تعیین حجم نمونه در مطالعات فاقد تخصیص تصادفی بقا با داده‌های بدون سانسور و دارای سانسور

دکتر سقراط فقیه زاده[1]، مهدی رهگذر

چکیده مقاله

مقدمه. در پژوهشهایی که ازروشهای آنالیز بقاء بهره گرفته می‌شود، تعیین حجم نمونه کافی برای دستیابی به توان آماری مناسب دارای اهمیت است. درروشهای پارامتری وناپارامتری آمار کلاسیک تخصیص تصادفی نمونه هااز شرایط اساسی محسوب می‌شود. تحقق چنین شرطی دربسیاری ازپژوهشهای کارآزمایی بالینی و بهداشتی امکان پذیرنیست از آنجاکه مطالعات مشاهده‌ای بقاء نمی‌توان افراد را بطورتصادفی به تیمارها منتسب کردلذا می‌توان ازرگرسیون خطی شامل متغیر مستقل بااثرثابت درتحلیل استفاده نمود، نوشتار فوق حاصل تعیین حجم نمونه دراینگونه مطالعات آنالیز بقاءباداده‌های فاقد سانسور ودارای سانسور می‌باشد.

روشها. در مطالعات فاقد تخصیص تصادفی بقاء، می‌توان از رگرسیون خطی که شامل متغیر دارای اثرثابت است، استفاده نمود. چنین رگرسیونی در واقع امید ریاضی متغیر وابسته به شرط متغیر مستقل را بیان می‌کند. با در نظر گرفتن متغیر دوجمله‌ای برای تعیین عضویت فرد دریکی ازدوگروه مورد مقایسه و منظور کردن توزیع نمایی برای تابع مخاطره در رگرسیون بقاء، تابع درستنمایی تشکیل گردید. باتوجه بااینکه توزیع مجانبی برآورد گرپارامتر درتابع درستنمایی نرمال می‌باشد؛ فرمول واریانس برآوردگرضریب رگرسیون بدست آمد. سپس بابیان واریانس ضریب رگرسیون متغیر دارای اثرثابت برحسب ضریب تعیین رگرسیون، فرمولهایی برای تعیین حجم نمونه هم در داده‌های فاقد سانسور وهم دارای سانسور بقاء فراهم شده است.

نتایج. برای مقایسه تابع مخاطره مربوط به دوگروه که فاقد سانسوردر داده‌ها میباشند، برآوردی ازضریب تعیین رگرسیون بقا، نسبت مخاطره و نسبت عضویت درگروه و واریانس‌های آنهااز تابع درستنمایی بدست آمد. درصورتیکه داده‌ها دارای سانسورباشندعلاوه برعناصر فوق برآوردی از احتمال سانسورباید مدنظرقرارگیرد. بایافتن واریانس برآوردگر ماکزیمم درستنمایی وباتوجه به توزیع مجانبی نرمال آن وبه کمک ضریب تعیین رگرسیون فرمولهایی برای تعیین حجم نمونه بدست آمده است. حجم نمونه تعیین شده توان لازم جهت آزمونی که برای اثر سایر متغیرهای مستقل تطبیق یافته است رادربردارد.

بحث. کاربرد مدل رگرسیونی درمطالعات فاقد تخصیص تصادفی بقاء موجب شده است که فرمولهای مناسبی برای برآورد حجم نمونه، هم درمطالعات فاقد سانسور و هم درمطالعات دارای سانسور فراهم شودکه درسطح اشتباه α دسترسی به توان آماری لازم نیز ممکن می‌گردد. درمدل‌های نیمه پارامتری متناسب کاکس که واریانس پارامتر به حالت ساده‌تری تبدیل نمی‌شود باید ازیک مدل شبیه سازی استفاده شود، وقتی ضریب تعیین رگرسیون نسبتاً بزرگ باشد توانی که براساس آزمون‌لگ - رتبه‌ای تطبیق نشده محاسبه می‌شود، توان رابیش ازمقدار واقعی برآوردمی کند ولی وقتی ضریب تعیین نزدیک به صفرباشد تفاوت درتوان‌ها کمتر می‌شود. باافزایش ضریب تعیین رگرسیون، تفاوت بین توانهای حاصل از آزمون لگ - رتبه‌ای وضریب تعیین تطبیق شده باروش مورد نظر افزایش می‌یابد.

