مجله علوم آماری، بهار و تابستان ۱۳۸۸
جلد ۳، شماره ۱، ص ۵۹-۷۸

برآورد تفاضل مخاطره‌های کولبک-لیبلر برای مشاهدات سانسوریده از راست نوع II تحت مدل‌های غیر آشیانه‌ای

عبدالرضا سیاره، پریسا ترکمان
گروه آمار، دانشگاه رازی کرمانشاه

تاریخ دریافت: ۱۳۸۸/۲/۲۶ تاریخ آخرین بازنگری: ۱۳۸۸/۶/۲۵

چکیده : معیار آکائیک به طور گسترده در تئوری انتخاب مدل برای داده‌های کامل به کار گرفته می‌شود، اما برای داده‌های ناقص وقتی مدل‌ها غیر آشیانه‌ای و بد توصیف شده هستند کمتر مورد توجه قرار گرفته است. در این مقاله به انتخاب یک مدل مناسب از بین مدل‌های رقابتی برای داده‌های سانسوریده از راست نوع II پرداخته می‌شود و اقدام به برآورد تفاضل مخاطره‌های بین دو مدل غیر آشیانه‌ای می‌گردد. سپس نشان داده می‌شود استنباط براساس داده‌های مشاهده شده و سانسوریده به طور همزمان به جای در نظر گرفتن فقط داده‌های مشاهده شده به نتایج بهتری منتهی خواهد شد. فاصله ردیابی مناسب برای تفاضل امید کولبک-لیبلر مشاهدات سانسوریده با احتمال مشخص معرفی می‌شود و از آنجا که هر فاصله اطمینان مجموعه‌ای از فرض‌های پذیرفتنی تحت فرض صفر است، فاصله به دست آمده برای انتخاب مدل مناسب به کار گرفته می‌شود.

واژه‌های کلیدی : سانسور راست نوع II، مدل‌های غیر آشیانه‌ایی، معیار آکائیک، معیار اطلاع کولبک-لیبلر.

آدرس الکترونیک مسئول مقاله: عبدالرضا سیاره، asayyareh@razi.ac.ir
کد موضوع‌بندی ریاضی (۲۰۰۰):۶۲N۰۱

۱ مقدمه

وقتی که از توزیع واقعی جامعه بی اطلاع هستیم، انتخاب مدل مناسب برای داده‌ها به منظور پیش بینی، کنترل، تصمیم‌گیری و استخراج اطلاعات، اهمیت فراوانی پیدا می‌کند. انتخاب مدل توسط آزمون فرض یا به کمک معیارهای انتخاب مدل انجام می‌گیرد. اطلاع کولبک-لیبلر[۱] (۱۹۵۱)، بر پایه تابع مخاطره، برای بررسی واگرایی مدل رقابتی از مدل درست نامعلوم مطرح شده است. هر چه مقدار کمیت کولبک-لیبلر کوچک‌تر باشد، مدل رقابتی به مدل درست نزدیک‌تر است. فرض کنید $X_1, \ldots, X_n$ یک نمونه تصادفی از جامعه‌ای با تابع چگالی نامعلوم f باشد، $G^\beta = \{g^\beta(x), \beta \in B \subset R^p\}$ مدل رقابتی برای f است و سعی می‌شود به گونه ای انتخاب شود که برازش مناسبی به داده‌ها داشته باشد. اطلاع کولبک-لیبلر به صورت

$$KL(f, g^\beta) = E_f[\log \frac{f(X)}{g^\beta(X)}] = E_f[\log f(X)] - E_f[\log g^\beta(X)]$$

تعریف می‌شود. آکائیک (۱۹۷۳) برآوردی سازگار برای جملهٔ دوم اطلاع کولبک-لیبلر به صورت

$$\frac{1}{n}\sum_{i=1}^{n} \log g^{\hat{\beta}_n}(X_i)$$

معرفی کرد، که در آن $\hat{\beta}_n$ برآوردگر شبه ماکسیمم درست‌نمایی[۲] (QMLE) برای β است. در عمل تعدادی مدل پارامتری رقابتی به عنوان جایگزین مدل درست در نظر گرفته می‌شوند. اگر $G^\beta \cap H^\gamma = \emptyset$ باشد، دو مدل رقابتی G^β و $H^\gamma = \{h^\gamma(x), \gamma \in \Gamma \subset R^q\}$ غیر آشیانه‌ایی و در غیر این صورت متداخل یا آشیانه‌ایی نامیده می‌شوند. اگر یک مدل رقابتی شامل مدل درست باشد، آن را خوب-توصیف شده[۳] و در غیر این صورت آن را بد-توصیف شده[۴] می‌نامیم. برای

---

۱ $Kullback - Leibler$

۲ Qausi Maximum Likelihood Estimator

۳ Well-specified

۴ Miss-specified

تابع چگالی رقابتی $g^{\beta}(x)$ تابع شبه لگاریتم درست‌نمایی به صورت

$$L_n^g(\beta) = \sum_{i=1}^{n} \log g^{\beta}(x_i)$$

تعریف می‌شود. وونگ (۱۹۸۹) آزمون انتخاب مدل غیرآشیانه‌ای برای داده‌های کامل را معرفی کرد، که در آن فرضیه صفر $H_0 : E_f \left[\log \frac{g^{\beta*}(X)}{h^{\gamma*}(X)}\right] = 0$ معادل بودن دو مدل رقابتی غیر آشیانه‌ای از نظر نزدیکی به مدل درست داده‌ها است و فرضیه مقابل آن به یکی از صورت‌های

$$H_g : E_f \left[\log \frac{g^{\beta*}(X)}{h^{\gamma*}(X)}\right] > 0, \quad H_h : E_f \left[\log \frac{g^{\beta*}(X)}{h^{\gamma*}(X)}\right] < 0$$

هستند. انتخاب مدل مناسب از میان مدل‌های رقابتی غیرآشیانه‌ایی، وقتی که داده‌ها سانسوریده باشند از جنبه‌های مختلف تئوری و کاربردی مورد توجه آمارشناسان قرار دارد. در چنین وضعیتی داده‌های قابل مشاهده X_o به همراه داده‌های سانسوریده X_c به جای داده‌های کامل $X = (X_1, \ldots, X_n)$، در اختیار هستند. باتاچاریا (۱۹۸۵) برآوردگر ماکسیمم درست‌نمایی تعمیم یافته برای داده‌های سانسوریده نوع II را بررسی و توزیع مجانبی آن را به دست آورد. لینهارت و زوکچینی (۱۹۸۶) به بررسی خطاهای انتخاب مدل برای داده‌های کامل پرداختند. تیکو (۱۹۶۸) تقریبی برای برآوردگرهای درست‌نمایی تعمیم یافته پارامترهای مدل لگ-نرمال سانسوریده را بدست آورد. سگوان و همکاران (۲۰۰۵) براساس معیار کولبک متقارن، معیاری برای انتخاب مدل معرفی کردند تا به کمک آن اطلاعات از دست رفتهٔ حاصل از داده های گمشده را بازیابی کنند. هافیدی و همکاران (۲۰۰۷) به تعمیم معیار آکائیک برای انتخاب مدل با داده‌های گمشده پرداختند. کومانژ و همکاران (۲۰۰۸) شکل وزنی تفاضل معیارهای آکائیک را به منظور برآورد تفاضل‌های مقدار مورد انتظار توابع مخاطره کولبک-لیبلر بین برآوردگرهای ماکسیمم درست‌نمایی توزیع در دو مدل بررسی کردند و فاصله ردیابی بااحتمال از پیش مشخص شده‌ای را برای مشاهدات کامل به دست آوردند. سیاره و همکاران (۲۰۱۱) معیارها و آزمون‌های انتخاب مدل را مقایسه نموده نشان دادند که پس از انتخاب مدل‌های معادل توسط یکی از معیارها یا آزمون‌ها به منظور دست یابی به

مدل بهینه، لازم است بررسی بیشتری روی مدل‌ها به عمل آید.

