دانشور

تعیین حجم نمونه در مطالعات رگرسیون خطی معمولی

نویسندگان: دکتر انوشیروان کاظم‌نژاد* و سید مهدی سادات هاشمی**

* استادیار آمار زیستی دانشگاه تربیت مدرس

** دانشجوی دکترای آمار زیستی دانشگاه تربیت مدرس

چکیده

در تحقیقات گوناگون، نمونه‌گیری و تعیین حجم مناسب برای دستیابی به نتایج توانمند آماری از اهمیت ویژه‌ای برخوردار است. در حیطه کار با مدل‌های رگرسیون خطی معمولی مسأله مهم این است که برای بررسی رابطه معنادار بین یک متغیر وابسته با چندین متغیر مستقل باید به چه صورت عمل کرد؛ زیرا نمی‌توان با تعیین حجم نمونه بر مبنای رابطه متغیر وابسته با تنها یک متغیر مستقل، امید داشت که این حجم تعیین شده توان آزمون‌ها برای سایر متغیرها را نیز تضمین کند. بنابراین باید با تعدیل‌های مناسب بر روی حجم نمونه اولیه، شرایط مناسب را ایجاد کرد. در این مقاله سعی شده بدون پرداختن به مسائل استنباط آماری و مشکلاتی که در صورت کافی نبودنِ حجم نمونه پیش خواهد آمد، روش ساده‌ای برای تعیین حجم نمونه در مطالعات رگرسیونی چندگانه ارائه گردد.

واژه‌های کلیدی: رگرسیون چندگانه، ضریب همبستگی، حجم نمونه

دوماهنامه علمی ـ پژوهشی
دانشگاه شاهد
سال هشتم ـ شماره ۳۲
اردیبهشت ۱۳۸۰

مقدمه

در مطالعات گوناگون مواردی پیش می‌آیند که محقق به ارتباط بین دو یا چند متغیر علاقه‌مند است و همچنین تمایل دارد که در صورت وجود روابط معنادار بین متغیرها، مدلی ریاضی برای پیش‌بینی یک یا چند متغیر برحسب سایر آن‌ها به‌دست آورد. برای مثال فرض کنید یک پزشک علاقه‌مند است ارتباط معنادار بین وزن یک بیمار با تعداد هفته‌هایی که از یک رژیم غذایی خاص پیروی می‌کرده و تعداد کل ساعاتی که در طول این مدت به ورزش می‌پرداخته را نشان دهد و از طرف دیگر، تمایل دارد تا میزان و جهت این ارتباط را نیز تعیین و یک مدل ریاضی ارائه کند تا به‌وسیله آن، توانایی پیش‌بینی وزن بیماران را در آینده برحسب این دو متغیر داشته باشد.

در این‌گونه موارد به‌طور طبیعی با مدل‌های رگرسیونی که ساده‌ترین آن‌ها مدل رگرسیون خطی معمولی است، مواجهیم. به‌طورکلی در این مدل‌ها، هدف، پیش‌بینی یک

یا چند متغیر Y (وابسته) - که فرض می‌شود دارای توزیع نرمال و پیوسته‌اند - براساس یک یا چند متغیر تبیینی (مستقل) X1 ,....., Xp است. در این حال، معادله‌ای که Y و X را به‌صورت خطی به یکدیگر پیوند می‌دهد عبارت است از:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+....+\beta_pX_p+e \quad (1)$$

که در آن e متغیری است که خطاها را نشان می‌دهد. معمولاً فرض می‌شود خطاها دارای توزیع نرمال با میانگین صفر و واریانس σ_e^2 و مقادیرشان مستقل از یکدیگر هستند. اهدافی که محقق دنبال می‌کند هم ارز با معنادار بودن ضریب هر X_j (یعنی β_j) در مدل رگرسیون خطی ۱ است. به عبارت دیگر، معنادار بودن β_j نشان‌دهنده رابطه X_j با Y، $|\beta_j|$ بیانگر میزان تأثیر X_j بر Y و علامت β_j نشان‌دهنده جهت رابطه بین X_j و Y است.

