مقایسه آزمون کاکس با آزمون وونگ و معیارهای انتخاب مدل در مدل‌های رگرسیونی خطی غیر آشیانه‌ای (غیر تودرتو) با خطای نرمال

اسداله فقیهی، عبدالرضا سیاره
گروه آمار، دانشگاه رازی

یکی از اساسی ترین مسائل بنیادی در انتخاب مدل بررسی دوری و نزدیکی مدل‌های پیشنهادی به مدل مولد داده‌ها یعنی $h(.)$ است. برای حالتی که دو مدل پیشنهادی وجود دارد بر اساس آزمون‌ها و معیارهای موجود تصمیم می‌گیریم که آیا این دو مدل را به‌عنوان دو مدل خوب در نظر بگیریم یا به عنوان دو مدل که برازش خوبی به داده‌ها ندارند. در فرآیند تصمیم‌گیری گاه دو مدل پیشنهادی را رد می‌کنیم. روش تصمیم‌گیری آزمون کاکس، رد یا پذیرش مدل‌های پیشنهادی بر اساس تعلق $h(.)$ به خانوادهٔ چگالی‌های پیشنهادی است. سوالی که مطرح می‌شود آن است که آیا این دو مدل، که توسط آزمون کاکس رد شده‌اند، دو مدل نزدیک به $h(.)$ بوده‌اند یا دو مدل دور از $h(.)$. از طرفی آزمون وونگ براساس دوری و نزدیکی مدل‌ها به $h(.)$ طراحی شده است. بر این اساس در این مقاله با شبیه‌سازی مدل‌های رگرسیونی خطی غیرآشیانه‌ای با خطای نرمال به بررسی موضوعاتی از این نوع پرداخته‌ایم که از دو مدل پیشنهادی که توسط آزمون کاکس پذیرفته نشده‌اند کدام یک به $h(.)$ نزدیک‌تر بوده‌اند. از طرفی چنانچه آزمون وونگ یا معیارهای انتخاب مدل مانند BIC،AIC،KIC،$KICc$، یکی از مدل‌های پیشنهادی را برگزینند آیا آن مدل توسط آزمون کاکس به عنوان یک مدل خوب-توصیف شده برگزیده می‌شود یا خیر. با بررسی این موضوعات می‌توان نتیجه گرفت که به دلیل حساسیت زیاد آزمون کاکس به ساختار مدل‌های پیشنهادی پذیرش مدل توسط آزمون کاکس به معنی پذیرش توسط همه معیارهاست، و معادل بودن دو مدل توسط آزمون وونگ به مفهوم رد هر دو توسط آزمون کاکس است. اما در مواردی که هر دو مدل توسط آزمون کاکس رد می‌شوند، آزمون وونگ و دیگر معیارها مدل نزدیکتر را برمی‌گزینند.

واژه‌های کلیدی: آزمون کاکس، آزمون وونگ، BIC، AIC، KIC، $KICc$، مدل‌های غیر تودرتو، مدل‌های درست-توصیف شده، مدل‌های خطی.

۱ مقدمه

در بحث انتخاب مدل، آزمون‌ها، معیارها، و روش‌های متنوعی تا به حال پیشنهاد و به کار گرفته شده است.در واقع ضعف‌هاو توانائی‌های این معیارها و آزمون‌ها، پژوهشگران این حوزه را همواره برآن داشته تا به ملاک یا معیار جامع تری دست یابند. بر این اساس در این مقاله دو نوع آزمون متفاوت از لحاظ ساختار و چهار معیار انتخاب مدل به طور هم‌زمان برای انتخاب مدل مورد بررسی قرار گرفته‌اند. همچنین استفاده‌های روزافزون از مدل‌های رگرسیونی خطی غیر تودرتو با خطای نرمال در مدل بندی پدیده‌های اقتصادی ما را برآن داشت تا در بررسی ها و شبیه سازی های این مقاله از این نوع مدل ها استفاده کنیم. شبیه سازی آماره کاکس به دو روش صورت گرفته است. روش اول با استفاده از آماره ارائه شده توسط پسران(۱۹۷۴) انجام گرفته که با نام شبیه سازی از آن یاد خواهیم کرد و دیگری به روشی که پسران و پسران(۱۹۹۳) ارائه کردند و دربخش های بعد به نام شبیه سازی تصادفی از آن نام برده شده است. در بخش (۲) به دلایل و ایده انجام این مقایسه‌ها پرداخته‌ایم. مدل‌های مورد بررسی و آماره آزمون کاکس در حالت رگرسیونی خطی در بخش (۳) آورده شده است. بخش (۴) شامل محاسبه آماره آزمون وونگ در حالت رگرسیونی خطی غیرتودرتو است. محاسبه معیارهای BIC، AIC، KIC، $KICc$ در بخش (۵) انجام گرفته است.در بخش (۶) نحوه شبیه سازی تصادفی و استفاده از آن در مدل‌های رگرسیونی خطی آورده شده است.نحوه شبیه‌سازی مدل‌ها در بخش (۷) مورد بحث قرار گرفته‌اند. بخش (۸) شامل بحث و نتیجه‌گیری در مورد جداول حاصل از شبیه سازی‌ها است و جداول حاصل از شبیه سازی‌ها در بخش (۹) آورده شده‌اند.

