افزایش دقت فواصل اطمینان بوت‌استرپی برای خط رگرسیون

محسن طبیبیان، سیدمحمدابراهیم حسینی‌نسب
گروه آمار، دانشگاه تربیت مدرس
گروه آمار، دانشگاه شهید بهشتی

در این مقاله روش‌های بوت‌استرپ را برای تشکیل فاصله اطمینان به روش $t-$صدکی برای خط رگرسیون استفاده می‌کنیم. هدف ما از این مقاله، بهبود مرتبه دقت فواصل اطمینان می‌باشد که برای این منظور، احتمال پوشش و خطای پوشش این فاصله را محاسبه می‌کنیم که خطای پوشش برای فاصله بوت‌استرپی منفرد به روش $t-$صدکی به صورت $O(n^{-1})$ و برای فاصله بوت‌استرپی دوگانه به روش $t-$صدکی برابر $O(n^{-2})$ می‌باشد.

واژه‌های کلیدی: احتمال پوشش، خطای پوشش، فاصله اطمینان بوت‌استرپی.

۱ مقدمه

اولین استفاده از بوت‌استرپ برای برآورد توزیع برآوردهای ضرایب رگرسیون خطی توسط افرون (۱۹۷۹) مطرح شد، و توسط فریدمن (۱۹۸۱) ادامه یافت. این روش بر تقریب توزیع خطاهای مشاهده نشده با توزیع تجربی مانده‌های مرکزی شده استوار است. نویدی (۱۹۸۹) نشان داد زمانی که عضوهای ماتریس طرح بدون محدودیت افزایش یابند، بوت‌استرپ به طور مجانبی بهتر از تقریب نرمال است. وی همچنین مطرح کرد که در موارد دیگر، همواره بوت‌استرپ به خوبی تقریب نرمال می‌باشد. هال (۱۹۹۰) و چرنیک (۲۰۰۸) نشان دادند زمانی که توزیع مانده‌ها در مدل نرمال نباشد و به خصوص اینکه توزیع آنها دارای دم‌های کلفت باشد می‌توان از روش‌های بوت‌استرپ در رگرسیون استفاده کرد. مارتین و رابرتز (۲۰۰۶) و چرنیک (۲۰۰۸) نیز نشان دادند اگر تعدادی مشاهده‌ی پرت وجود در مدل داشته باشند آنگاه می‌توان از روش‌های بوت‌استرپ استفاده کرد. فیشر و هال (۱۹۹۱) نشان دادند که روش بوت‌استرپ برای نمونه‌های با حجم کوچک، عملکرد مناسبی دارد. همچنین هال (۱۹۹۲) فواصل اطمینان بوت‌استرپی را به روش‌های صدکی و $t-$صدکی برای پارامتر شیب در مدل رگرسیونی خطی ساده بدست آورد و احتمال و خطای پوشش این فواصل را

بررسی کرد.

در این مقاله ابتدا، فواصل اطمینان بوت‌استرپی را به روش t–صدکی برای خط رگرسیونی بدست می‌آوریم و سپس احتمال پوشش و خطای پوشش این فواصل را نیز محاسبه می‌کنیم. در انتها دقت فواصل اطمینان بوت‌استرپی منفرد و دوگانه را به طور نظری بدست آورده و روش‌های عددی برای تشکیل فواصل اطمینان بوت‌استرپی منفرد و دوگانه مطرح می‌کنیم.

