بررسی مقایسه ای مدل پاسخ تصادفی بر روی سوال حساس بر اساس رگرسیون لجستیک

علی شمس پور ، دانشجوی کارشناسی ارشد گروه آمار، دانشگاه شهید چمران اهواز، ali_shamspour@yahoo.com
دکتر سید محمد رضا علوی ،عضو هیأت علمی،گروه آمار، دانشگاه شهید چمران اهواز alavi_m@mscu.ac.ir
دکتر محمدرضا آخوند ،عضو هیأت علمی ،گروه آمار،،دانشگاه شهید چمران اهواز mr.akhoond@scu.ac.ir

چکیده: مدل پاسخ تصادفی (RM) [1] معرفی شده توسط وارنر(۱۹۶۵) ، برای جلوگیری از پاسخ نادرست به سوالات حساس و محافظت از حریم خصوصی مصاحبه شونده طراحی شد. در این مقاله یک رویکرد از مدل نمونه مورد بررسی قرار می گیرد و برآوردی از کل افراد با مشخصه حساس را ارائه می دهد.همچنین مدل های پاسخ تصادفی، مدل رگرسیون لجستیک برای داده های پاسخ تصادفی شده وبرآورد_π مورد مطالعه قرار می گیرد و با استفاده از شبیه سازی کارایی آن مقایسه می شود.

کلمات کلیدی: پاسخ تصادفی ، برآورد رگرسیون لجستیک ، برآورد_ π.

مقدمه

در مبحث نمونه گیری پاسخ درست به سوال حساس دارای اهمیت زیادی می باشد. زیرا بسیاری از افراد در هنگام مواجه با سوال حساس، پاسخ واقعی نمی دهند و یا بطور کلی از پاسخ دادن به سوال امتناع می کنند. از جمله مسائل حساسی مانند: استفاده از مواد مخدر، ترجیحات جنسی، مالیات، مشروبات الکلی و...این موضوع باعث می شود که، نتایج بدست آمده با استفاده از روش مستقیم در پاسخ به سوال حساس، دارای اریبی باشند. روش پاسخ تصادفی این امکان را می دهد، تا علاوه بر امنیت پاسخ دهنده، برآوردی با کارایی مناسب تر در پاسخ به سوال حساس مطرح شود. این روش ابتدا توسط وارنر(۱۹۶۵) برای پاسخ به سوال حساس طراحی شد، به منظور بهبود این روش هارویتس و همکاران(۱۹۶۷) ؛ گرینبرگ وهمکاران (۱۹۶۹) طراحی سوال نامربوط را نیز ارائه دادند، در همین حوزه به مدل های پاسخ تصادفی دیور(۱۹۷۶) و مگنت وسینگ (۱۹۹۰) می توان اشاره کرد. همچنین در همین راستا لهتونن_ ویجین(۱۹۹۸) ، با استفاده از مدل رگرسیون لجستیک به منظور بهبودکارایی روش پاسخ تصادفی به تحقیق

[1] Randomized response technique

پرداختند.مدلW(وارنر۱۹۶۵) این مدل با استفاده از ابزار تصادفی دو سوال با یک پاسخ بله وخیر را مطرح می کند، سوال اول: آیا شما متعلق به گروهی با ویژگی A می باشید؟ سوال دوم: آیا شما متعلق به گروهی با ویژگیA^c می باشید؟ A مشخصه حساس مورد توجه می باشد.به عنوان مثال: سوال اول اینگونه مطرح می شود: آیا شما مالیات سال گذشته خود را پرداخت کرده اید؟ سوال دوم: آیا مامور مالیات برای ثبت مشخصات شما در سال گذشته مراجعه کرده است؟ مدلU (گرینبرگ و همکاران ۱۹۶۹) مدل پاسخ تصادفی در طراحی سوال حساس نامربوط می باشد و همانند مدل (W) ، با استفاده از ابزار تصادفی دو سوال با یک پاسخ بله و خیر مطرح می کند، تفاوت این روش با وارنر نامربوط بودن سوال غیر حساس نسبت به سوال حساس می باشد. برای مثال: آیا شما در سال گذشته مالیات خود را پرداخت کرده اید؟ سوال دوم : آیا شما علاقمند به تماشای سریال های تلویزیونی می باشید؟ مدلD (مدل دیور وهمکاران ۱۹۷۷) همانند مدل U می باشد، با این تفاوت که پاسخ مربوط به سوال غیر حساس نامربوط، پاسخی کاملا مشخص می باشد. مدل H (هارویتس وهمکاران ۱۹۷۶) این مدل برای طراحی سوال نامربوط می باشد، در این مدل سه سوال مطرح می شود: سوال اول، یک سوال حساس سوال دوم، سوالی غیرحساس که پاسخ آن مثبت می باشدوسوال سوم، سوالی غیرحساس که پاسخ آن منفی می باشد.که به ترتیب احتمال انتخاب سوالات $P_۱, P_۲, P_۳$ می باشد. بطوری که $P_۱ + P_۲ + P_۳ = ۱$. مدل M (مگنت و سینگ ۱۹۹۰) این مدل از دو ابزار تصادفی استفاده می کند، ابتدا دو سوال مطرح می شود: سوال اول، سوالی حساس و سوال دوم سوال غیر حساس نامربوط می باشد، در صورتی که پاسخ مصاحبه کننده، مربوط به سوال نامربوط باشد، مصاحبه کننده وارد ابزار تصادفی دوم می شود، که در اینجا نیز دو سوال با محوریت حساس و غیر حساس مطرح می شود. در این مقاله مدل رگرسیون لجستیک ارائه خواهد شد و فرض بر این خواهد بود، که متغییر کمکی حساس نمی باشد.سپس برآورد واریانس برای مدل های مورد نظر ارائه خواهد شد و با استفاده از شبیه سازی کارایی مدل ها مورد بررسی قرار می گیرد.

