دانشگاه صنعتی اصفهان

دانشکده علوم ریاضی

آزمون‌های استوار در مدل رگرسیون لجستیک

پایان‌نامه کارشناسی ارشد (آمار اقتصادی و اجتماعی)

الهام همایونی

اساتید راهنمای پایان‌نامه

دکتر سروش علیمرادی
دکتر علی رجالی

۱۳۹۰

دانشگاه صنعتی اصفهان

دانشکده علوم ریاضی

پایان‌نامه کارشناسی ارشد (آمار اقتصادی و اجتماعی)خانم الهام همایونی

تحت عنوان

آزمون‌های استوار در مدل رگرسیون لجستیک

در تاریخ توسط کمیته تخصصی زیر مورد بررسی و تصویب نهائی قرار گرفت.

۱ — استاد راهنمای پایان‌نامه دکتر سروش علیمرادی

۲ — استاد راهنمای پایان‌نامه دکتر علی رجالی

۳ — استاد داور ۱

()

۴ — استاد داور ۲

سرپرست تحصیلات تکمیلی دانشکده دکتر اعظم اعتماد

فهرست مطالب

چکیده:

کاربرد مدل رگرسیون لجستیک در طول دهه‌های اخیر پیشرفت‌های بسیاری داشته است. این مدل نخست در تحقیقات اپیدمیولوژی به کار می‌رفت، اما امروزه در بسیاری از زمینه‌ها از جمله مدیریت، اقتصاد، جرم‌شناسی، مهندسی پزشکی به کاربرده می‌شود.

در این پایان‌نامه ابتدا مدل رگرسیون لجستیک معرفی می‌شود و برآوردگرهای کلاسیک به منظور برآورد ضرائب این مدل بیان می‌شود. با توجه به این‌که برآوردگرهای کلاسیک به شدت تحت تأثیر مشاهدات پرت قرار می‌گیرند، برآوردگرهای استوار معرفی می‌شوند. سطح و توان آزمون‌های کلاسیک با ورود مشاهده پرت فروریزش می‌کند، بدین ترتیب آزمون‌های استوار تعریف می‌شود. برای انجام آزمون فرض پارامتری، آماره آزمون والد استوار بر اساس برآوردگر بیانکو و یوهی وزنی بیان می‌شود و توزیع مجانبی آماره آزمون مورد مطالعه قرار می‌گیرد. با ارائه یک مثال عددی ابتدا برآوردگرهای استوار و کلاسیک و $p-$مقدار آزمون‌های استوار و کلاسیک محاسبه می‌شوند و سپس پایداری $p-$مقدار هر دو آزمون مورد بررسی قرار می‌گیرند. در پایان با انجام یک مطالعه شبیه‌سازی و استفاده از روش مونت‌کارلو پایداری سطح و توان آزمون مورد مطالعه قرار می‌گیرد.

رده‌بندی موضوعی: ۶۲h۳۵.

کلمات کلیدی: رگرسیون لجستیک، استواری، آزمون فرض‌پارامتری، توزیع مجانبی

فصل ۱

مقدمه

مشکل دنیای امروز کمبود داده و اطلاعات کافی برای تصمیم‌گیری‌های علمی نیست، بلکه محققان در بیشتر زمینه‌های مطالعاتی با سیلی انبوه از داده‌های خام مواجه هستند که برای ارائه تحلیل‌های مفید و کارآمد نیازمند روشی مناسب برای استخراج اطلاعات از آن داده‌ها هستند. در بسیاری از موارد با داده‌هایی رو به رو می‌شویم که از نوع دودویی هستند. یک مدل متداول برای تحلیل داده‌ها، رگرسیون لجستیک است که در کلاس مدل‌های خطی تعمیم‌یافته[۱] قرار می‌گیرد و مدل‌سازی متغیر پاسخ غیر نرمال را ممکن می‌سازد. این مدل در بسیاری از زمینه‌ها مانند مطالعات اجتماعی و علوم پزشکی کاربرد دارد. از جمله متغیرهای پاسخ دودویی می‌توان به (مرگ و حیات)، (بهبود یا عدم بهبود بیماری)، (موافق یا مخالف بودن با یک موضوع) اشاره کرد.