در مطالعات آماری با توجه به اهداف مورد بررسی باید نمونه‌ای را با حجم لازم از جامعه استخراج کرد تا براساس آن بتوان با خطا و توان مناسب، استنباط‌هایی را درباره پارامترها انجام داد. از آن‌جا که بررسی‌های رگرسیون نیز در حیطه مطالعات آماری قرار دارد و به‌طور گسترده نیز در همه زمینه‌های علمی به‌کار گرفته می‌شود باید این موضوع در آن‌ها مورد توجه قرار بگیرد. در این مقاله سعی می‌کنیم در ابتدا فرمولی ساده که حجم نمونه را برای نشان دادن رابطه بین یک متغیر که می‌تواند وابسته باشد و یک متغیر دیگر که می‌تواند مستقل باشد ارائه دهیم. در مرحله بعد این فرمول را برای به‌دست آوردن حالاتی که در آن، هدف مطالعه نشان دادن روابط معنادار بین یک متغیر وابسته و چند متغیر مستقل است، تعمیم می‌دهیم.

۲- مواد و روش‌ها

الف) حالت رگرسیونی ساده یا بررسی ارتباط بین فقط دو متغیر

کار را با دو متغیر X و Y آغاز و فرض می‌کنیم که هدف، بررسی رابطه بین این دو و همچنین در صورت وجود ارتباط، پیش‌بینی Y برحسب X باشد. با این فرض که تغییرات Y برحسب X خطی است (مثلاً نمودارهای پراکنش نگار تجارب قبلی این را نشان داده باشند) هدف عبارت خواهد بود از برآورد کردن ضریب β_1 و آزمون کردن آن در معادله زیر:

$$Y=\beta_0+\beta_1X+e \quad (2)$$

در حالت کلی برای نشان دادن رابطه خطی بین دو متغیر که فرض می‌شود دارای توزیع نرمال توأم هستند، از ضریب همبستگی پیرسون (ρ) استفاده می‌شود که عبارت است از:

$$\rho_{xy}=\frac{COV(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} \quad (3)$$

و می‌توان نشان داد که رابطه‌ای بین β_1 و ρ_{xy} برقرار است:

$$\rho_{xy}=\frac{\beta_1\sigma_x}{\sigma_y} \quad (4)$$

که در آن $\sigma_x^2=Var(X)$ و $\sigma_y^2=Var(y)$ و بنابراین $\rho_{xy}=0$ اگر و فقط اگر $\beta_1=0$. چنانچه X و Y استاندارد شده باشند (یعنی $\mu_x=\mu_y=0$ و $\sigma_x=\sigma_y=1$) آنگاه آزمون فرض $\beta_1=0$ و $\rho_{xy}=0$ معادل یکدیگر می‌شوند و حجم‌های نمونه مورد نیاز برای این آزمون برابر یکدیگر خواهند بود [۱].

حال اگر $\rho_{xy}=r$ برآوردی از ضریب همبستگی بین X و Y باشد، فرمول حجم نمونه برای آزمون کردن فرض $H_1: \rho_{xy}=r$ مطابق رابطه زیر است [۲]:

$$n=\frac{\left(Z_{1-\frac{\alpha}{2}}+Z_{1-\beta}\right)^2}{[C(r)]^2}+3 \quad (5)$$

که در آن $C(r)=\frac{1}{2}\text{Ln}\frac{1+r}{1-r}$ که همان تبدیل فیشر، و α سطح معناداری آزمون و $1-\beta$ توان آن است.

