نامه به سردبیر

رگرسیون خطی، نرمال بودن توزیع مقادیر خطا یا نرمال بودن توزیع متغیر وابسته؟

رضا بهنام‌فر*، اعظم راستی

مجله ایرانی آموزش در علوم پزشکی/ ۱۳۹٤، ۱۵(۳۳): ۲۶۳ تا ۲۶۵

سردبیر محترم مجله ایرانی آموزش در علوم پزشکی

یکی از پرکاربردترین روش‌های آماری برای تجزیه و تحلیل داده‌ها در علوم مختلف، رگرسیون خطی ساده یا چندگانه است. در تحلیل رگرسیون نوع روابط متغیرها و این که آیا یک متغیر می‌تواند در متغیر دیگر تأثیرگذار باشد یا خیر، بررسی می‌شود(۱). به عبارتی چنین بیان شده که "کاربرد اصلی رگرسیون خطی، تعیین عوامل مؤثر بر یک متغیر عددی است که توزیع نرمال دارد"(۲). برای استفاده از این روش آماری، پیش فرض‌هایی ذکر گردیده است: ۱. خطی بودن رابطه متغیرهای مستقل و وابسته ۲. نرمال بودن توزیع مقادیر خطا ۳. استقلال مقادیر خطاها و ٤. نرمال بودن توزیع متغیر وابسته(۱تا۳).

مسأله چالش برانگیز، پیش فرض نرمال بودن است. سؤال این است که در واقع کدام یک باید به عنوان "پیش‌فرض اولیه" استفاده از رگرسیون خطی مد نظر قرار گیرد: نرمال بودن توزیع متغیر وابسته یا نرمال بودن توزیع مقادیر خطا؟

همان‌گونه که عنوان شد، در بعضی از منابع، نرمال بودن توزیع "متغیر وابسته"به عنوان پیش شرط استفاده از رگرسیون خطی بیان شده است. اما، کیانی(۱) نرمال بودن توزیع متغیر وابسته را "شرط لازم" برای استفاده از رگرسیون خطی ندانسته و نرمال بودن توزیع مقادیر خطا را مد نظر دانسته است. به نظر می‌رسد این تحلیل به واقعیت نزدیک‌تر باشد. در منابع دیگر نیز به نرمال بودن توزیع مقادیر خطا به عنوان یکی از پیش شرط‌های"اساسی" استفاده از رگرسیون خطی اشاره گردیده و همگی موافق هستند که "در صورت عدم برقراری این پیش‌گزیده، نمی‌توان از رگرسیون استفاده نمود"(۳).

اما بحث نرمال بودن توزیع متغیر وابسته را چگونه می‌توان تحلیل نمود؟ بار دیگر باید تأکید نمود که نرمال بودن توزیع مقادیر خطا، شرط اولیه (در کنار استقلال خطاها و هم خط نبودن متغیرهای مستقل) برای استفاده از رگرسیون خطی ساده یا چندگانه است. نرمال بودن توزیع متغیر به عنوان یک شرط ثانویه و در زمان نرمال نبودن توزیع مقادیر خطا مطرح می‌شود و هدف از طرح آن، تلاش برای دستیابی به توزیع نرمال مقادیر خطا است. کما این که چنین ذکر شده است که: "در صورتی مقادیر خطا توزیع نرمال نداشته باشند، آنگاه ممکن است انجام تبدیل در مورد متغیر وابسته با روش‌های سنتی و یا روش باکس-کاکس بتواند این مشکل را حل نماید"(۱).

* **نویسنده مسؤول:** رضا بهنام‌فر، دانشجوی دکترای مدیریت آموزشی، کارشناس مرکز مطالعات و توسعه آموزش علوم پزشکی، دانشگاه علوم پزشکی شهید صدوقی یزد، یزد، ایران.
reza82br@yahoo.com
اعظم راستی، کارشناس ارشد ژنتیک انسانی، دانشگاه علوم پزشکی شهید صدوقی یزد، یزد، ایران. (rasti_azam@yahoo.com)

تاریخ دریافت: ۹۴/۴/۱۶، تاریخ پذیرش: ۹۴/۴/۳۱

همان‌گونه که مشخص است، در اینجا از عبارات "ممکن" و "متغیر وابسته" استفاده شده است. به این ترتیب ممکن است حتی با وجود نرمال بودن توزیع متغیر وابسته (چه از ابتدا و چه از طریق استفاده از تبدیل) امکان استفاده از رگرسیون خطی (به واسطه نبود یکی از سه شرط نرمال بودن توزیع مقادیر خطا، نبود هم خطی بین متغیرهای مستقل و استقلال خطاها)فراهم نباشد. بنابراین برای استفاده از رگرسیون خطی باید حتما توزیع مقادیر خطا نرمال باشد. اگر این پیش شرط برقرار نباشد و متغیر وابسته از توزیع نرمال برخوردار باشد، شانس استفاده از این روش آماری کاملاً از بین می‌رود زیرا دیگر امکان استفاده از تبدیل‌ها وجود ندارد. در صورت نرمال نبودن توزیع متغیر وابسته، این شانس هنوز وجود دارد که با نرمال کردن آن، احتمال نرمال شدن توزیع مقادیر خطا نیز پدید آید و بتوانیم از رگرسیون خطی استفاده نماییم. البته در صورتی که تبدیل‌های مختلف موفق به نرمال کردن توزیع متغیر وابسته شوند، باز هم تضمینی برای نرمال بودن مقادیر خطا و امکان استفاده از روش آماری مدنظر وجود ندارد. به این ترتیب می‌توان گفت که نرمال بودن توزیع متغیر وابسته، یک شرط اولیه نیست و صرفا می‌تواند به عنوان یک شرط ثانویه و با هدف ایجاد یک شانس مجدد (با فرایند یاد شده) مدنظر قرار داشته باشد.

