
بسم الله الرحمن الرحيم

حمایت از حقوق پدیدآورندگان

پایان نامهٔ حاضر، حاصل پژوهشهای نگارنده در دورهٔ **کارشناسی ارشد** رشتهٔ **آمار** گرایش **آمار ریاضی** است که در بهمن ۱۳۹۲ در **دانشکده علوم پایه** دانشگاه **یاسوج** به راهنمایی **دکتر حیدرعلی مردانی‌فرد** و مشاورهٔ **دکتر انوشیروان غفاری‌پور** از آن دفاع شده است و کلیهٔ حقوق مادی و معنوی آن متعلق به دانشگاه یاسوج است.

دانشکده علوم پایه
گروه ریاضی

پایان نامه کارشناسی ارشد رشته آمار گرایش آمار ریاضی

برآورد کمترین مربعات جریمه‌دار در مدل‌های جزئاً خطی با ابعاد بالا

استاد راهنما

دکتر حیدرعلی مردانی‌فرد

پژوهشگر

زهرا فریدونی

بهمن ۱۳۹۲

برآورد کمترین مربعات جریمه‌دار در مدل‌های جزئاً خطی با ابعاد بالا

به وسیلهٔ

زهرا فریدونی

پایان نامه

ارائه شده به تحصیلات تکمیلی دانشگاه به عنوان بخشی از فعالیت‌های تحصیلی لازم برای اخذ درجهٔ کارشناسی ارشد

در رشتهٔ:

آمار

در تاریخ ..... توسط هیأت داوران زیر بررسی و با درجهٔ ..... به تصویب نهایی رسید.

۱- استاد راهنما:	دکتر حیدرعلی مردانی‌فرد	با مرتبهٔ علمی استادیار	امضاء
۲- استاد مشاور:	دکتر انوشیروان غفاری‌پور	با مرتبهٔ علمی استادیار	امضاء
۳- استاد داور داخل گروه:	دکتر آرش اردلان	با مرتبهٔ علمی استادیار	امضاء
۴- استاد داور خارج گروه:	دکتر علیرضا نعمت‌الهی	با مرتبهٔ علمی استاد	امضاء
۵- نمایندهٔ تحصیلات تکمیلی دانشگاه:	دکتر زهرا رفیعی	با مرتبهٔ علمی استادیار	امضاء

تقدیم به:

همسرم، اسطوره زندگی‌ام، پناه خستگی‌ام

و

فرزندم، امید بودنم

قدردانی

سپاس خدای را که سخنوران در ستودن او بمانند و شمارندگان شمردن نعمت‌های او ندانند و کوشندگان حق او را گزاردن نتوانند و سلام و درود بر محمد و خاندان پاک او، طاهران معصوم، هم‌آنان که وجودمان وامدار وجودشان است. وظیفه خود می‌دانم از استاد راهنمای خود، جناب آقای دکتر حیدرعلی مردانی‌فرد، که در کمال سعه صدر، حسن خلق و فروتنی از هیچ کمکی در این عرصه بر من دریغ ننمودند صمیمانه تشکر و قدردانی کنم. ازاستاد صبور، جناب آقای دکترانوشیروان غفاری پورکه زحمت مشاوره این پایان‌نامه را برعهده گرفتند کمال تشکر و قدر دانی را دارم. از استاد گرامی دکتر آرش اردلان سپاسگزارم ، چرا که زحمت داوری این پایان‌نامه را درحالی بر عهده گرفتند که بدون راهنمائی‌های ارزنده ایشان این مجموعه به انجام نمی‌رسید.

برخود لازم می‌دانم از استاد عالی‌قدر دانشگاه شیراز جناب آقای دکتر نعمت‌اللهی قدر دانی نمایم. به پاس قدردانی با قلبی آکنده از عشق و معرفت از همسرم و پسرم که صبورانه و صادقانه مرا همراهی نمودند تا بتوانم در کمال آرامش و آسایش به تهیه و تنظیم این اثر بپردازم تشکر و قدردانی می‌نمایم.