در بخش دوم نشان داده می‌شود که استفاده از داده‌های مشاهده شده و سانسوریده به طور همزمان بهتر از استفاده از داده‌های ناقص در تحلیل این گونه داده‌هاست. در بخش سوم به بررسی رفتار برآوردگرهای ماکسیمم درست‌نمایی در دو حالت که مدل رقابتی داده های سانسوریده نوع II خوب-توصیف شده یا بد توصیف شده هستند، پرداخته شده است و آزمونی برای انتخاب مدل برای داده‌های سانسوریده معرفی می‌شود. در بخش چهارم فاصله ردیابی برای تفاضل ریسک‌های کولبک-لیبلر دو مدل رقابتی به دست آمده و استنباط برای انتخاب مدل بر اساس فاصلهٔ ردیابی در سطح اطمینان مشخص شده برای این نوع سانسور مورد مطالعه قرار گرفته است. بخش چهارم شامل مطالعه شبیه‌سازی برای بررسی توانایی فاصله ردیابی در انتخاب مدل مناسب است.

۲ انتخاب مدل براساس داده‌های کامل یا داده‌های ناقص

بسیاری از مدل‌های آماری ویژگی‌های داده‌های کامل را توصیف می‌کنند. این در حالی است که گاهی اوقات تنها دستیابی به زیر مجموعه‌ای از مشاهدات امکان‌پذیر است. از طرفی بسیاری از مدل‌ها حساسیت زیادی به فرض کامل بودن مشاهدات دارند و هنگامی که برای تحلیل داده‌های ناقص به کار برده می‌شوند، خواص مطلوب خود را از دست می‌دهند. اگر چه روش‌هایی برای بازسازی یا تقریب داده‌های ناقص، از جمله الگوریتم EM، معرفی شده‌اند، اما شیمودایرا (1994) معیارهای انتخاب مدل بر پایهٔ داده‌های کامل را برای بررسی داده‌های ناقص مورد استفاده قرار داده است. در این بخش نشان داده می‌شود وقتی که داده‌ها سانسوریده هستند عملکرد معیار کولبک-لیبلر براساس داده‌های مشاهده شده و سانسوریده، بهتر از حالتی است که داده‌های سانسوریده در نظر گرفته نمی‌شوند. تفاضل کولبک-لیبلرهای دو مدل G^β و H^γ را برای $\theta = (\beta, \gamma)$ به صورت

$$KL(\theta, f) = E_{f(X)}\left[\log \frac{h^\gamma(X)}{g^\beta(X)}\right] \qquad (۱)$$

تعریف می‌شود، که در آن $E_{f(X)}$ برای امید ریاضی تحت چگالی درست f است. بطور مشابه

$$KL_o(\theta, f) = E_{f(X_o)}\left[\log\frac{h^\gamma(X_o)}{g^\beta(X_o)}\right],$$

و

$$KL(\hat{\theta}_n, f) = E_{f(X_o)}\left[\log\frac{h^\gamma(X_o)}{g^\beta(X_o)}\right]|_{\theta=\hat{\theta}_n}.$$

امید ریاضی KL را هنگامی که در $\hat{\theta}_n$ ارزیابی می‌شود با روابط

$$\Delta_X(d, f) = E_{f(X_o)}\left[E_{f(X)}\{\log\frac{h^\gamma(X)}{g^\beta(X)}\}|_{\theta=\hat{\theta}_n}\right]$$

و

$$\Delta_{X_o}(d, f) = E_{f(X_o)}\left[E_{f(X_o)}\{\log\frac{h^\gamma(X_o)}{g^\beta(X_o)}\}|_{\theta=\hat{\theta}_n}\right]$$

نشان می‌دهیم. لذا ریسک کولبک-لیبلر ارزیابی شده در $\hat{\theta}_n$ عبارت خواهد بود از

$$\Delta(\hat{\theta}_n, f) = E_{f(X)}\left[\log\frac{h^\gamma(X)}{g^\beta(X)}\right]|_{\theta=\hat{\theta}_n}.$$

چگالی داده‌های کامل به حاصل‌ضرب چگالی داده‌های مشاهده شده و چگالی شرطی داده‌های سانسوریده به شرط داده‌های مشاهده شده به صورت $f(X) = f(X_o)f(X_c|X_o)$ قابل تفکیک است. لذا

$$KL(\theta, f) \quad = \quad E_{f(X)}\left[\log\frac{h^\gamma(X_o)}{g^{\beta(X_o)}}\right] + E_{f(X)}\left[\log\frac{h^\gamma(X_c|X_o)}{g^\beta(X_c|X_o)}\right].$$

جملهٔ اول عبارت سمت راست رابطه اخیر برابر $KL_o(\theta, f)$ است. پس از رابطه (۱) خواهیم داشت

$$KL(\theta, f) \quad = \quad KL_o(\theta, f) + E_{f(X)}\left[\log\frac{h^\gamma(X_c|X_o)}{g^\beta(X_c|X_o)}\right].$$

ریسک کولبک-لیبلر شرطی را به صورت

$$KL_{c|o}(\theta, f|X_o) = E_{f(X_c|X_o)}\left[\log\frac{h^\gamma(X_c|X_o)}{g^\beta(X_c|X_o)}\right]$$

تعریف می‌کنیم. لذا

$$KL(\theta, f) = KL_o(\theta, f) + E_{f(X_o)}\left[KL_{c|o}(\theta, f|X_o)\right].$$

بنابر نامساوی جنسن

$$KL_{c|o}(\theta, f|X_o) \geq KL_{c|o}(f, f|X_o).$$

در نتیجه

$$KL(\theta, f) \geq KL_o(\theta, f) + E_{f(X_o)}\left[KL_{c|o}(f, f|X_o)\right].$$

و بنابراین

$$E_{f(X_o)}\left\{E_{f(X)}\left[\log\frac{h^\gamma(X)}{g^\beta(X)}\right]|_{\theta=\hat{\theta}_n}\right\} \geq E_{f(X_o)}\left\{E_{f(X_o)}\left[\log\frac{h^\gamma(X)}{g^\beta(X)}\right]|_{\theta=\hat{\theta}_n}\right\}.$$

سمت چپ نامساوی اخیر امید ریاضی تفاضل ریسک‌های کولبک-لیبلر برای داده‌های کامل و سمت راست نامساوی امید ریاضی تفاضل ریسک‌های کولبک-لیبلر برای داده‌های ناقص است. بنابراین در مقایسه با داده‌های سانسوریده، معیار واگرایی برای داده‌های کامل حساسیت بیشتری از خود نشان می‌دهد و اخذ تصمیم بر اساس داده کامل منجر به تصمیم محافظه کارانه‌تری می‌شود؛ لذا در بخش‌های بعدی استنباط بر اساس تجزیه داده‌های کامل صورت می‌پذیرد.

