مرکزتهران
دانشکدهٔ علوم پایه

پایان نامه

برای دریافت درجهٔ کارشناسی ارشد

در رشتهٔ آمار ریاضی

عنوان پایان نامه:

برآورد پارامترهای مدلهای رگرسیون مرتبه ای با روشهای متفاوت

استاد راهنمای اول:

دکتر علی شادرخ

استاد راهنمای همکار:

دکتر پرویز نصیری

نگارش:

علی خاشعی ورنامخواستی

تیر ماه ۱۳۸۹


بسم الله الرحمن الرحيم

مرکزتهران
دانشکده علوم پایه

پایان نامه

برای دریافت درجهٔ کارشناسی ارشد

در رشتهٔ آمار ریاضی

عنوان پایان نامه:

برآورد پارامترهای مدلهای رگرسیون مرتبه ای با روشهای متفاوت

استاد راهنمای اول:

دکتر علی شادرخ

استاد راهنمای همکار:

دکتر پرویز نصیری

نگارش:

علی خاشعی ورنامخواستی

تیر ماه ۱۳۸۹

جمهوری اسلامی ایران
وزارت علوم تحقیقات و فناوری

مجتمع علوم پایه و کشاورزی

دانشگاه پیام نور
دانشگاه پیام نور استان تهران

شماره: .....
تاریخ: .....
پیوست: .....

تصویب نامه

پایان نامه کارشناسی ارشد دررشته: آمار ریاضی

تحت عنوان:

"برآورد پارامترهاي مدلهاي رگرسيون مرتبه اي باروشهاي متفاوت"

تاریخ دفاع : ۸۹/۰۴/۲۹ **ساعت:** ۱۰-۱۱

نمره پایان نامه: ۱۸,۵ **درجه ارزشیابی:** عالی

اساتیدهیات داوران	نام و نام خانوادگی	مرتبه علمي	امضاء
استادراهنما	دکتر علي شادرخ		
استادمشاور	دکترپرویزنصیري		
استاد داور	دکتراحسان جمالي		
نماینده علمي گروه	دکترمسعودیارمحمدي		

تهران، خیابان استاد نجات الهی، خیابان شهید فلاح پور، پلاک ۲۷
تلفن: ۸۸۸۰۰۲۵۲
دورنگار: ۸۸۳۱۹۴۷۵
www.tpnu.ac.ir
science.agri@tpnu.ac.ir

باتشکر و قدردانی از اساتید گرامی

جناب آقای دکتر علی شادرخ

و

جناب آقای دکتر پرویز نصیری

که مرا در تهیه و تنظیم این تحقیق یاری نموده و اینجانب را مورد لطف و عنایت خود قرار دادند.

نام خانوادگی دانشجو: خاشعی ورنامخواستی **نام:** علی
عنوان پایان نامه: برآورد پارامترهای مدلهای رگرسیون مرتبه ای باروشهای متفاوت
استاد راهنما: دکتر علی شادرخ
استاد راهنمای همکار: دکتر پرویز نصیری
مقطع تحصیلی: کارشناسی ارشد **رشته:** آمار **گرایش:** آمار ریاضی **دانشگاه:** پیام نور مرکز تهران
دانشکده: علوم پایه **تاریخ فارغ التحصیلی:** **تعداد صفحه:**

چکیده:

در بسیاری از مطالعات علمی و پژوهشی مایل به بررسی تأثیرگذاری متغیرها بر هم و یا پیش بینی یک متغیر به وسیلهٔ گروهی از متغیرهای دیگر هستیم. در مباحث آماری این عمل اغلب با استفاده از مدلهای رگرسیونی امکان پذیر است. امّا برای استفاده از این مدلهای رگرسیونی پیش فرضهائی نیاز است. از جملهٔ این پیش فرضها استقلال مشاهدات و داده هاست، که باید مدنظر گرفته شود. البته ممکن است که این فرض برای همهٔ مشاهدات بر قرار نباشد. یعنی مشاهدات نسبت به هم وابسته بوده و یا ساختاری تودرتو داشته باشند. پس در این شرایط نیاز به استفاده از روشهای رگرسیونی مناسب دیگری غیر از روشهای استاندارد می باشد.

