رگرسیون خطی موضعی با داده‌های سری زمانی

رضا رستا

۱۳۹۰/۴/۱۲

چکیده

نام خانوادگی: رستا **نام:** رضا

عنوان پایان‌نامه: رگرسیون خطی موضعی با داده‌های سری زمانی

استاد راهنما: دکتر رحیم چینی پرداز، دکتر علی اکبر راسخی

استاد مشاور: دکتر سید محمد رضا علوی

درجه تحصیلی: کارشناسی ارشد **رشتهٔ:** آمار **گرایش:** ریاضی

محل تحصیل: دانشگاه شهید چمران اهواز **دانشکده:** علوم ریاضی و کامپیوتر

تاریخ فارغ التحصیلی: ۱۳۹۰/۴/۵ **تعداد صفحه:** ۹۴ + هشت‌ص

کلید واژه‌ها: پارامتر هموار سازی، پهنای باند، تاوان قدر مطلق انحرافات جمع شده‌ی هموار، کمترین توان‌های دوم تاوان‌داده، کمترین توان‌های دوم نیم‌رخی تاوان‌داده، ملاک اطلاع بیزی

چکیده:

در بسیاری از مدل‌های آماری که در طول زمان جمع آوری می‌شوند، داده‌های آماری همبسته هستند. در این مدل‌ها روش معمول برای برآورد پارامترهای مدل، استفاده از اطلاعات حاصل از همبستگی بین خطاها در مدل سازی است. در این پایان‌نامه هدف، بررسی نوعی رگرسیون خطی موضعی است که در آن فرض می‌شود خطاهای آماری از یک سری زمانی پیروی می‌کنند. در ابتدا به برخی مفاهیم پایه در رگرسیون و سری‌های زمانی پرداخته‌ایم. همچنین ملاک اطلاع بیزی را به منظور تعیین مقدار مناسب پارامتر میزان سازی در روش کمترین توان‌های دوم تاوان‌داده با تاوان قدر مطلق انحرافات جمع شده‌ی هموار معرفی کرده‌ایم. پس از آن به برآورد توابع رگرسیونی به روش ناپارامتری پرداخته و در پایان‌نامه نوعی رگرسیون خطی موضعی را مورد استفاده قرار داده‌ایم. روش کمترین توان‌های دوم تاوان‌داده رامعرفی کرده و در ادامه با استفاده از مطالب پیشین به هموار سازی Y_t توسط رگرسیون خطی موضعی پرداخته‌ایم. و از روش کمترین توان‌های دوم تاوان داده با تاوان قدر مطلق انحرافات جمع شده‌ی هموار به منظور تعیین مرتبه‌ی فرآیند خطاها که یک فرآیند خودبازگشت بود استفاده کرده‌ایم. سپس با شبیه‌سازی گسترده‌ای برتری این روش را نسبت به روش‌های زیائو و همکاران و روش نیم‌رخی نشان داده‌ایم. یک مثال واقعی در رابطه با اقتصاد کشورمان را با این روش تحلیل کرده‌ایم و به بررسی رابطه‌ی بین شاخص کل بورس و نرخ تورم پرداخته‌ایم. در آخر نیز به نتیجه گیری از پایان‌نامه و ترسیم سر خط پژوهش‌های آینده یک می‌پردازیم.

فهرستِ مطالب

فصلِ ۱

پیشینه،تعاریف و مفاهیم اولیه

۱.۱ مقدمه

روش‌های رگرسیونی کلاسیک همواره به صورت گسترده‌ای مورد استفاده‌ی محققین بوده‌اند و اهمیت و توانایی این‌گونه روش‌ها در مدل‌سازی بر هیچ کس پوشیده نیست اما با گسترش مسائلی که محققین با آن‌ها روبه‌رو هستند به مسائلی برمی‌خوریم که روش‌های کلاسیک قادر به پاسخگویی مناسب به آن‌ها نیستند و آماردانان باید به فکر ابداع روش‌های جدید رگرسیونی باشند. برای مثال در رابطه با روش‌های کلاسیک رگرسیونی با مفروضات زیر مواجهیم:

الف. برای مدل ساختار پارامتری در نظر گرفته می‌شود.