۳ توسعه آزمون وونگ برای داده‌های سانسوریده

در یک آزمون طول عمر فرض کنید طول عمرهای n مؤلفه، متغیر های تصادفی مستقل $X_1, \ldots, X_n$ از یک توزیع پیوسته و $Y_1 \leq \ldots \leq Y_n$ آماره‌های مرتب نمونه تصادفی باشند. در سانسور راست نوع II مطالعه تا زمان شکست r امین مؤلفه ادامه می‌یابد و تعداد شکست‌ها از قبل تعیین شده است. مثالی از اهمیت کاربرد نمونه

های سانسوریده از راست نوع II این است که به دلیل صرفه جویی در زمان، به جای انتظار کشیدن تا زمان شکست تمام n نمونه با ثبت زمان شکست r مؤلفهٔ اول آزمون متوقف می‌شود. اگر g تابع چگالی، G و $\bar{G} = ۱ - G$ به ترتیب تابع توزیع و تابع بقا یک مدل رقابتی در نظر گرفته شوند، تابع شبه لگاریتم درست‌نمایی داده‌های سانسوریده از راست نوع II به صورت

$$L_n^g(\beta) = \sum_{i=۱}^{r} \log g^\beta(y_i) + (n-r)\log \bar{G}^\beta(y_r)$$

است. برآوردگر شبه درست‌نمایی ماکسیمم، $\hat{\beta}_n$، باید در رابطه

$$L_n^g(\hat{\beta}_n) = \sup_{\beta \in B} L_n^g(\beta)$$

صدق کند. در فضای پارامتر، B، یک $\beta_* = \arg\max_{\beta\in B}\{\frac{۱}{n}L_n^g(\beta)\}$ وجود دارد که مقدار شبه درست پارامتر نامیده می‌شود و $KL(f, g^\beta(X))$ به ازای آن می‌نیمم می‌شود. وایت (۱۹۸۲) مرجع مناسبی برای مطالعهٔ رفتارهای مجانبی برآوردگرهای (شبه) ماکسیمم درست‌نمایی است. در حالتی که مدل رقابتی شامل چگالی درست نباشد $\hat{\beta}_n$ در احتمال به مقدار شبه درست پارامتر، β_*، همگرا می‌شود که از رابطه

$$\beta_* = \arg\max_{\beta\in B}\left\{pE_{\frac{f}{F(\zeta)}}\left[\log g^\beta(Y)\right] + (۱-p)\log \bar{G}^\beta(\zeta)\right\}.$$

به دست می‌آید. این همگرایی منتج از این واقعیت است که $n^{-۱}\sum_{i=۱}^{r}\frac{\partial}{\partial\beta}\log g^\beta(y_i)$ و $n^{-۱}\sum_{i=۱}^{r}\frac{\partial}{\partial\beta}\log \bar{G}^\beta(y_r)$ به ترتیب به $pE_{\frac{f}{F(\zeta)}}\left[\log g^\beta(Y)\right]$ و $(۱-p)\log \bar{G}^\beta(\zeta)$ همگرا خواهند بود. بنابراین برآوردکننده شبه ماکسیمم درست‌نمایی در احتمال به مقدار شبه درست پارامتری همگرا می‌شود که $pE_{\frac{f}{F(\zeta)}}\left[\log g^\beta(Y)\right] + (۱-p)\log \bar{G}^\beta(\zeta)$ را ماکسیمم کند. این پارامتر شبه درست با β_* نشان داده شده است. تفاضل تابع شبه لگاریتم درست‌نمایی دو مدل رقابتی بصورت

$$L_n^{g/h}(\hat{\beta}_n, \hat{\gamma}_n) = L_n^g(\hat{\beta}_n) - L_n^h(\hat{\gamma}_n) = \sum_{i=۱}^{r}\log\frac{g^{\hat{\beta}_n}(y_i)}{h^{\hat{\gamma}_n}(y_i)} + (n-r)\log\frac{\bar{G}^{\hat{\beta}_n}(y_r)}{\bar{H}^{\hat{\gamma}_n}(y_r)}$$

است. توزیع بریده شده از راست یک توزیع شرطی است که به ازای مقادیر کوچکتر از نقطه ζ_n دارای تابع چگالی $f(x|x \leq \zeta_n) = \frac{f(x)}{F(\zeta_n)}$ است. فرض کنید $Y_1 \leq \ldots \leq Y_{r-1}$ متغیرهای تصادفی از یک توزیع بریده شده در Y_r با تابع چگالی $\frac{f(x)}{F(\zeta_n)}$ برای $X \leq \zeta_n$ باشد، که در آن ζ چندک pام توزیع اصلی و $p \in (0,1)$ احتمال ثابت و به این معناست که در جامعه مورد بررسی $100p$ درصد از مؤلفه‌ها شکست خورده و طول عمر آن‌ها ثبت شده است و بقیهٔ مؤلفه‌ها طول عمر بیشتری از طول عمر مؤلفه r ام دارند. r تعداد شکست ها و برابر با جزء صحیح np است. $\zeta_n = Y_r$ چندک نمونه‌ای است که $F(p_n) = \zeta_n$ و $p_n = \frac{r}{n}$ در احتمال همگرا به p است. توزیع مجانبی چندک نمونه‌ای با استفاده از قضیه دلتا به صورت

$$n^{1/2}(\zeta_n - \zeta) \xrightarrow{L} N(0, pqf^{-2}(\zeta))$$

است، که در آن $p = 1 - q$ و $Y_r = \zeta_n \xrightarrow{P} \zeta$. وقتی که مدل خوب توصیف شده است، باتاچاریا (۱۹۸۵) با اثبات دو قضیه توزیع مجانبی برآوردگر درست‌نمایی را برای داده‌های سانسوریده به دست آورد. در ادامه به کمک این دو قضیه و تحت شرایط وایت (۱۹۸۲)، برای حالتی که مدل رقابتی شامل چگالی درست داده‌ها نباشد، توزیع مجانبی برآوردگر شبه ماکسیمم درست‌نمایی به دست آورده شده است و آزمونی بر اساس معادل بودن دو مدل رقابتی یا بهتر بودن یکی از آنها معرفی خواهد شد.