ب) حالت رگرسیونی چندگانه یا بررسی ارتباط بین یک متغیر وابسته یا چند متغیر مستقل

در این مرحله فرض می‌کنیم که بیش از یک متغیر تبیینی (مستقل) داشته باشیم؛ یعنی هدف بررسی ارتباط بین متغیر Y و متغیرهای $X_1, X_2, ...X_p$ و پیش‌بینی Y با استفاده از مدل رگرسیونی آن برحسب Xها باشد. به

عبارت دیگر، فرض مورد علاقه در هر حالت عبارت است از تأثیر داشتن یک متغیر تبیینی بر روی متغیر وابسته در حضور دیگر متغیرهای تبیینی.

مجموعه فرض‌ها برای اُامین پارامتر ($1\le i\le P$) عبارتند از:

$$H_{i0}:[\beta_1,.....,\beta_{i-1},\beta_i,\beta_{i+1},.....,\beta_p]=[\beta_1,.....,\beta_{i-1},0,\beta_{i+1},.....,\beta_p]$$

$$H_{i1}:[\beta_1,.....,\beta_{i-1},\beta_i,\beta_{i+1},.....,\beta_p]=[\beta_1,.....,\beta_{i-1},\beta_i^*,\beta_{i+1},.....,\beta_p]$$

که در آن β_i^* مقداری معلوم و تعیین شده است. باید ببینیم که حجم نمونه در این حالت چه خواهد شد. برای روشن‌تر شدن حالت بالا فرض کنید که می‌خواهید تأثیر میزان دوز مصرفی یک داروی خاص (X_1) را بر پرفشار خونی (Y) نشان دهید و از طرف دیگر، نشان دادن رابطه بین پرفشار خونی با تری‌گلیسیرید (X_2) و کلسترول آن (X_3) نیز مورد نظر است. از مطالعات قبلی نیز چنین نتیجه‌گیری شده است که رابطه معناداری بین میزان دوز مصرفی این دارو و میزان تری‌گلیسیرید و کلسترول خون وجود دارد. در این‌جا با مسأله‌ای مشابه با آنچه در آزمون فرض‌های بالا ترتیب داده شد روبه‌رو هستیم؛ یعنی:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+e$$

به دلیل همبستگی بین Xها، حجم نمونه‌ای که از رابطه ۶ براساس فقط یک متغیر (مثلاً دوز مصرفی) به‌دست می‌آید و اگر با این تعداد مدل، برازش شود همه ضرایب رگرسیونی در سطح معناداری α لزوماً توان $1-\beta$ را به‌دست نخواهد داد [۳].

بنابراین باید تعدیلی را بر روی آن انجام داد تا اثر سایر متغیرها نیز در آن ظاهر شود.

بدون این‌که از کلیت مسأله کاسته شود β_1 را در نظر می‌گیریم و فرض می‌کنیم که b_1 برآورد کم‌ترین مربعات آن باشد. وایت مور (Whittemore) نشان داده که برای Xهای پیوسته و نرمال می‌توان واریانس b_1 ($Var_1(\beta_1)$) را از مدل رگرسیونی تک متغیره ۲ به‌دست آورد و سپس با استفاده از عامل تورم واریانس (Variance Infuation Factor=VIF)، که در رگرسیون چندگانه قابل محاسبه است (پیوست ۱) واریانس β_1 را در مجموعه چند متغیره P متشکل از متغیر تبیینی ($Var_p(b_1)$) به‌دست آورد. پس از آن، حجم نمونه اولیهٔ به‌دست آمده از رابطه ۵ برای بررسی رگرسیون چندگانه Y بر روی Xها به‌صورت زیر تبدیل می‌شود [۱]:

$$n_p=n\times VIF \qquad (۶$$

که در آن n و n_p به‌ترتیب عبارتند از حجم نمونه مورد نیاز برای مدل رگرسیونی با 1 و p متغیر تبیینی. مشکلی که باقی می‌ماند این است که چگونه VIF را محاسبه کنیم. ما نحوه این محاسبه را در پیوست ۱ آورده‌ایم.

۳- کاربرد

حال با مثالی کاربردی نحوهٔ استفاده از این فرمول را نشان می‌دهیم.