ب. فرض می‌شود خطاها از توزیع نرمال پیروی می‌کنند.

پ. $E(\epsilon) = ۰$

ت. $var(\epsilon) = \sigma^۲$

ث. فرض می‌شود خطاها ناهمبسته هستند.

اما ممکن است در مسائل کاربردی با مواردی برخورد کنیم که برخی از این مفروضات در مورد آن‌ها صادق نباشد. برای مثال خطاها ناهمبسته نباشند و یا یافتن ساختار خطی و یا حتی پارامتری به سادگی مقدور نباشد. هدف ما در این پایان‌نامه پرداختن به مدل رگرسیونی است که در آن برخی از مفروضات مدل‌های کلاسیک نقض شده‌اند.

مدل رگرسیونی $y = m(x) + \epsilon$ را در نظر بگیرید، در این مدل با فرض این‌که خطاها همبسته هستند و یافتن ساختار پارامتری برای تابع رگرسیونی مقدور نیست به این طریق عمل می‌کنیم که:

۱. $m(x)$ را که همان تابع رگرسیونی است به روش ناپارامتری برآورد می‌کنیم.

۲. فرض می‌کنیم خطاها از یک فرآیند خودبازگشت تبعیت می‌کنند.

در رابطه با مدلهای رگرسیونی که در آن‌ها خطاها همبسته هستند پیش‌تر کارهایی صورت پذیرفته است که به صورت مختصر در بخش بعد به آن‌ها اشاره می‌گردد.

۲.۱ پیشینه

در بسیاری از موارد داده‌های آماری که در طول زمان جمع آوری می‌شوند همبسته هستند. در این گونه موارد مایلیم که با استفاده از اطلاعات حاصل از این همبستگی به برآوردهای کاراتری برای پارامترهای مدل دست پیدا کنیم. عده‌ی زیادی از آماردانان در گذشته بدون در نظر گرفتن همبستگی بین خطاها به مدل‌سازی پرداخته‌اند ولی به دلیل در نظر نگرفتن همبستگی بین خطاها نتایج مطلوبی حاصل

نشده است. در گذشته استفاده از اطلاعات حاصل از وجود همبستگی در بین داده‌ها به منظور به دست آوردن برآوردهای دقیق‌تری از پارامترهابه خوبی ذر مدل‌های مربوط به داده‌های طولی و داده‌های پنلی مورد استفاده قرار گرفته است. روش گشتاور تعمیم یافته[۱]، روش برآوردیابی تعمیم یافته‌ی معادلات[۲] و تابع استنباط درجه‌ی دوم [۳] روش‌های متعارف استفاده از اطلاعات حاصل از همبستگی داده‌ها در برآورد پارامترها در مدل‌های رگرسیون پارامتری با داده‌های طولی است.

لین[۴] و کارول[۵] نشان دادند که روش GEE قادر به دخالت دادن اطلاعات حاصل از همبستگی داده‌ها در برآورد هسته‌ی تابع ناپارامتری با داده‌های طولی خوشه‌ای نیست. وانگ روش هسته‌ی حاشیه‌ای را برای داده‌های طولی پیشنهاد داد، این روش کارایی خود را با استفاده از ساختار واقعی همبستگی بهبود می‌بخشد. فن[۶] و همکاران ایده‌ی کمینه ساختن واریانس تعمیم یافته[۷] را برای بهبود کارایی برآورد تابع رگرسیون ناپارامتری در داده‌های طولی مطرح کردند.

جدای از حوزه‌ی داده‌های طولی نیز، بسیاری از آماردانان، رگرسیون ناپارامتری با خطاهای همبسته را مورد مطالعه و بررسی قرار داده‌اند. آپسومر[۸] و همکاران در مقاله‌ای مروری کلی روی این مبحث داشته‌اند که در آن یک مدل رگرسیون ناپارامتری با ماتریس طرح ثابت مورد توجه قرار گرفته است. آن‌ها همچنین در

[1] GMM

[2] GEE

[3] QIF

[4] GMM

[5] GMM

[6] GMM

[7] GMM

[8] GMM

مقاله‌ای به بحث برآورد پارامترها در مدل‌های با داده‌های همبسته پرداختند و نشان دادند که عدم استفاده از اطلاعات حاصل از همبستگی داده‌ها در برآورد پارامترها ممکن است به نتایج نامطلوبی منجر شود. کارهای درخشانی توسط آلتمن[۹] و هارت[۱۰] به صورت جداگانه صورت پذیرفته است که هر دو فرض کرده‌اند همبستگی بین ϵ_t و ϵ_s به شکل $\rho_n(|t-s|)$ باشد.