۲ مدل خطی ساده

مدل خطی ساده زیر را در نظر بگیرید:

$$Y_i = \beta_0 + X_i\beta_1 + \epsilon_i, \qquad i = 1, \cdots, n,$$

که در آن X_iها نقاط طرح‌اند و ثابت فرض می‌شوند و جملات خطا، ϵ_iها، مستقل و هم‌توزیع با میانگین صفر و واریانس σ^2 هستند. برآوردهای $\hat{\beta}_0$، $\hat{\beta}_1$ و $\hat{\sigma}^2$ به ترتیب برای پارامترهای β_0، β_1 و σ^2 به صورت زیر تعریف می‌شوند:

$$\hat{\beta}_0 = \bar{Y} - \bar{X}\hat{\beta}_1, \quad \hat{\beta}_1 = n^{-1}\sigma_x^{-2}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}), \quad \hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2, \quad (1)$$

که در آن داریم:

$$\sigma_x^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

فرض کنید چولگی و کشیدگی جامعه به ترتیب به صورت زیر باشند:

$$\gamma = E(\epsilon/\sigma)^3, \qquad \kappa = E(\epsilon/\sigma)^4 - 3. \quad (2)$$

همچنین فرض کنید که برآورد تجربی آنها را با روابط زیر نشان دهیم:

$$\hat{\gamma} = n^{-1}\sum_{i=1}^{n}(\hat{\epsilon}_i/\hat{\sigma})^3, \qquad \hat{\kappa} = n^{-1}\sum_{i=1}^{n}(\hat{\epsilon}_i/\hat{\sigma})^4 - 3.$$

با استفاده از روش بازنمونه‌گیری از مانده‌ها در رگرسیون داریم:

$$Y_i^* = \hat{\beta}_0 + X_i\hat{\beta}_1 + \epsilon_i^*, \qquad 1 \le i \le n,$$

که در آن ϵ_i^*ها به طور تصادفی و با جایگذاری از مانده‌های $\hat{\epsilon}_i$ استخراج شده‌اند. به علاوه برای محاسبه‌ی $\hat{\beta}_0^*$، $\hat{\beta}_1^*$ و $\hat{\sigma}^*$ کافی است که به ترتیب در فرمول‌های $\hat{\beta}_0$، $\hat{\beta}_1$ و $\hat{\sigma}$، که در (۱) معرفی شده‌اند به جای استفاده از Y_i، از مقدار Y_i^* استفاده شود.

برآورد مقدار میانگین

$$d_0 = E(Y|X = x_0) = \beta_0 + x_0\beta_1,$$

عبارت است از:

$$\hat{d}_0 = \hat{\beta}_0 + x_0\hat{\beta}_1. \qquad (۳)$$

اگر قرار دهیم $x_0 = 0$، آنگاه $\hat{\beta}_0$ برآورد عرض از مبدا β_0 بدست می‌آید. بنابراین مساله‌ی برآورد d_0، یک مساله کلی است.

همچنین داریم:

$$Var(\hat{d}_0) = n^{-1}\sigma^2\sigma_d^2,$$

که در آن

$$\sigma_d^2 = 1 + \sigma_x^{-2}(x_0 - \bar{x})^2. \qquad (۴)$$

در نتیجه آماره زیر (به طور مجانبی) یک کمیت محوری است:

$$T = n^{1/2}(\hat{d}_0 - d_0)/(\hat{\sigma}\sigma_d). \qquad (۵)$$

با روش بازنمونه‌گیری از مانده‌ها در مدل‌های رگرسیونی، بازنمونه‌ی $\{(X_1, Y_1^*), (X_2, Y_2^*), ..., (X_n, Y_n^*)\}$ را تولید می‌کنیم. فرض کنید که $\hat{\beta}_0^*$، $\hat{\beta}_1^*$ و $\hat{\sigma}^{*2}$ به ترتیب دارای فرمول یکسانی با $\hat{\beta}_0$، $\hat{\beta}_1$ و $\hat{\sigma}^2$ که در (۱) معرفی شده‌اند باشند به جز اینکه برای محاسبه‌ی آنها به جای استفاده از Y_i از مقدار Y_i^* استفاده شود. نسخه‌های بوت‌استرپی $\hat{d}_0$ و T به ترتیب عبارتند از:

$$\hat{d}_0^* = \hat{\beta}_0^* + x_0\hat{\beta}_1^*,$$
$$T^* = n^{1/2}(\hat{d}_0^* - \hat{d}_0)/(\hat{\sigma}^*\sigma_d). \qquad (۶)$$