فرض کنیم پزشکی می‌خواهد رابطهٔ بین فشار خون بیماران و میزان کلسترول خون آن‌ها را بررسی کند. با سطح معناداری ۵ درصد و توان ۹۵ درصد، حجم نمونه لازم در هر یک از موارد زیر چه تعداد بیمار خواهد بود؟

الف) در صورتی‌که از مطالعات مشابه حدس زده شود که ضریب همبستگی پیرسون بین فشار خون و میزان کلسترول آن حدود ۰/۵ است.

ب) از مطالعات قبلی مشخص شده باشد که بین میزان کلسترول خون با دوز مصرفی یک نوع داروی ضد فشار خون و تری‌گلسیرید آن رابطهٔ رگرسیونی با ضریب تعیینی $R^2=0/6$ وجود دارد و هدف از مطالعه نیز به‌دست آوردن الگوی پیش‌بینی فشار خون برحسب دوز مصرفی دارو و میزان کلسترول خون و تری‌گلیسیرید آن باشد.

با در نظر گرفتن:

$$z_{1-\beta}=1/64,\ z_{1-\alpha/2}=1/96$$

در حالت اول با استفاده از رابطهٔ ۵، تعداد بیماران را حداقل ۴۶ نفر و در حالت دوم با استفاده از رابطهٔ ۶ تعداد بیماران را حداقل ۱۱۵ نفر به‌دست می‌آوریم. از آن‌جا که اساس استنباط‌هایی که در رگرسیون صورت می‌گیرد، آزمون t و F است، در صورتی‌که هدف، مدل‌بندی رگرسیونی فشار خون برحسب سایر متغیرها بوده، و حجم نمونه کفایت لازم را نداشته باشد، توان آزمون‌های t و F تحت تأثیر قرار خواهند گرفت.

بحث و نتیجه‌گیری

مطالبی که به بحث از آن‌ها پرداختیم از جمله مسائلی هستند که به وفور در تحقیقات گوناگون با آن‌ها مواجه می‌شویم. به دلیل اهمیتی که مدل‌های رگرسیونی در نشان دادن روابط معنادار بین متغیرها دارند و همچنین بنابر تفاوت فاحشی که بین دو حجم نمونه‌ای که به‌طور مثال ارائه شد وجود دارد، برای دستیابی به نتایج توانمند آماری باید بیش از پیش به ارتباطات خطی بین متغیرهای مستقل در تعیین حجم نمونه توجه گردد؛ زیرا در کلیه مطالعات آماری در سطح معناداری ثابت α با کاهش پیدا کردن حجم نمونه از توان آزمون‌ها کاسته خواهد شد؛ یعنی احتمال خطای نوع دوم افزایش می‌یابد. همچنین با اعمال تعدیل‌های مناسب می‌توان این روش را به حالتی که با متغیر پاسخ گسسته نیز سر و کار داریم تعمیم داد. البته در این خصوص به‌دلیل نبودن تعریف مشخص و استاندارد برای ضریب تعیین R^2 با انواع مختلفی از R^2ها مواجه هستیم که کمی مشکل‌ساز است؛ ولی به‌هرحال باز هم می‌توان عامل تورم واریانس را با تعاریف مختلف R^2 محاسبه و توان آزمون‌ها را براساس این R^2ها با استفاده از شبیه‌سازی‌های کامپیوتری برآورد کرد.

پیوست ۱

روش محاسبه VIF

۱- حالتی که مدل از حالت یک متغیر X_1 به حالت p متغیره $X_p,.....,X_1$ افزایش می‌یابد.