باشد که این خردترین ، بخشی از زحمات آنان را سپاس گوید.

زهرا فریدونی

بهمن ۱۳۹۲

چکیده

در مطالعات رگرسیونی، زمانی که بین متغیرهای مستقل همبستگی بالائی وجود داشته باشد استفاده از روش‌های معمول از جمله روش کمترین مربعات معمولی باعث ناپایداری واریانس برآوردها می‌شود. یک راه حل معمول، استفاده از روش کمترین مربعات جریمه‌دار است که در آن برای مقادیر بزرگ برآوردگرها، جریمه بالائی در نظر گرفته می‌شود و به نوعی تغییرات برآوردگر تحت کنترل در می‌آید.

مورد دیگر استفاده از رگرسیون جریمه‌دار در مدل‌های با ابعاد بالا یعنی مدل‌هائی با تعداد زیادی متغیر مستقل است. در این مدل‌ها تلاش می‌شود از ضرائب " نزدیک به صفر" حتی‌الامکان صرف‌نظر گردد تا فقط متغیرهائی در مدل باقی بمانند که تاثیر کاملا معنی‌داری در متغیر پاسخ دارند.

در این پژوهش تلاش شده است ضمن مرور مختصری بر روش کمترین مربعات جریمه‌دار، رگرسیون جریمه‌دار و نحوه عمل این روش در هموارسازی نمودارهای رگرسیونی، استفاده از این روش در برازش مدل‌هائی با ابعاد بالا مورد مطالعه و بررسی قرار گیرد. با استفاده از شبیه‌سازی ، درستی بعضی از روابط و برتری این روش در مقایسه با سایر روش‌ها تحقیق شده است که با ارائه چند سری داده واقعی، این روش در تحلیل آنها مورد استفاده قرار گرفته است.