در حالت وجود همبستگی بین مشاهدات و یا داشتن ساختاری تودرتو، مدلهای جدیدی به نام مدلهای رگرسیون مرتبه ای و یا به عبارت دیگر رگرسیون سلسله مراتبی می تواند برای برازش داده ها و کسب نتایج قابل قبولتر، مفید باشد. در این تحقیق به معرفی مدلهای رگرسیون مرتبه ای (سلسله مراتبی) و پارامترهای موجود در آن پرداخته ایم. سپس با استفاده از روشهای مختلف آماری از قبیل برآوردهای درستنمائی ماکسیمم، کمترین مربعات، بیز تجربی، نیم بیز و تمام بیز پارامترها را برآورد کرده و مورد بررسی و مقایسه قرار داده ایم و سعی در بهبود و اصلاح آنها نموده ایم.

در نهایت با ارائهٔ چند نمونهٔ عملی و کاربردی از داده های واقعی ، تلاش کرده ایم که با مدلسازی رگرسیون مرتبه ای و برآورد پارامترهای آن، نتایج قابل قبولتری نسبت به مدلهای رگرسیون معمولی نشان دهیم. نتایج کامل این تحقیق در متن پایان نامه مورد توجه قرار گرفته است.

واژه های کلیدی: رگرسیون سلسله مراتبی– داده های تودرتو– برآورد کمترین مربّعات – برآورد درستنمائی ماکسیمم– برآورد بیزتجربی– برآورد نیم بیز – برآورد تمام بیز.

فهرست مطالب

فهرست جداول

فهرست اشکال و نمودارها

۱

مدلهای رگرسیونی

۱-۱ مقدمه

تحلیل رگرسیونی ما را قادر می سازد که بین یک متغیر مورد نظر که آن را متغیر وابسته یا پاسخ می نامیم و یک یا چند متغیر مستقل یا پیش بین رابطه ای را معین کنیم و از آن استفاده نمائیم. مثلاً، در یک مطالعهٔ رگرسیونی، متغیر پاسخ، قد افراد برحسب سانتیمتر و متغیر مستقل، وزن افراد برحسب کیلوگرم در یک زمان معین است.

تحلیل رگرسیونی را غالباً برای پیش بینی متغیر پاسخ و یا اساساً برای بررسی وابستگی بین متغیرهای مستقل و متغیر پاسخ به کار می برند. مثلاً، مایلیم قد افراد را با استفاده از آگاهیی که از وزن آنها داریم پیش بینی کنیم و یا بررسی کنیم که آیا با افزایش وزن افراد، قد آنها نیز افزایش می یابد یا نه.

عموماً متغیر پاسخ با Y و متغیر مستقل با X نشان داده می شود. همانطور که بیان شد تحلیل رگرسیونی به ما این امکان را می دهد که بین دو متغیر X وY رابطه ای برقرار کنیم. وقتی مقدار X مشخص باشد و مقدارY به صورتی یکتا معین شود، بین X وY رابطه ای دقیق برقرار شده است که این رابطه، یک رابطهٔ تابعی بین دو متغیر X وY است. اگر براساس مقادیر X وY نمودار پراکنشی رسم شود، بدلیل اینکه مقدار Y به صورتی یکتا از روی مقدار X تعیین شده، تمام مشاهدات بر خط رابطهٔ بین X وY می افتند. حال اگر مقدار X مشخص باشد و مقدارY به صورتی یکتا معین نشود، رابطهٔ بین X وY دقیق نخواهد بود که این رابطه یک رابطهٔ آماری بین دو متغیر X وY است. برای توصیف این رابطه اگر نمودار پراکنش مشاهدات رسم شود و از محل تمرکز و تجمع نقاط خطی رسم شود، دیده می شود که این رابطه، رابطه ای کامل نیست و

مشاهدات حول این خط پراکنده اند. این دو نوع رابطهٔ تابعی و آماری بین دو متغیر، نسبت به مشاهدات X و Y به صورت خطی یا غیر خطی نشان داده می شوند. در این فصل، از برخی مفاهیم پایه ای تحلیل رگرسیونی بحث خواهیم کرد.

۱-۲ رگرسیون خطی ساده

اگر رابطهٔ بین دو متغیر X و Y آماری باشد، مدل رگرسیونی برای مشاهدهٔ Y_i به صورت زیر است:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad (۱-۱)$$

مؤلفهٔ $\alpha + \beta X_i$ که رابطهٔ آماری را منعکس می کند، مؤلفهٔ رگرسیونی می خوانند و مؤلفهٔ ε_i که پراکنش تصادفی را منعکس می کند، خطا می نامند. Y_i هم متغیری تصادفی است، به این دلیل که هر مشاهدهٔ Y_i ، شامل یک مؤلفهٔ تصادفی ε_i است.