فرض کنیم متغیر $Var_1(b1)$ و $Var_p(b_1)$ به‌ترتیب واریانس برآورد پارامتر β_1 و β_p باشند که از مدل‌های رگرسیونی با یک و P متغیر تبیینی به‌دست آمده‌اند. در این صورت با اضافه شدن P-1 متغیر $X_p,...,X_2$ به مدل فقط شامل متغیر X_1، واریانس پارامتر برآورد شده $Var_1(b_1)$ تبدیل $Var_p(b_1)$ شده است. به عبارت دیگر، $Var_1(b_1)$ با عامل تورم واریانس VIF به‌صورت

$$Var_p(b_1)=Var_1(b_1)\times VIF \qquad (۷$$

متورم شده است. این رابطه از همخطی بین متغیرها نتیجه می‌گردد [۴و۵]. نشان می‌دهیم که در بیش‌تر مواقع که در آن

$$VIF=\frac{Var_p(b_1)}{Var_1(b_1)}\approx\frac{1}{1-R^2_{X_1|X_2.....X_P}} \qquad (۸$$

$R^2_{x_1|x_2.....x_p}$ ضریب تعیینی است که از رگرسیون X_1 بر مجموعه $X_p,.....X_2$ به‌دست می‌آید. یعنی معادله‌ای به‌صورت $X_1=\alpha_0+\alpha_1X_2+.....,+\alpha_pX_p+\varepsilon'$ داریم و معادله رگرسیونی تک متغیره به‌صورت ۲ است. برای سادگی فرض می‌کنیم $\mu_{x1}=0$ (در غیراین‌صورت با تبدیلات مناسب می‌توان این کار را انجام داد [۴]). واریانس برآورد کم‌ترین مربعات پارامتر β_1 (یعنی b_1) از رابطه زیر به‌دست می‌آید [۵]:

$$Var_1(b_1)=\frac{\sigma_1^2}{\Sigma x_1^2} \qquad (۹$$

حال فرض کنید که متغیر دیگر، یعنی X_2 با $\mu_{x2}=0$ نیز به مدل اضافه می‌شود. برآورد ماتریس واریانس کوواریانس پارامترهای برآورد شده به‌صورت زیر است:

$$Var(b_1,b_2)=\sigma_2^2\left(X^TX\right)^{-1}$$
$$=\sigma_2^2\begin{bmatrix}\Sigma X_1^2 & \Sigma X_1X_2\\ \Sigma X_1X_2 & \Sigma X_2^2\end{bmatrix}^{-1} \qquad (۱۰$$

که در آن X ماتریس متغیرهای تبیینی است ($[X_1\ X_2]$). از بین عناصر ماتریس ۲×۲ $Var(b_1, b_2)$ می‌توان واریانس b_1 را به‌صورت زیر به‌دست آورد [۱]:

$$Var_2(b_1)=\frac{\sigma_2^2\Sigma X_2^2}{\Sigma X_1^2\Sigma X_2^2-\left(\Sigma X_1X_2\right)^2}$$
$$=\frac{\left(\frac{\sigma_2^2}{\sigma_1^2}Var_1(b_1)\right)}{1-R^2_{X_1|X_2}} \qquad (۱۱$$

مقدار $\left(\frac{\sigma_2^2}{\sigma_1^2}\right)$ در بیش‌تر مواقع کوچک‌تر یا نزدیک به یک است و از آن‌جا که با اضافه شدن یک متغیر تبیینی به مدل، یک درجه آزادی از درجه آزادی مربوط به عبارت مجموع مربعات خطا کم می‌شود، ممکن است برآورد

واریانس $\left(\frac{\sigma_2^2}{\sigma_1^2}\right)$ گاهی اوقات نیز به‌طور مختصر از یک بیش‌تر شود.

لذا با تعمیم رابطهٔ ۱۱ به P متغیر تبیینی (به جای ۲ متغیر) رابطهٔ ۸ به‌دست خواهد آمد؛ یعنی:

$$Var_p(b_1,.....,b_p)=\sigma_p^2(X^TX)^{-1}=\sigma_p^2\Sigma \qquad (۱۲$$

که در آن Σ ماتریس واریانس - کوواریانس متغیرهای مستقل است.