فهرست مطالب

فهرست علائم اختصاری iii

فهرست تصاویر iv

فصل ۱: مقدمه و تاریخچه ۱

فصل ۲: مفاهیم و تعاریف ۴

فصل ۳: رگرسیون جریمه‌دار ۱۹

۳-۱ برآورد کمترین مربعات جریمه‌دار ۱۹

۳-۲ مدل‌های برآورد کمترین مربعات جریمه‌دار ۲۱

۳-۲-۱ رگرسیون جریمه‌دار ریج ۲۲

۳-۲-۲ رگرسیون جریمه‌دار لاسو ۳۱

۳-۲-۳ رگرسیون جریمه‌دار اسکاد ۳۶

۳-۲-۴ انتخاب پارامتر همواری ۳۸

فصل ۴: چندجمله‌ایها و رگرسیون تکه تکه ۴۲

۴-۱ چندجمله‌ایها ۴۲

۴-۲ چندجمله‌ای تکه‌ای ۴۳

۴-۳ رگرسیون چندجمله‌ای ۴۴

۴-۴ رگرسیون شکسته ۵۰

۴-۵ مدلهای رگرسیون تکه تکه ۵۱

۴-۶ کاربرد رگرسیون جریمه‌دار ریج در هموارسازی اسپلاین ۵۳

فصل ۵: برآوردهای جریمه دار در مدل‌های جزئاً خطی و ابعاد بالا ۵۷

۵-۱ مدل‌های نیمه‌پارامتری ۵۷

۵-۱-۱ برآورد پارامترها در مدل‌های جزئاً خطی ۶۲

۵-۱-۲ هموارسازی در مدل‌های جزئاًخطی ۶۳

۵-۲ برآورد در مدل‌های جزئاً خطی با ابعاد بالا ۷۰

فصل ۶: شبیه‌سازی ۷۴

۶-۱ برآورد پارامتر برای داده‌های شبیه‌سازی شده ۷۴

۶-۱-۱ برآورد پارامتر توسط روش انتخاب بهترین زیرمجموعه ۷۴

۶-۱-۲ برآورد پارامترها توسط رگرسیون ریج و لاسو ۷۵

۶-۱-۳ برآوردپارامترها به روش ریج در ابعاد بالا ۷۸

پیوست آ: تعاریف و مفاهیم ریاضی ۷۹

پیوست ب:برنامه‌های نوشته شده برای برآورد پارامترها با استفاده از نرم‌افزار R ۹۱

ب-۱ برنامه شبیه‌سازی شده ۹۱

ب-۲ برنامه نمودارهای فصل‌های قبل ۹۲

واژه‌نامه فارسی به انگلیسی ۱۱۰

واژه‌نامه انگلیسی به فارسی ۱۱۱

مراجع ۱۱۲

فهرست علائم اختصاری

میانگین مربع خطا ..... MSE
خطای پیش‌بینی ..... PE
میانگین قدر مطلق خطا ..... MAE
مجموع توان دوم خطا ..... SSE
کمترین مربعات معمولی ..... OLS
مجموع توان دوم رگرسیونی ..... SS_R
مجموع توان دوم کل ..... SS_T
میانگین توان دوم رگرسیونی ..... MSR
میانگین توان دوم خطا ..... MSE
اعتبارسنجی متقابل CV
کمترین مربعات جریمه‌دار ..... $PRSS$
مدل جزئاً خطی ..... PLM

فهرست تصاویر

۲-۱ نمودار پراکنش برای داده‌های cars ۹
۲-۲ تابع موجدار برای داده‌های cars ۹
۲-۳ پایه رگرسیون خطی ۱۰
۲-۴ رگرسیون خطی ۱۱
۲-۵ پایه مدل رگرسیون درجه دو ۱۱
۲-۶ مدل رگرسیون درجه دو ۱۲
۲-۷ مدل با یک شکستگی ۱۳
۲-۸ پایه مدل با یک شکستگی ۱۳
۲-۹ پایه مدل با چندین شکستگی ۱۴
۲-۱۰ R^2 برای داده‌های پروستات ۱۷
۲-۱۱خطای پیش‌بینی کمترین مربعات و بهترین زیرمجموعه ۱۷
۲-۱۲بهترین زیر مجموعه ۱۸

۳-۱ رگرسیون خطی و تابع موجدار ۲۰
۳-۲ فضای کانتور ریج ۲۶
۳-۳ خطای پیش‌بینی ریج و کمترین مربعات ۲۸
۳-۴ برآورد ریج در برابر کمترین مربعات ۳۰
۳-۵ فضای رگرسیون لاسو ۳۲
۳-۶ مقایسه لاسو و کمترین مربعات ۳۳
۳-۷ رگرسیون لاسو ۳۵
۳-۸ خطای پیش‌بینی لاسو و کمترین مربعات ۳۵
۳-۹ برآورد اسکاد ۳۷

۳-۱۰ GCV ۴۰
۳-۱۱ نمودار CV ۴۱

۴-۱ چندجمله‌ای تکه‌ای ۴۳
۴-۲ چندجمله‌ای درجه دو برای داده‌های اتانول ۴۵
۴-۳ رگرسیون خطی برای داده‌های لیدار ۴۷
۴-۴ رگرسیون درجه دو برای داده‌های لیدار ۴۷
۴-۵ رگرسیون درجه چهار برای داده‌های لیدار ۴۸
۴-۶ رگرسیون درجه ده برای داده‌های لیدار ۴۸
۴-۷ اسپلاین با یک گره برای داده‌های لیدار ۴۹
۴-۸ اسپلاین درجه یک با۳ گره برای cars ۵۲
۴-۹ اسپلاین درجه دو ۵۳
۴-۱۰ اسپلاین خطی با ۴ گره ۵۴
۴-۱۱ اسپلاین خطی با ۲۰ گره ۵۴
۴-۱۲ اسپلاین با $\lambda = ۱۰$ ۵۶