● *واژه‌های کلیدی: آنالیز بقاء، تعیین حجم نمونه ،عدم تخصیص تصادفی، داده‌های فاقد سانسور، داده‌های دارای سانسور.*

مقدمه

در مطالعات آماری وتحلیل‌های بقاء باتوجه به اهداف موردبررسی لازم است که نمونه‌ای رابا حجم لازم انتخاب نمائیم تابراساس آن بتوان در سطح اشتباه معین و توان مناسب استنباطهایی رادرمورد پارامترهای مورد نظر به عمل آوریم .تخصیص تصادفی افراد به تیمارها، لازمه بسیاری از روش‌ها و تحلیل‌های آماری ازجمله درآنالیزبقاء است. امامطالعاتی نیز وجود دارند که تخصیص تصادفی امکان پذیرنمی باشد.فرض کنید دریک مرکز آموزشی بخواهیم زمان تاوقوع سرطان ریه را، دردوگروه سیگاری وغیر سیگاری‌ها مقایسه کنیم، دراینجا امکان تخصیص تصادفی افراد به دو گروه فوق امکان پذیرنیست یادرمقایسه روند آخرین باروری زنان شهری بازنان روستائی، تخصیص زنان به دوگروه شهری وروستایی به صورت تصادفی صورت نمی‌پذیرد. دراین مثال ممکن است زمان تاوقوع سرطان ریه به عوامل دیگری غیرازکشیدن سیگارنیزبستگی داشته باشد. بنابراین در رگرسیون بقاء رابطه متغیر موردنظر Z که تخصیص به گروهها را تعیین می‌کندو دارای اثر ثابت است، بامتغیرهای کمکی (Covariates) W و متغیرتصادفی زمان بقاء، T نه تنها درمطالعه بلکه دربرآورد حجم نمونه نیز باید مدنظر قرارگیرد.

جهت برآورد حجم نمونه برای مطالعات تصادفی درآنالیز بقاء روشهایی پیشنهاد شده است از جمله توسط فریدمن (Freedman) (۱۹۸۲) (۱)، شوئن فلد (Schoenfeld) (۱۹۸۳) (۲)، لاچین و فولکس (Lachin and Foulkes) (۱۹۸۶) (۳)، لاگاتوس ولن (Lagatos and Lan) (۱۹۹۲) (۴). برای مطالعات فاقد تخصیص تصادفی آنالیز بقاء، فرمولی برای برآورد

۱- دانشیار گروه آمارزیستی ،دانشگاه تربیت مدرس

حجم نمونه توسط برناردو و همکاران (M.V.patricia Bernardo) (۲۰۰۰) ارائه شده (۵) و درآن فرض شده است که توزیع زمان تا وقوع حادثه یک توزیع نمایی است. آنها به تبعیت ازروش رگرسیونی که توسط لیپزیتس و پارزن (Lipsitz and parzen) (۱۹۹۵) (۶) برای مطالعات غیر مشاهده‌ای ارائه شده ازرگرسیون نمایی جهت مدل سازی مخاطره Zنسبت به W استفاده نموده‌اند. معمولاً درمقایسه‌های دارای تخصیص تصادفی بقاء توان آزمون براساس فرمول آزمون لگ -رتبه‌ای (Log-rank) محاسبه می‌شود. برناردو وهمکاران (۲۰۰۰) (۵) نشان داده‌اند که توان بدست آمده براساس فرمول آزمون لگ- رتبه‌ای، بیش ازمقدار واقعی آن برآورد می‌شود. وایت مر (Whittemore) (۱۹۸۱) (۷) مساله مشابه‌ای رابرای توزیع دوجمله‌ای وسیگنورینی (Signorini) (۱۹۹۱) (۸) برای داده‌های شمارشی موردمطالعه قرار داده‌اند. دراین نوشتار بااستفاده از روش برنارد و همکاران (۲۰۰۰) (۵) ولاپیزیتس و پارزن (۱۹۹۵) (۶) فرمول‌های تعیین حجم نمونه در مطالعات آنالیز بقاء باداده‌های فاقد سانسوروودارای سانسور و کاربردی ازآنها ارائه می‌شود.