۳ دقت فواصل اطمینان بوت‌استرپی برای خط رگرسیون

فاصله اطمینان دو طرفه با دم‌های برابر به روش t–صدکی برای خط رگرسیون (d_0) با سطح پوشش اسمی $1-\alpha$ عبارت است از:

$$J_2 = \Big(\hat{d}_0 - n^{-1/2}\hat{\sigma}\sigma_d v_{1-\alpha/2},\ \hat{d}_0 - n^{-1/2}\hat{\sigma}\sigma_d v_{\alpha/2}\Big),$$

که در آن مقادیر $\hat{\sigma}$، $\hat{d}_0$ و σ_d به ترتیب در روابط (۱)، (۳) و (۴) معرفی شده‌اند. چندک‌های $v_{1-\alpha/2}$ و $v_{\alpha/2}$ به ترتیب جواب‌های معادلات $P(T \le v_{1-\alpha/2}) = 1-\alpha/2$ و $P(T \le v_{\alpha/2}) = \alpha/2$ هستند و آماره T در رابطه (۵) معرفی شده است.

نسخه بوت‌استرپی این فاصله به صورت زیر می‌باشد:

$$\hat{J}_2 = \Big(\hat{d}_0 - n^{-1/2}\hat{\sigma}\sigma_d \hat{v}_{1-\alpha/2},\ \hat{d}_0 - n^{-1/2}\hat{\sigma}\sigma_d \hat{v}_{\alpha/2}\Big), \qquad (7)$$

که در آن چندک‌های $\hat{v}_{1-\alpha/2}$ و $\hat{v}_{\alpha/2}$ به ترتیب جواب‌های معادلات $P(T^* \le \hat{v}_{1-\alpha/2}|\chi) = 1-\alpha/2$ و $P(T^* \le \hat{v}_{\alpha/2}|\chi) = \alpha/2$ هستند و آماره T^* در رابطه (۶) معرفی شده است.

لازم بذکر است که فواصل J_2 و $\hat{J}_2$، به هر دم احتمال مساوی تخصیص می‌دهند، یعنی برای مثال برای فاصله J_2 داریم:

$$P(d_0 \le \hat{d}_0 - n^{-1/2}\hat{\sigma}\sigma_d v_{1-\alpha/2}) = P(d_0 \ge \hat{d}_0 - n^{-1/2}\hat{\sigma}\sigma_d v_{\alpha/2}) = \alpha/2.$$

قضیه ۱ . فرض کنید که فاصله اطمینان بوت‌استرپی با سطح پوشش اسمی $1-\alpha$ که از روش t–صدکی بدست آمده است به صورت فاصله $\hat{J}_2$ باشد که در رابطه‌ی (۷) معرفی شده است، آنگاه احتمال پوشش آن برابر است با:

$$\begin{aligned} P(d_0 \in \hat{J}_2) &= 1-\alpha - \frac{n^{-1}}{3}\sigma_d^{-1}(\kappa - \frac{3}{2}\gamma^2)z_{1-\alpha/2} \\ &\quad \times\Big\{(\gamma_d - 3\sigma_z^{-1})d_{1-\alpha/2}^2 + \gamma_d\Big\}\phi(z_{1-\alpha/2}) + O(n^{-2}), \end{aligned}$$

که در آن مقادیر γ، κ و σ_d در روابط (۲) و (۴) معرفی شده‌اند و $\phi(.)$ تابع چگالی نرمال استاندارد می‌باشد.

اثبات: برای اثبات به حسینی‌نسب و طبیبیان (۲۰۰۹) رجوع شود.

بنابراین با توجه به قضیه ۱ می‌توان گفت که فاصله اطمینان بوت‌استرپی $\hat{J}_2$ دارای خطای پوشش از مرتبه $O(n^{-1})$ است.