۵-۱ نمودار پراکنش شوری تالاب ۵۸
۵-۲ باقی‌مانده در برابر شوری آب تالاب ۵۹
۵-۳ مقایسه نمودارهای خطی باحضور داده پرنفوذ ۶۰
۵-۴ مقایسه منحنی‌ها باحضور داده پرنفوذ ۶۰
۵-۵ نمودار پراکنش داده‌های ragweed ۶۵
۵-۶ برازش با استفاده از اسپلاین در مدل نیمه‌پارامتری ۶۷
۵-۷ نمودار اعتبار سنجی متقابل با استفاده از برازش لاسو ۷۰
۵-۸ نمودار اعتبارسنجی متقابل با استفاده از برازش اسکاد ۷۳

آ-۱ تصویر بردار روی بردار دیگر ۸۶

فصل ۱

مقدمه و تاریخچه

هدف استنباط آماری، نتیجه‌گیری در مورد یک یا چند جمعیت بر اساس اطلاعات موجود در نمونه است. یکی از مسائل مهم در استنباط آماری ، بحث برآوردیابی می‌باشد که شامل سه شاخه

۱ - برآوردهای کلاسیک

۲ - برآوردهای بیزی

۳ - برآوردهای دنباله‌ای است

که هرکدام از برآوردهای کلاسیک و برآوردهای بیزی شامل برآوردهای نقطه‌ای و برآورد فاصله‌ای است. در برآورد نقطه‌ای برآوردگر باید تا حد امکان به پارامتر نامعلوم نزدیک باشد. از ساده‌ترین و در عین حال کاراترین روش‌های برآورد، روش کمترین مربعات معمولی است که اولین بار توسط کارل فردریش گاوس[۱] در سال ۱۷۹۴ میلادی مطرح شد. این روش برای برآورد پارامترهای رگرسیون خطی مرتبه‌های بالاتر از یک نیز قابل استفاده می‌باشد و ازجمله کسانی که به استفاده از این روش در رگرسیون چندجمله‌ای پرداخت جرگونه [۲] می‌باشد. یکی از ایرادهای وارد به روش کمترین مربعات معمولی، توانائی تفسیر آن است . هنگامیکه با وضعیتی روبه‌رو هستیم که به دلیل تعداد زیاد پیش‌بینی کننده‌ها، شناسائی گروه کوچک و

---

[۱] gauss

[۲] Gergonne

موثری از متغیرها مفید می‌باشد، رگرسیون خطی چنین ویژگی را ندارد و همچنین رگرسیون خطی هنگامیکه تعداد متغیرها بیشتر از مشاهدات باشد ،تعریف نشده است و چون در بسیاری از شاخه‌های علوم زیست‌شناسی و پزشکی با مواردی روبه‌رو هستیم که در آن تعدادمتغیرها بیش از نمونه است اهمیت برآورد در چنین نمونه‌هائی آشکار می‌شود.

بنابراین رگرسیون جریمه‌دار، به عنوان یک توسیع از پیش‌بینی کننده‌های خطی معرفی می‌شود.رگرسیون جریمه‌دار ریج توسط هورل [۳] در سال ۱۹۶۲ معرفی شد که در آن اظهار داشت، با وجود همبستگی بین متغیرهای توضیحی، بکارگیری روش حداقل مربعات باعث ایجاد خطا در برآورد می‌شود و رگرسیون ریج را به عنوان جایگزینی برای این روش، توسعه داد که اجازه می‌دهد که برآوردها با واریانس کمتراز روش حداقل مربعات با داشتن اریبی، محاسبه شوند. اندازه‌گیری اصلی این روش در آمار توسط هورل و کنالد[۴] در سال ۱۹۷۰ مطرح شد.

در سال ۱۹۹۸ ولیلا [۵] و هسیج [۶] نشان دادند که اگر متغیرهای پیش‌بینی کننده از همبستگی بالائی برخوردار باشند، برآورد ریج عملکرد خوبی نخواهد داشت بنابراین سعی شد از نوع دیگر رگرسیون جریمه‌دار که یک نوع روش انتخاب متغیرها ست وتوسط تیبشیرانی [۷] در سال ۱۹۹۶معرفی و توسیع داده شد استفاده نمایند.

