۱ در جایگاه اِم قرار دارد.
- $R=[0\,1\,1\,0 \ldots 0] \ \& \ r=1 \ \& \ q=1$.
- $R=[0\,0\,1\,-1\,0 \ldots 0] \ \& \ r=0 \ \& \ q=1$
- $R=[o \ \ I_{k-1}] \ \& \ r=o \ \& \ q=k-1$ که در آن o برداری است که $k-1$ صفر را در بر گرفته است.
- $R=[o_{k_2 \times k_1} \ \ I_{k_2}] \ \& \ r=o \ \& \ q=k_2$

رویه آزمون فرضیه‌های خطی

برای فهم بهتر، رویه آزمون را برای فرضیه خطی عمومی مرور کنیم:

$$H_0 = R\beta - r = 0$$

برای اجرای این آزمون لازم است برآوردکننده نمونه‌ای آن و توزیع این برآورد کننده مشخص شود. از آنجا که می‌دانیم برآورد کننده حداقل مربعات برای β، b است؛ بر این اساس برآوردکننده ترکیب خطی آن ($R\beta$) هم ترکیب خطی برآورد کننده‌اش (Rb) خواهد بود.

برای ارائه توزیع این برآوردکننده به میانگین و واریانس آن نیاز است. با توجه به میانگین و واریانس b، میانگین و واریانس ترکیب خطی آن (Rb) هم به صورت زیر خواهد بود:

$$E(Rb) = R\beta$$

$$\mathrm{var}(Rb) = E[R(b-\beta)(b-\beta)'R']$$

$$= R\,\mathrm{var}(b)R'$$

$$= \sigma^2 R(X'X)^{-1}R'$$

توزیع آزمون فرضیه‌های خطی

از آمار به یاد داریم که ترکیب خطی هر متغیر، توزیع متغیر را حفظ می‌کند. بنابراین Rbهمان توزیع b را خواهد داشت با میانگین و واریانس مشخص.

$$b \sim N[\beta, \sigma^2 (X'X^{-1})]$$

$$Rb \sim N[R\beta, \sigma^2 R(X'X^{-1})R']$$

پس می‌توان نوشت:

$$R(b-\beta) \sim N[0, \sigma^2 R(X'X^{-1})R']$$

و اگر فرضیه صفر برقرار باشد ($Rb = r$) خواهیم داشت:

$$(Rb-r) \sim N[0, \sigma^2 R(X'X^{-1})R']$$

از آمار کارشناسی به یاد داریم که اگر متغیر X به صورت زیر دارای توزیع نرمال باشد:

$$X \sim N[\mu, \Sigma]$$

در آن صورت عبارت زیر دارای توزیع χ^2 خواهد بود:

$$X'\Sigma^{-1}X \sim \chi^2_n$$

توزیع آزمون فرضیه‌های خطی (ادامه)

بر این اساس می‌توان توزیع زیر را تعریف کرد:

$$(Rb-r)'[\sigma^2 R(X'X^{-1})R']^{-1}(Rb-r) \sim \chi^2_q$$

اما اینجا مسئله‌ای که وجود دارد این است که واریانس جامعه (σ^2) مجهول است و باید برآورد شود. بر این اساس رابطه بالا دارای توزیع مشخصی نخواهد بود.

از اقتصادسنجی کارشناسی به یاد داریم که:

$$\frac{e'e}{\sigma^2} \sim \chi^2_{n-k}$$

(درستی رابطه فوق را به عنوان تمرین نشان دهید.) ضمن اینکه می‌دانیم این رابطه مستقل از b است. با استفاده از روابط فوق خواهیم داشت:

$$\frac{(Rb-r)'[R(X'X^{-1})R']^{-1}(Rb-r)/q}{e'e/(n-k)} \sim F_{q,n-k}$$

این رابطه را می‌توان به صورت زیر نیز بازنویسی کرد:

$$(Rb-r)'[s^2 R(X'X^{-1})R']^{-1}(Rb-r)/q \sim F_{q,n-k}$$