روش‌ها

لیپزیتس و پارزن (۱۹۹۵) محاسبه حجم نمونه را درمطالعات غیر تصادفی مورد بحث قراردادەبیان نموده‌اند که درمقایسه دوتیمار وقتی افراد به صورت تصادفی به دوتیمار تخصیص داده نمی‌شوند، می‌توان ازرگرسیون چندگانه متغیر وابسته به شرط متغیر کمکی استفاده نمود (۶). برناردو و همکاران (۲۰۰۰) این روش را به آنالیز رگرسیون بقاء، تعمیم داده فرمولهایی را اثبات نموده‌اند (۵).

زمینه موضوع درآنالیز بقاء

فرض کنید $i=1,\ldots,n, T_i$ یک متغیر تصادفی نامنفی باشد که برای n فرد مورد نظر، زمان راازلحظه ورود درمطالعه تاوقوع حادثه مورد نظر بیان می‌نماید و C_i متغیر تصادفی مبین سانسور شدن باشد که توزیع آن دلخواه است و X_i به صورت $\min (C_i, T_i)$ در نظر گرفته شود. متغیر مورد نظر یعنی Z_i، یک متغیر کمکی دوجمله‌ای است که عضویت در گروه را در یک مطالعه غیر تصادفی مشخص می‌کند. اگر فرد i در گروه ۱ باشد مقدار آن به صورت $z_i=1$ و اگر فردi درگروه ۲ قرار گیرد، $z_i=0$ است. سایر متغیرهای کمکی بانام W_i برای هرفرد به صورت بردار منظور و درپایه (base-line) قرار می‌گیرند.

فرض می‌کنیم که T_iبه صورت شرطی مستقل از C_iبه شرط Z_iو W_i باشد. هدف اصلی مقایسه تابع مخاطره مربوط دوگروه است، دررگرسیون نمایی، تابع مخاطره فردi ام به صورت زیراست:

$$exp(\hat{\beta}+\beta_1 Z_i+\beta_2' W_i)$$

بنابراین توان آزمونی به صورت $H_0:\beta_1=0$ بستگی به واریانس $\hat{\beta}_1$ خواهد داشت. با استفاده از نظریه درستنمایی استاندارد و برای مقادیر به اندازه کافی بزرگ n، $\hat{\beta}_1$ دارای توزیع تقریباً نرمال با میانگین β_1 و واریانس $var(\hat{\beta}_1)$ می‌باشد. آزمونی باسطح اشتباه α به ازای مقادیر زیر، فرض H_0 رارد می‌کند:

$$\left|\hat{\beta}_1\right| > Z_{1-\alpha/2}\sqrt{var(\hat{\beta}_1)}$$

که درآن $Z_{1-\alpha}$ چندک $100(1-\alpha)\%$ توزیع نرمال استاندارد است. برای داشتن توان $\gamma - 1$ تحت فرض

$$pr\left\{|\beta_1| > Z_{1-\alpha/2}\sqrt{var(\hat{\beta})}\right\}=1-\gamma \quad ,H_a:\beta_1\neq 0$$

یعنی:

واریانس مجانبی $\hat{\beta}_1$ می‌شود:

$$var(\hat{\beta}_1)=\left[\frac{\beta_1}{Z_{1-\alpha/2}+Z_{1-\gamma}}\right]^2 \qquad (1)$$

که درآن $Z_{1-\gamma}$ چندک $100(1-\gamma)\%$ توزیع نرمال استانداردمی باشد. توجه شود که $var(\hat{\beta}_1)$بستگی به تعداد افراد درمطالعه یعنیn دارد ولی این وابستگی درمعادله (۱) محوشده است. هم برای داده‌های دارای سانسور وهم درداده‌های فاقد سانسور نشان داده می‌شودکه $var(\hat{\beta}_1)$ می‌تواند به صورت تابعی ازضریب تعیین رگرسیون خطی Zبرحسب W، $R^2_{z*/w*}$ بیان شود. بامساوی قراردادن واریانس حاصل ازرگرسیون بامعادله (۱) فرمولی برای n، یعنی حجم نمونه مورد لزوم برای آزمونی در سطح اشتباه α با توان $\gamma - 1$ نتیجه خواهدشد.