در سال ۲۰۰۱ فان و لی انتخاب متغیر از طریق درستنمائی جریمه‌شده غیرمقعر و ویژگی‌های آن را ارائه کردند[۱]. که ایکسی [۸] و هونگ [۹] برای انتخاب متغیر و برآورد ضرائب رگرسیون در یک مدل جزئاً خطی از آن استفاده نمودند. فان و پنگ در سال ۲۰۰۴ درستنمائی جریمه‌دار غیرمقعر با مشتق‌پذیری تعدادی از پارامترها را معرفی نمودند[۲].همچنین هانگ و هورویتز در سال ۲۰۰۸ ویژگی مجانبی از برآوردهای بریج را در مدل‌های رگرسیونی بحث کردند [۳] .رگرسیون جریمه‌دار اسکد در مدل‌های جزئاً خطی توسط ایکس و هانتر بحث

---

[۳] Horel

[۴] Kenald

[۵] velilla

[۶] Hsich

[۷] Tibshirani

[۸] Xie

[۹] Huang

شد[۴] .

فصل ۲

مفاهیم و تعاریف

در این فصل مفاهیم و تعاریف موردنیاز ارائه خواهند شد.

فرض می‌کنیم $\mathbf{X}_i = (x_{i1}, ...x_{ip})$ متغیرهای مستقل و Y_i که $1 \leq i \leq n$ هستند متغیرهای وابسته باشند و $\epsilon_i \overset{iid}{\sim} (0, \sigma^2)$. بررسی اثر متغیرهای مستقل روی متغیرهای وابسته، هدف تحلیل رگرسیون است و رگرسیون را از لحاظ نوع تابعی که رابطه‌ی بین متغیرهای مستقل و وابسته را تعیین می‌کند به دو گروه رگرسیون پارامتری و رگرسیون ناپارامتری دسته‌بندی می‌کنیم.

در رگرسیون پارامتری رابطه بین متغیرهای مستقل و وابسته از طریق تابع پارامتری f به صورت $Y_i = f(x_i) + \epsilon_i$ بیان می‌شود. $f(x_i)$ می‌تواند یک تابع تک متغیره یا چندمتغیره از متغیرهای مستقل و بردار پارامترهای مجهول$\boldsymbol{\beta}' = (\beta_1, ...\beta_p)$ باشد.

اما در رگرسیون ناپارامتری رابطه بین متغیرهای مستقل و وابسته از طریق تابع ناپارامتری $m(.)$ به صورت $Y_i = m(x_i) + \epsilon_i$ نشان داده می‌شود که $m(.)$تابعی نامعلوم است. برای برآورد توابع پارامتری و ناپارامتری روش‌های بسیاری وجود دارد به طوریکه برای برآورد $m(x)$ می‌توان به رگرسیون چندجمله‌ای موضعی، هموارسازی اسپلاین و روش هسته‌ای هموارسازی اشاره کرد [۶].

در این پایان‌نامه با تمرکز بر رگرسیون پارامتری ، انواعی از روش‌های برآورد تابع پارامتری بیان می‌شود. به همین منظور ابتدا مروری کوتاه بر برآورد پارامترها در رگرسیون پارامتری

خواهیم داشت .

رگرسیون پارامتری

فرض کنید برای n فرد از افراد یک جامعه آماری مقادیر $(x_i, y_i), \ldots \quad i = ۱, \ldots . n$ مشاهده شده باشد. مدل $Y_i = f(x_i) + \epsilon_i$ رادر نظر می‌گیریم. هدف برآورد تابع $f(x_i)$ می‌باشد که با $\hat{f}(x_i)$ نمایش می‌دهیم.اگر $f(x_i)$ را بصورت $\sum_{j=۰}^{q} \beta_j x_i^j + \epsilon_i$ باشد آنگاه $Y_i = \sum_{j=۰}^{q} \beta_j x_i^j + \epsilon_i$خواهدبود وبرآورد $f(x_i)$ به برآورد پارامترهای مجهول منجر خواهد شد.