۴ طریقه تشکیل فواصل اطمینان بوت‌استرپی برای خط رگرسیون

۴-۱ بوت‌استرپ منفرد

فرض کنید که نمونه $\chi = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$ را در اختیار داریم و برآوردهای $\hat{\sigma}$ و $\hat{d}_0$ که در روابط (۱) و (۳) معرفی شده‌اند را از این نمونه محاسبه می‌کنیم. بازنمونه $\chi^* = \{(X_1, Y_1^*), (X_2, Y_2^*), \cdots, (X_n, Y_n^*)\}$ که به صورت تصادفی و باجایگذاری از نمونه χ انتخاب می‌شود را در نظر بگیرید. همچنین فرض کنید که $\hat{d}^*_{0,b}$ و $\hat{\sigma}^*$ به ترتیب نسخه‌ای از $\hat{d}_0$ و $\hat{\sigma}$ باشند که مقدار آنها به جای استفاده از نمونه χ، از بازنمونه‌ی $\chi^*_b = \{(X_1, Y^*_{1b}), \cdots, (X_n, Y^*_{nb})\}$ $(b = 1, 2, \cdots, B)$، به شرط نمونه χ محاسبه شده است، آنگاه چندک‌های $\hat{v}_{1-\alpha/2}$ و $\hat{v}_{\alpha/2}$ به وسیله رابطه زیر بدست می‌آیند:

$$\frac{1}{B}\sum_{b=1}^{B} I\Big(\frac{n^{1/2}(\hat{d}^*_{0,b} - \hat{d}_0)}{\sigma_d \hat{\sigma}^*} \le \hat{v}_\alpha | \chi\Big) = \alpha,$$

که در آن $I(\mathcal{A})$ تابع نشانگر مجموعه $\mathcal{A}$ می‌باشد. لازم بذکر است که اصل بوت‌استرپ این برابری را تضمین می‌کند، یعنی $P(\hat{d}^*_0 - \hat{d}_0/\hat{\sigma}^* \le \hat{v}_\alpha|\chi) \simeq P(\hat{d}_0 - d_0/\hat{\sigma} \le v_\alpha)$. در نتیجه فاصله اطمینان بوت‌استرپی منفرد همان فاصله‌ی $\hat{J}_2$ می‌باشد که در رابطه (۷) معرفی شده است.

قضیه ۱ . اگر فاصله اطمینان بوت‌استرپی منفرد $\hat{J}_2$ مانند بالا بدست آید آنگاه احتمال پوشش آن برابر است با:

$$P(d_0 \in \hat{J}_2) = 1 - \alpha + O(n^{-1}),$$

که در آن $O(n^{-1})$ خطای پوشش می‌باشد.

اثبات: برای اثبات به حسینی‌نسب و طبیبیان (۲۰۰۹) رجوع شود.

۴-۲ بوت‌استرپ دوگانه

حال اگر از هر یک از B بازنمونه‌ای که در بوت‌استرپ منفرد استخراج کردیم، C بازنمونه به طور تصادفی و با جایگذاری انتخاب کنیم، آنگاه از حل معادله زیر، چندک

$\hat{v}^*_{b,\alpha} = \hat{v}^*_{b,\alpha/۲}, \hat{v}^*_{b,۱-\alpha/۲}$ را بدست می‌آوریم:

$$\frac{۱}{C}\sum_{c=۱}^{C} I\Big(\frac{n^{۱/۲}(\hat{d}^{**}_{\circ,bc} - \hat{d}^{*}_{\circ,b})}{\sigma_d \hat{\sigma}^{**}} \le \hat{v}^*_{b,\alpha}\Big) = \alpha,$$