داده‌های فاقد سانسور

درداده‌های فاقدسانسور $E(X_i)=\frac{1}{\lambda_i}$ است پس از یافتن $var(\hat{\beta}_1)$ به صورت زیر:

$$var(\hat{\beta}_1)=\left\{\left(1-R^2_{z*|w*}\right) nP_1(1-p_1)\right\}^{-1}$$

باتوجه به توزیع مجانبی نرمال $\hat{\beta}_1$،تعداد نمونه لازم برای انجام آزمونی درسطح اشتباه α باتوان $\gamma - 1$ خواهدشد:

$$n=\left[\frac{Z_{1-\alpha/2}+Z_{1-\gamma}}{\beta_1}\right]^2 \frac{1}{p_1(1-p_1)\left\{1-R^2_{z*/w*}\right\}} \qquad (2)$$

که در آن P_1 نسبت افراد درگروه ۱ می‌باشد. باتغییر ازحالت آزمون فرض به سمت برآورد فاصله‌ای، حجم نمونه لازم برای برآورد β_1 توسط یک فاصله اطمینان باعرض 2d می‌شود:

$$var(\hat{\beta}_1)=\left[\frac{\beta_1}{Z_{1-\alpha/2}+Z_{1-\gamma}}\right]^2 \qquad (3)$$

که درآن d مقدارثابتی است که ازپیش معین می‌شود.

داده‌های دارای سانسور

درداده‌های دارای سانسور $E(X_i)=\frac{1}{\lambda_i}(1-\pi_i)$ می‌باشد،

بنابراین، واریانس خواهد شد:

$$var(\hat{\beta}_1)=\left\{(1-\pi)\left(1-R^2_{z*|w*}\right) np_1(1-p_1)\right\}^{-1}$$

لذا حجم نمونه nبرای آزمونی درسطح اشتباه α باتوان $\gamma - ۱$ دراین حالت می‌شود:

$$n=\left[\frac{z_{۱-\alpha/۲}+z_{۱-\gamma}}{\beta_۱}\right]^۲ \frac{۱}{p_۱(۱-p_۱)\left[۱-R^۲_{z^*/w^*}\right](۱-\pi)} \quad (۴)$$

که π عبارت است از احتمال اینکه یک فرد به طور تصادفی ضمن نادیده انگاشتن متغیرهای کمکی دچار سانسور شود.

یافته‌ها

حال بااستفاده از داده‌های مطالعاتی قبلی بقاء (۹). کاربرد فرمولهای تعیین حجم نمونه مذکور ارائه می‌شود.

الف) داده‌های فاقد سانسور

محققی می‌خواهد تابع مخاطره تسکین استروئید- ایندیوسد (Steroid-induced) راکه دراثر استفاده از داروی مرکاپتوپورین - ۶ (Mercaptoporine) پدیده می‌آید دربین زنان ومردانی که دچار سرطان گلبولهای سفید هستند مقایسه نماید. در صورتیکه بداند سانسور داده‌ها اتفاق نخواهدافتاد وتوزیع داده‌ها نمایی است ومطالعات قبلی مقدار ضریب تعیین را $R^2_{z/w}$ = ۰/۴۵ و β_1 = ۰/۸۵ نشان داده باشد. حجم نمونه لازم جهت کسب توان ۸۵ درصد درسطح‌اشتباهα =۰/۰۵ چقدر است؟ درصورتیکه نسبت مردان درجامعه دارندگان سرطان گلبولهای سفید P_1=۰/۶۵ باشد باتوجه به اینکه $Z_{1-\gamma}$ =۱/۰۳ و $Z_{1-\alpha/2}$ = ۱/۹۶است بااستفاده ازفرمول (۲) حجم نمونه برابر ۹۹ نفر خواهدشد.