روش کمترین مربعات[۱]

روشی برای برازش یک مدل به داده‌ها است و از طریق مینیمم کردن

$$Q = \sum_{i=۱}^{n} (\sum_{j=۰}^{q} \beta_j x_i^j - y_i)^۲$$

بدست می‌آید.با بکارگیری نماد ماتریسی، مدل رگرسیونی بصورت $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ بازنویسی می‌شود.

$$\mathbf{X} = \begin{bmatrix} ۱ & x_۱ & x_۱^۲ & \cdots & x_۱^q \\ ۱ & x_۲ & x_۲^۲ & \cdots & x_۲^q \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_۱ & x_n & x_n^۲ & \cdots & x_n^q \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} y_۱ \\ y_۲ \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_۰ \\ \beta_۱ \\ \vdots \\ \beta_q \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_۱ \\ \epsilon_۲ \\ \vdots \\ \epsilon_n \end{bmatrix}$$

بامشتق‌گیری از Q نسبت به β_j ها و به کارگیری نماد ماتریسی برآورد پارامترهای مجهول بصورت

$$\hat{\beta}^{ols} = (X^T X)^{-۱} X^T Y$$

وبرآورد مدل بصورت $\hat{Y} = X\hat{\beta}$ خواهد شد.

تعریف ۲-۱. باقی‌مانده(e_i): اختلاف بین i امین مقدار مشاهده شده (y_i) و i امین مقدار برازش شده $(\hat{y}_i)$ می‌باشد و بصورت ماتریسی $e = Y - \hat{Y}$ نوشته می‌شود.

□

[۱] ordinary least-square

تعریف ۲-۲. اریبی: فرض کنید $T(\mathbf{X})$ یک برآوردگر برای $\gamma(\theta)$ باشد ، میزان اریبی $T(\mathbf{X})$ با b نشان داده می‌شود.

$$b(\theta) = E[T(\mathbf{X}) - \gamma(\theta))$$

□

تعریف ۲-۳. نااریبی: برآوردگر $T(\mathbf{X})$ را برای $\gamma(\theta)$ نااریب گویند اگر وتنها اگر

$$E_\theta[T(\mathbf{X})] = \gamma(\theta) \qquad b(\theta) = ۰ \quad \forall\theta \in \Theta$$

□

تعریف ۲-۴. میانگین مربع خطا [۲]: فرض کنید $T(\mathbf{X})$ یک برآوردگر برای $\gamma(\theta)$ باشد در این صورت میانگین توان دوم خطای برآورد$T(\mathbf{X})$ به صورت زیر تعریف می‌شود

$$MSE_\theta = E_\theta(T(\mathbf{X}) - \gamma(\theta))^۲ = Var(T(\mathbf{X})) + \underbrace{[E(T(\mathbf{X}) - \gamma(\theta)]^۲}_{b}$$

□

میانگین مربع خطا شامل دو مقدار، واریانس $T(\mathbf{X})$ و یک مقدار نامنفی می‌باشد.

تعریف ۲-۵. مقدار پیش‌بینی شده: فرایند تعیین مقدار تابع در یک نقطه جدید مانند $x_۰$ به کمک تابع $\hat{Y}(x_۰)$ را مقدار پیش‌بینی شده گویند.

□

تعریف ۲-۶. اختلاف مقدار واقعی و مقدار پیش‌بینی شده در یک نقطه جدید مانند $x_۰$ که با عبارت $\hat{Y} - \hat{Y}(x_۰)$ نشان داده می‌شود را خطای پیش‌بینی گویند و با نماد PE [۳]نشان داده می‌شود.

