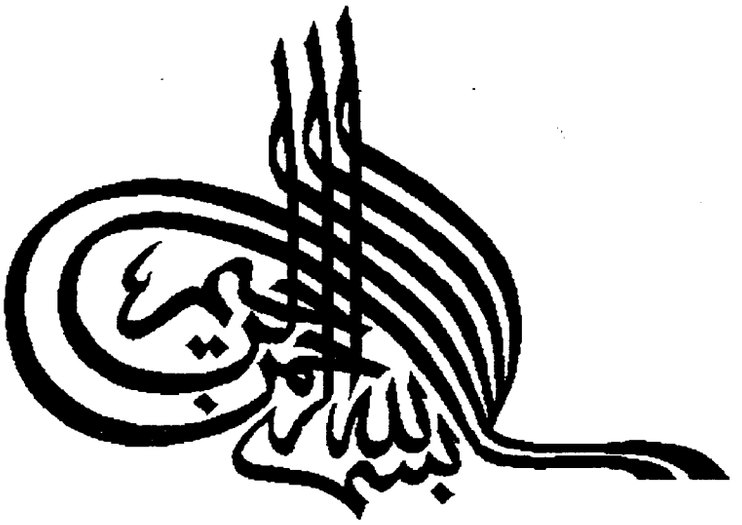

دانشگاه شهید بهشتی

دانشکده علوم ریاضی

گروه آمار

پایان نامه کارشناسی ارشد آمار

عنوان:

نیرومندی در رگرسیون بیزی

استاد راهنما:

دکتر سیامک نوربلوچی

نگارش:

علی بابایی

شهریور ۱۳۷۸

بسمه تعالی

تاریخ
ردیف
پیوست

صورتجلسه دفاع از پایان نامه

جلسه هیئت داوران ارزیابی پایان‌نامه آقای/ خانم/ علی بابائی

به شناسنامه شماره ۴۰ صادره از اراک متولد ۱۳۵۲

دانشجوی دوره کارشناسی‌ارشد ناپیوسته رشته آمار گرایش محض

باعنوان نیرومندی بیزی در رگرسیون بیزی

به راهنمائی آقای دکتر نوربلوچی طبق دعوت قبلی درتاریخ ۷۸/۶/۳۱

تشکیل گردید وبراساس رأی هیأت داوری و با عنایت به ماده ۲۰ آئین‌نامه کارشناسی‌ارشد مورخ ۷۳/۱۰/۲۵ پایان‌نامه مزبور با نمره ۱۸ تمام (هیجده تمام) و درجه عالی مورد تصویب قرارگرفت.

۱- آقای دکتر سیامک نوربلوچی استاد راهنما

۲- " " عبدالرحیم شهلایی استاد مشاور

۳- " " علی عمیدی استاد داور

۴- " " محمد ذکایی استاد داور

۵-

تقدیم:

به **مادرم** که در لحظه لحظه دوران تحصیلم چشمهای نگرانش بدرقه راهم بود، او که همواره امید را در دلم زنده نگه میداشت و فردایی روشن را نوید میداد.

و **پدرم** که همواره چون کوهی استوار تکیه گاهم بودند.

و **برادرم** که همه تحصیلاتم را مدیون محبت‌های ایشان هستم.

و **خواهرانم** که همواره غمخوارم بودند.

سپاس و قدردانی

دکتر شریعتی در وصیت به پسرش می‌نویسد" پسرم تنها چیزی که برای تو در این مسیری که عمر نام دارد آرزو می‌کنم ، برخورد با یکی دو دل بزرگ، یکی دو روح عظیم است، یک بالاترین عدد ممکن است، دو را برای وزن کلام آورده‌ام " و من چقدر خوشبختم که به آن یک، برخوردم، آن هم در بهترین ایام زندگی، جوانی. آری آن یک، کسی نیست جز استادی بزرگوار، انسانی متعهد نیک سرشت و دوست داشتنی، جناب آقای دکتر نوربلوچی که در تمام ایام که در خدمت ایشان بودم نه تنها در زمینهٔ علمی از محضر ایشان کسب فیض نمودم بلکه به مثابهٔ شاگردی که تازه پا در کلاس درس می‌گذارد قدم در کلاس درس اخلاق ایشان گذاشتم و بهره‌های فراوان از حسن خلق و حس انسان دوستی ایشان بردم. امیدوارم این عزیز همواره پاینده باشند تا فرزندان این مرز و بوم از وجود ایشان بهره‌مند شوند.