در واقع به نظر می‌رسد بیان شرط نرمال بودن توزیع متغیر وابسته برای افزایش شانس نرمال بودن توزیع مقادیر خطا باشد. هر چند که بیان آن به شکل "مطلق" باعث می‌شود تا پژوهشگران کمتر آشنا به مباحث آماری (در صورتی که تبدیل‌ها هم به آنها کمکی نکند) از رگرسیون خطی صرف نظر نموده و از روش‌های دیگری استفاده نمایند. در حالی که می‌توانستند با بررسی سه پیش شرط اصلی و در صورت برقراری آنها (حتی با وجودتوزیع غیرنرمال متغیر وابسته) از رگرسیون خطی استفاده نمایند.

متأسفانه بعضی از اساتید آمار و اپیدمیولوژی نیز نرمال بودن توزیع متغیر وابسته را شرط لازم برای استفاده از رگرسیون خطی و مدل سازی از این طریق می‌دانند؛ حال آن که همان طور که توضیح داده شد، این یک برداشت اشتباه و گمراه کننده است و بحث نرمال بودن، صرفا برای توزیع مقادیر خطا "لازم" است.

در اینجا ذکر دو نکته کوتاه ولی مهم دیگر برای محققین عزیز که تمایل به استفاده از رگرسیون خطی و مدل یابی از این طریق را دارند، خالی از لطف نیست.

اول این که، در سراسر این نوشته به نرمال بودن توزیع متغیرهای مستقل اشاره نشد. زیرا این امر، پیش شرط و لازمه رگرسیون خطی نیست.

دوم این که، باید توجه داشت که بین رگرسیون خطی چندگانه و رگرسیون چند متغیره تفاوت وجود دارد. حال آن که به اشتباه در بسیاری از کتب و مقالات به جای استفاده از رگرسیون خطی چندگانه از رگرسیون خطی چند متغیره استفاده می‌شود. "در بحث تخصصی، موقعی از رگرسیون چندمتغیره صحبت می‌کنیم که چند متغیر وابسته داشته باشیم. به عبارت دیگر می‌خواهیم بین یک یا چند متغیر مستقل با چند متغیر وابسته رابطه‌ای توأم برقرار کنیم"(٤). در حالی که در رگرسیون خطی چندگانه، تأثیر یا رابطه چند متغیر مستقل و یک متغیر وابسته بررسی می‌شود.

برای بررسی استقلال خطاها از آزمون دوربین واتسون استفاده می‌گردد. چنانچه مقدار آن در بازه ۱.۵ تا ۲.۵ قرار بگیرد به معنای عدم همبستگی بین خطاها است(۳). برای بررسی هم خطی (که نشان‌دهنده آن است که یک متغیر مستقل تابعی خطی از سایر متغیرهای مستقل است)، می‌توان عامل تورم واریانس و تولرانس را محاسبه نمود. به عنوان یک قاعده کلی، تولرانس کم‌تر از ۰/۱ و عامل تورم واریانس بزرگ‌تر از ۱۰ نشان‌دهنده مشکل ساز بودن هم خطی هستند(۵).

به طور خلاصه، استفاده از رگرسیون خطی منوط به نرمال بودن توزیع خطا است. در صورتی که توزیع مقادیر خطا

نرمال نباشد، حتی با وجود نرمال بودن توزیع متغیر وابسته، امکان استفاده از رگرسیون خطی وجود ندارد. زمانی که هم توزیع مقادیر خطا و هم توزیع متغیر وابسته نرمال نباشد، با استفاده از تبدیل‌های مختلف برای توزیع متغیر وابسته، سعی در ایجاد شانس برای نرمال کردن توزیع مقادیر خطا داریم. در واقع در این شرایط، هدف اصلی از نرمال کردن توزیع متغیر وابسته، نرمال کردن توزیع خطا است.

در پایان نویسندگان از دریافت نظرات صاحب‌نظران در این زمینه استقبال نموده و امیدوارند تا این نوشتار کوتاه و نظرات احتمالی سایر نویسندگان در روشن شدن نکات مبهم استفاده از رگرسیون خطی گره گشا باشند. به هرحال، تفاسیر مبهم یا نادرست سبب می‌شوند تا طیف گسترده‌ای از پژوهش‌گران نتوانند از روش‌های آماری موردنظر خود استفاده نمایند.

منابع

1. Kiani B. [Applying Modern Statistics in Natural Resources]. Yazd: Yazd University; 2014.[Persian]
2. Chehrei A, Haghdoost AA, Fereshtehnejad M, Bayat A. [Statistical Methods in Medical Science Researches Using SPSS Software]. 2nded. Tehran: PejvakeElme Arya; 2011. [Persian]
3. Momeni M, Ghayoumi AF. [Statistical Analysis with SPSS]. 6thed. Tehran: Mo'alef; 2012.[Persian]
4. Wagheie Y, Chahkandi M. [kankashidarregreseyonechandmotagayere]. Students Statistical Journal (NEDA). 2007; 4(2): 28-34. [Persian]
5. Toghraei Z. [Barrasiyeasratehamkhatidarmodelhayeregresiyonechandgane].Students Statistical Journal (NEDA). 2007; 5(1): 31-39.[Persian]