۲ انگیزه مقایسه‌ی آزمون‌ها و معیارهای انتخاب مدل

در آزمون‌های کلاسیک رد کردن فرض $H_۰$ به معنی پذیرش فرض مقابل است. در حالیکه در آزمون‌های انتخاب مدل ما با تصمیم گیری‌های متفاوتی در مورد دو مدل روبرو هستیم. در مواردی مایلیم مدلی را به عنوان مدل خوب بپذیریم ولی آزمون کاکس هر دو مدل را رد می‌کند، سوالی که مطرح می‌شود این است که در چنین مواردی چه باید کرد؟ می‌دانیم آزمون کاکس برتری نسبی مدل‌ها نسبت به یکدیگر را مشخص نمی‌کند. رد هر دو مدل توسط

آزمون کاکس نشان می‌دهد که هر دومدل بد_توصیف[1] شده هستند. در این موارد برای انتخاب یک مدل می توان ازآزمون وونگ و معیارهای دیگر انتخاب مدل استفاده کرد. استفاده از آزمون کاکس ما را قادر می‌سازد که هر دو مدل پیشنهادی را پذیرفته و یا رد کنیم. در این مقاله پیشنهاد می‌کنیم اگر آزمون کاکس مدل‌های پیشنهادی را رد کرده باشد بر اساس آزمون وونگ و سایر معیارهای مناسب، مانند AIC از بین مدل‌های پیشنهادی براساس نتیجه آزمون کاکس مدل بهینه‌ای که به توزیع مولد داده‌ها نزدیکتر است را برگزینیم. تفاوت بین آزمون‌های کاکس و وونگ‌در این است که آزمون کاکس یکی از مدل‌های پیشنهادی را در فرض صفر قرار داده و به عنوان مدل درست می‌پذیرد در صورتی که فرض صفر در آزمون وونگ عدم اختلاف بین مدل‌ها از لحاظ نزدیکی به مدل مولد داده‌ااست. برگزیده شدن یک مدل توسط آزمون وونگ به معنی نزدیکتر بودن آن مدل به مدل مولد داده‌ها نسبت به مدل پیشنهادی دیگراست. معیارAIC به‌عنوان نمونه‌ای از معیارهایی که توسط آنها تنها یک عدد مطلق برای مقایسه ارائه می‌شود، در بحث انتخاب مدل فارغ از سطح α بوده و مستقل از فرض صفر به شمار می‌روند. قابلیت دیگر آزمون وونگ نسبت به آزمون کاکس سادگی محاسبه آماره این آزمون است. از طرف دیگربا توجه به آزمون‌های یاد شده و معیاری مانند AIC این سوال مطرح می‌شود، زمانی که یک مدل توسط آزمون وونگ یا معیار AIC به عنوان مدل بهینه انتخاب شده است آیا این مدل انتخابی نسبت به دیگر مدلهای پیشنهادی یک مدل خوب_توصیف[2] شده هست یانه؟ اگر تعلق مدل درست به خانواده‌ی مدل پیشنهادی[3] به عنوان معیاری برای خوبی مدل در نظر گرفته شود، آنگاه آیا مدل انتخابی توسط معیارAIC و آزمون وونگ نیز مدل خوب محسوب می‌شود؟ برای پاسخ به این سوال آزمون کاکس را پیشنهاد می‌کنیم.

۳ آماره آزمون کاکس برای آزمون مدل‌های رگرسیونی خطی غیرتودرتو

تعریف.۱دو مدل رگرسیونی با فرم تابعی وساختارهای خطای یکسان غیرتودرتو هستند اگر وفقط اگر حداقل یک ستون از ماتریس طرح یک مدل رگرسیونی را نتوان به صورت ترکیب خطی از ستون ماتریس طرح مدل رگرسیونی دیگر نوشت.(کلارک،۲۰۰۲)

[1] $Mis - Specified$

[2] $CorrectlySpecify$

[3] $Correctly\ Specification$

مثال ۱ دو مدل رگرسیونی خطی به این شکل در نظر می‌گیریم $H_f: Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_0$ و $H_g: Y = \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3 + \epsilon_1$ همانطور که ملاحظه می‌شود با اعمال محدودیت بر روی بردار پارامترهای مدل g نمی‌توان به مدل f رسید. در واقع ماتریس طرح در مدل f زیر مجموعه‌ای از ماتریس طرح در مدل g نیست.
در اینجا فرض های صفر و مقابل به شکل زیر هستند.

$$H_0: \quad Y = Xb_0 + u_0; \; u_0 \sim N(0, \sigma_0^2 I)$$
$$H_1: \quad Y = Zb_1 + u_1; \; u_1 \sim N(0, \sigma_1^2 I)$$

که درآن Y بردار $n \times 1$ از مشاهدات متغییر وابسته، Z و X ماتریس‌های طرح بترتیب $n \times k_0, n \times k_1$ و b_0, b_1 بردارهای ضرایب رگرسیونی $k_0 \times 1, k_1 \times 1$، u_0, u_1 بردارهای خطای $n \times 1$ و I ماتریس واحد مرتبه n است.

بر اساس مقاله پسران(۱۹۷۴) آماره آزمون کاکس در حالت رگرسیونی خطی برای مدل‌های تحت H_0 و H_1 با فرض درستی مدل تحت H_0 برابر است با $N_0 = T_0/[V_0(T_0)]^{\frac{1}{2}}$ که درآن $T_0 = \frac{n}{2} \log\left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2 + \frac{1}{n}\hat{b}_0' X' M_z X \hat{b}_0}\right)$ همچنین $\hat{V}_0(T_0) = \frac{\hat{\sigma}_0^2}{\hat{\sigma}_{10}^4} \hat{b}_0 X' M_z M_x M_z X \hat{b}_0$ به‌طوری‌که $M_z = I - Z(Z'Z)^{-1}Z'$ و $M_x = I - X(X'X)^{-1}X'$ است. $\hat{b}_0$ و $\hat{\sigma}_0^2$ به ترتیب برآوردگرهای ML، b_0 و σ_0^2 هستند و $\hat{\sigma}_{10}^4$ عبارت است از توان دوم $\hat{E}_0(\hat{\sigma}_1^2)$. کاکس نشان می‌دهد با فرض درستی مدل تحت H_0، N_0 به‌طور مجانبی دارای توزیع نرمال استاندارد است.
حال اگرآماره را با فرض درستی مدل تحت H_1 به دست آوریم $N_1 = T_1/[V_1(T_1)]^{\frac{1}{2}}$ است. در سطح معناداری α آزمون کاکس به یکی از حالت های زیر منتج خواهد شد.