جا دارد از کلیه اساتید گروه آمار خصوصاً آقای دکتر عبدالرحیم شهلایی بخاطر قبول زحمت مشاورت و از آقای دکتر ذکایی که مرحمت نموده و داوری پایان نامه اینجانب را پذیرفته‌اند تشکر و قدردانی نمایم.

همچنین آقای دکتر علی عمیدی و آقای دکتر محمد رضا مشکانی که براستی تجسم عینی معلم بودن هستند چه بخاطر زحماتی که در مطالعه و داوری پایان نامه و چه در دوران تحصیل اینجانب متحمل شده‌اند صمیمانه تشکر و قدردانی می‌نمایم.

بر خود لازم میدانم که از کمکهای بی شائبهٔ دوست عزیزم آقای محمدرضا اناری صمیمانه تشکر کنم. همچنین از مساعدت دوستان عزیزم در دوران تحصیل خصوصاً آقایان قدرت روشنایی، حبیب اسماعیلی، مهدی امانی، سعدا.. مرادی و مهدی ابراهیم زاده قدرانی می‌نمایم .

پیشگفتار

با پیدایش مکتب بیزین‌ها در علم آمار، همه شاخه‌های آماری تحت تأثیر این مکاتب قرار گرفت در این رهگذر مساله رگرسیون نیز از این تأثیر مصون نماند و مسأله رگرسیون بیزی مطرح شد.

عموماً در رگرسیون بیزی از دو نوع توزیع پیشین برای پارامترهای مدل استفاده شده است: پیشین جفریز و پیشین نرمال - گاما یا نرمال - گامای وارون. در مساله رگرسیون بیزی به خاطر تعداد زیاد پارامترها انتخاب پیشین مشکل است، این امر انگیزه‌ای برای ورود به بحث نیرومندی بیزی در رگرسیون شده است. در این رساله چند صورت از توزیعهای پیشین که نتایج رگرسیون بیزی نسبت به آنها نیرمند است را مورد مطالعه قرار داده‌ایم. ساختار این رساله به شرح زیر است:

در فصل اول با مروری کوتاه بر رگرسیون کلاسیک وارد رگرسیون بیزی می‌شویم ابتدا مساله رگرسیون بیزی یک متغیره با پیشین ناآگاهنده (جفریز) را مورد بررسی قرار داده و سپس مساله را در حالت رگرسیون چند گانه با پیشین مذکور مطرح می‌کنیم. در ادامه فصل با در نظر گرفتن پیشین نرمال - گاما برای پارامترهای مدل تحلیل‌های مربوطه را ارائه کرده‌ایم.

در فصل دوم برای کامل بودن بحث، نیرمندی بیزی را بطور مستقل مطرح کرده‌ایم بدین صورت که ابتدا مفهوم نیرومندی بیزی را بیان کرد، و سپس نتایج نیرومندی بیزی را در رده‌های خاص مثل ردهٔ ε - آلوده‌ها با انواع آلاینده‌ها بررسی کرده‌ایم.

در فصل سوم مساله نیرومندی در رگرسیون بیزی در ردهٔ بخصوص از پیشین‌ها را مطرح کرده‌ایم. ابتدا رگرسیون بیزی را با پیشین آمیخته نرمال - گامای

واردن و نا آگاهنده عنوان کرده و نیرومندی نتایج مدل نسبت به انتخاب پیشین پایه را نشان داده‌ایم. در قسمت بعد مساله را با انتخاب رده ء پیشین ε - آلوده وقتی توزیع پایه و آلاینده یک نوع خاص از g- پیشین‌هاست پی‌گیری کرده و نتایج نیرومندی را ارائه داده‌ایم.

امید است که این رساله بتواند مبحث نیرومندی بیزی در رگرسیون بیزی را به جامعه آماری کشور معرفی کرده باشد.