چارچوب

جامعه محدود، $U = \{۱, ۲, \ldots, N\}$ در نظر گرفته می شود. y متغییر دوحالتی می باشد، که اشاره به این دارد، فرد پاسخ دهنده به کدام گروه تعلق دارد(گروه حساسA یا گروه غیر حساس A^c) y_k مقداری از y برای k^{th} عنصر از جامعه می باشد.$y_k = ۱$، فرد kام پاسخ دهنده متعلق به ویژگی حساس A می باشد و $y_k = ۰$، فرد kام پاسخ دهنده متعلق به ویژگی غیر حساس A^c می باشد.هدف بدست آوردن برآورد $t_A = \sum_U y_k$ می باشد.t_A تعدادکل افراد با مشخصه حساس می باشد. روند نمونه گیری شامل دو مرحله می باشد: مرحله اول شامل انتخاب نمونه، نمونه ای به اندازه n ، ازجامعه مورد نظربا توجه به طرح نمونه گیری$p(s)$ با احتمال شمول مثبت و مرحله دوم مصاحبه از افراد نمونه مطابق با مدل تصادفی صورت می گیرد و برای هر $k \in s$، مدل پاسخ تصادفی را با متغییر z_k معرفی می کنیم. چنان که: $\widehat{z_k} = az_k + b_k$ برآوردی نا اریبی برای y_k می باشد. a, b_k ثابت های شناخته شده می باشند.

برآورد رگرسیون لجستیک برروی مدل های پاسخ تصادفی

$t_{AG;L\ RGEG}$[۲]برآوردرگرسیون لجستیک تعمیم یافته t_A می باشد.فرض می شود: $\underline{y} = \{y_۱, y_۲, \ldots, y_N\}$ برگرفته شده ازجامعه $\underline{Y} = \{Y_۱, Y_۲, \ldots, Y_N\}$

[۲] Generalized Logestic Regression Model Estimator

می باشد.این بردارمشاهدات متغییر تصادفی مستقل هستندوهمچنین دارای توزیعی می باشند که در زیر معرفی می کنیم:

$$P(Y_k = ۱ \mid \underline{x_k}, \underline{\beta}) = \frac{exp(\underline{x_k}, \underline{\beta})}{۱ + exp(\underline{x_k}, \underline{\beta})}$$

$$k = \{۱, ۲ \ldots, N\}$$

$$t_A = \sum_U Y_K = \sum_U \mu_k + \sum_U (Y_K - \mu_k)$$

معادله زیر برآورد(t_A) را معرفی می کند:

$$\widehat{(t_A)}_{G,L\ RGEG} = \sum_U \widehat{\mu_k} + \sum_s \frac{\widehat{z_k} - \widehat{\mu_k}}{\widehat{\mu_k}}$$

$$\widehat{\mu_k} = \mu\{\underline{x_k^۲}, \underline{\widehat{\beta}}\} = \frac{۱}{۱ + exp(-\underline{x_k^t}, \underline{\widehat{\beta}})}$$

$\underline{\widehat{\beta}}$ برآورد درستنمایی ماکزیمم$\underline{\beta}$می باشد.که با الگوریتم نیوتن رافسون محاسبه می شود.همچنین در زیر معرفی می کنیم: برآورد$-\pi$، برآوردی از واریانس تعداد کل افرادی که به سوال حساس پاسخ می دهند.[۳] (سارندال و همکاران ۱۹۹۲؛ ولتنن و ویجانن ۱۹۹۸)

$$\widehat{V}(\widehat{t}_{AG,L\ RGEG}) = a^۲ \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}} (\frac{z_k - \widehat{\lambda}_k}{\pi_k})(\frac{z_l - \widehat{\lambda}_l}{\pi_l})$$

$$= \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}} (\frac{\widehat{z_k} - \widehat{\lambda}_k}{\pi_k})(\frac{\widehat{z_l} - \widehat{\lambda}_l}{\pi_l})$$

$$cov(I_K, I_l) = \Delta_{kl} = \pi_{kl} - \pi_k \pi_l$$

$$\widehat{\lambda_k} = \frac{\widehat{\mu_k} - b_k}{a}$$

$$\widehat{\lambda_l} = \frac{\widehat{\mu_l} - b_l}{a}$$

$$k, l = ۱, ۲, \ldots, N$$

پیوستار

طرح نمونه گیری با احتمال شمول مثبت π_{kl}،π_k در زیر معرفی می کنیم:

$$\pi_k = \sum_{k \in s} p(s), \pi_{kl} = \sum_{k,l \in s} p(s)$$

نشانگر شمول: فرض کنید $k \in s$، نشان دهنده این پیشامد باشد، که نمونه s واحد i را در بر داشته باشد.