۱. پذیرش H_0 و رد H_1، هرگاه، $|N_1| \geq z_{\alpha/2}$, $|N_0| < z_{\alpha/2}$

۲. رد H_0 و پذیرش H_1، هرگاه، $|N_1| < z_{\alpha/2}$, $|N_0| \geq z_{\alpha/2}$

۳. رد هر دو مدل، هرگاه، $|N_1| \geq z_{\alpha/2}$, $|N_0| \geq z_{\alpha/2}$

۴. پذیرش هر دو مدل، هرگاه، $|N_1| < z_{\alpha/2}$, $|N_0| < z_{\alpha/2}$

۴ آماره‌ی آزمون وونگ برای آزمون مدل‌های رگرسیونی خطی نرمال غیرتودرتو

آزمون وونگ که از لحاظ ساختار با آزمون کاکس متفاوت است، برای بررسی نزدیکی مدل‌های پیشنهادی به مدل درست توسط وونگ(۱۹۸۹) معرفی شد. این آزمون بر پایه اطلاع کولبک_لیبلربنا شده است و آماره‌ی این آزمون برای مدل‌های رگرسیونی خطی غیرتودرتو با خطای نرمال محاسبه شده است. در این آزمون با فرض این که مدل $h(x)$ مدل مولد داده‌ها بوده و $\mathcal{F}_\theta$ و $\mathcal{G}_\gamma$ دو خانواده از مدل‌های پیشنهادی باشند، نزدیکی این دو مدل به مدل $h(.)$ مورد آزمون قرار می‌گیرد. فرض‌های صفر و مقابل در این آزمون به صورت زیر هستند.

$$H_0 \quad : E_h\left\{\log\frac{f(X;\theta_*)}{g(X;\gamma_*)}\right\}=0 \qquad (1)$$

$$H_f \quad : E_h\left\{\log\frac{f(X;\theta_*)}{g(X;\gamma_*)}\right\}>0 \qquad (2)$$

$$H_g \quad : E_h\left\{\log\frac{f(X;\theta_*)}{g(X;\gamma_*)}\right\}<0 \qquad (3)$$

رابطه (۱) به‌این معنی است که $f(x,\theta)$ و $g(x,\gamma)$ معادلند.
رابطه (۲) به‌این معنی است که $f(x,\theta)$ بهتر است از $g(x,\gamma)$.
رابطه (۳) به‌این معنی است که $f(x,\theta)$ بدتر است از $g(x,\gamma)$.
که در آن θ_* و γ_* مقدارهای شبه_درست θ و γ هستند. اگر مدل‌های رگرسیونی را که در بخش قبل معرفی شدند در نظر بگیریم، تحت فرض H_0 آزمون وونگ، توزیع آماره‌ی آزمون به توزیع نرمال استاندارد میل می‌کند. اگر آماره‌ی آزمون وونگ را با V_{fg} نشان دهیم داریم $V_{fg}=n^{-1/2}\hat{L}_{10}(\hat{\alpha}_0,\hat{\alpha}_1)/\hat{\omega}_n$. بنابراین برای محاسبه V_{fg} به محاسبه $\hat{L}_{10}$ و $\hat{\omega}_n$ نیاز داریم. لذا $\hat{L}_{10}$ را می‌توان به صورت $\hat{L}_{10}=\frac{n}{2}\log\frac{\left(\sum_{i=1}^{n}\hat{e}_{1i}^2\right)}{\left(\sum_{i=1}^{n}\hat{e}_{0i}^2\right)}$ نوشت که در آن e_{0i} ها درآیه‌های بردار $e_0=M_x y$ و $e_1=M_z y$ هستند. براساس آماره‌ی وونگ(۱۹۸۹) $\hat{\omega}_n^2$ برابر

است با $\hat{\omega}_n^2 = \frac{n}{4}\sum_{i=1}^{n}\left[\frac{\hat{e}_{1i}^2}{\sum_{i=1}^{n}\hat{e}_{1i}^2} - \frac{\hat{e}_{0i}^2}{\sum_{i=1}^{n}\hat{e}_{0i}^2}\right]^2$، و لذا V_{fg} برابر خواهد شد با

$$V_{fg} = \frac{\log\frac{\left(\sum_{i=1}^{n}\hat{e}_{1i}^2\right)}{\left(\sum_{i=1}^{n}\hat{e}_{0i}^2\right)}}{\left[\sum_{i=1}^{n}\left[\frac{\hat{e}_{1i}^2}{\sum_{i=1}^{n}\hat{e}_{1i}^2} - \frac{\hat{e}_{0i}^2}{\sum_{i=1}^{n}\hat{e}_{0i}^2}\right]^2\right]^{1/2}}$$

طبق وونگ (۱۹۸۹)، تحت فرض صفر، $V_{fg}\xrightarrow{d}N(0,1)$. برای آزمون در سطح معنی داری α اگر $V_{fg} > z_{\alpha/2}$ باشد فرض صفر در مقابل فرض H_f، برتری مدل f بر g رد می‌شود. و اگر $V_{fg} < -z_{\alpha/2}$ آنگاه فرض صفر در مقابل فرض H_g، برتری مدل g بر f پذیرفته می‌شود. در نهایت اگر $|V_{fg}| \le z_{\alpha/2}$ آنگاه نمی‌توان تفاوت معنی داری بین دو مدل قائل شد و فرض صفر به مفهوم معادل بودن دو مدل پذیرفته می‌شود.

۵ محاسبه معیارهای BIC، AIC، KIC، $KICc$ در مدل های رگرسیونی خطی غیرتودرتو

آکائیک (۱۹۷۳) معیار AIC را به صورت $AIC = -2\log f(y;\hat{\alpha}_0) + 2k_0$ تعریف کرد این معیار در واقع برآورد اطلاع کولبک-لیبلر است که به طور مجانبی می توان آن‌را نااریب در نظر گرفت.