فهرست

عنوان **صفحه**

عنوان **صفحه**

عنوان **صفحه**

فصل اول

رگرسیون خطی بیزی پارامتری

مقدمه:

رگرسیون کلاسیک مبتنی بر روش کمترین توانهای دوم و نرمال بودن توزیع خطاها گسترش یافته است. نظریهٔ بیزی در برخورد با مسأله رگرسیون نیز حداقل سابقه‌ای ۳۰ ساله دارد. مسأله نیرومندی بیزی در رگرسیون بیزی موضوع اصلی این رساله است. در فصل حاضر با رگرسیون کلاسیک و همچنین نظریهٔ بیزی آن آشنا می‌شویم. مراجع اصلی ما در این فصل زلنر[۱] (۱۹۷۱)، برومللینگ[۲] (۱۹۸۵) و پرس[۳] (۱۹۸۹) است.

این فصل دارای ساختار زیر است. بخش اول را با توجه به دانش عامی که درباره رگرسیون کلاسیک موجود است صرفاً جهت کامل بودن بحث و معرفی نمادها معرفی کرده‌ایم و در بخش دوم با توجه به نمادهای بخش قبل، رگرسیون بیزی پارامتری را مطرح نموده‌ایم. به صورت دقیقتر این بخش‌ها عبارتند از:

I: رگرسیون کلاسیک

۱) رگرسیون یک متغیرهٔ کلاسیک

۲) رگرسیون چندگانهٔ کلاسیک

II: رگرسیون بیزی

الف: وقتی σ^2 معلوم است.

1- Zellner

2- Bromelling

3- Press

ب : وقتی σ^2 مجهول است.

۱) رگرسیون سادهٔ بیزی (با پیشین نا آگاهنده)

۲) رگرسیون چندگانه بیزی پارامتری (ناتجربی) { با پیشین نا آگاهنده / با پیشین مزدوج نرمال - گاما }

در بخش اخیر مباحث مربوطه به روشهای کلاسیک استنباطی یعنی مسألهٔ برآورد کردن و آزمون فرضیه‌ها و همچنین مسألهٔ پیشگوئی مقادیر آینده را مطرح می‌کنیم و نتایج بدست آمده در این حوزه‌ها را مرور می‌نمائیم. قضایا و روشهای مطرح در این فصل ابزاری ضروری برای فصلهای آینده این رساله است.

۱-۱- رگرسیون کلاسیک ساده و چندگانه

۱-۱-۱- رگرسیون کلاسیک ساده

در مدل رگرسیون خطی ساده یک متغیر وابسته و یک متغیر مستقل داریم و مدل رگرسیونی به صورت زیر مطرح است.

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \quad , \quad i = 1, 2, \ldots, n \qquad (1-1)$$

Y_i مقدار متغیر وابسته از i-امین مشاهده و x_i مقدار i-امین متغیر مستقل است. فرض می‌شود که متغیر X تحت کنترل آزمایشگر است. ε_i ها را خطای تصادفی یا ماندهٔ مدل رگرسیونی می‌نامند.

فرضهای مدل رگرسیونی ساده

۱- متغیر تصادفی ε_i دارای توزیعی با میانگین صفر و واریانس ثابت σ^2 است.

۲- ε_i ها دوبدو ناهمبسته‌اند: $\forall_{i\neq j}\ \ Cov(\varepsilon_i, \varepsilon_j) = 0$

β_1 را عرض از مبدأ و β_2 را شیب خط رگرسیونی می‌نامند. اگر $x=0$ در حوزهٔ مدل باشد β_1 مقدار Y به ازای $x=0$ است و β_2 نتغییر Y به ازای یک واحد تغییر در x است.

از مقدمات فوق توزیع y به شرط x عبارتست از:

$$Y_i \mid x_i \sim N(\beta_1 + \beta_2 X_i, \sigma^2) \qquad , \qquad i = 1, \ldots, n$$

که در آن داریم: $Cov(Y_i \mid x_i, Y_i \mid x_i) = 0 \qquad \forall\, i \neq j$

برآورد پارامترهای مدل به روش کمترین توانهای دوم خطا[1] (LSE)

در این روش ایدهٔ اصلی این است که خط رگرسیونی را طوری برآورد کنیم که فاصله عمودی مشاهدات از خط بدست آمده کمترین باشد. معیاری که بدین منظور تعریف می‌کنیم به صورت زیر است:

$$Q(\beta_1, \beta_2) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$

Q تابعی از β_1 و β_2 است. مقادیر $\hat{\beta}_1$ و $\hat{\beta}_2$ رایج در رگرسیون یعنی:

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (۲-۱)$$

$$\hat{\beta}_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

1- Least Square Estimation

با مشتق گیری از $Q(\beta_1,\beta_2)$ نسبت به β_1 و β_2 بدست آمده است.