ب) داده‌های دارای سانسور

پژوهشگری می‌خواهدتابع مخاطره دریافت کنندگان کلیه را دریک مرکز تحقیقاتی کلیه درمیان دوگروه شهری وروستایی مقایسه نماید. اگر توزیع بقاء نمایی باشد و احتمال سانسور براساس تجربه‌های قبلی π=۰/۳ و مقدار ضریب تعیین $R^2_{z/w}$ = ۰/۰۵ و β_1 = ۰/۹۲ باشد، حجم نمونه لازم جهت کسب توان ۹۰ درصد در سطح اشتباه α = ۰/۰۵ چقدر است؟ در صورتی که نسبت شهری‌ها به روستائیان در دریافت کنندگان کلیه در مراکز برابر p_1 = ۰/۷۵ باشد. با توجه به این که $Z_{1-\gamma}$ =۱/۲۸ و $Z_{1-\alpha/2}$= ۱/۹۶ است با استفاده از فرمول (۴) حجم نمونه برابر ۱۸۸ نفر خواهدشد. دردادههای موجود بامنظور کردن نسبت مخاطره β_1 = ۰/۲۲ در رگرسیون مخاطره متناسب یانمایی، توانهامحاسبه شدند، جدول ۱ توان حاصل ازروش آزمون تطبیق نشده (Unadjusted test) دو نمونه‌ای لگ - رتبه‌ای و روش آزمون والد (Wald) در رگرسیون تطبیق شده رابه ازای نسبت مخاطره فوق نشان می‌دهد. ضرایب تعیین با انجام $R^2_{z/w}$ رگرسیون هرمتغیر بر حسب سایر متغیرهای موجود بدست آمده است و P_1 عبارت از نسبت تعلق افراد به گروه اول می‌باشد.

جدول ۱: محاسبه توان براساس آزمون دونمونه‌ای تطبیق نشده لگ - رتبه‌ای وروش تطبیق شده برای آزمون دوطرفه باخطای ۰/۰۵

نوع داده	متغیر مورد نظر	مقدار			توان	
فاقد سانسور	سن	۰/۰۱۹	۰/۴۴	۰/۲۲	۸۱/۳۲	۸۲/۱۸
	لگاریتم تعداد گلبولهای سفید	۰/۰۶۴	۰/۳۵	۰/۲۲	۷۲/۸۹	۷۵/۷۱
دارای سانسور	وضعیت انجام عمل	۰/۲۰	۰/۲۹	۰/۲۲	۶۰/۰۸	۶۹/۴۸
	مرحله بیماری	۰/۱۸۵	۰/۷۸	۰/۲۲	۷۲/۲۰	۸۰/۹۰
	داشتن علائم سیستمیک	۰/۲۲	۰/۴۹	۰/۲۲	۷۲/۵۲	۸۳/۰۹

بحث

هرچنددرروشهای پارامتری و ناپارامتری درآمار کلاسیک انتخاب تصادفی نمونه‌ها دربسیاری از مدلهای آماری ازشرایط اساسی محسوب می‌شود. لیکن دربسیاری از پژوهشهای پزشکی و بهداشتی امکان تخصیص تصادفی نمونه‌ها فراهم نیست وهمواره این سوال مطرح بوده است که درتحلیل چنین داده هایی باید ازچه روشهایی استفاده نمود. بانقش اساسی که آنالیز رگرسیون خطی ایفاء می‌نماید اکنون به این نیاز پاسخ مناسبی داده شده است. ازآنجا که درتحلیل مدل رگرسیون خطی چند گانه بامتغیر مستقل دارای اثر ثابت، متغیر پاسخ (وابسته) وباقیمانده‌ها دارای توزیع تصادفی نرمال بوده و امید ریاضی متغیر وابسته به شرط متغیر مستقل به صورت رگرسیون خطی بیان می‌شود، بنابراین برای تحلیل داده‌های فاقد تخصیص تصادفی، درتحلیل رگرسیون خطی ،متغیر مستقل نقش عدم تخصیص تصادفی را ایفاء می‌نماید. این مطلب به آنالیز بقاء نیز تعمیم داده می‌شود.به طوریکه اگر بخواهیم تابع مخاطره دوگروه راکه امکان تخصیص تصادفی برای آنها وجودندارد باهم مقایسه نمائیم کافی است یک متغیر Z راطوری درنظربگیریم که مقادیر آن مبین گروهها باشد وسایر متغیرهای تبیینی (مستقل) رابه صورت بردار W درمدل رگرسیون منظور می‌نمائیم. استفاده از مدل رگرسیونی درمطالعات غیر تصادفی آنالیز بقاء موجب می‌شودکه فرمولهای مناسبی برای برآورد حجم نمونه هم برای مطالعات فاقد سانسور وهم دارای سانسور فراهم شود. دراین صورت درسطح اشتباهα پیداکردن حجم نمونه برای رسیدن به توان آماری لازم نیز ممکن می‌گردد. اگر $\hat{\beta}_1$ ضریب متغیر Z در مدل رگرسیون بقاء باشد. باتوجه به ارتباط موجود بین