کاوانوگ (۱۹۹۹) معیار KIC را به عنوان معیاری متقارن که برآورد حاصل جمع دو اطلاع کولبک-لیبلر است را به صورت $KIC = -2\log f(y;\hat{\alpha}_0) + 3k_0$ معرفی کرد همچنین او در سال (۲۰۰۴) معیار $KICc$ را به عنوان KIC تصحیح شده برای مدل های رگرسیونی خطی نرمال به صورت زیر ارائه کرد

$$KICc = -2\log f(y;\hat{\alpha}_0) + n\log\left(\frac{n}{n-k_0}\right) + \frac{((n-k_0)(2k_0+3)-2)}{(n-k_0-2)(n-k_0)}.$$

معیار اطلاع بیزی BIC (معیار اطلاع شوارتز SIC) توسط شوارتز (۱۹۷۸) ارائه گردید. می توان گفت، BIC معیارارزیابی مدل هایی است که بر حسب احتمال پیشین خود تعریف شده‌اند و بر خلاف بقیه معیارهای گفته شده بر اساس اطلاع کولبک-لیبلر به‌دست نیامده است و به صورت $BIC = -2\log f(y;\hat{\alpha}_0) + k_0\log n$ تعریف می‌شود.

۶ شبیه سازی تصادفی

محدودیت کاربرد آزمون کاکس در نتیجه مشتق‌های پیچیده‌ای است که در محاسبه‌های مربوط به صورت آماره کاکس در موقعیت‌های غیر رگرسیونی وجود دارد. پسران[۴] و پسران[5](۱۹۹۳) روش شبیه سازی تصادفی برای آماره کاکس را که راه حلی برای مسأله مذکور است ارائه کردند. این روش نیازی به محاسبه مشتق های تحلیلی برای به دست برآوردهای شبه-درست در آماره کاکس ندارد. ما از این روش شبیه‌سازی برای پاسخ به سوال‌های مطرح شده در بخش (۲) در مدل‌های رگرسیونی خطی استفاده کرده‌ایم.

۶-۱ آماره آزمون کاکس

در این بخش آماره کاکس به گونه‌ای آورده شده است که بتوان شبیه سازی را ساده‌تر تشریح کرد. فرض کنید $Y_1,\ldots,Y_N$، N مشاهده مستقل و هم‌توزیع باشند. دو توزیع احتمال غیر تودرتو $H_f : Y \sim f(y,\theta)$ و $H_g : Y \sim g(y,\gamma)$ برای Y_tها در نظر گرفته شده به طوری که θ و γ به ترتیب بردارهای به ترتیب $p \times 1$ و $q \times 1$ از پارامترهای نامعلوم تحت H_f و H_g هستند. طبق آماره کاکس(۱۹۶۱،۱۹۶۲) $T_f(\hat\theta,\hat\gamma) = \bar d - \hat E_f(\bar d)$ به طوری که $E_f(.)$ نشان دهنده امید ریاضی تحت H_f، $\bar d = N^{-1}\sum_{t=1}^{N} d_t$ و $d_t = \log f(Y_t,\hat\theta) - \log g(Y_t,\hat\gamma)$ هستند. $\hat\theta$ و $\hat\gamma$ به ترتیب برآوردگرهای (ML)، θ و γ تحت H_f و H_g هستند، $T_f(\hat\theta,\hat\gamma)$ برای نمونه های بزرگ دارای میانگین صفراست، (مک آلر و پسران، ۱۹۸۶).
اگر واریانس مجانبی $\sqrt{n}T_f(\hat\theta,\hat\gamma)$ تحت فرض H_f را با $\nu_f^2(\theta_0,\gamma_*)$ نشان دهیم تحت شرایط نظم (وایت، ۱۹۸۲) آماره استاندارد شده کاکس $N_f(\hat\theta,\hat\gamma) = \sqrt{N}\left(\dfrac{T_f(\hat\theta,\hat\gamma)}{\hat\nu_f(\hat\theta,\hat\gamma)}\right)$ به طور مجانبی دارای توزیع $N(0,1)$ است و $\hat\nu_f(\hat\theta,\hat\gamma)$ برآورد سازگار $\nu_f(\theta_0,\gamma_*)$ است. $\hat\nu_f(\hat\theta,\hat\gamma)$ را می‌توان به صورت $\hat\nu_f^2 = N^{-1}d'\left\{I_N - R(\hat\theta)\left[R'(\hat\theta)R(\hat\theta)\right]^{-1}R'(\hat\theta)\right\}d$ نوشت که در آن $d' = (d_1,d_2,\ldots,d_N)$ و $R(\theta)$ ماتریس $N\times(p+1)$ هستند که برای مدل‌های رگرسیونی

[4] M.Hashem Pesaran

[5] Bahram Pesaran

گفته شده در بخش (۲) به شکل زیر به دست می‌آیند

$$R(\alpha_\circ) = \begin{bmatrix} 1 & -\frac{1}{2\sigma_\circ^2} + \frac{1}{2\sigma_\circ^4}(y_1 - x_1 b_\circ)^2 & \cdots & \frac{x_{k_\circ 1}}{\sigma_\circ^2}(y_1 - x_{k_\circ} b_\circ) \\ 1 & -\frac{1}{2\sigma_\circ^2} + \frac{1}{2\sigma_\circ^4}(y_2 - x_2 b_\circ)^2 & \cdots & \frac{x_{k_\circ 1}}{\sigma_\circ^2}(y_2 - x_{k_\circ} b_\circ) \\ \vdots & \vdots & & \vdots \\ 1 & -\frac{1}{2\sigma_\circ^2} + \frac{1}{2\sigma_\circ^4}(y_n - x_n b_\circ)^2 & \cdots & \frac{x_{k_\circ 1}}{\sigma_\circ^2}(y_n - x_{k_\circ} b_\circ) \end{bmatrix},$$

که در آن x_i، نشان دهنده iـامین سطر ماتریس طرح X است. سطر دوم ماتریس $R(\alpha_\circ)$، $\partial \log f(Y_i,\theta)/\partial\sigma_\circ^2$ است و از سطر دوم به بعد $\partial \log f(Y_i,\theta)/\partial b_{\circ i}$ برای $i = 1, 2, \ldots, k_\circ$ است. در این حالت $\bar{d} = -\frac{1}{2}\log(\frac{\sigma_1^2}{\sigma_\circ^2}) - \frac{1}{2n}\sum_{t=1}^{n}\left[\frac{1}{\sigma_\circ^2}(y_t - x_t b_\circ)^2 - \frac{1}{\sigma_1^2}(y_t - z_t b_1)^2\right]$ و $d_t = -\frac{1}{2}\left[\log(\frac{\sigma_1^2}{\sigma_\circ^2}) + \frac{1}{\sigma_\circ^2}(y_t - x_t b_\circ)^2 - \frac{1}{\sigma_1^2}(y_t - z_t b_1)^2\right]$ هستند. جهت محاسبه $T_f(\hat{\theta}, \hat{\gamma})$ نیاز به محاسبه $\hat{E}_f(\bar{d})$ داریم. با استفاده از معیار کولبک‌ـلیبلر داریم

$$E_f(\bar{d}) = C(\theta_\circ, \gamma_*) = \int \log\left\{\frac{f(y,\theta_\circ)}{g(y,\gamma_*)}\right\} f(y,\theta_\circ)\,dy,$$

که برای مدل‌های رگرسیونی به‌این صورت به‌دست می‌آید

$$C_R\left(\hat{\alpha}_\circ, \hat{\alpha}_{1*}(R)\right) = -\frac{1}{2}\log\left[\frac{\hat{\sigma}_\circ^2}{\hat{\sigma}_{1*}^2(R)}\right] - \frac{1}{2}\left[\frac{\tilde{\sigma}_\circ^2(R)}{\hat{\sigma}_\circ^2} - \frac{\tilde{\sigma}_{1*}^2(R)}{\hat{\sigma}_{1*}^2(R)}\right]$$

وقتی که $\tilde{\sigma}_{1*}^2(R) = \frac{1}{nR}\sum_{j=1}^{R}\left(y_j - Z\hat{b}_{1*}(R)\right)'\left(y_j - Z\hat{b}_{1*}(R)\right)$ و $\tilde{\sigma}_\circ^2(R) = \frac{1}{nR}\sum_{j=1}^{R}\left(y_j - X\hat{b}_\circ\right)'\left(y_j - X\hat{b}_\circ\right)$ هستند. جهت محاسبه $R(\hat{\alpha}_\circ)$ و برآورد آماره‌های گفته شده نیاز به برآورد پارمترهای مدل رگرسیونی داریم که‌آن‌ها را به روش شبیه سازی تصادفی محاسبه می‌کنیم. در روش شبیه سازی جهت محاسبه $\hat{b}_{1*}(\hat{b}_{1\circ})$ و $\hat{\sigma}_{1*}^2(\hat{\sigma}_{1\circ}^2)$ ابتدا نمونه‌های مستقل $n \times 1$ از y با بردار میانگین $(X\hat{b}_\circ)$ و واریانس $(\hat{\sigma}_\circ^2 I)$ تولید می‌کنیم و اگر jـامین نمونه را با y_j، $(j = 1, 2, \ldots, R)$ نشان دهیم و برآوردگرهای شبیه سازی شده از b_{1*} و σ_{1*}^2 برابر هستند با $\hat{b}_{1*}(R) = \frac{1}{R}\sum_{j=1}^{R}\hat{b}_{1j} = (Z'Z)^{-1}Z'\left(\frac{1}{R}\sum_{j=1}^{R}y_j\right)$ و $\hat{\sigma}_{1*}^2(R) = \frac{1}{R}\sum_{j=1}^{R}\hat{\sigma}_{1j}^2 = \frac{1}{nR}\left(\sum_{j=1}^{R}\hat{y}_j' M_z y_j\right)$ که $\hat{b}_{1j} = (Z'Z)^{-1}Z'y_j$ و $\hat{\sigma}_{1j}^2 = n^{-1}(\hat{y}_j' M_z y_j)$

۷ شبیه سازی

در این شبیه سازی فرض شده است که مدل $h: Y = Hb_h + u_h$ با $u_h \sim N(0, \sigma_h^2)$ مدل مولد داده‌ها است. فرض می‌کنیم $b'_h = (2, 5, 6, -8, 7)$ ، $\sigma_h^2 = 1$ و ماتریس طرح H ماتریسی تصادفی و $n \times 5$ در نظر گرفته شده است. همچین دو مدل $f: Y = Xb_0 + u_0$ برای $u_0 \sim N(0, 2)$ و $Y = Zb_1 + u_1$ با $u_1 \sim N(0, 3)$ مدل های پیشنهادی با ماتریس های طرح X و Z هستند که آن‌ها نیز به طور تصادفی تولید شده وبه ترتیب دارای بعد $n \times k_0$ و $n \times k_1$ هستند. این شبیه سازی در محیط S-plus برای آماره آزمون ها و معیار هایی که در بخش های قبل محاسبه شده‌اند انجام گرفته است و برای تعداد نمونه های مختلف کد نویسی شده است. در برنامه نوشته شده تعداد تکرارها برابر $r = 1000$ و تعداد نمونه های تولید شده برابر $n = 10, 30, 100, 200, 300, 500$ در نظر گرفته شده‌اند. فراوانی نسبی حالت های حاصل به عنوان خروجی در جدول های بخش (۹) ارائه شده است. همچنین لازم به ذکر است سطح معناداری برای آزمون‌های کاکس و وونگ ۵٪ در نظر گرفته شده است.