همگرایی جملهٔ اول و دوم، تفاضل تابع شبه لگاریتم درست‌نمایی دو مدل رقابتی به صورت

$$\frac{1}{n}\sum_{i=1}^{r} \log \frac{g^{\hat{\beta}_n}(y_i)}{h^{\hat{\gamma}_n}(y_i)} \xrightarrow{P} pE_{\frac{f}{F(\zeta)}}\left[\log \frac{g^{\beta_*}(Y)}{h^{\gamma_*}(Y)}\right],$$

است و

$$\frac{1}{n}(n-r)\log \frac{\bar{G}^{\hat{\beta}_n}(y_r)}{\bar{H}^{\hat{\gamma}_n}(y_r)} \xrightarrow{P} (1-p)\log \frac{\bar{G}^{\beta_*}(\zeta)}{\bar{H}^{\gamma_*}(\zeta)}.$$

در نتیجه رابطه همگرایی

$$\frac{1}{n}L_n^{g/h}(\hat{\beta}_n, \hat{\gamma}_n) \xrightarrow{P} pE_{\frac{f}{F(\zeta)}}\left[\log \frac{g^{\beta_*}(Y)}{h^{\gamma_*}(Y)}\right] + (1-p)\log \frac{\bar{G}^{\beta_*}(\zeta)}{\bar{H}^{\gamma_*}(\zeta)} \qquad (2)$$

به دست می‌آید. برای یافتن توزیع مجانبی برآوردگر ماکسیمم درست‌نمایی در حالتی که مدل بد- توصیف شده باشد، بسط تیلور $n^{-۱/۲}\frac{\partial L_n^g(\hat{\beta}_n)}{\partial \beta}$ حول β_* را به صورت

$$۰ = n^{-۱/۲}\frac{\partial L_n^g(\hat{\beta}_n)}{\partial \beta} = n^{-۱/۲}\frac{\partial L_n^g(\beta_*)}{\partial \beta} + n^{-۱/۲}(\hat{\beta}_n - \beta_*)'\frac{\partial^۲ L_n^g(\beta_*)}{\partial \beta \partial \beta'} + o_p(۱)$$

در نظر بگیرید. از طرفی

$$n^{-۱/۲}\frac{\partial L_n^g(\beta_*)}{\partial \beta} \xrightarrow{L} N(۰, B_{gc}(\beta_*)),$$

که در آن

$$\begin{aligned} B_{gc}(\beta) &= pE_{\frac{f}{F(\zeta)}}\left[\frac{\partial \log g^\beta(Y)}{\partial \beta}.\frac{\partial \log g^\beta(Y)}{\partial \beta'}\right] \\ &- p\left\{E_{\frac{f}{F(\zeta)}}\left[\frac{\partial \log g^\beta(Y)}{\partial \beta}\right]\right\}\left\{E_{\frac{f}{F(\zeta)}}\left[\frac{\partial \log g^\beta(Y)}{\partial \beta}\right]\right\}' + pqbb', \end{aligned}$$

و

$$b = \frac{\partial \log g^\beta(\zeta)}{\partial \beta} - E_{\frac{f}{F(\zeta)}}\left[\frac{\partial \log g^\beta(Y)}{\partial \beta}\right] + (۱-p)\frac{۱}{f(\zeta)}.\frac{\partial}{\partial y}.\frac{\partial log\bar{G}^\beta(y)}{\partial \beta}|_{y=\zeta}.$$

همچنین

$$-n^{-۱}\frac{\partial^۲ L_n^g(\beta_*)}{\partial \beta \partial \beta'} \xrightarrow{P} A_{gc}(\beta_*),$$

بطوری که

$$A_{gc}(\beta) = -\left\{pE_{\frac{f}{F(\zeta)}}\left[\frac{\partial^۲ \log g^\beta(Y)}{\partial \beta \partial \beta'}\right] + (۱-p)\frac{\partial^۲ \log \bar{G}^\beta(\zeta)}{\partial \beta \partial \beta'}\right\},$$

و

$$n^{-۱/۲}\frac{\partial L_n^g(\beta_*)}{\partial \beta} = A_{gc}(\beta_*)n^{۱/۲}(\hat{\beta}_n - \beta_*).$$

لذا توزیع مجانبی برآوردگر شبه ماکسیمم درست‌نمایی β_* به صورت

$$n^{۱/۲}(\hat{\beta}_n - \beta_*) \xrightarrow{L} N(۰, A_{gc}^{-۱}(\beta_*)B_{gc}(\beta_*)A_{gc}^{-۱}(\beta_*))$$

به دست می‌آید. در صورتی که مدل خوب- توصیف شده باشد، $\beta_* = \beta_\circ$ ($\beta_\circ$ مقدار درست پارامتر است)، $B_{gc} = A_{gc}$ و توزیع مجانبی $\hat{\beta}_n$ به صورت

$$n^{۱/۲}(\hat{\beta}_n - \beta_\circ) \xrightarrow{L} N(\circ, A_{gc}^{-۱}(\beta_\circ))$$

خواهد بود. آماره سمت راست رابطه (۲) دارای واریانسی به صورت

$$W_{*c}^{۲} = Var_{\frac{f}{F(\zeta)}}[\log \frac{g^{\beta_*}(Y)}{h^{\gamma_*}(Y)}] + (۱-p)^{۲} Var_f[\log \frac{\bar{G}^{\beta_*}(Y_r)}{\bar{H}^{\gamma_*}(Y_r)}]$$

است. برای هر $\epsilon > \circ$

$$\begin{aligned} \hat{w}_{nc}^{۲} &= \frac{۱}{r}\sum_{i=۱}^{r}[\log \frac{g^{\hat{\beta}_n}(y_i)}{h^{\hat{\gamma}_n}(y_i)}]^{۲} - [\frac{۱}{r}\sum_{i=۱}^{r}[\log \frac{g^{\hat{\beta}_n}(y_i)}{h^{\hat{\gamma}_n}(y_i)}]]^{۲} \\ &+ (۱-\frac{r}{n})^{۲}\{\frac{۱}{n}\sum_{i=۱}^{n}[\log \frac{\bar{G}^{\hat{\beta}_n}(X_i)}{\bar{H}^{\hat{\gamma}_n}(X_i)} I_{[Y_r-\varepsilon, Y_r+\varepsilon]}(X_i)]^{۲} \\ &- [\frac{۱}{n}\sum_{i=۱}^{n}\log \frac{\bar{G}^{\hat{\beta}_n}(X_i)}{\bar{H}^{\hat{\gamma}_n}(X_i)} I_{[Y_r-\varepsilon, Y_r+\varepsilon]}(X_i)]^{۲}\} \end{aligned}$$

برآوردی برای $W_{*c}^{۲}$ است. اگر سانسور وجود نداشته باشد ($p = ۱$)، با احتمال ۱ طول عمر تمام n نمونه مشاهده خواهد شد. بنابراین $F(\zeta) = p = ۱$ و $r = n$. لذا رابطه (۱) معادل لم ۳.۱ وونگ (۱۹۸۹) و $W_{*c}^{۲}$ و $B_{gc}(\beta)$ و $A_{gc}(\beta)$ به ترتیب واریانس، $B_g(\beta)$ و $A_g(\beta)$ معرفی شده در وونگ خواهند بود که بر اساس داده‌های کامل به دست آمده‌اند.