ضریب تعیین $R^2_{z/w}$ و واریانس پارامتر $\hat{\beta}_1$، فرمول هایی برای برآورد حجم نمونه به اثبات می‌رسد. فرمول آزمون لگ-رتبه‌ای برای مقایسه تابع مخاطره بین گروهها، زمانی مورد استفاده قرارمی گیرد که تخصیص افرادبه گروهها به صورت تصادفی باشد. درصورت فقدان تخصیص تصادفی ممکن است که تابع مخاطره دربین گروهها یکسان نباشد لذا روش این مقاله مناسب است. درمدل‌های رگرسیون مخاطره‌های نیمه پارامتری‌متناسب کاکس که واریانس $\text{var}(\hat{\beta}_1)$ به حالت ساده‌تری تبدیل نمی‌شود باید ازیک مدل شبیه سازی استفاده شود.

جدول ۱ نشان می‌دهد که وقتی $R^2_{z/w}$ برای تعیین فرمول حجم نمونه مناسب نسبتاً بزرگ می‌باشد توانی که براساس آزمون لگ - رتبه‌ای تطبیق نشده محاسبه می‌شود، توان را بیش ازمقدار واقعی برآوردمی کند ولی وقتی $R^2_{z/w}$نزدیک به صفرباشد مثل متغیرهای سن ولگاریتم تعداد گلبولهای سفید، تفاوت درتوان کمتر می‌شود. به طور مثال درمقایسه متغیر مرحله بیماری ، توان براساس لگ - رتبه‌ای برابر ۸۰/۹۰ درصد است ولی توان واقعی ۷۲/۲۰ درصد می‌باشد. براساس جدول ۱، وقتی مقدار $R^2_{z/w}$ افزایش می‌یابد تفاوت بین توانهای حاصل‌ازمحاسبات لگ-رتبه‌ای و $R^2_{z/w}$تطبیق شده نیز افزایش می‌یابد.

ضمیمه

اکنون اثبات فرمولها که بیشتر براساس مطالب برناردو وهمکارانش (۲۰۰۰) (۵) ولیپزیتس و پارزن (۶) استوار است به اختصار بیان می‌شود:

مدل رگرسیون نمایی

اگرتمام متغیرهای کمکی به صورت تصحیح شده بامیانگین درنظرگرفته شوند، بعضی از فرمولهای بعدی ساده‌تر می‌شوند. بنابراین مدل رگرسیون نمایی رابه صورت زیربازنویسی می‌کنیم.

$$\lambda_i = exp\left[\beta^*_0 + \beta_1 Z_i^* + \beta_2' W_i^*\right]$$

$$\beta^*_0 = \beta_0 + \beta_1 \bar{Z} + \beta_2' \bar{W}, Z_i^* = Z_i - \bar{Z}, W_i^* = W_i - \bar{W}$$

که درآن: $\beta = (\beta^*_0 = (\beta^*_0, \beta_1, \beta_2'),$

درستنمایی برای β بصورت:

$$L(\beta) = \prod_{i=1}^{N} \left\{\lambda_i exp(-\lambda_i x_i)\right\}^{\delta_i} exp(-\lambda_i x_i)^{1-\delta_i}$$

$$= \prod_{i=1}^{N} exp\left(\beta^*_0 + \beta_1 Z_i^* + \beta'_2 W_i^*\right)^{\delta_i} exp\left\{-exp\left(\beta^*_0 + \beta_1 Z_i^* + \beta'_2 W_i^*\right) x_i\right\},$$

است که درآن، اگر $T_i \le C_i$ باشد۱ $= \delta_i$ است ودرغیر اینصورت $\delta_i = 0$ می‌باشد. نخست ماتریس اطلاع مشاهده شده $I(\beta)$اذیربط راازتابع درستنمایی بدست، می‌آوریم.

طبق نظریه درستنمایی ، که ازحل معادلات امتیاز (score equations) بدست می‌آید. برآوردگری سازگار ومجانباً نرمال برای β وجود خواهد داشت که دارای ماتریس کوواریانس $E\{I(\beta)\}^{-1}$می‌باشد.