۸ بحث و نتیجه گیری

تحلیل این جداول در جهت پاسخ به سوال‌هایی است که در بخش (۲) مطرح کردیم. در جداول $R.$ مخفف کلمه $Reject$ به معنی رد شدن مدل و $equi.$ مخفف $equivalent$ به منظور معادل بودن دو مدل به کار رفته است. تمامی مدل های مورد بررسی در بخش (۷) تشریح شده‌اند، در این جدول‌ها تعداد پارامترهای هر مدل نسبت به تعداد پارامترهای مدل درست تغییر کرده است. همان‌طور که در بخش (۷) گفته شد تعداد پارامتر مدل درست، پنج است . مدل‌های مورد بررسی با تعداد پارامترهای نزدیک به پنج انتخاب شده است تابه‌توان علاوه بر پاسخ به سوالات مطرح شده، تاثیر افزایش و کاهش تعداد پارامترها را در پذیرش و رد مدل‌ها بررسی کرد. در جدول (۱) مدل‌هایی با $k_0 = 3$ و $k_1 = 8$ مورد بررسی قرار گرفته‌اند به طوری که هیچکدام درست_توصیف شده نیستند. این جدول برای پاسخ به این سوال آورده شده است که اگر آزمون کاکس هر دو مدل را رد کرد کدام مدل به مدل مولد داده‌ها نزدیکتر است؟ در جدول (۱) به دلیل این که هیچ‌کدام از مدل‌ها درست_توصیف شده نیستند و همچنین تعداد پارامتر آن‌ها با تعداد پارامتر مدل درست متفاوت است با افزایش تعداد نمونه از $n = 10$ به $n = 500$ احتمال رد شدن هر دو مدل از ۰/۵۱۷ در $n = 10$ به ۰/۸۳۶ در $n = 500$ می‌رسد. اگر به سطر دوم جدول آزمون کاکس یعنی رد H_f نه H_g توجه کنید در $n = 10$ این احتمال برابر ۰/۴۴۳ است که با افزایش نمونه به $n = 500$ به مقدار ۰/۱۶۲ کاهش می‌یابد و این اتفاق در بقیه آماره‌ها نیز به نوعی رخ می‌دهد در آماره

وونگ در $n =$ ۱۰ احتمال برتری f (۰/۲۹۲) کمتر از احتمال برتری g (۰/۴۹۲) است که با افزایش نمونه حتی از ۱۰ به ۳۰ این احتمالات تفاوت فاحشی پیدا می‌کنند، به‌طوری‌که در $n =$ ۵۰۰ با احتمال ۰/۹۴۷ برتری f محسوس است. برای آماره AIC نیز روند مشابه‌ای تکرار می‌شود. اما در $KICc$ حتی در $n =$ ۱۰ فراوانی نسبی تعداد دفعاتی که $KICc$ برای چگالی f کمتر از $KICc$ برای چگالی g می‌شود، برابر ۰/۹۹۶ است. اما با افزایش تعداد نمونه و کاهش اثر جمله جریمه، مقدار این فراوانی نسبی کم می‌شود. اما باز اختلاف زیادی بین آن دو وجود دارد که موجب پذیرش مدل f می‌شود. در مابقی جدول‌های ارائه شده در این فصل مدل f درست_توصیف شده انتخاب شده است و برای مدل پیشنهادی دو حالت ۳ پارامتری و ۸ پارامتری درنظر گرفته شده است. این کار به این دلیل انجام گرفته است تا علاوه بر پاسخ به سوالات مطرح شده در بخش (۲) بتوان تأثیر کاهش یا افزایش تعداد پارامتر را نیز بررسی کرد. متأسفانه آزمون کاکس در این حالت نیز هر دو مدل را با احتمال بالاتری رد می‌کند، در صورتی که می‌بایست مدل درست_توصیف شده را بپذیرد به همین دلیل در این حالت برای نتیجه‌گیری و مقایسه از نتایج حاصل از شبیه سازی تصادفی آماره آزمون کاکس استفاده شده است. در جدول (۲) نتایج حاصل از شبیه سازی مدل درست_توصیف شده f در مقابل مدل ۳ پارامتری g آورده شده است. آماره آزمون کاکس در این حالت نتایج مورد انتظار را به دست نمی‌دهد. به همین دلیل برای نتیجه‌گیری و مقایسه از جدول (۳) که نتایج حاصل از شبیه سازی تصادفی است استفاده شده است که در همهٔ آماره‌ها و آزمون‌ها برتری مدل f با احتمال بسیار بالا مشخص است. برای مثال در آزمون وونگ حتی برای $n =$ ۱۰ احتمال برتری f برابر ۰/۸۷۲ است.