در مدل رگرسیون لجستیک ضرائب مدل مجهول است. به منظور استفاده از مدل، لازم است ضرائب، برآورد شوند. یک روش متداول، برآوردگر ماکزیمم درستنمایی[۲] است، این برآوردگر در صورت نرمال بودن توزیع خطاها مناسب است اما اگر خطاها دارای توزیعی با دم طولانی باشد(داده پرت وجود داشته باشد) چندان کارا نیست. کروکس[۳] و همکاران [۱۷] رفتار فروریزشی MLE در مدل رگرسیون لجستیک مطالعه کرده‌اند و نشان داده‌اند که اضافه شدن داده پرت باعث فروریزش آن می‌شود. بنابراین آماردانان برآوردگرهای استوار را معرفی کردند که برای تأثیر مشاهدات پرت در تحلیل داده‌ها به کار می‌رود.

---

[۱] Generalized Linear Model(GLM)

[۲] Maximum Likelihood Estimator (MLE)

[۳] Croux

در چند دهه گذشته مطالعات زیادی برای به دست آوردن برآوردگرهای استوار برای پارامترها در مدل رگرسیون لجستیک و به طور کلی در چارچوب مدل های خطی تعمیم یافته، انجام شده است که می‌توان به کارهای پریجیبن[۴] [۳۲]، استفانسکی[۵] و همکاران [۴۰]، کانچ[۶] و همکاران[۲۹]، مورجنتلر[۷] [۳۰]، کارل[۸] و پدرسن[۹] [۱۰]، کریسمن[۱۰] [۱۴]، بیانکو[۱۱] و یوهی[۱۲] [۴]، کروکس و هائسبرک[۱۳] [۱۶] و باندل[۱۴] [۷] اشاره کرد.

بیانکو و یوهی[۴] کلاس جدیدی از برآوردگرهای M برای مدل رگرسیون لجستیک معرفی کردند. این برآوردگر نسبت به مشاهدات پرت در جهت y (متغیر پاسخ) استوار است. آن‌ها سازگاری و نرمال مجانبی این برآوردگر را نیز اثبات کردند.

کروکس و هائسبرک[۱۶] حالت وزنی برآوردگر بیانکو و یوهی را به منظور کاهش اثر مشاهدات پرت در جهت x (متغیر مستقل) ارائه دادند و الگوریتم سریع را برای محاسبه برآوردگر استوار برای مدل رگرسیون لجستیک معرفی کردند و شرایط وجود این برآوردگر را در نمونه‌های متناهی بیان کردند. همچنین نتایج استفاده از برآوردگر بیانکو و یوهی وزنی نیز مورد بررسی قرار داده‌اند و با شبیه‌سازی و مثال، این برآوردگرهارا تشریح کردند.

برای آزمون پارامترهای رگرسیون لجستیک، آزمون‌هایی که بر اساس برآوردگرهای کلاسیک انجام شود تحت تأثیر مشاهدات پرت قرار می‌گیرد و سطح و توان آزمون پایدار نیست ولی آزمون‌هایی که بر اساس برآوردگرهای استوار هستند، سطح و توان پایدارتری نسبت به آزمون کلاسیک دارند. در این زمینه هریتر[۱۵] و رنچتی[۱۶] [۲۵] آزمون های استوار برای مدل پارامتری تعمیم یافته را، شامل مدل رگرسیون لجستیک، معرفی کردند.

کانتونی[۱۷] و رنچتی[۱۸] [۹] انحراف های استوار بر اساس تابع شبه درستنمایی را تعریف کرده‌اند و یک

[۴] Pregibon
[۵] Stefanski
[۶] kunsch
[۷] Morgenthaler
[۸] Carroll
[۹] Pederson
[۱۰] Christmann
[۱۱] Bianco
[۱۲] Yohai
[۱۳] Haesbroeck
[۱۴] Bondell(2005,2008)
[۱۵] Heritier
[۱۶] Roncheti
[۱۷] Cantoni
[۱۸] Roncheti