در ادامه مدلی به صورت $y = m(x) + \epsilon$ را در نظر می‌گیریم، در این مدل $m(x)$ را به صورت ناپارامتری برآورد کرده و خطاها را هم همبسته در نظر گرفته و فرض می‌کنیم از یک فرآیند خودبازگشت تبعیت کنند و با این شرایط به مدل‌سازی متغیر پاسخ در مقابل متغیر توضیحی می‌پردازیم.

در ادامه در فصل اول به برخی مفاهیم و تعاریف اولیه می‌پردازیم. در فصل دوم روش‌های رگرسیون ناپارامتری از جمله روش میانگین موضعی، رگرسیون خطی موضعی، هموارساز هسته‌ای و برآوردگر نادارایا–واتسون رامعرفی می‌کنیم. در فصل سوم روش کمترین توان‌های دوم تاوان‌داده را معرفی کرده و ضمن مروری کلی بر روش‌های مختلف تاوان‌داده، تاوان قدر مطلق انحرافات جمع شده‌ی هموار را معرفی می‌کنیم. در فصل چهارم با استفاده از مفاهیم ارائه شده در سه فصل قبل به برآورد مولفه‌های مدل $y = m(x) + \epsilon$ می‌پردازیم به این صورت که با استفاده از رگرسیون خطی موضعی به برآورد $m(x)$ می‌پردازیم و با به کارگیری روش کمترین توان‌های دوم تاوان‌داده با تاوان قدرمطلق انحرافات جمع شده‌ی هموار به مدل‌سازی خطاها به صورت یک فرآیند خودبازگشت با مرتبه‌ی پایین اقدام می‌کنیم، پس از آن با شبیه‌سازی‌های مفصلی برتری این روش را نسبت به روش‌های مشابه در مدل‌های رگرسیونی ناپارامتری با خطاهای همبسته نشان می‌دهیم. در فصل پنجم نیز با

[9] GMM

[10] GMM

ارائه‌ی یک مثال واقعی از اقتصاد کشورمان، به مدل‌سازی شاخص کل بورس در مقابل نرخ تورم می‌پردازیم، و در پایان، فصل ششم به نتیجه‌گیری و بیان سرخط پژوهش‌های آینده اختصاص یافته است.

۳.۱ رگرسیون

در بسیاری از بررسی‌های علمی، تغییرات یک متغیر به طور وسیعی به متغیرهای دیگری وابسته است. اغلب امکان دارد که با درک و استفاده از طبیعت این متغیرها، به توصیف و استنباط درباره‌ی متغیر اولیه پرداخت. متغیر اولی، معمولا متغیر وابسته یا پاسخ و متغیرهای سری دوم نیز متغیرهای مستقل یا توضیحی نامیده می‌شود. رگرسیون قسمت مهمی از روش‌های آماری است که روابط این متغیرها را فرمول بندی می‌کند. به عبارت دیگر رگرسیون ابزاری است که از آن برای مدل‌سازی متغیر پاسخ به وسیله‌ی یک یا چند متغیر توضیحی دیگر استفاده می‌شود.

۴.۱ سری زمانی

سری‌های زمانی مشاهداتی هستند که در طول زمان جمع آوری می‌شوند. فراوانی چنین مشاهداتی تحلیل سری‌های زمانی را به یکی از کاربردی‌ترین شاخه‌های علم آمار تبدیل کرده است. به عنوان مثال، متوسط درجه‌ی حرارت، رطوبت یا سرعت باد که بسته به نیاز به صورت روزانه، هفتگی یا ماهانه ثبت می‌شوند، قیمت سهامی خاص یا شاخصی کلی در بازار بورس، مقدار تقاضا، تولید یا فروش محصولات یک شرکت، درآمد یک شرکت و مبلغی که این شرکت بابت تبلیغات محصولات

خود صرف می‌کند، تعداد توریست‌ها و درآمد حاصل از این صنعت نمونه‌ای از سری‌های زمانی هستند. به طور کلی یک سری زمانی دنباله‌ای از مشاهدات است که در زمان‌های معلوم ثبت می‌شوند.