در جدول (۴) نتایج حاصل از شبیه‌سازی مدل درست_توصیف شده f در مقابل مدل ۸ پارامتری g آورده شده است. لازم به ذکر است به دلیل غیرمنطقی بودن نتایج آزمون کاکس از نتایج جدول(۵) استفاده می‌کنیم. با توجه به مقادیر جدول (۴) و (۵) دوباره برتری مدل f کاملاً مشخص است. می‌توان این نتیجه‌گیری را نیز داشت که افزایش تعداد پارامتر مدل رقیب، از ۳ به ۸ تأثیری در برگزیده شدن مدل درست_توصف شده ندارد. به طور کلی در مورد مدل‌های رگرسیونی خطی غیرتودرتو با خطای نرمال می‌توان این نتیجه را گرفت که زمانی که آزمون کاکس هر دو مدل را به عنوان بد_توصیف شده تشخیص می‌دهد و هر دو مدل را رد می‌کند برای انتخاب مدل باید به نتایج آزمون وونگ و بقیه معیارهای موجود توجه کرد و مدل برگزیده شده توسط آن‌ها را به عنوان مدل بهتر انتخاب کرد. اگر سایر معیارها و آزمون وونگ نیز مدل ها را معادل تشخیص دادند تفاوت معناداری بین مدل ها وجود ندارد. اما باید به یاد داشت که در این حالت هر دو مدل بد_توصیف شده‌اند. در حالتی که آزمون کاکس مدلی را برمی‌گزیند آن مدل به طور حتم خوب_توصیف شده است و باید از آن در استنباط استفاده کرد. اگر چه آزمون وونگ و بقیه معیارها نیز آن مدل را با احتمال بالاتری انتخاب نمایند اما توجه ما در وحله اول به نتیجه آزمون کاکس خواهد بود. آزمون کاکس را می‌توان برای انتخاب مدل های خوب_توصیف شده به کار برد. اگر چه محاسبه آماره آزمون کاکس بسیار مشکل است امامعمولاً به کار گرفته می‌شود و در واقع می‌توان چنین نتیجه‌گیری کرد که مدل انتخابی توسط آزمون کاکس حداقل به خوبی سایر مدل‌های رقابتی است. محاسبه‌ی آماره آزمون وونگ ساده‌تر از محاسبه آماره آزمون کاکس است و به همین

دلیل هم فراگیری بیشتری نسبت به آزمون کاکس دارد. استفاده از معیارهای AIC، BIC، KIC، $KICc$ در کنار این دو آزمون نزدیکی مدل‌های انتخابی را به مدل مولد داده‌ها نشان می‌دهد و با توجه به محاسبه‌ی ساده این معیارها و نتایج مناسب آن ها حتی برای نمونه‌های کوچک می‌توان آن ها را به عنوان معیارهای مناسبی در انتخاب مدل به کار برد.

جدولِ ۱: جدول حاصل از شبیه سازی ($k_۰ = ۳\,, k_۱ = ۸$)

$Statistics$	$Conclusion$	n=۱۰	n=۳۰	n=۱۰۰	n=۲۰۰	n=۳۰۰	n=۵۰۰
Cox	$R.\ H_g$ but not H_f	۰/۰۳۸	۰/۰۷۱	۰/۰۲	۰/۰۰۶	۰/۰۰۱	۰/۰۰۲
	$R.\ H_f$ but not H_g	۰/۴۴۳	۰/۵۱۲	۰/۳۷۲	۰/۲۳۷	۰/۲۱۲	۰/۱۶۲
	$R.$ both H_f and H_g	۰/۵۱۷	۰/۴۱۲	۰/۶۰۵	۰/۷۵۷	۰/۷۸۷	۰/۸۳۶
	$R.$ neither H_f or H_g	۰/۰۰۲	۰/۰۰۵	۰/۰۰۱	۰	۰	۰
$Vuong$	f and g are equi.	۰/۲۱۶	۰/۰۸۸	۰/۰۲۶	۰/۰۳۲	۰/۰۲۸	۰/۰۵۳
	f is better than g	۰/۲۹۲	۰/۸۹۵	۰/۹۷۴	۰/۹۶۸	۰/۹۷۲	۰/۹۴۷
	g is better than f	۰/۴۹۲	۰/۰۱۷	۰	۰	۰	۰
AIC	$AICf < AICg$	۰/۲۹۴	۰/۷۸۹	۰/۸۶۹	۰/۸۶۱	۰/۸۵۱	۰/۸۴۶
	$AICg < AICf$	۰/۷۰۶	۰/۲۱۱	۰/۱۳۱	۰/۱۳۹	۰/۱۴۹	۰/۱۵۴
KIC	$KICf < KICg$	۰/۴۹۸	۰/۹۲۱	۰/۹۷۲	۰/۹۳۷	۰/۹۶۸	۰/۹۷۱
	$KICg < KICf$	۰/۵۰۲	۰/۰۷۹	۰/۰۲۸	۰/۰۲۷	۰/۰۳۲	۰/۰۲۹
$KICc$	$KICcf < KICcg$	۰/۹۹۶	۰/۹۹۴	۰/۹۸۲	۰/۹۷۶	۰/۹۷۲	۰/۹۷۴
	$KICcg < KICcf$	۰/۰۰۴	۰/۰۰۶	۰/۰۱۸	۰/۰۲۴	۰/۰۲۸	۰/۰۲۶
BIC	$BICf < BICg$	۰/۳۶۸	۰/۹۵۷	۱	۱	۰/۹۹۹	۰/۹۹۹
	$BICg < BICf$	۰/۶۳۲	۰/۰۴۳	۰	۰	۰/۰۰۱	۰/۰۰۱

مراجع

[1] Akaike, H. (1973). Information Theory and on Extension of the Likelihood Ratio Princple, *In 2nd International Symposium on Information Theory*, eds B. N. Petrov & F. Csaki, Budapest: Akademiai Kiodo, 267-281.

جدولِ ۲: جدول حاصل از شبیه سازی زمانی که مدل f درست-توصیف شده است. ($k_۰ = ۵\,,\, k_۱ = ۳$)