۱-۲-۱ مدل رگرسیون خطی[۱]

حال مدل رگرسیونی را که در آن رابطهٔ آماری خطی است مورد بررسی قرار می دهیم:

$$Y_i = \alpha + \beta x_i + \varepsilon_i \qquad i = ۱,۲, \dots, n \qquad (۲-۱)$$

که در آن:

- Y_i، پاسخ در i امین مورد است.
- x_i، مقدار متغیر مستقل در i امین مورد، مقداری ثابت و معلوم فرض شده است.
- α و β ، پارامترند.

---

۱) Linear regression model

- ε_iها، متغیرهای مستقل دارای توزیع نرمال $N(۰, \sigma^۲)$ هستند.

مدل (۱-۲) را مدل رگرسیون خطی ساده می نامند. این مدل شامل سه پارامتر α ،β و$\sigma^۲$ می باشد.

۱-۲-۲ تابع رگرسیون خطی[۱]

می خواهیم مقدار امید Y_i را بدست آوریم. با توجه به شرایطی که در بالا ذکر کردیم، داریم:

$$E(Y_i|x) = E(\alpha + \beta x_i + \varepsilon_i) = E(\alpha + \beta x_i) + E(\varepsilon_i) = \alpha + \beta x_i + E(\varepsilon_i)$$

از آنجایکه $E(\varepsilon_i) = ۰$ است، در پایان با توجه به معلوم بودن x ،به نتیجهٔ زیر می رسیم.

$$E(Y_i|x) = \alpha + \beta x_i \qquad (۳-۱)$$

رابطهٔ بین x و$E(Y_i)$ را تابع رگرسیونی می نامند. پارامتر α ، عرض از مبدأ خط رگرسیونی و پارامتر β ، هم شیب خط رگرسیونی است.

حال می خواهیم واریانس $\mathrm{Y_i}$ را بیابیم.

$$Var(Y_i|x) = Var(\alpha + \beta x_i + \varepsilon_i)$$

از آنجایکه $\alpha + \beta x_i$ مقداری ثابت است، داریم:

$$Var(Y_i|x) = Var(\varepsilon_i) = \sigma^۲ \qquad (۴-۱)$$

پس صرفنظر از مقادیر x_iها، Y_iها تغییرپذیری یکسانی دارند.

چون Y_i تابعی خطی از متغیر تصادفی نرمال ε_i است، می توان گفت که هر Y_i هم به صورت نرمال توزیع شده است و سرانجام، چون ε_iها برای مشاهدات مختلف مستقل فرض شده بودند، Y_iها نیز چنین خواهند بود. بنابراین Y_iها متغیرهای تصادفی نرمال مستقلی هستند.

---

۱) Linear regression function

$$Y_i \sim N(\alpha + \beta x_i , \sigma^2) \qquad (۵-۱)$$

۱-۲-۳ برآورد α و β و σ^2

معمولاً پارامترهای رگرسیونی α و β و σ^2 مجهول اند و باید از داده های نمونه برآورد شوند. بنابراین برای برآورد این پارامترها از برآوردگرهایی مانند برآوردگر کمترین مربّعات[۱] و برآوردگر درستنمائی ماکسیمم[۲] استفاده خواهیم کرد.

الف) برآوردگرهای کمترین مربّعات

اگر نمودار پراکنشی از مشاهدات را در نظر بگریم و بخواهیم خط راستی را به عنوان بهترین برازش از این داده ها رسم کنیم، روش کمترین مربّعات، مجموع توانهای دوم انحرافهای مشاهدات Y از خط راست را بعنوان معیار برازش در نظر می گیرد. در اینجا این مجموع را با Q نشان می دهیم. روش کمترین مربّعات، خطی را بعنوان بهترین خط راست برازش داده شده در نظر می گیرد که برای آن مقدار Q یعنی مجموع توانهای دوم انحرافها کمترین مقدار ممکن را نشان دهد.

برآوردهای کمترین مربّعات α و β را به ترتیب با a و b نشان می دهیم، بنابراین خط رگرسیونی برازش داده شده با برآوردهای کمترین مربّعات برابر خواهد بود با:

$$\hat{Y}_i = a + bX_i$$

پس برای مقدار Q داریم:

$$Q = \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2 \qquad (۶-۱)$$

مطابق با اصل کمترین مربّعات ، مقادیر a و b را باید طوری تعیین کنیم که Q کمینه شود.