بررسی مناسب بودن مدل

برای بررسی مناسب بودن مدل رگرسیونی ، به مجموعه بزرگی از تکنیک‌ها نیازمندیم که مجموعه این تکنیک‌ها درستی تشخیص [۴] نام دارد. دو مولفه اصلی این تکنیک ها،

[۲] Mean Sqaure Error
[۳] prediction error
[۴] diagnostic checking

مقادیر برازش‌شده و باقی‌مانده‌ها می‌باشند. بیشترین اطلاعات راجع به مناسب بودن مدل، در باقی‌مانده‌ها است. زیرا به کشف هر نوع ناسازگاری بین داده‌ها و مدل برازش‌شده کمک می‌کند. اگر مدل مناسب باشد باقی‌مانده‌ها الگوی خاصی نخواهند داشت. در صورت وجود یک الگوی خاص در بین باقی‌مانده‌ها می‌توانیم روی داده‌ها تبدیل مناسبی قرار دهیم .برای اطلاعات بیشتر می‌توان به [۷] مراجعه کرد.

معیارهای سنجش مناسب بودن یک مدل را می توان از دو دیدگاه مورد بررسی قرار داد:

۱ -میزان دوری یا نزدیکی مقدار y برازش‌شده ومقدار واقعی.

۲ - پیش‌بینی مقدار y برای مقادیر جدید توسط $\hat{f}$.

سنجش دوری یا نزدیکی مقدار y برازش‌شده ومقدار واقعی از طریق خطای برآورد (e_i) صورت می‌گیرد. اگر $\rho(t)$ هر تابع غیر نزولی در $|t|$ باشد آنگاه $\Sigma\rho(e_i)$ را می‌توان به عنوان یک معیار برای مقایسه دو مدل بکار برد. از رایج‌ترین توابع غیرنزولی می‌توان تابع درجه‌دوم و تابع قدرمطلق را نام برد که به ترتیب معیارهای MSE و MAE را تولید می‌کنند.

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$
$$MAE = \frac{1}{n}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

برازش‌هائی که دارای خطای برآورد مینیمم باشند برازش مناسب محسوب می‌شوند. خطای پیش‌بینی در یک نقطه خاص را زمانی می‌توان محاسبه کرد که مقدار y آن نقطه را مشاهده کرده باشیم.

ویژگی‌های برآورد کمترین مربعات

۱-$\hat{\beta}^{ols}$ نااریب با حداقل واریانس می‌باشد.

۲- واریانس برآوردگر بصورت زیر است:

$$Var(\hat{\beta}) = \hat{\sigma}^2(X^TX)^{-1}$$

۳-میانگین توان دوم خطا ،(MSE) نیز به روش زیر محاسبه می‌شود:

$$\begin{aligned} MSE(\hat{\beta}^{ols}) &= \hat{\sigma}^{۲} trace(X^T X)^{-۱} \\ &= \hat{\sigma}^{۲} \sum_{i=۱}^{p} \frac{۱}{d_j} \end{aligned}$$

که در آن d_jمقادیر ویژه ماتریس $(X^T X)$ می‌باشد.

نمودار پراکنش:

در مطالعه رابطه بین دو متغیر، اولین وساده‌ترین قدم رسم داده‌ها به صورت نقاطی روی یک صفحه نمودار است. شکل حاصله که نمودار پراکنش [۵] نامیده می‌شود، یک نمودار ریاضی با استفاده از مختصات دکارتی است وبا استفاده از این نمودار نحوه و میزان پراکندگی و هم‌پراشی (پراکندگی توام) متغیرها آشکار می‌شود. یکی از جنبه‌های قدرتمند نمودار پراکنش، توانائی نشان دادن رابطه غیر خطی است. همچنین با استفاده از این نمودار داده‌های پرت شناسائی می‌شوند.

مثال ۲-۷. نمودار پراکنش داده‌های Cars : این داده‌ها در بسته نرم‌افزاری datasets در نرم‌افزار R موجود است و مربوط به سرعت و فاصله طی شده از لحظه ترمز کردن تا توقف کامل برای ۵۰ ماشین می‌باشد که در سال ۱۹۲۰ ثبت شده‌اند.

[۵] scatter plot