تعریف: فرض کنید $(\Omega, \mathcal{F}, P)$ یک فضای احتمال است و T نیز یک مجموعه‌ی شاخص باشد. یک سری زمانی حقیقی مقدار (یا یک فرآیند تصادفی)، یک تابع حقیقی مقدار $X(t, \Omega)$ است که روی $T \times \Omega$ به قسمی تعریف شده است که به ازای هر t ثابت، $X(t, \omega)$ یک متغیر تصادفی روی $(\Omega, \mathcal{F}, P)$ است. تابع $X(t, \omega)$ اغلب به صورت $X_t(\omega)$ یا X_t نوشته می‌شود. یک سری زمانی را می‌توان به صورت مجموعه‌ی $\{X_t : t \in T\}$ از متغیرهای تصادفی در نظر گرفت. به ازای مقادیر ثابت ω، $X(t, \omega)$ یک تابع حقیقی مقدار از t است. این تابع از t یک تحقق نامیده می‌شود. ما نیز در ادامه هرگاه از سری‌های زمانی صحبت می‌کنیم منظورمان سری‌هایی است که با تعریف بالا هم‌خوانی دارند.

۱.۴.۱ سری زمانی ایستا

در تحلیل سری زمانی X_t معمولا بخشی از تحقق این سری در اختیار ما است. در صورتی که خواص سری زمانی X_t در گذر زمان تغییر نکند انتظار داریم که اطلاعات حاصل از این بخش را بتوان در برآورد مشخصه‌های X_t و پیش‌بینی آینده‌ی آن به کار برد. یک سری زمانی با این مطلوبیت که خواصش با انتقال زمان تغییر نکند را ایستا[۱۱] می‌نامیم. در ادامه دو نوع ایستایی را معرفی می‌کنیم.

سری زمانی ایستای قوی: تعریف: سری $\{X_t\}$ را اکیدا ایستا[۱۲] (یا ایستای

[11] Stationary

[12] Strictly stationary

قوی[۱۳]) می‌گوییم هر گاه برای هر n و $k \in \mathcal{R}$ بردارهای تصادفی $(X_{t_1}, \ldots, X_{t_n})$ و $(X_{t_1+k}, \ldots, X_{t_n+k})$ هم توزیع باشند یعنی:

$$F_{X_{t_1},\ldots,X_{t_n}}(x_1, \ldots, x_n) = F_{X_{t_1+k},\ldots,X_{t_n+k}}(x_1, \ldots, x_n)$$

به عبارت دیگر توزیع‌های توام n بعدی تحت انتقال زمان پایا باشند.

نوع ضعیف‌تری از ایستایی نیز وجود دارد. در این نوع ایستایی گشتاورها را ملاک قرار داده و فرض می‌کنیم که بردار میانگین و ماتریس کوواریانس بردار تصادفی $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})'$ با انتقال زمان تغییر نکند.

سری زمانی ایستای ضعیف: تعریف: سری $\{X_t\}$ را ایستای کوواریانس [۱۴] (یا ایستای ضعیف [۱۵]) می‌گوییم هرگاه برای هر t، $E(X_t^2) < \infty$ و تابع میانگین X_t ثابت و تابع اتوکوواریانس تابعی از فاصله‌ی زمانی باشد. یعنی:

$$\mu_{X_t} = \mu$$

و

$$C_x(t, t+h) = g_x(|h|) = \gamma_x(h)$$

که در روابط بالا $\mu_{X_t} = E(X_t) = \mu$ و $Var(X_t) = E(X_t - (\mu_{X_t}))^2$ و $Cov(X_t, X_s) = E(X_t - \mu_{X_t})(X_s - \mu_{X_s})$.

تعریف: فرض کنید Z_t یک فرآیند نوفه‌ی سفید باشد به این معنی که متغیری تصادفی و مستقل بامیانگین صفر و واریانس ثابت σ^2 باشد. از این فرآیند در مدل سازی سری‌های $ARMA$ استفاده می‌شود.

[13] Strongly stationary

[14] Covariance stationary

[15] Weakly stationary

۵.۱ سری‌های $ARMA$

سری‌های $ARMA$ که از ترکیب جملات اتورگرسیو و میانگین متحرک حاصل می‌شوند غالبا با مرتبه‌ای پایین (p و q کوچک) می‌توانند رفتار بسیاری از سری‌های زمانی را که مشاهده می‌کنیم تبیین کنند. این سری‌ها به صورت زیر تعریف می‌شوند.

تعریف: سری ایستای $\{X_t\}$ صادق در معادله‌ی تفاضلی زیر را سری اتورگرسیو-میانگین متحرک از مرتبه‌ی (p,q) می‌نامند

$$X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

که در آن $\{Z_t\} \sim WN(0, \sigma^2)$ فرآیند نوفه‌ی سفید است و به اختصار با نماد $X_t \sim ARMA(p,q)$ نشان می‌دهیم. سری $\{X_t\}$ را $ARMA(p,q)$ با میانگین μ می‌گوییم هرگاه $X_t - \mu \sim ARMA(p,q)$ باشد.

مدل فوق مدل $ARMA$ی مرتبه‌ی (p,q) نامیده می‌شود. با استفاده از عملگر پسبرنده می‌توان این مدل را به شکل

$$\varphi_p(B) X_t = \theta_q(B) Z_t$$

نوشت، که در آن عملگرهای

$$\varphi_p(B) = 1 - \varphi_1 B - \cdots - \varphi_p B^p$$

و

$$\theta_q(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$$

به ترتیب عملگرهای اتورگرسیو مرتبه‌ی p و میانگین متحرک مرتبه‌ی q نامیده می‌شوند.

۱.۵.۱ سری خودبازگشت

در صورتی که $q = ۰$، $(\theta_q(B) = ۱)$ باشد مدل $ARMA$ به صورت زیر در می آید:

$$X_t - \varphi_۱ X_{t-۱} - \cdots - \varphi_p X_{t-p} = Z_t$$

یا

$$\varphi_p(B) X_t = Z_t$$

در این حالت سری $\{X_t\}$ را سری خودبازگشت محض مرتبه‌ی p نامیده و با نماد $\{X_t\} \sim AR(p)$ نشان داده می‌شود.

فرآیند خطی وارون‌پذیر: سری $ARMA$ی $\{X_t\}$ را (نسبت به $\{Z_t\}$) عکس پذیر گوییم هر گاه دنباله‌ی به طور مطلق جمع پذیر $\{\pi_j\}$ موجود باشد به طوری که برای هر t،

$$Z_t = \sum_{j=۰}^{\infty} \pi_j X_{t-j} = \pi(B) X_t.$$

شرط لازم و کافی برای عکس پذیری X_t این است که ریشه‌های $\theta(B) = ۰$ خارج دایره‌ی واحد باشند. در این صورت ضرایب π_j از رابطه‌ی $\pi(B) = \frac{\varphi(B)}{\theta(B)}$ یا $\pi(B)\theta(B) = \varphi(B)$ به دست می آیند.

۶.۱ مدل‌های طرح تصادفی و طرح ثابت

- مدل ثابت

در این گونه مدل‌ها متغیرهای توضیحی X ثابت و قابل کنترل می‌باشند. این گونه مدل‌ها مربوط به مشاهداتی هستند که در آن‌ها متغیرهای X قابل کنترل هستند.

• مدل تصادفی

در این گونه مدل‌ها متغیرهای X، متغیرهای تصادفی هستند. این مدل ها مربوط به مشاهداتی هستند که متغیرهای X غیر قابل کنترل باشند.

• مدل آمیخته

در بسیاری از موارد کاربردی از مشاهداتی استفاده می‌شود که در آن‌ها ترکیبی از متغیرهای ثابت و تصادفی وجود دارد. برای مثال در مطالعه‌ی تمرکز لایه‌ی اوزون در جو بعضی از متغیرهای پیش‌گو ممکن است اثر فصل، روز هفته، زمان در شبانه روز و موقعیت (مکان) باشند که تمامی این‌ها معلوم و ثابت هستند در حالی‌که سایر متغیرهای توضیحی مانند متغیرهای مربوط به هواشناسی و تمرکز آلاینده‌ها مانند اکسید نیتروژن و هیدروکربن‌ها متغیرهای تصادفی هستند.