---

۱) Least squares estimator

۲) Maximum likelihood estimator

برای بدست آوردن برآوردهای a و b ، یک راه تحلیلی وجود دارد. برای ارائهٔ ساده تر این راه حل مجموعهای زیر را تعریف می کنیم:

$$\boldsymbol{S}_X^2 = \sum(X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2 \qquad (۷-۱)$$

$$S_Y^2 = \sum(Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2 \qquad (۸-۱)$$

$$S_{XY} = \sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - n\bar{X}\bar{Y} \qquad (۹-۱)$$

که در آن،

- $\bar{X}$ و $\bar{Y}$ ، میانگینهای مقادیر X_iها و Y_iها،
- $\boldsymbol{S}_X^2$ و $\boldsymbol{S}_Y^2$ ، مجموع توانهای دوم انحرافها از میانگینها،
- $\boldsymbol{S}_{XY}$ ، مجموع حاصلضربهای برداری انحرافها،

ابتدا می نویسیم

$$Y_i - (a + bX_i) = (Y_i - \bar{Y}) - b(X_i - \bar{X}) + (\bar{Y} - a - b\bar{X})$$

با مربع کردن دو طرف داریم:

$$(Y_i - a - bX_i)^2 = (Y_i - \bar{Y})^2 + b^2(X_i - \bar{X})^2 + (\bar{Y} - a - b\bar{X})^2$$

$$-2b(X_i - \bar{X})(Y_i - \bar{Y}) - 2b(X_i - \bar{X})(\bar{Y} - a - b\bar{X}) + 2(Y_i - \bar{Y})(\bar{Y} - a - b\bar{X})$$

هر دو طرف رابطه را بر روی مقادیر $i = 1,2, \dots, n$ جمع بندی می کنیم. چون

$$\sum(X_i - \bar{X}) \quad = 0$$

$$\sum(Y_i - \bar{Y}) \quad = 0$$

با توجه به معادلات (۱-۷)،(۱-۸) و (۱-۹) بنابراین داریم:

$$Q = \boldsymbol{S}_Y^2 + b^2\boldsymbol{S}_X^2 + n(\bar{Y} - a - b\bar{X})^2 - 2b\boldsymbol{S}_{XY}$$

می توانیم جملات را اینگونه نیز مرتب کنم،

$$Q = n(\bar{Y} - a - b\bar{X})^2 + (b^2\boldsymbol{S}_X^2 - 2b\boldsymbol{S}_{XY} + \frac{\boldsymbol{S}_{XY}^2}{\boldsymbol{S}_X^2}) + \boldsymbol{S}_Y^2 - \frac{\boldsymbol{S}_{XY}^2}{\boldsymbol{S}_X^2}$$

$$Q = n(\bar{Y} - a - b\bar{X})^2 + (b^2 \boldsymbol{S}_X^2 - 2b\boldsymbol{S}_{XY} + \frac{\boldsymbol{S}_{XY}^2}{\boldsymbol{S}_X^2}) + \boldsymbol{S}_Y^2 - \frac{\boldsymbol{S}_{XY}^2}{\boldsymbol{S}_X^2}$$

اگر a و b موجود در رابطه را برابر با $a = \bar{Y} - b\bar{X}$ و $b = \frac{\boldsymbol{S}_{XY}}{\boldsymbol{S}_X^2}$ قرار دهیم، دو جملهٔ اول را به کوچکترین مقدار، یعنی صفر کاهش خواهیم داد.

بنابراین برآوردکننده های کمترین مربّعات برای α و β که از اینجا به بعد آنها را با $\hat{\alpha}$ و $\hat{\beta}$ نشان می دهیم، عبارت اند از:

$$\hat{\beta} = \frac{S_{XY}}{S_X^2} \qquad (۱۰-۱)$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \qquad (۱۱-۱)$$

که با توجه به تعریف S_{XY} و S_X^2، برای (۱-۱۰) داریم

$$\hat{\beta} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

سرانجام خط رگرسیون کمترین مربّعات عبارت خواهد بود از:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i \qquad (۱۲-۱)$$

معادلهٔ (۱-۱۲) برآوردی از تابع رگرسیونی $E(Y_i|X) = \alpha + \beta X_i$ است که آن را تابع رگرسیونی برآورد شده می نامند.

بطور کلی در روش کمترین مربّعات ، برآوردکننده های $\hat{\alpha}$ و $\hat{\beta}$ را به قسمی پیدا می کنیم که مجموع توانهای دوم انحرافها $Y_i - \hat{Y}_i$ را مینیمم سازد. برای این مینیم سازی روش دیگری نیز وجود دارد که در اینجا به ذکر آن می پردازیم.

برای یافتن مقادیر $\hat{\alpha}$ و $\hat{\beta}$ که Q را مینیمم سازد، نسبت به $\hat{\alpha}$ و $\hat{\beta}$ مشتق می گیریم و حاصل را برابر صفر قرار می دهیم: