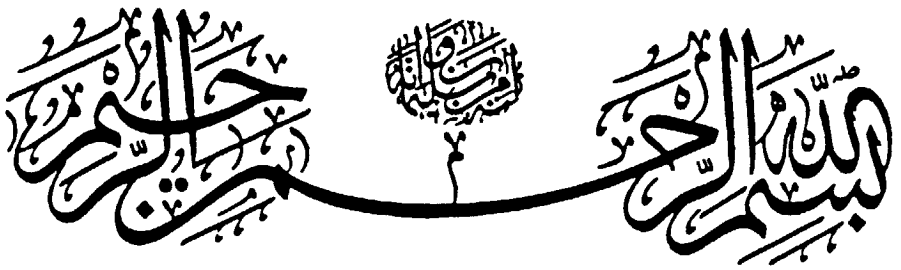


دانشگاه تربیت مدرس

دانشکده علوم پایه

پایان نامه دوره کارشناسی ارشد آمار

مقایسه اسپلاین با رگرسیون چندجمله‌ای و کاربرد آن

توسط

روشنک علی‌محمدی

استاد راهنما

دکتر محسن محمدزاده

آذر ۱۳۷۷

تأییدیه اعضای هیأت داوران حاضر در جلسه دفاع از پایان نامه کارشناسی ارشد

اعـضـای هیـئـت داوران نـسخـهٔ نـهایی پایان نامه خـانم/ آقای روشنک علی محمدی

تحت عنوان: **مقایسه اسپلاین با رگرسیون چندجمله‌ای و کاربرد آن**

را از نظر فرم و محتوی بررسی نموده و پذیرش آنرا برای تکمیل درجه کارشناسی‌ارشد پیشنهاد می‌کنند.

اعضای هیأت داوران	نام و نام خانوادگی	رتبهٔ علمی	امضاء
۱- استاد راهنما	آقای دکتر محمدزاده	استادیار	
۲- نماینده شورای تحصیلات تکمیلی	آقای دکتر محمدزاده	استادیار	
۳- استاد ناظر	آقای دکتر اسماعیل خرم	استادیار	
۴- استاد ناظر	آقای دکتر عین ا... پاشا	دانشیار	
۵- استاد ناظر	آقای دکتر عباس گرامی	استادیار	

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

آیین‌نامهٔ چاپ پایان‌نامه (رساله)های دانشجویان دانشگاه تربیت مدرّس

نظر به اینکه چاپ و انتشار پایان‌نامه (رساله)های تحصیلی دانشجویان دانشگاه تربیت مدرّس مبیّن بخشی از فعالیتهای علمی - پژوهشی دانشگاه است بنابراین به منظور آگاهی و رعایت حقوق دانشگاه، دانش‌آموختگان این دانشگاه نسبت به رعایت موارد ذیل متعهد می‌شوند:

مادهٔ ۱ در صورت اقدام به چاپ پایان‌نامه (رساله)ی خود، مراتب را قبلاً به طور کتبی به مرکز نشر دانشگاه اطلاع دهد.

مادهٔ ۲ در صفحهٔ سوم کتاب (پس از برگ شناسنامه)، عبارت ذیل را چاپ کند:

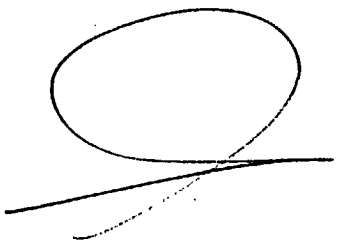
«کتاب حاضر، حاصل پایان‌نامهٔ کارشناسی ارشد/رسالهٔ دکتری نگارنده در رشتهٔ ..... است که در سال ..... در دانشکدهٔ ..... دانشگاه تربیت مدرّس به راهنمایی سرکار خانم/جناب آقای دکتر ..... و مشاورهٔ سرکار خانم/جناب آقای دکتر ..... از آن دفاع شده است.»

مادهٔ ۳ به منظور جبران بخشی از هزینه‌های نشریات دانشگاه تعداد یک درصد شمارگان کتاب (در هر نوبت چاپ) را به مرکز نشر دانشگاه اهدا کند دانشگاه می‌تواند مازاد نیاز خود را به نفع مرکز نشر در معرض فروش قرار دهد.

مادهٔ ۴ در صورت عدم رعایت مادهٔ ۳، ۵۰٪ بهای شمارگان چاپ‌شده را به عنوان خسارت به دانشگاه تربیت مدرّس، تأدیه کند.

مادهٔ ۵ دانشجو تعهد و قبول می‌کند در صورت خودداری از پرداخت بهای خسارت، دانشگاه می‌تواند خسارت مذکور را از طریق مراجع قضایی مطالبه و وصول کند؛ به علاوه به دانشگاه حق می‌دهد به منظور استیفای حقوق خود، از طریق دادگاه، معادل وجه مذکور در مادهٔ ۴ را از محلّ توقیف کتابهای عرضه‌شدهٔ نگارنده برای فروش، تأمین نماید.

مادهٔ ۶ اینجانب رویا محمدی دانشجوی رشته آمار مقطع کارشناسی ارشد تعهد فوق و ضمانت اجرایی آن را قبول کرده، به آن ملتزم می‌شوم.



قدردانی

در این‌جا لازم است از کلیه‌ی افرادی که مرا در انجام این پایان‌نامه کمک نموده‌اند، خصوصاً استاد گرامی جناب آقای دکتر محسن محمدزاده که در تمام مراحل انجام کار با مساعدت‌ها و راهنمایی‌های بی‌دریغ خود مرا یاری کردند، تشکر کنم.

همچنین از اساتید محترم آقایان دکتر پاشا، دکتر گرامی و دکتر خرّم که حق تعلیم براینجانب دارند و به‌سبب حضور در جمع داوران سپاسگزاری و قدردانی می‌نمایم.

سلامت، سعادت و موفقیت روزافزون اساتید ارجمند را از درگاه خداوند متعال خواستارم.

مقایسه اسپلاین با رگرسیون چندجمله‌ای و کاربرد آن

چکیده

رگرسیون چندجمله‌ای روشی پارامتری برای برازش منحنی به داده‌ها می‌باشد که در آن ارتباط بین متغیر پاسخ و متغیرهای مستقل به صورت چندجمله‌ای برآورد می‌گردد. در بسیاری از مواقع در مورد نحوه ارتباط بین متغیرها اطلاع زیادی در دست نیست. در این صورت بهتر است به جای مفروض داشتن یک الگوی پارامتری خاص (مانند چندجمله‌ای) برای داده‌ها از روشی استفاده شود که داده‌ها ماهیت روند خود را بهتر نشان دهند. اسپلاین هموارى روشی برای برآورد منحنی است که در آن در مورد شکل منحنی فرضهای قوی اعمال نمی‌شود و تنها فرض هموارى منحنی در نظر گرفته می‌شود. در این پایان‌نامه برازش منحنی به دو روش رگرسیون چندجمله‌ای و اسپلاین هموارى صورت گرفته و نهایتاً میزان دقت این دو روش بر اساس معیار مجموع مربعات باقیمانده‌ها مورد مقایسه عددی قرار گرفته‌اند. این مطالعه نشان می‌دهد که برای نمونه‌های باحجم کوچک و برای هرمقدارانحراف معیار، اسپلاین هموارى منحنی بهتری از رگرسیون چندجمله‌ای به داده‌ها برازش می‌دهد، اما برای نمونه‌های باحجم بزرگ رگرسیون چندجمله‌ای روشی آسانتر و سریعتر است.

واژه‌های کلیدی : اسپلاین هموارى، رگرسیون چندجمله‌ای و اعتبار متقابل تعمیم‌یافته

فهرست مندرجات

۱ مقدمه ۱

۱.۱ آنالیز رگرسیون ۱

۲.۱ هموارسازی ۳

۲ رگرسیون چندجمله‌ای ۸

۱.۲ رگرسیون خطی ۸

۱.۱.۲ روش کمترین مربعات ۹

۲.۲ رگرسیون چندجمله‌ای ۱۱

۱.۲.۲ مدلهای چندجمله‌ای یک متغیره ۱۲

۲.۲.۲ مدلهای چندجمله‌ای دو یا چند متغیره ۱۵

۳.۲ چندجمله‌ای‌های متعامد ۱۸

۳ هموارسازها و اسپلاین ۲۵

۱.۳ مقدمه ۲۵

۲.۳ هموارسازی ۲۶
۱.۲.۳ فنون هموارسازی ۲۶
۲.۲.۳ رهیافت جریمه ناهمواری ۳۰
۳.۲.۳ اسپلاین رگرسیونی ۳۳
۴.۲.۳ اسپلاین درجه سه ۳۴
۵.۲.۳ اسپلاین درجه سه طبیعی ۳۵
۶.۲.۳ محاسبه g ۳۵
۷.۲.۳ الگوریتم رینش ۳۸
۸.۲.۳ هموارسازها برای بیش از یک متغیر مستقل ۴۰
۹.۲.۳ تقابل اریبی—واریانس ۴۲
۱۰.۲.۳ انتخاب پارامتر همواری ۴۳
۱۱.۲.۳ روش اعتبار متقابل ۴۴
۱۲.۲.۳ تقابل اریبی—واریانس برای هموارسازهای خطی ۴۵
۱۳.۲.۳ اعتبار متقابل برای هموارسازهای خطی ۴۵
۱۴.۲.۳ اعتبار متقابل تعمیم‌یافته (GCV) ۴۷
۱۵.۲.۳ درجه آزادی هموارساز ۴۷

۳.۳ برازش مدل به داده‌ها ۴۸

۴ مقایسه اسپلاین و رگرسیون چندجمله‌ای ۵۰

۱.۴ مقدمه ۵۰

۲.۴ روش محاسبه و کاربرد ۵۰

۳.۴ برازش رگرسیون چندجمله‌ای ۵۳

۴.۴ شیوه شبیه‌سازی ۵۳

۵.۴ نتایج شبیه سازی ۵۴

۶.۴ بحث و نتیجه‌گیری ۶۴

فصل ۱

مقدمه

در بسیاری از مسائل کاربردی، تعیین ارتباط بین عوامل مختلف از اهمیت خاصی برخوردار است. در آمار روشهای مختلفی برای تعیین نوع ارتباط و ارائهٔ مدل ارتباطی بین عوامل وجود دارد. یکی از این فنون، آنالیز رگرسیون و دیگری روشهای هموارسازی است که در این فصل مختصراً مرور خواهد شد.

۱.۱ آنالیز رگرسیون

آنالیز رگرسیون یک فن آماری برای تعیین نحوه ارتباط بین دو یا چند متغیر است به‌طوریکه بتوان مقدار یک متغیر Y را با استفاده از مقادیر متغیرهای $X_1, \ldots, X_p$ برآورد یا پیش‌بینی کرد. بدلیل آنکه مقدار متغیر Y تحت تاثیر مقادیر متغیرهای $X_1, \ldots, X_p$ می‌باشد، Y متغیر وابسته (پاسخ) و $X_1, \ldots, X_p$ متغیرهای مستقل (پیشگو) نامیده می‌شوند.

معمولاً منظور از آنالیز رگرسیون، برازش یک مدل ریاضی به مجموعه داده‌ها به عنوان الگوی ارتباطی بین متغیرها است. به عبارت دیگر ابتدا خانواده‌ای از توابع در نظر گرفته شده و

سپس تعیین یک عضو خانواده مورد نظر، به عنوان مدل ارتباط بین متغیرها تعیین می‌شود. پس از تعیین مدل، با بررسی نمودار باقیمانده‌ها و برآورد، آزمون فرض در مورد پارامترهای مدل و نیکویی برازش مدل مورد تحقیق قرار می‌گیرد. این شیوه تحلیل رگرسیون، رگرسیون پارامتری نامیده می‌شود. جنبه دیگر رگرسیون، رگرسیون ناپارامتری است.

ابتدا حالتی را که تنها یک متغیر مستقل (پیشگو) و یک متغیر وابسته (پاسخ) موجود است، در نظر می‌گیریم.

فرض کنید که مشاهدات متغیر تصادفی Y در n مقدار X به دست آمده است. می‌توان مدل

$$y_i = g(x_i) + e_i \quad ; i = ۱, \ldots, n \qquad (۱.۱.۱)$$

را در نظر گرفت که در آن e_i ها متغیرهای تصادفی مستقل از توزیع $N(۰, \sigma^۲)$ و g تابعی نامعلوم است که تابع یا منحنی رگرسیون نامیده می‌شود.

بکارگیری رگرسیون پارامتری یا ناپارامتری در مدل (۱.۱.۱) بستگی به فرضهایی دارد که درباره تابع g می‌شود. در رگرسیون پارامتری شکل تابع g معلوم فرض می‌شود. به عبارت دیگر، خانواده‌ای از توابع را درنظر گرفته و با استفاده از فنون تحلیل رگرسیون، تابع مناسب برای برآورد منحنی تعیین می‌شود.

مدلهای پارامتری می‌توانند به طور خطی یا غیرخطی به پارامترها وابسته باشند. برای مدلهایی که نسبت به پارامترها خطی‌اند، می‌توان با استفاده از یک روش مناسب برآورد مانند حداقل مربعات خطا یا حداکثر درستنمایی، پارامترها را برآورد نموده و برآورد منحنی g را بدست آورد. این مدلها شامل رگرسیون چندجمله‌ای نیز می‌باشد. بعضی از مدلهایی که نسبت به پارامترها غیرخطی‌اند ممکن است با کمک یک تبدیل، خطی شوند. منحنی‌های لگاریتمی و توانی مثالهایی از این نوع هستند. تحلیل این نوع از مدلها را می‌توان مانند مدلهای خطی انجام داد.

آنالیز آن دسته از مدلها که حتی توسط تبدیلات، خطی نمی‌شوند، با بکارگیری فنون

رگرسیون غیرخطی ممکن می‌گردد. برای مطالعه در این زمینه می‌توان به گالانت (۱۹۷۵)، دراپر و اسمیت (۱۹٦٦)، بیتس و واتس (۱۹۸۸)، مایرز (۱۹۹۰) و روتکوسکی (۱۹۹۰) مراجعه نمود.

رگرسیون ناپارامتری شیوه کلی‌تر برای استنباط درباره g است. برآوردهایی از منحنی رگرسیون که با این شیوه حاصل می‌شود، انعطاف پذیرند و به یک الگوی پارامتری خاص محدود نمی‌شوند. تفاوت عمده رگرسیون پارامتری و ناپارامتری درجه اتکای آنها به اطلاعاتی است که از داده‌ها و تجربیات آماردان برای برآورد g وجود دارد. در مدل ناپارامتری فرد بنابر اعتقاد خود، یک فضای تابعی[1] مناسب برای g برمی‌گزیند. سپس از داده‌ها جهت تعیین عضوی از این فضای تابعی برای برآورد منحنی نامعلوم رگرسیون استفاده می‌کند. اما در یک الگوی پارامتری، یک خانواده از منحنی‌ها انتخاب شده و آنالیز رگرسیونی بر مبنای این خانواده از منحنی‌ها قرار می‌گیرد. بنابراین فنون رگرسیون ناپارامتری بیشتر از رگرسیون پارامتری به داده‌ها وابسته هستند و باعث می‌شوند که برآورد حاصل روند داده‌ها را بهتر تبیین نماید.

به طور کلی هنگامی که مدل پارامتری معتبر باشد باید از رگرسیون پارامتری بهره گرفت و تنها در صورتی که اطلاعات در مورد g ناچیز است، رگرسیون ناپارامتری مناسب می‌باشد. زیرا در صورتی که مدل پارامتری مناسب باشد کارایی برآوردگرهای ناپارامتری کمتر است و در ضمن برآوردهای ناپارامتری اریب هستند.

۲.۱ هموارسازی

برای بررسی نحوه ارتباط متغیر پاسخ Y و متغیر مستقل X مجدداً مدل (۱.۱.۱) را در نظر می‌گیریم. تابع g نامعلوم است و هدف برآورد این تابع می‌باشد. در حالت کلی برای

[1] Functional Space

تعیین برآوردی منحصربفرد از تابع g باید شرایطی را برای آن در نظر گرفت. به عنوان مثال، در روش رگرسیون خطی فرض می‌شود g تابعی خطی از X است و با استفاده از روش کمترین توانهای دوم[۲] ضرایب مدل خطی بطور یکتا برآورد می‌گردند. در مواردی که برازش مدل خطی به داده‌ها مناسب نباشد، ممکن است رگرسیون چندجمله‌ای برآورد بهتری از g فراهم نماید و بسته به اینکه یک یا چند متغیر مستقل موجود باشد، مدل چندجمله‌ای یک یا چند متغیره برازش داده می‌شود. برای تعیین درجه و ضرایب چندجمله‌ای از روش پیشرو[۳] یا پسرو[۴] استفاده می‌شود که برای مطالعه آن می‌توان به نتر (۱۹۷۴) و مونت گومری (۱۹۹۱) مراجعه نمود. ساویلین (۱۹۹۱) به طور هندسی به مطالعه مدلهای چندجمله‌ای و چندجمله‌ای‌های متعامد پرداخته است.

اگر چه رگرسیون چندجمله‌ای برای برخی از مجموعه داده‌ها مناسب می‌باشد لیکن کاستی‌هایی نیز دارد. یکی از اشکالات این روش این است که تک مشاهدات دورافتاده به‌طور قابل توجهی روی برآورد تابع g تاثیر می‌گذارند. هر چند چندجمله‌ای برازش شده به داده‌ها از درجه بالاتری باشد به رابطه واقعی داده‌ها نزدیکتر می‌شود اما با افزایش درجه چندجمله‌ای برازاندن مدل پیچیده‌تر می‌شود. این اشکالات از مفروض داشتن الگویی مشخص برای داده‌ها ایجاد می‌شود. برای رهایی از مشکلات فوق می‌توان g را تابعی دلخواه در نظر گرفت و اقدام به برآورد آن نمود. اگر مدل (۱.۱.۱) با روش کمترین مربعات باقیمانده‌ها برازش داده شود حالات متفاوتی را می‌توان در نظر گرفت. اگر شرطی برای تابع نامعلوم g منظور نشود و نقاط با خطوط راست به‌هم وصل شوند، مجموع مربعات باقیمانده‌ها در نقاط مشاهده شده صفر می‌شود، منحنی حاصل ناهموار خواهد بود. برای رفع این نقیصه می‌توان شرط هموار g را روی مدل (۱.۱.۱) اعمال نمود. میزان هموار تابع g را می‌توان توسط $\int_a^b \{g''(x)\}^2 dx$ اندازه‌گیری نمود. بنابراین برای برآورد منحنی g، علاوه بر عامل مجموع مربعات باقیمانده‌ها،

---

[۲] Mean Square Error

[۳] forward

[۴] backward

میزان ناهمواری را نیز در نظر گرفته و ملاک مجموع مربعات جریمه‌ای بصورت

$$S(g,\lambda)=\sum_{i=1}^{n}(y_i-g(x_i))^2+\lambda\int_a^b\{g''(x)\}^2dx$$

تعریف می‌شود که در آن $\lambda>0$ پارامتر همواری[5] نامیده می‌شود. تابعی مانند $\hat{g}$ که مجموع مربعات جریمه‌ای $S(g,\lambda)$ را کمینه نماید، اسپلاین همواری[6] نامیده می‌شود. این شیوه برآورد منحنی را رهیافت جریمه ناهمواری [7] نامند. در مورد روشهای جریمه ناهمواری در اوبانک (۱۹۸۸)، واهبا (۱۹۹۰)، هاردل (۱۹۹۰) و راسنبلات (۱۹۹۱) مطالب مبسوطی آمده است.

تعمیمی از اسپلاینها حالتی است که دو متغیر مستقل وجود دارد که در اینصورت اسپلاین رویه نازک[8] نامیده می‌شود. در این مورد می‌توان به فصل دوم واهبا (۱۹۹۰) مراجعه نمود.

در روش اسپلاین همواری مقدار λ نقش مهمی در برآورد منحنی ایفا می‌نماید. اگر λ بزرگ باشد، $\hat{g}$، خیلی کم‌انحنا خواهد شد (شکل ۳.۲.۳). در حد وقتی λ به سمت بی‌نهایت میل کند، برآورد منحنی همان رگرسیون خطی خواهد بود. از طرف دیگر، اگر λ نسبتاً کوچک باشد برآورد g تا حد زیادی روند داده‌ها را دنبال می‌کند، حتی اگر به بهای تغییرات سریع آن تمام شود (شکل ۴.۲.۳). در وضعیت حدی اگر λ به صفر نزدیک شود، برآورد حاصل به اسپلاین درونیاب[9] میل می‌کند (شکل (۲.۲.۳)). بنابراین محاسبه مقدار مناسب λ در اسپلاین همواری مساله مهمی است که در مورد آن می‌توان به هاردل، مارون و هال (۱۹۸۸) و گرین و سیلورمن (۱۹۹۴) مراجعه نمود. هاردل (۱۹۹۱ و ۱۹۹۰) چگونگی انتخاب پارامتر همواری بر اساس مینیمم کردن متوسط مربع خطای برآوردگر هسته‌ای را ارائه داده

---

[5] Smoothing Parameter

[6] Smoothig Spline

[7] Roughness Penalty

[8] Thin plate

[9] Interpolating Spline

است. دو روش اعتبار متقابل [۱۰] و اعتبار متقابل تعمیم‌یافته [۱۱] برای تعیین مقدار λ ، در استون (۱۹۷۳)، کریون و واهبا (۱۹۷۹)، گرین و سیلورمن (۱۹۹۴) و محمدزاده (۱۹۹۸) آمده است.

از جمله کاربردهای روش اسپلاین همواری در مدل‌بندی نیمه پارامتری [۱۲] است . مدلهای خطی تعمیم‌یافته (GLM) [۱۳] که اولین بار نلدر و ودربرن (۱۹۷۲) آنرا ارائه نمودند و بعدها توسط افراد دیگری از جمله مک کانگ و نلدر (۱۹۸۹) و اوسالیوان (۱۹۸٦) مورد مطالعه قرار گرفته است، از دیگر مواردی است که در آنها روش اسپلاین همواری بکار گرفته می‌شود.

هر دو روش اسپلاین همواری و رگرسیون چندجمله‌ای به منظور برآورد منحنی رگرسیون بکار می‌رود. هر یک از دو روش فوق‌الذکر ممکن است برای مجموعه داده‌های خاصی برآورد خوبی از منحنی ایجاد نمایند.

هدف ما از این تحقیق، مقایسه دو روش اسپلاین همواری و رگرسیون چندجمله‌ای در برآورد منحنی رگرسیون می‌باشد. نکته قابل توجه این است که لزوماً هیچیک از دو روش اسپلاین همواری یا رگرسیون چندجمله‌ای همواره نسبت به یکدیگر برآورد بهتری از منحنی را ایجاد نمی‌کنند و برای مجموعه داده‌های مختلف روش برآورد برتر قابل تغییر است. در این پژوهش برای مقایسهٔ این دو روش از شیوه شبیه‌سازی استفاده می‌شود. در این شیوه به تولید داده‌های تصادفی و برازش منحنی به آنها می‌پردازیم. بدین منظور ابتدا یک نمونه تصادفی n تایی از توزیع یکنواخت استخراج نموده تا مشاهدات $x_n, \ldots, x_1$ بدست آیند و سپس از توزیع $N(0, \sigma^2)$ نمونه تصادفی n تایی گرفته می‌شود که مقادیر $e_n, \ldots, e_1$ را تشکیل می‌دهند و از رابطه $y = x + x^2 + x^3 + e$ مقادیر y از یک مدل درجه سه حاصل می‌شوند.

---

[۱۰] Cross Validation

[۱۱] Generalized Cross Validation

[۱۲] Semiparametric

[۱۳] Genaralized Linear Mosdel