در مثال زیر به بررسی این موضوع پرداخته می‌شود که برای داده‌های سانسوریده اگر مدل رقابتی شامل چگالی درست داده‌ها نباشد، برآوردگر شبه ماکسیمم درست‌نمایی به مقدار شبه درست همگرا می‌شود. حالت غیر سانسوریده در کاکس (۱۹۶۲) مطالعه شده است.

مثال ۱ : فرض کنید $X_۱, \ldots, X_n$ متغیرهای تصادفی و مستقل از تابع چگالی لگ-نرمال

$$f^{(\mu,\sigma)}(x) = \frac{۱}{\sqrt{۲\pi}x\sigma} e^{-\frac{(\log x - \mu)^{۲}}{۲\sigma^{۲}}}; \quad x > \circ, \quad \mu \in R, \quad \sigma > \circ. \qquad (۳)$$

باشند. مدل رقابتی نمایی با تابع چگالی

$$g^{\beta}(x) = \frac{1}{\beta}e^{-\frac{x}{\beta}}; \quad x > 0, \quad \beta > 0 \qquad (۴)$$

را در نظر بگیرید. لذا

$$pE_f[\frac{\partial}{\partial\beta}(-\log\beta - \frac{Y}{\beta})] + (1-p)\frac{\partial}{\partial\beta}\log(-e^{-\frac{\zeta}{\beta}}) = 0$$

$$pE_{\frac{f}{F(\zeta)}}[-\frac{1}{\beta} + \frac{Y}{\beta^2}] - (1-p)\frac{\zeta}{\beta^2} = 0$$

و

$$-\frac{1}{\beta}\{pE_{\frac{f}{F(\zeta)}}[Y] - (1-p)\zeta\} = p$$

$$\beta_{(\mu,\sigma)} = E_{\frac{f}{F(\zeta)}}[Y] - \frac{(1-p)}{p}\zeta$$

برای محاسبه $E_{\frac{f}{F(\zeta)}}[Y]$ ابتدا انتگرال

$$E_{\frac{f}{F(\zeta)}}[Y] = \frac{1}{p}\int_0^{\zeta} x\frac{1}{x\sqrt{(2\pi)}\sigma}e^{\frac{-(\log x-\mu)^2}{2\sigma^2}}dx$$

را به دست می‌آوریم. پس از تغییرمتغیر مناسب جواب انتگرال به صورت

$$E_{\frac{f}{F(\zeta)}}[Y] = \frac{1}{p}\{e^{\mu+\frac{\sigma^2}{2}}[F(log(\zeta))]\}$$

به دست می‌آید، که در آن F تابع توزیع نرمال با میانگین $\mu+\sigma^2$ و واریانس σ^2 است. در نتیجه مقدار شبه درست برای پارامتر β در حالت سانسور از راست نوع II به صورت

$$\beta_{(\mu,\sigma^2)} = \frac{1}{p}\{e^{\mu+\frac{\sigma^2}{2}}[F(\log(\zeta))] - (1-p)\zeta\} \qquad (۵)$$

خواهد بود. اگر تمام X_iها مشاهده شوند و سانسوری وجود نداشته باشد، $p = 1$ و $\zeta = +\infty$. در این صورت مقدار شبه درست پارامتر، $e^{\mu+\frac{\sigma^2}{2}}$ خواهد بود که

کاکس (۱۹۶۲) برای مشاهدات کامل آن را محاسبه کرد. برای آزمون فرض‌های غیرآشیانه‌ای ($g^{\beta_*}(.) \neq h^{\gamma_*}(.)$)، وقتی داده‌ها سانسوریده هستند باتوجه به نتایج بدست آمده

$$n^{-1/2}\{L_n^{g/h}(\hat{\beta}_n, \hat{\gamma}_n)\} - n^{1/2}\left\{pE_{\frac{f}{F(\zeta)}}\left[\log\frac{g^{\beta_*}(Y)}{h^{\gamma_*}(Y)}\right] + (1-P)\log\frac{\bar{G}^{\beta_*}(\zeta)}{\bar{H}^{\gamma_*}(\zeta)}\right\}$$

در توزیع همگرا به $N(0, W_{*c}^2)$ است. با استفاده از فرضیات وونگ (۱۹۸۹) به منظور بررسی معادل بودن دو مدل رقابتی یا بهتر بودن یکی از مدل‌ها نسبت به مدل دیگر، آزمون فرضیه‌های

$$H_0 : E_{\frac{f}{F(\zeta)}}\left[\log\frac{g^{\beta_*}(Y)}{h^{\gamma_*}(Y)}\right] = 0 \qquad (\log\frac{\bar{G}^{\beta_*}(\zeta)}{\bar{H}^{\gamma_*}(\zeta)} = 0),$$

$$H_g : E_{\frac{f}{F(\zeta)}}\left[\log\frac{g^{\beta_*}(Y)}{h^{\gamma_*}(Y)}\right] > 0 \qquad (\log\frac{\bar{G}^{\beta_*}(\zeta)}{\bar{H}^{\gamma_*}(\zeta)} > 0),$$

$$H_h : E_{\frac{f}{F(\zeta)}}\left[\log\frac{g^{\beta_*}(Y)}{h^{\gamma_*}(Y)}\right] < 0 \qquad (\log\frac{\bar{G}^{\beta_*}(\zeta)}{\bar{H}^{\gamma_*}(\zeta)} < 0).$$

برای داده‌های سانسوریده پیشنهاد می‌شود.

تحت فرض H_0:

$$n^{-1/2}L_n^{g/h}(\hat{\beta}_n, \hat{\gamma}_n)/\hat{w}_{nc} \xrightarrow{L} N(0, 1)$$

تحت فرض H_g:

$$n^{-1/2}L_n^{g/h}(\hat{\beta}_n, \hat{\gamma}_n)/\hat{w}_{nc} \xrightarrow{L} +\infty$$

و تحت فرض H_h:

$$n^{-1/2}L_n^{g/h}(\hat{\beta}_n, \hat{\gamma}_n)/\hat{w}_{nc} \xrightarrow{L} -\infty$$

برای انجام آزمون های فرضیه‌های بالا در سطح معناداری مشخص، اگر مقدار $n^{-1/2}L_n^{g/h}(\hat{\beta}_n, \hat{\gamma}_n)/\hat{w}_{nc}$ بزرگ‌تر از چندک $(1-\alpha)$ام توزیع نرمال

استاندارد، یعنی $Z_{1-\alpha}$ باشد، آنگاه فرض H_0 در مقابل فرض H_g رد می‌شود و مدل g از لحاظ نزدیکی به مدل درست داده‌ها بهتر از مدل h است، اگر مقدار $n^{-1/2}L_n^{g/h}(\hat{\beta}_n,\hat{\gamma}_n)/\hat{w}_{nc}$ کوچک‌تر از $-Z_{1-\alpha}$ باشد، فرض H_0 در مقابل H_h رد می‌شود و مدل h بهتر از مدل g خواهد بود. اگر $|n^{-1/2}L_n^{g/h}(\hat{\beta}_n,\hat{\gamma}_n)/\hat{w}_{nc}| < Z_{1-\alpha}$، نتیجه خواهیم گرفت که هیچ تمایزی بین دو مدل وجود ندارد، یعنی دو مدل رقابتی ازنظر نزدیکی به مدل درست معادل هستند.

۴ انتخاب مدل بر اساس فاصله ردیابی

امید ریسک کولبک-لیبلر قابل تفکیک به دو ریسک، بد-توصیف شدگی مدل و ریسک آماری به صورت

$$EKL(f,g^{\hat{\beta}_n}) = EKL(f,g^{\beta_*}) + EKL(g^{\beta_*},g^{\hat{\beta}_n})$$

است. در صورتی که مدل خوب-توصیف شده باشد $EKL(f,g^{\beta_*}) = 0$. وقتی که دو مدل رقابتی غیر آشیانه‌ایی هستند، کومانژ و همکاران (۲۰۰۸) فاصله ردیابی بااحتمال از پیش مشخص شده‌ای برای مشاهدات کامل به دست آوردند. از بسط تیلور $n^{-1}L_n^g(\beta_*)$ حول $\hat{\beta}_n$ داریم

$$\begin{aligned} n^{-1}L_n^g(\beta_*) &= n^{-1}L_n^g(\hat{\beta}_n) + (\hat{\beta}_n-\beta_*)'n^{-1}\frac{\partial L_n^g(\hat{\beta}_n)}{\partial\beta} \\ &+ \frac{1}{2}(\hat{\beta}_n-\beta_*)'n^{-1}\frac{\partial^2 L_n^g(\hat{\beta}_n)}{\partial\beta\partial\beta'}(\hat{\beta}_n-\beta_*)+o_p(1) \end{aligned}$$

از آنجا که $n^{-1}\frac{\partial L_n^g(\hat{\beta}_n)}{\partial\beta}=0$ و $n^{-1}\frac{\partial^2 L_n^g(\hat{\beta}_n)}{\partial\beta\partial\beta'} \xrightarrow{P} -A_{gc}$ داریم

$$n^{-1}L_n^g(\beta_*) = n^{-1}L_n^g(\hat{\beta}_n) - \frac{1}{2}(\hat{\beta}_n-\beta_*)'A_{gc}(\hat{\beta}_n-\beta_*)+o_p(1).$$

که از امید ریاضی دو طرف آن نسبت به چگالی درست f داریم

$$\begin{aligned} E_f(n^{-1}L_n^g(\beta_*)) &= E_f(n^{-1}L_n^g(\hat{\beta}_n)) - \frac{1}{2}E_f((\hat{\beta}_n-\beta_*)'A_{gc}(\hat{\beta}_n-\beta_*)) \\ &= E_f(n^{-1}L_n^g(\hat{\beta}_n)) - \frac{1}{2}Tr(A_{gc}E((\hat{\beta}_n-\beta_*)(\hat{\beta}_n-\beta_*)') \\ &= E_f(n^{-1}L_n^g(\hat{\beta}_n)) - \frac{1}{2n}Tr(B_{gc}A_{gc}^{-1})+o_p(1). \end{aligned}$$

در نتیجه

$$EKL(f, g^{\beta_*}) = H(f) - E_f(n^{-1} L_n^g(\hat{\beta}_n)) + \frac{1}{2n} Tr(B_{gc} A_{gc}^{-1}) + o_p(1).$$

با استفاده از بسط تیلور $E_f(n^{-1} L_n^g(\hat{\beta}_n))$ حول β_*

$$\begin{aligned} E_f(n^{-1} L_n^g(\hat{\beta}_n)) &= E_f(n^{-1} L_n^g(\beta_*)) + (\hat{\beta}_n - \beta_*)' E_f(n^{-1} \frac{\partial L_n^g(\beta_*)}{\partial \beta}) \\ &+ \frac{1}{2}(\hat{\beta}_n - \beta_*)' E_f(\frac{\partial^2 L_n^g(\beta_*)}{\partial \beta \partial \beta'}).(\hat{\beta}_n - \beta_*) \end{aligned}$$

حال اگر از دو طرف رابطه اخیر نسبت به چگالی درست f امید ریاضی گرفته شود، رابطه‌ای به صورت

$$E_f[E_f(n^{-1} L_n^g(\hat{\beta}_n)) - E_f(n^{-1} L_n^g(\beta_*))] = \frac{1}{2} Tr(A_{gc} E((\hat{\beta}_n - \beta_*)(\hat{\beta}_n - \beta_*)')$$

خواهیم داشت و در نتیجه

$$E_f[E_f(n^{-1} L_n^g(\hat{\beta}_n)) - E_f(n^{-1} L_n^g(\beta_*))] = \frac{1}{2n} Tr(B_{gc} A_{gc}^{-1})$$

و

$$EKL(g^{\beta_*}, g^{\hat{\beta}_n}) = \frac{1}{2n} Tr(B_{gc} A_{gc}^{-1}).$$

امید ریسک کولبک-لیبلر مشاهدات سانسوریده از راست نوع II برای مدل g^{β} به صورت

$$EKL(f, g^{\beta_*}) = H(f) - E_f(n^{-1} L_n^g(\hat{\beta}_n)) + \frac{1}{n} Tr(B_{gc} A_{gc}^{-1}) + o_p(n^{-1})$$

به دست می‌آید. بطور مشابه برای مدل h^{γ} خواهیم داشت

$$EKL(f, h^{\gamma_*}) = H(f) - E_f(n^{-1} L_n^h(\hat{\gamma}_n)) + \frac{1}{n} Tr(B_{hc} A_{hc}^{-1}) + o_p(n^{-1}).$$

امید ریاضی ریسک کولبک-لیبلر به توزیع نامعلوم وابسته است، اما در برآورد تفاضل امید ریسک های کولبک-لیبلر دو مدل رقابتی این وابستگی از بین می‌رود.

اختلاف امید ریسک های کولبک-لیبلر مشاهدات سانسوریده از راست نوع II به صورت

$$\Delta_c(\hat{\beta}_n, \hat{\gamma}_n) = EKL(f, g^{\hat{\beta}_n}) - EKL(f, h^{\hat{\gamma}_n})$$

است، لذا

$$\Delta_c(\hat{\beta}_n, \hat{\gamma}_n) = -E_f[\frac{1}{n}\{L_n^{g/h}(\hat{\beta}_n, \hat{\gamma}_n) - \{Tr(B_{gc}A_{gc}^{-1}) - Tr(B_{hc}A_{hc}^{-1})\}\}]. \quad (۶)$$

با استفاده از تقریب آکائیک، $Tr(B_{gc}A_{gc}^{-1}) \approx p$ و برآوردی برای $\Delta_c(\hat{\beta}_n, \hat{\gamma}_n)$ به صورت

$$D_c(\hat{\beta}_n, \hat{\gamma}_n) = -\frac{1}{n}\{L_n^{g/h}(\hat{\beta}_n, \hat{\gamma}_n) - (p-q)\} \quad (۷)$$

به دست می‌آید، که در آن $p-q$ تفاضل پارامترهای دو مدل رقابتی است.

لم ۱ تحت شرایط عمومی وونگ (۱۹۸۹) فاصله ردیابی با ضریب اطمینان $۱۰۰(۱-\alpha)\%$ برای $\Delta_c(\hat{\beta}_n, \hat{\gamma}_n)$ به صورت (A_n, B_n) است، که در آن $A_n = D_c(\hat{\beta}_n, \hat{\gamma}_n) - Z_{\alpha/۲}n^{-1/۲}\hat{w}_{nc}$ و $B_n = D_c(\hat{\beta}_n, \hat{\gamma}_n) + Z_{\alpha/۲}n^{-1/۲}\hat{w}_{nc}$.

برهان اگر $g^{\beta_*} \neq h^{\gamma_*}$ با توجه به روابط (۲) و (۳)،

$$n^{1/۲}\{D_c(\hat{\beta}_n, \hat{\gamma}_n) - \Delta_c(\hat{\beta}_n, \hat{\gamma}_n)\} \xrightarrow{L} N(۰, W_{*c}^{۲}).$$

لذا

$$\frac{n^{1/۲}\{D_c(\hat{\beta}_n, \hat{\gamma}_n) - \Delta_c(\hat{\beta}_n, \hat{\gamma}_n)\}}{\hat{w}_{nc}} \xrightarrow{L} N(۰, ۱).$$

با توجه به این که $۱ - \Phi(Z_{۱-\alpha/۲}) = \alpha/۲$ (Φ تابع توزیع تجمعی نرمال استاندارد است)، مقادیر A_n و B_n بدست می‌آیند. فاصله به دست آمده دارای این ویژگی است که $P_f(A_n < \Delta_c(\hat{\beta}_n, \hat{\gamma}_n) < B_n) \to ۱ - \alpha$.

فرع ۱ فاصله ردیابی کمک خواهد کرد که مدل‌های رقابتی نسبت به یکدیگر ارزیابی شوند، به این معنا که اگر فاصله به دست آمده شامل صفر باشد، نتیجه گرفته می‌شود که با اطمینان از قبل تعیین شده، دو مدل رقابتی معادل هستند.

۵ مطالعه شبیه‌سازی

برای بررسی توانایی فاصله ردیابی در پیدا کردن مدل بهینه، دو مدل غیر آشیانه‌ایی وایبل و لگ-نرمال را در نظر گرفته‌ایم. مدل وایبل خوب- توصیف شده، فرض شده است. پارامترهای دو مدل برآورد تحت سانسور راست نوع II به شکل زیر برآورد شده‌اند. توزیع وایبل دو پارامتری با تابع چگالی

$$g(x) = \frac{\gamma}{\theta}(\frac{x}{\theta})^{\gamma-1}e^{(-\frac{x}{\theta})^{\gamma}} \qquad x \geq 0, \gamma > 0, \theta > 0$$

در نظر بگیرید. تابع لگاریتم درست‌نمایی برای مشاهدات سانسوریده از راست نوع II به صورت

$$L_n^g = r\log\frac{\gamma}{\theta} + (\gamma-1)\sum_{i=1}^{r}\log\frac{y_i}{\theta} + \sum_{i=1}^{r}(\frac{y_i}{\theta})^{\gamma} - (n-r)(\frac{y_r}{\theta})^{\gamma}$$

است. بعد از حل معادلات $\frac{\partial L_n^g}{\partial\gamma}$ و $\frac{\partial L_n^g}{\partial\theta}$، برآورد پارامترهای توزیع وایبل به صورت

$$\hat{\theta}_n = (\frac{\sum_{i=1}^{r} y_i^{\hat{\gamma}_n} + (n-r)y_r^{\hat{\gamma}_n}}{r})^{1/\hat{\gamma}_n}$$

و

$$\frac{1}{\hat{\gamma}_n} + \frac{\sum_{i=1}^{r}y_i}{r} = \frac{\sum_{i=1}^{r} y_i^{\hat{\gamma}_n}\log y_i + (n-r)y_r^{\hat{\gamma}_n}\log y_r}{\sum_{i=1}^{r} y_i^{\hat{\gamma}_n} + (n-r)y_r^{\hat{\gamma}_n}}$$

به دست می‌آیند. تیکو (۱۹۶۸) نشان داد که حل معادلات درست‌نمایی برای تابع چگالی لگ-نرمال سه پارامتری سانسوریده از راست نوع II مشکل است. با استفاده از تقریب تیکو (۱۹۶۸) برای این تابع چگالی با احتمال سانسور مشاهدات، $q(=1-p)$، برآورد پارامترهای چگالی لگ-نرمال دو پارامتری به صورت

$$\mu = \frac{\frac{1}{n}\sum_{i=1}^{r}\log y_i + q\beta_2\log y_r + q\alpha_2\sigma}{1-q+q\beta_2}$$

و

$$(1-q)\sigma^2 - (q\alpha(\log y_r - \mu))\sigma - (\frac{1}{n}\sum_{i=1}^{r}(\log y_i - \mu)^2 + q\beta_2(\log y_r - \mu)^2) = 0$$

به دست می‌آید، که در آن σ ریشه مثبت معادله دوم است. از طرفی $\beta_۲ = \frac{g(k)-g(h)}{k-h}$ و $Q(h) = q + \sqrt{\frac{۱}{n}q(۱-q)}$ که درآن k و h از معادلات $\alpha_۲ = g(h) - h\beta_۲$ و $Q(k) = q - \sqrt{\frac{۱}{n}q(۱-q)}$ به دست می‌آیند. $g(x)$ عبارت است از $\frac{f(x)}{Q(x)}$، که در آن $f(x) = \frac{۱}{\sqrt{(۲\pi)}}e^{-\frac{x^۲}{۲}}$ و $Q(x) = \int_x^\infty f(x)dx$ است.

جدول ۱: فاصله ردیابی و طول فاصله برای مدل وایبل در مقابل مدل لگ-نرمال برای $p = ۰/۲$ وقتی چگالی درست است.

(γ, θ)	(n, r)		
	(۵۰، ۱۰)	(۱۰۰، ۲۰)	(۱۵۰، ۳۰)
(۱، ۲)	(−۰/۷۶۱، −۰/۴۱۴) ۰/۳۴۶	(−۰/۹۸۴، −۰/۴۴۰) ۰/۵۴۴	(−۰/۸۵۷، −۰/۴۱۵) ۰/۴۴۲
(۱/۴، ۲)	(−۱/۲۶۱، −۰/۳۰۹) ۰/۸۵۱	(−۰/۷۷۹، −۰/۳۵۶) ۰/۴۲۳	(−۰/۹۱۱، −۰/۲۲۵) ۰/۶۸۶
(۴، ۲)	(−۵/۱۹۱، ۷/۳۱۰) ۱۲/۵۰۱	(−۳/۹۱۵، ۵/۰۷۳) ۸/۹۸۹	(−۴/۰۷۶، ۵/۲۵۶) ۹/۳۳
(۱/۷، ۱)	(−۰/۵۹۹، −۰/۳۷۰) ۰/۲۲۹	(−۰/۷۵۰، −۰/۳۷۹) ۰/۳۷۱	(−۰/۸۶۲، −۰/۲۸۶) ۰/۵۷۶
(۰/۸، ۱/۸)	(−۰/۸۲۱، −۰/۱۵۳) ۰/۶۶۸	(−۱/۰۳۷، −۰/۲۴۷) ۰/۷۹۰	(−۱/۰۷۵، −۰/۳۲۱) ۰/۷۵۴
(۰/۸، ۲/۵)	(−۱/۶۴۶، ۱/۰۱۱) ۲/۲۵۷	(−۱/۳۲۱، ۰/۵۲۴) ۱/۸۴۵	(−۱/۳۵۱، ۰/۵۵۶) ۱/۹۰۸
(۰/۸، ۳)	(−۱/۹۳۶، ۱/۵۹۱) ۳/۵۲۷	(−۱/۹۹۷، ۱/۶۰۳) ۳/۶۰۰	(−۲/۸۱۷، ۳/۰۹۳) ۵/۹۱۱
(۴، ۳)	(−۵/۸۰۱، ۸/۳۰۲) ۱۴/۱۰۳	(−۵/۹۲۲، ۸/۶۳۷) ۱۴/۵۶۰	(−۵/۲۶۴، ۷/۳۶۳) ۱۲/۶۲۷

در جدول ۱ برای $(n, r) = (۵۰, ۱۰), (۱۰۰, ۲۰), (۱۵۰, ۳۰)$ در سطح اطمینان ۰/۹۵ فاصله ردیابی برای مقادیر مختلف پارامترهای توزیع وایبل به دست آمده است. داده‌ها از چگالی وایبل با پارامترهای $(\gamma_۰, \theta_۰) = (۰/۸, ۲)$ تولید و مقادیر پارامترهای مدل لگ-نرمال (μ, σ) با استفاده از داده‌ها محاسبه شده‌اند. مقادیری که در متن جدول و در زیر فواصل ردیابی آمده‌اند، طول فاصله متناظر به هر فاصله است. برای مقادیر (γ, θ) نزدیک به $(\gamma_۰, \theta_۰)$ حدود بالا و پایین فاصله منفی است. منفی بودن حدود فاصله به این معنا است که مدل وایبل بهتر از مدل

لگ-نرمال است. وقتی پارامترهای مدل وایبل رقابتی از (γ_0, θ_0) دور می‌شوند، فاصله شامل صفر می‌گردد و طول آن افزایش می‌یابد. برای بررسی دقت فاصله ردیابی با استفاده از 10^4 مشاهده، پارامترهای مدل‌ها برآورد و مشخص شد که همواره مدل درست توسط این فاصله انتخاب می‌شود. در مواردی که منحنی چگالی وایبل توسط منحنی چگالی لگ-نرمال احاطه می‌شود و دو مدل پراکنش تقریباً یکسانی دارند، فاصله ردیابی به، درستی دو مدل را معادل یکدیگر در نظر می‌گیرد.

بحث و نتیجه‌گیری

یک فاصله ردیابی برای داده‌های سانسوریده از راست نوع II معرفی و به کمک شبیه‌سازی نشان داده شد که در سطح اطمینان خوبی این فاصله برای مدل‌های رقابتی غیرآشیانه‌ای، به درستی مدل بهینه را انتخاب می‌کند. در واقع نشان داده شده است که شکل وزنی تفاضل معیارهای آکائیک یک برآوردگر برای مخاطره‌های کولبک-لیبلر تحت سانسور است. تعابیری که برای یک فاصله اطمینان داریم، برای فاصله ردیابی معتبر می‌ماند، جز این که این فاصله برای کمیتی همگرا به تفاضل مخاطره‌های کولبک-لیبلر ساخته شده است که تابعی از n نیز هست. یکی از مشکلاتی که در بررسی مدل‌های غیرآشیانه‌ای وجود دارد، برآورد پارامترهای یک مدل رقابتی تحت مدل رقابتی دیگر است. هنگامی که داده‌ها سانسوریده نیز هستند، موضوع برآورد پارامترها پیچیده‌تر می‌شود. به همین دلائل، ساختن مثال های متنوع فرآیندی پیچیده و در مواردی غیرممکن است.

تقدیر و تشکر

نویسندگان از اصلاحات پیشنهادی داوران محترم که موجب بهبود این مقاله گردید، کمال تشکر و قدردانی را دارند.

مراجع

Akaike, H. (1973), Information Theory and an Extension of Maximum Likelihood Principle, *In Second International Symposium on Information Theory*, 267-281.

Battacharyya, G. K. (1985), The Asymptotics of Maximum Likelihood and Related Estimators based on Type II Censored Data, *Journal of the American Statistical Association*, **80**, 398-404.

Commenges, D. Sayyareh, A., Letenneur, L., Guedj, J. and Bar-Hen, A. (2008), Estimating a Difference of Kullback-Lebler Risk Using a Normalized Difference of AIC, *Annals of Applied Statistics*, **2**, 1123-1142.

Cox, R. (1962), Further Results on Test of Separate Families of Hypothesis. *Journal of the Royal Statistical Society*, **24**, 406-424.

Hafidi, B. and Mkhadri, A. (2007), An Akaike Criterion based on Kullback Symmetric Divergence in the Presence of Incomplete Data, *Afrika Statistika*, **2**, 1-21.

Kullback, S. and Leibler R. A. (1951), On Information and Sufficiency, *Annals of Mathematical Statistics*, **22**, 79-86.

Linhart, H and Zucchini, W. (1986), Model Selection, *Wiley*, New-York.

Sayyareh, A., Obeidi, R. and Bar-Hen, A. (2011), Empirical Comparison between Some Model Selection Criteria, *Communications in Statistics-Simulation and Computation*, **89**, 72-86.

Seghouane, A. K., Bekara, M. and Fleury, G. (2005), A Criterion for Model Selection in the Presence of Incomplete Data based on Kullback's Symmetric Divergence, *Signal Processing*, **85**, 1405-1417.

Shimodaira, H. (1994), A New Criterion for Selecting Models from Partially Observed Data, In Cheeseman, Oldford, R. W. (Eds). *Lecture Notes in Statistics*, **89**, Springer, New York, 21-29.

Tiku, M. L. (1968), Estimating the Parameters of Log-normal Distribution from Censored Sample. *Journal of the American statistical Association*, **63**, No. 321. 134-140.

Vuong, Q. H. (1989), Likelihood Ratio Test for Model Selection and Non-Nested Hypotheses, *Econometrica*, **57**, 307-333.

White, H. (1982), Maximum Likelihood Estimation of Misspecified Models, *Econometrica*, **50**, 1-25