داده‌های فاقد سانسور

درداده‌های بدون سانسور $E(Xi)=\frac{1}{\lambda_i}$ است.پس از بدست آوردن ماتریس مربوط به $E\{I(\beta)\}$ واریانس مجانبی بستگی به عنصر (۲و۲)، درمعکوس ماتریس مذکور دارد.لذا باکمک قانون ماتریسهای مجانبی ،وطبق روابط ماتریسها،می‌توان $\text{var}(\hat{\beta}_1)^{-1}$رابه صورت زیرنوشت (۱۰):

$$Z^{*\prime}\left\{I - W^*\left(W^{*\prime}W^*\right)^{-1}W^{*\prime}\right\}Z^* = Z^{*\prime}\left(1 - \frac{1}{n}J\right)Z^* - Z^{*\prime}\left\{W^*\left(W^{*\prime}W^*\right)^{-1}W^{*\prime} - \frac{1}{n}J\right\}Z^*$$

که درآن Z^*برداری $1 \times n$است که عنصر iام آن Z_i^*می‌باشد و W^* یک ماتریس $P \times n$است که سطر iام آن $W_i^{*\prime}$است.و J ماتریس $n \times n$است که تمام عناصر آن ۱ هستند. باتوجه به رابطه اساسی موجود درنظریه رگرسیون است وقتی که رگرسیون Z^* برحسب W^*بیان شود [۱۱]یعنی:

رگرسیون SS- کل SS= خطا SS

نتیجه می‌شود که: **(۵)**

$$Z^{*\prime}\left\{I - W^*\left(W^{*\prime}W^*\right)^{-1}W^{*\prime}\right\}Z^* = \left(1 - R^2_{z^*|w^*}\right)\sum_{i=1}^{N}(Z_i - \bar{Z})^2,$$

که درآن $R^2_{z^*/w^*}$ ضریب تعیین حاصل ازرگرسیون Z^* به شرط W^* است. چون Z_iیک متغیر تصادفی دوجمله‌ای است، معادله (۵) ساده شده به صورت زیر درمی آید:

(۶)

$$\text{var}(\hat{\beta}_1) = Z^{*\prime}\left\{I - W^*\left(W^{*\prime}W^*\right)^{-1}W^{*\prime}\right\}z^* = \left(1 - R^2_{z^*|w^*}\right)nP_1(1 - p_1)$$

بامساوی قراردادن معادلات (۱) و (۶) وحل آن برای n، تعداد افراد لازم برای انجام آزمونی درسطح اشتباه α با توان γ-۱ به صورت رابطه (۲) درمی‌آید.

بافرض اینکه (Z^*_i, W^*_i)مشاهدات مستقل باتوزیع یکسان باشند.$R^2_{z^*/w^*}$ دراحتمال به مربع ضریب همبستگی چندگانه $\rho^2_{z^*/w^*}$ همگراست (کریستنسن (Christensen) ، ۱۹۷۲) (۱۲). بنابراین، درمطالعات آینده نگر که درآن دارای ماتریس طرح (Design matrix) نیستیم درواقع $\rho^2_{z^*/w^*}$ رابرآورد می‌کنیم ونه مقدار نمونه‌ای آن $R^2_{z^*/w^*}$ را.

داده‌های دارای سانسور

فرض کنید که افراد بعداز T_aسال وارد مطالعه می‌شوند، و تعقیب (Fllow -up) آنها T_Cسال دیگر پس ازپایان حضورشان ادامه می‌یابد. فرض کنید متغیر تصادفی سانسورشدن Cدارای یک چگالی دلخواه f(c) می‌باشد ازآنجاکه T ازیک توزیع نمایی پیروی می‌کند.لذا:

(۷)

$$\pi_i = Pr(T_i > C_i) = \int_0^\infty pr\left(T_i > C_i \mid C_i = c_i\right) f(c_i)dc_i$$

$$= \int_0^\infty exp(-\lambda_i c_i) f(c_i)\, dc_i$$

که درآن: $\lambda_i = exp(\beta^*_0 + \beta_1 Z_i^* + \beta' W_i^*)$

یادآوری می‌شودکه دراینجا:

$$X_1 = min\,(T_i, C_i)$$

$$E(X_i) = E\left\{T_1 I\left[T_i \quad C_i\right) + CI(T_i \quad C_i\right]\right\}$$

$$= \int_0^\infty \int_0^{C_i} t_i \lambda_i exp(-\lambda_i t_i) dt_i f(c_i) dc_i + \int_0^\infty C_i exp\,(-\lambda_i c_i) f(c_i) dc_i$$

$$= \frac{1}{\lambda_i}(1 - \pi_i)$$

که درآن I(A) یک تابع مشخصه (Indicator function) است که اگر A درست باشد مقدار ۱ ودرغیر این صورت مقدار۰ رااختیار می‌کند. بنابراین مقدارمنتظره ماتریس اطلاع مشاهده شده نظیر $I(\beta)$ حاصل از تابع درستنمایی است منتها x_i ازتمام عناصر ماتریس حذف وتمام عناصر درمقدار$(1 - \pi_i)$ ضرب می‌شوند.