۷.۱ ملاک اطلاع بیزی

با توجه به این که در این پایان‌نامه به منظور محاسبه‌ی پارامتر میزان‌سازی λ [۱۶] در روش کمترین توان‌های دوم تاوان‌داده[۱۷] با تاوان قدر مطلق انحرافات جمع شده‌ی هموار، ملاک اطلاع بیزی مورد استفاده قرار گرفته است لذا توضیح مختصری در ارتباط با این روش ارائه می‌شود. روش کمترین توان‌های دوم تاوان‌داده با تاوان قدر مطلق انحرافات جمع شده‌ی هموار و پارامتر میزان‌سازی در فصل سوم مورد بررسی قرار خواهند گرفت.

ملاک اطلاع بیزی [۱۸] یا به اختصار BIC به وسیله‌ی شوآرتز[۱۹] (۱۹۷۸) به

[16] Tuning parameter

[17] Penalized least square

[18] Bayesian information criterion

عنوان رقیبی برای ملاک اطلاع آکائیک [۲۰](۱۹۷۳) ارائه شد. شوآرتز ملاک اطلاع بیزی خود را به عنوان یک تقریب مجانبی برای احتمال پسین بیزی[۲۱] یک مدل آماری معرفی کرد.

در نمونه‌های بزرگ، مدلی که با به کار گیری ملاک اطلاع بیزی ترجیح داده می‌شود همانند مدلی است که بیشترین مقدار توزیع پسین را دارد، یعنی مدلی که بر مبنای داده‌های موجود بیشترین احتمال درست بودن را داراست. محاسبات مربوط به BIC بر مبنای لگاریتم تابع درست‌نمایی تجربی هستند. بنا بر این نیازی به مشخص کردن توزیع‌های پیشین ندارند. معمولا مقایسه‌ی دو به دوی مدل‌ها توسط روش بیز بر مبنای به کار گیری عامل بیز[۲۲] صورت می‌گیرد.

دو مدل را در نظر بگیرید که دارای توزیع پیشین برابر باشند، عامل بیز نسبت احتمال‌های پسین این دو مدل است. مدلی که با توجه به توزیع پسین محتمل‌تر است به وسیله‌ی این که عامل بیز آن کوچک‌تر یا بزرگ‌تر از یک است مشخص می‌گردد. لذا انتخاب مدل بر اساس ملاک اطلاع بیزی معادل با انتخاب مدل بر اساس عامل بیز است (کاس و رفتری، ۱۹۹۵). بنابراین ملاک اطلاع بیزی در بسیاری از مسائل مدل‌سازی بیزی که قرار دادن پیشین‌ها به صورت دقیق مشکل است، از مطلوبیت خاصی برخوردار است.

ملاک اطلاع بیزی را به صورت مختصر در دو قسمت مورد بررسی قرار می‌دهیم.

الف. مروری بر ملاک اطلاع بیزی

[19]Schwarz

[20]Akaike information criterion(AIC)

[21]Bayesian posterior probability

[22]Bayes factor

ب. به دست آوردن ملاک اطلاع بیزی

به منظور مروری کلی بر ملاک اطلاع بیزی ابتدا باید برخی اصطلاحات را معرفی کنیم.

مدل مولد یا مدل واقعی که با$g(y)$ نشان داده می شود. مدل کاندیدا یا مدل تقریبی که با$f(y|\theta_k)$ نمایش می دهیم. کلاس مدل‌های کاندیدا که با $\{f(y|\theta_k)|\theta_k \in \Theta(k)\}$ نمایش داده می‌شود و در آخر مدل برازش داده شده که با $f(y|\hat{\theta}_k)$ نمایش می‌دهیم.

همان طور که در پایین ملاحظه می‌گردد، عبارت مربوط به نیکویی برازش یعنی $-۲\ln f(y|\hat{\theta}_k)$ در هر دوی ملاک‌های AIC و BIC یکسان است.