متغییر تصادفی

$$I_k(s) = \begin{cases} ۱ & ۱ \le k \le N, K \in s \\ ۰ & \text{otherwise} \end{cases}$$

نشانگر شمول معرفی می کنیم. مدل (W):

$$Z_k = \begin{cases} y_k & p \\ ۱ - y_k & 1 - \text{p} \end{cases}$$

به ازای هر$k \in s$

$$\widehat{z_k} = \frac{z_k - (۱ - p)}{۲p - ۱}$$

مدل (D):

$$Z_k = \begin{cases} y_k & p \\ ۱ & 1 - \text{p} \end{cases}$$

$$E_R = (Z_K) = y_k p + ۱ - p$$

$$\widehat{z_k} = \frac{z_k - (۱ - p)}{p}$$

مدل (U):

$$Z_k = \begin{cases} y_k & p \\ w_k & 1 - \text{p} \end{cases}$$

$$E_R = (Z_K) = y_k p + w_k(۱ - p)$$

$$\widehat{z_k} = \frac{z_k - w_k(۱ - p)}{p}$$

[۳] Estimator of The Variance Estimator

نتایج

در این مقاله سه روش W، U و D از نظر کارایی مورد مقایسه قرار می گیرد و نتایج شبیه سازی به شرح زیر خواهد بود.

	$\widehat{t_A}$	$\sqrt{var(\widehat{t_A})}$	$\sqrt{\widehat{var}(\widehat{t_A})}$
W	۴۷۳,۲۳۱	۷۶,۰۲۵	۶۵,۴۶
D	۴۷۵,۸۸	۷۶,۰۲	۳۱,۴۴

با توجه به جدول بالا که بررسی مقایسه ای دو مدل وارنر ودیور می باشد، وهمان طور که ملاحظه می کنید: مدل Dنسبت به مدلW برآوردی با کارایی مناسب تری برای تعداد کل افراد جامعه با صفت حساس دارد.ودارای اریبی کمتری نسبت به مدل وارنر می باشد .درادامه مقایسه مدل وارنر ومدل گرینبرگ وهمکاران مطرح شده است و همانظور که در جدول زیر ملاحظه می کنید: مدل U نسبت به مدل W برآوردی باکارایی مناسبتری برای تعدادکل افراد جامعه با صفت حساس دارد و دارای اریبی کمتری نسبت به مدل وارنر می باشد.

	$\widehat{t_A}$	$\sqrt{var(\widehat{t_A})}$	$\sqrt{\widehat{var}(\widehat{t_A})}$
W	۴۷۳,۲۳۱	۷۶,۰۲۵	۶۵,۴۶
U	۴۷۴,۸۵	۷۴,۱۴۴	۳۰,۸۳

همچنین از مقایسه دو جدول بالا به این نتیجه نیز خواهیم رسید که، مدل U نسبت به مدلD دارای اریبی کمتری خواهد بود .دراین نتیجه گیری به این موضوع پی خواهیم برد، که پاسخ به سوال حساس با پرسش مستقیم با مشکل جدی مواجه هست، این مقاله برای بررسی این موضوع، به برآورد رگرسیون لجستیک با توجه به مدل پاسخ تصادفی اشاره کرده است و همچنین، روش مناسبی برای برآورد تعدادکل افراد با مشخصه حساس را نشان می دهد، که نتایج حاکی از این است که مدل U دارای کارایی بهتری نسبت به دو روش ذکر شده می باشد و این روش کاهش قابل توجهی دربرآورد واریانس دارد.

مراجع

[1] Soberanis-Cruz, Víctor Hugo, and Víctor Miranda-Soberanis. ”The Generalized Logistic Regression Estimator in a Finite Population Sampling without Replacement Setting with Randomized Response.” Revista Colombiana de Estadística 34.3 (2011): 451-46 Enviromet-rics.

[2] LEHTONEN, R., and VEIJANEN, A., (1998): Logistic Generalized Regression Estimators. Survey. New York: Springer Verlag.

[3] SÄRNDAL, C.E., SWENSSON, B., and WRETMAN, J. (1992): Model Assisted Survey Sampling. New York: Springer Verlag.

[4] HORVITZ,D.C., GREENBERG, B. G., and ABERNATHY, J. R.(1976): RR: a data gathering device for sensitive questions. Internat.Statist.Rev. 44,181-196

[5] Devore, Jay L. ”A note on the randomized response technique.”Communications in Statistics-Theory and Methods 6.15 (1977): 1525-1529. Methodology, 24, 51-55.

[6] Warner, Stanley L. ”Randomized response: A survey technique for eliminating evasive answer bias.” Journal of the American Statistical Association 60.309 (1965): 63-69.

[7] Mangat, N. S., and Ravindra Singh. ”An alternative randomized response procedure.” Biometrika 77.2 (1990): 439-442.