Statistics	*Conclusion*	n=۱۰	n=۳۰	n=۱۰۰	n=۲۰۰	n=۳۰۰	n=۵۰۰
Cox	$R.\, H_g$ *but not* H_f	۰/۳۶۴	۰/۳۶۴	۰/۲۱۷	۰/۱۶۸	۰/۱۲۴	۰/۱۲۳
	$R.\, H_f$ *but not* H_g	۰/۰۰۴	۰	۰	۰	۰	۰
	$R.$ *both* H_f *and* H_g	۰/۶۲۸	۰/۶۳۶	۰/۷۸۳	۰/۸۳۲	۰/۸۷۶	۰/۸۷۷
	$R.$ *neither* H_f *or* H_g	۰/۰۰۴	۰	۰	۰	۰	۰
Vuong	f *and* g *are equi.*	۰/۱۲۸	۰/۰۱۹	۰/۰۱۰	۰/۰۰۶	۰/۰۰۸	۰/۰۰۶
	f *is better than* g	۰/۸۷۲	۰/۹۸۱	۰/۹۹۰	۰/۹۹۴	۰/۹۹۲	۰/۹۹۴
	g *is better than* f	۰	۰	۰	۰	۰	۰
AIC	$AICf < AICg$	۱	۱	۱	۱	۱	۱
	$AICg < AICf$	۰	۰	۰	۰	۰	۰
KIC	$KICf < KICg$	۱	۱	۱	۱	۱	۱
	$KICg < KICf$	۰	۰	۰	۰	۰	۰
KICc	$KICcf < KICcg$	۰/۷۵۵	۱	۱	۱	۱	۱
	$KICcg < KICcf$	۰/۲۴۵	۰	۰	۰	۰	۰
BIC	$BICf < BICg$	۱	۱	۱	۱	۱	۱
	$BICg < BICf$	۰	۰	۰	۰	۰	۰

جدولِ ۳: جدول حاصل از شبیه سازی تصادفی زمانی که مدل f درست-توصیف شده است. ($k_۰ = ۵\,,\, k_۱ = ۳$)

Statistics	*Conclusion*	n=۱۰	n=۳۰	n=۱۰۰	n=۲۰۰	n=۳۰۰	n=۵۰۰
Cox	$R.\, H_g$ *but not* H_f	۰/۹۷۲	۰/۹۹۷	۱	۱	۱	۱
	$R.\, H_f$ *but not* H_g	۰	۰	۰	۰	۰	۰
	$R.$ *both* H_f *and* H_g	۰/۰۲۸	۰/۰۰۳	۰	۰	۰	۰
	$R.$ *neither* H_f *or* H_g	۰	۰	۰	۰	۰	۰

جدولِ ۴: جدول حاصل از شبیه سازی زمانی که مدل f درست-توصیف شده است. $(k_0 = 5\,,\ k_1 = 8)$

Statistics	Conclusion	n=۱۰	n=۳۰	n=۱۰۰	n=۲۰۰	n=۳۰۰	n=۵۰۰
Cox	R. H_g but not H_f	۰/۴۴۰	۰/۵۷۵	۰/۳۸۸	۰/۲۴۴	۰/۲۴۸	۰/۱۷۹
	R. H_f but not H_g	۰/۰۷۱	۰	۰	۰	۰	۰
	R. both H_f and H_g	۰/۳۹۰	۰/۴۲۵	۰/۶۱۲	۰/۷۵۶	۰/۷۵۲	۰/۸۲۱
	R. neither H_f or H_g	۰/۰۹۹	۰	۰	۰	۰	۰
Vuong	f and g are equi.	۰/۱۹۹	۰/۰۱۴	۰/۰۰۹	۰/۰۱۴	۰/۰۰۸	۰/۰۰۷
	f is better than g	۰/۸۰۱	۰/۹۸۶	۰/۹۹۱	۰/۹۸۶	۰/۹۹۲	۰/۹۹۳
	g is better than f	۰	۰	۰	۰	۰	۰
AIC	$AICf < AICg$	۰/۹۹۱	۱	۱	۱	۱	۱
	$AICg < AICf$	۰/۰۰۹	۰	۰	۰	۰	۰
KIC	$KICf < KICg$	۰/۹۹۳	۱	۱	۱	۱	۱
	$KICg < KICf$	۰/۰۰۷	۰	۰	۰	۰	۰
KICc	$KICcf < KICcg$	۱	۱	۱	۱	۱	۱
	$KICcg < KICcf$	۰	۰	۰	۰	۰	۰
BIC	$BICf < BICg$	۰/۹۹۱	۱	۱	۱	۱	۱
	$BICg < BICf$	۰/۰۰۹	۰	۰	۰	۰	۰

جدولِ ۵: جدول حاصل از شبیه سازی تصادفی زمانی که مدل f درست-توصیف شده است. $(k_0 = 5\,,\ k_1 = 8)$

Statistics	Conclusion	n=۱۰	n=۳۰	n=۱۰۰	n=۲۰۰	n=۳۰۰	n=۵۰۰
Cox	R. H_g but not H_f	۰/۹۷۴	۱	۱	۱	۱	۱
	R. H_f but not H_g	۰	۰	۰	۰	۰	۰
	R. both H_f and H_g	۰/۰۲۶	۰	۰	۰	۰	۰
	R. neither H_f or H_g	۰	۰	۰	۰	۰	۰

[2] Cavanaugh, J.E. (1999). A Large-Sample Model Selection Criterion Based on Kullback's Symmetric Divergence, *Statistics & Probability Letters*, **42**, 333-343.

[3] Cavanaugh, J.E. (2004). Criterion for Linear Model Selection Based on Kullback's Symmetric Divergence, *Statistics & Probability Letters*, **63**, 262-279.

[4] Cox, D.R. (1962). Further Results on Tests of Separate Families of Hypotheses, *Journal of the Royal Statistical Society*, **24**, 406-424.

[5] Lien, D. and Vuong, Q.H. (1989). Selection the Best Linear Regression Model: A Classical Approach, *Social Science Working Paper 606* .

[6] Pesaran, M.H. (1974). On the General Problem of Model Selection, *The Review of Economic Studies*, Vol. 41, NO. 2, pp: 153-171.

[7] Pesaran, M.H. and Pesaran, B. (1993). A Simulation Approach to the Problem of Computing Cox's Statistic for Testing Nonnested Models, *Journal of Econometrics* , **57**, 377-392.

[8] Vuong, Q.H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses, *Econometrica* , Vol. 57, NO. 2, pp: 307-333.