که در آن مقادیر $\hat{d}^{**}_{\circ,bc}$ و $\hat{\sigma}^{**}$ از بازنمونه $\chi^{**}_{bc} = \{(X_۱, Y^{**}_{۱bc}), \cdots, (X_n, Y^{**}_{nbc})\}$، $(c = ۱, ۲, \cdots, C)$، به شرط بازنمونه χ^*_b، $(b = ۱, ۲, \cdots, B)$ محاسبه شده‌اند. حال از حل معادلات زیر، $\tilde{v} = \tilde{v}_{\alpha/۲}, \tilde{v}_{۱-\alpha/۲}$ بدست می‌آید:

$$\frac{۱}{B}\sum_{b=۱}^{B} I\Big(\frac{n^{۱/۲}(\hat{d}^{*}_{\circ,b} - \hat{d}_{\circ})}{\sigma_d \hat{\sigma}^{*}} \le \hat{v}^*_{b,\alpha/۲} + \tilde{v}\Big) = \alpha/۲,$$

$$\frac{۱}{B}\sum_{b=۱}^{B} I\Big(\frac{n^{۱/۲}(\hat{d}^{*}_{\circ,b} - \hat{d}_{\circ})}{\sigma_d \hat{\sigma}^{*}} \le \hat{v}^*_{b,۱-\alpha/۲} + \tilde{v}\Big) = ۱ - \alpha/۲.$$

بنابراین فاصله اطمینان بوت‌استرپ دوگانه با دم‌های برابر برای خط رگرسیونی با سطح پوشش اسمی $۱ - \alpha$ عبارت است از:

$$\hat{J}_{۲d} = \Big(\hat{d}_{\circ} - n^{-۱/۲}\sigma_d\hat{\sigma}(\hat{v}_{۱-\alpha/۲} + \tilde{v}_{۱-\alpha/۲}),\ \hat{d}_{\circ} - n^{-۱/۲}\sigma_d\hat{\sigma}(\hat{v}_{\alpha/۲} + \tilde{v}_{\alpha/۲})\Big).$$

قضیه ۲ . اگر فاصله اطمینان بوت‌استرپی دوگانه به روش t–صدکی برای خط رگرسیون مانند بالا بدست آید، آنگاه:

$$P(d_{\circ} \in \hat{J}_{۲d}) = ۱ - \alpha + O(n^{-۲}),$$

که در آن خطای پوشش به صورت $O(n^{-۲})$ می‌باشد. یعنی، در صورت استفاده از روش بوت‌استرپ دوگانه، خطای پوشش فاصله از $O(n^{-۱})$ به $O(n^{-۲})$ کاهش می‌یابد.

اثبات: برای اثبات به حسینی‌نسب و طبیبیان (۲۰۰۹) رجوع شود.

۵ نتیجه‌گیری

خطای پوشش برای فاصله اطمینان بوت‌استرپی منفرد به صورت $O(n^{-۱})$ می‌باشد در حالیکه خطای پوشش فاصله اطمینان بوت‌استرپی دوگانه برابر $O(n^{-۲})$ است، یعنی در صورت استفاده از روش بوت‌استرپ دوگانه، خطای پوشش کاهش می‌یابد.

مراجع

[1] Chernick, M. R. (2008). *Bootstrap Methods: A Guide for Practitioners and Researchers*, 2nd ed, United BioSource Corporation, Willey.

[2] Efron, B. (1979). Bootstrap methods: Another look at the jackknife.*Ann. Statist.* **7**, 1-26.

[3] Fisher, N. I. and Hall, P. (1991). Bootstrap algorithms for small samples. *Journal of Statistical Planning and Inference*, **27**, 157-169.

[4] Freedman, D. (1981). Bootstrapping regression models.*Ann. Statist.*, **9**, 1218-1228.

[5] Hall, P. (1990). Asymptotic Properties of the Bootstrap for Heavy-Tailed Distributions. *Annals of Probability*, **18**, No. 3, 1342-1360.

[6] Hall, P. (1992).*The Bootstrap and Edgeworth expansion*, Springer-Verlag, New York.

[7] Hosseininasab, S.M.E. and Tabibian, M. (2009). Accuracy of Bootstrap Prediction intervals based on simple regression model. Manuscript.

[8] Martin, M. A. and Roberts, S. (2006). An evaluation of bootstrap methods for outlier detection in least squares regression. *Journal of Applied Statisticd*, **33**, No. 7, 703-720.

[9] Navidi, W. (1989). Edgeworth expansion for Bootstrapping regression models.*Ann. Statist.* **17**, 1472-1478.