برآورد پارامترها به روش ماکسیمم درستنمایی[1] (ML) :

اگر فرض نرمال بودن ε_i ها را بپذیریم، لگاریتم تابع درستنمایی عبارتست از:

$$\log L(\beta_1,\beta_2,\sigma^2) = \frac{-n}{2}(\log 2\pi + \log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_1-\beta_2 X_i)^2$$

که چنانچه آن را نسبت به β_1 و β_2 و σ^2 ماکسیمم کنیم. خواهیم داشت:

$$\hat{\sigma}^2 = MSE = \frac{1}{n}\sum(y_i-\hat{\beta}_1-\hat{\beta}_2 x_i)^2 = \frac{SSE}{n}$$

(۳-۱)

$$\hat{\beta}_2 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})} \quad , \quad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}$$

که همان برآوردهای کمترین توانهای دوم قبلی است.

به علاوه می‌توان نشان داد که اگر σ^2 معلوم باشد داریم:

$$\hat{\beta}_2 \sim N(\beta_2, \frac{\sigma^2}{\sum(x_i-\bar{x})^2})$$

(۵-۱)

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}))$$

از نتایج فوق می‌توان برای انجام آزمون فرضیه و به دست آوردن فواصل اطمینان برای β_1 و β_2 استفاده کرد.

1- Maximum Liklihood

۱-۱-۲- رگرسیون کلاسیک چندگانه

در این مدل تغییرات متغیر وابستهٔ Y به چندین متغیر مستقل مربوط می‌شود. فرض کنید Y متغیر وابسته و $X_1, \ldots, X_{k-1}$ متغیرهای مستقل باشند. مدل رگرسیونی به صوت زیر است.

$$Y_i = \beta_1 + \beta_2 x_i + \ldots + \beta_{k-1} x_{i\,k-1} + \varepsilon_i \quad , \quad i = 1, \ldots, n \qquad (۱-۶)$$

که برای سهولت آن را به دو صورت ماتریسی زیر می‌نویسیم.

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon} \qquad (۱-۷)$$

$$\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad , \quad X_n = \begin{bmatrix} 1 & x_{11} & x_{12} \ldots x_{1k-1} \\ 1 & x_{21} & x_{22} \ldots x_{2k-1} \\ & & \vdots \\ 1 & x_{n1} & x_{n2} \ldots x_{n\,k-1} \end{bmatrix} \quad , \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad , \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{bmatrix}$$

فرض می‌کنیم:

$$\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I_n)$$

در نتیجه داریم: $E(\underline{y}) = X\underline{\beta} \quad , \quad Var(\underline{y}) = \sigma^2 I_n \quad , \quad \underline{\varepsilon} = \underline{y} - X\underline{\beta}$

گیریم: $Q(\underline{\beta}) = \underline{\varepsilon}'\underline{\varepsilon} = (\underline{y} - X\underline{\beta})'(\underline{y} - X\underline{\beta})$

که از حداقل کردن آن داریم:

$$\hat{\underline{\beta}} = (X'X)^{-1} X'\underline{y} \qquad (۱-۸)$$

$$SSE = \sum_{i=1} e_i^2 = (\underline{y} - X\hat{\underline{\beta}})'(\underline{y} - X\hat{\underline{\beta}}) = \underline{y}'\underline{y} - \hat{\underline{\beta}}\, X'\underline{y} \qquad (۱-۹)$$

با فرض نرمال بودن توزیع $\underline{\varepsilon}$ خواهیم داشت:

$$\hat{\underline{\beta}} \sim N(\underline{\beta}, \sigma^2 (X'X)^{-1}) \qquad (۱-۱۰)$$