با استفاده از این رابطه، اندرسون (Anderson) نشان داده است [۶] که اگر:

$$\Sigma_{11}^{-1}=\Sigma X_1^2(1-R^2_{X_1|X_2,.....X_p}) \qquad (۱۳$$

و از رابطه ۱۲ به‌دست می‌آید:

$$Var_p(b_1)=\sigma_p^2\Sigma_{11}=\frac{\sigma_p^2}{\Sigma_{11}^{-1}}$$

$$=\frac{\sigma_p^2}{\Sigma X_1^2(1-R^2_{X_1|X_2,.....X_p})}=$$

$$=\frac{\left(\frac{\sigma_p^2}{\sigma_1^2}\right)Var_1(b_1)}{1-R^2_{X_1|X_2,.....Xp}} \qquad (۱۴$$

باز هم در بیش‌تر حالات $\frac{\sigma_p^2}{\sigma_1^2}\le 1$.

$$Var_p(b_1)\le\frac{Var_1(b_1)}{1-R^2_{X_1|X_2,.....,X_p}} \qquad (۱۵$$

در نتیجه حد بالایی برای VIF عبارت است از:

$$\frac{1}{1-R^2_{X_1|X_2,.....,X_p}} \qquad (۱۶$$

وضعیت‌های بسیار نادری نیز پیش می‌آیند که در آن‌ها $\frac{\sigma_p^2}{\sigma_1^2}>1$؛ اما باز هم رابطه ۱۶ تقریب خوبی برای VIF است [۷].

۲- حالتی که مدل q متغیره به حالت p متغیره (p>q) افزایش می‌یابد.

در صورتی‌که نظیر حالت قبل عمل کنیم، واریانس برآورد (b_1) در مدل q متغیره $Var_q(b_1)$ و در مدل p متغیره $Var_p(b_1)$ توسط رابطه زیر با یکدیگر مرتبط خواهند بود:

$$\frac{Var_p(b_1)}{Var_q(b_1)}=\frac{\left(\frac{\sigma_p^2}{\sigma_q^2}\right)\left(1-R^2_{X_1|X_2,.....,X_p}\right)}{1-R^2_{X_1|X_2,.....X_q}}$$

$$\le\frac{1-R^2_{X_1|X_2,.....,X_p}}{1-R^2_{X_1|X_2,.....,X_q}}$$

$$=\frac{1}{R^2_{(X_1,X_{q+1},.....,X_p)|(X_2,X_3,.....,X_q)}} \qquad (۱۷$$

که در آن، ضریب تعیین چندگانه

(Multiple determination cofficient)

$R^2_{(X_1,X_{q+1},\dots,X_p)|(X_2,X_3,\dots,X_q)}$

میزان ارتباط خطی بین متغیرهای تبیینی $(X_1,X_{q+1},...,X_p)$ و $(X_2,...,X_q)$ را که از رگرسیون چندمتغیره $(X_p,...,X_{q+1},X_1)$ بر روی $(X_q,...,X_2)$ به‌دست می‌آید، تعیین می‌کند.

در این‌جا $\left(\frac{\sigma_p^2}{\sigma_q^2}\right)$ بیش‌تر از $\left(\frac{\sigma_p^2}{\sigma_1^2}\right)$ به مقدار یک نزدیک است.

منابع

1. **Whittemore, A.: Sample size for Logistic Regression with small Response probability, Journal of the American Statistical Association, 1981; 76, 27-32.**
2. **Sokal, R.R. & Rohlf, F. J. Biometry, Freeman & Company, New York, 1989; 56.**
3. **Liu. Gand, Liany. K.Y., Sample size Calculation for studies with Correlated Observations, Biometrics. 1997; 53, 537-547.**
4. **Neter, John. & Authors, Applied Linear Statistical Models, IRwin, USA. 1988.**

5. Weisberg, Sanford. Applied Linear Regression, Wiley, New York, 1988.
6. Anderson, T.W. An Introduction to Multivariate Statistical Analysis, Wiley, New York, 32, 1958.
7. Guenther, W. C. Sample size formulas for normal theory t tests american statistician, 1981; 35, 243-244.