ملاک اطلاع آکائیک: $$AIC = -۲\ln f(y|\hat{\theta}_k) + ۲k$$

ملاک اطلاع بیزی: $$BIC = -۲\ln f(y|\hat{\theta}_k) + k\ln n$$

بنا بر این وجه افتراق این دو روش تنها در عبارت مربوط به تاوان می‌باشد. عبارت تاوان مربوط به BIC در فرآیند گزینش مدل نسبت به AIC حساس‌تر است. چرا که برای $n \geq ۸$ عبارت $k.\ln(n)$ که مربوط به BIC است همواره از $۲k$ مقدار بیشتری به خود می‌گیرد. بنا بر این BIC نسبت به AIC گرایش بیشتری به برگزیدن مدل‌های کوچک‌تر دارد. ملاک اطلاع بیزی در بسیاری از موارد ملاک اطلاع شوآرتز نیز نامیده می‌شود و با حروف اختصاری BIC، SIC، SBC و SC نمایش داده می‌شود.

با انتخاب مدل کاندیدا با مقدار کمینه‌ی ملاک BIC، سعی در انتخاب مدل کاندیدا با بیشترین مقدار احتمال پسین بیزی داریم. ملاک اطلاع بیزی توسط شوآرتز (۱۹۷۸) برای مشاهدات مستقل و هم توزیع و مدل‌های خطی تحت این فرض که حداکثر درست‌نمایی از یک خانواده‌ی نمایی منظم آمده باشد ارائه شد. ما در این‌جا توجیهی برای BIC می‌آوریم که حالت عمومی دارد.

با فرض این که بردار $\boldsymbol{Y}$ داده‌های مشاهده شده باشند[۲۳]، فرض می‌کنیم که این داده‌ها توسط مدل M_k توصیف می‌شود و M_k از مجموعه‌ای از مدل‌های کاندیدای $\{M_{k_1}, \ldots, M_{k_L}\}$ برای برازش به داده‌ها انتخاب می‌شود. فرض کنید که هر M_k به صورت یکتا به وسیله‌ی بردار θ_k، پارامتر بندی می‌شود. که θ_k عنصری از فضای پارامتر $\Theta_k(k \in \{k_1, \ldots, k_L\})$ است. فرض کنید $L(\theta_k|y)$ بیانگر تابع درست‌نمایی Y بر مبنای M_k باشد که به صورت $L(\theta_k|y) = f(y|\theta_k)$ نمایش داده می‌شود. فرض کنید $\hat{\theta}_k$ بیانگر برآوردگر حداکثر درست‌نمایی θ_k باشد که به وسیله‌ی بیشینه سازی $L(\theta_k|y)$ روی $\Theta(k)$ حاصل شده است. همچنین فرض می‌کنیم که مشتقات $L(\theta_k|y)$ نسبت به θ_k تا مرتبه‌ی دو، وجود داشته باشد و به ازای تمامی مقادیر $\theta_k \in \Theta(k)$ پیوسته و کران دار باشند. فرض کنید $g(\theta_k|k)$ نشان دهنده‌ی توزیع احتمال پیشین θ_k در مدل $M_k(k \in \{k_1, \ldots, k_L\})$ باشد.

با استفاده از قضیه‌ی بیز می‌توان تابع چگالی توام پسین M_k و θ_k را به صورت

$$h((k, \theta_k)|y) = \frac{\pi(k)g(\theta_k|k)L(\theta_k|y)}{m(y)}$$

نوشت که در این رابطه $m(y)$ نشان دهنده‌ی توزیع حاشیه‌ای y است. بر مبنای روش گزینش مدل بیزی، مدل M_kای انتخاب می‌شود که دارای احتمال پسین بیشتری باشد.

احتمال پسین برای M_k به صورت زیر است:

$$P(k|y) = m^{-1}(y)\pi(k) \int L(\theta_k|y)g(\theta_k|y)d\theta_k.$$

حال با توجه به این که کمینه‌سازی عبارت $-2\ln P(k|y)$ معادل با بیشینه‌سازی عبارت $P(k|y)$ است، بنابراین:

$$-2\ln P(k|y) = 2\ln\{m(y)\} - 2\ln\{\pi(k)\} - 2\ln\{\int L(\theta_k|y)g(\theta_k|k)d\theta_k\}$$

[23] Observed data