برآورد احتمال شرطی سانسور شدن π_i بشرط آنکه متغیرهای کمکی افراد مشخص باشد از طریق داده‌ها مشکل است. لذا مقدار π_i توسط π برآورد می‌شود. می‌توان π رابه صورت نسبت افرادی که دریک مجموعه ازداده‌های معین سانسور شده‌اند بر آورد نمود.

بعد از جایگزینی π_i توسط π ماتریس اطلاع منتظره بدست خواهد آمد. که وقتی $\pi = 0$ باشد ماتریس حاصل به ماتریس حالت داده‌های فاقد سانسور تبدیل میشود.

بطور مشابه بامساوی قراردادن معادلات، حجم نمونهn برای آزمونی درسطح اشتباه αبا توان $\gamma - 1$ به صورت رابطه (۴) درخواهد آمد.

توجه نمایید که باانتقال $\pi - 1$ به سمت چپ معادله (۴) فرمولی برای تعداد شکست‌ها بدست (Failures) خواهدآمد.

برعکس، بامعلوم بودن n، می‌توانیم توان یک مقایسه رابرای آزمونهای درسطح α به صورت زیر بدست آوریم:

$$1 - \gamma = \Phi\left[\left\{\beta_1^2 n p_1 (1 - p_1)\left(1 - R^2_{z*Iw*}\right)\overset{(8)}{(1 - \pi)}\right\}^{1/2} - Z_{1-\alpha/2}\right]$$

که در آن Φ تابع توزیع نرمال استاندارد می‌باشد.

در مدل نیمه پارامتری متناسب کاکس مطالعه شبیه سازی که توسط برناردو و همکاران (۲۰۰۰) (۵) انجام شده نشان داده است که این برآورد بطورمناسبی عمل می‌کند.بابیان مدل رگرسیون کاکس توسط فرایندهای شمارشی که توسط اندرسن وهمکاران (۱۹۸۲) (۱۳) ارائه شده می توان به دست‌آوردهای جدیدی رسید.

مراجع

1- Freedman , L.S.Tables of the number of patients required in clinical trials using the log -rank test. Statist. Med. 1982, 1,121-129.

2- Schoenfeld ,D.A. Sample -size formula for the proportional -hazards regression model . Biometrics . 1983 ,39, 499-503.

3- Lachin , J.M.and Foulkes ,M.A. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry , losses to follow -up , noncompliance and stratification . Biometrics. 1986, 42,507-519.

4- Lakatos , E.and Lan ,K.K.G.A comparison of sample size methods for the logrank statistic. Statist. Med. 1992, 11, 179-191.

5- Bernardo patricia M.V., Lipsitz Stuart R .,Harrington David P.and Catalano paul J.. Sample size calculations for failure time random variables in non-randomized studies. The statistician , 49 , 2000,part 1,31-40.

6- Lipsitz, S.R and parzen,M.Sample size calculations for non-randomized studies Statistician 1995, 44,81-90

7- Whittemore , A.S. Sample size for logistic regression with small response probability .J.Am Statist .Ass. 1981.76,27-32.

8- Signorini, D,F.Sample size for poisson regression. Biometrika.1991,78,446-450.

9- Kleinbaum .D.G,. Survival analysis , A self -learning text. Springer-Verlag New York,Inc 1989.

10- Snedecor ,G.W.and Cochran,W.G. Statistical Methods. Ames : Iowa State University press 1980.

11- Myers , R.H.and Milton , J.S. A 1-1rst coures in the theory of Linear Models .Boston , 1999, PW,S-kent.

12- Christensen,R.Plan Answers to complex Questions.New York : Springer .1987.

13- Andersen ,P.K. and Gill ,R.D. Cox's regression model for counting processes:a large sample study .Anals of Statist.1982,10,1100-1120.