خانواده از آماره آزمون ها را برای انتخاب مدل در مدل های خطی تعمیم یافته، معرفی نموده‌اند.
در این پایان‌نامه مدل رگرسیون لجستیک معرفی می‌شود و نحوه‌ی برآورد ضرائب آن مدل (β) به روش کلاسیک و برآوردگر استوار بررسی می‌شود و به دلیل عدم استواری برآوردگر کلاسیک برآوردگرهای استوار معرفی می‌شود. از آنجا که به آزمون‌های استوار توجه کمتری نسبت به برآوردگر استوار شده است، آزمون فرض پارامتری، پارامترهای رگرسیون لجستیک بررسی می‌شود و آماره والد بر اساس نوع وزنی برآوردگر بیانکو و یوهی که توسط کروکس و هائسبرک معرفی شد[۱۶]، بیان می‌شود. هدف بیان شکل استوار آماره آزمون والد کلاسیک است. این آزمون به شکل مربعی و بر اساس برآوردگرهای استوار پارامترهای رگرسیونی و ماتریس کواریانس مجانبی آن است. رفتار مجانبی این آماره آزمون مشابه حالت کلاسیک است و دارای توزیع کای‌دو است.
مطالب گرد آوری شده در قالب فصل‌های زیر ارائه می‌شود.
در فصل ۲، بعضی از تعاریف و قضایایی که در فصل‌های بعد مورد نیاز است مطرح می‌شود.
در فصل ۳، ابتدا برآوردگر ماکزیمم درستنمایی معرفی می‌شود و با توجه به آن‌که این برآوردگر در مقابل مشاهدات پرت استوار نیست، برآوردگر بیانکو و یوهی معرفی خواهد شد. شرایط وجود این برآوردگر با استفاده از یک گزاره مطرح شده و بعد از آن نحوه‌ی محاسبه این برآوردگر توضیح داده می‌شود. این برآوردگر در مقابل مشاهدات پرت در جهت y استوار است، ولی در مقابل مشاهدات پرت در جهت x استوار نیست. بدین منظور برای کاهش اثر مشاهدات پرت در جهت x، برآوردگر وزنی بیانکو و یوهی مورد استفاده قرار گرفته و خصوصیات مجانبی آن بیان می‌شود. بعد از آن آماره آزمون برای این برآوردگر معرفی و رفتار مجانبی آن بررسی می‌شود. در پایان با ارائه یک مثال عددی برآوردگرهای استوار و کلاسیک و $p-$ مقدار آزمون کلاسیک و استوار محاسبه می‌شود و پایداری $p-$مقدار بررسی می‌شود.
در فصل ۴ روش‌های عددی بهینه‌سازی بیان می‌شود.
در فصل ۵، با استفاده از روش مونت کارلو، پایداری سطح و توان آزمون‌های استوار و کلاسیک مورد مقایسه قرار می‌گیرند. برنامه مربوط به تهیه جدول‌ها و نمودارها در پیوست آورده شده است.

فصل ۲

پیش نیازها

در این فصل به منظور یادآوری، برخی از تعاریف، قضایا و اصطلاحات مورد نیاز در این پایان‌نامه، بیان می‌شوند.

تعریف ۱.۲ سه‌تایی $(\Omega, \mathcal{A}, P)$ یک فضای احتمال است و $X : \Omega \longrightarrow \mathbb{R}$ یک متغیر تصادفی است (یعنی تابع اندازه‌پذیر روی $(\Omega, \mathcal{A}, P)$ است.)

قضیه ۱.۲ نامساوی مارکوف[۴۱]. اگر X یک متغیر تصادفی و g تابعی غیر منفی باشد و $E[g(X)] < \infty$، آن‌گاه برای هر $a > 0$،

$$P(g(X) \geq a) \leq \frac{E[g(X)]}{a}$$

قضیه ۲.۲ نامساوی چبیشف[۱] [۴۱]. با انتخاب $g(x) = |x|$، در نامساوی مارکوف و با توجه به این‌که $P(|X| \geq a) = P(X^2 \geq a^2)$، نامساوی زیر برقرار است.

$$P(|X| \geq a) \leq \frac{E[X^2]}{a^2} \quad a > 0$$

[۱] Chebyshev Inequality

حالت خاصی از نامساوی چبیشف با انتخاب $g(X) = |X - \mu|$ به صورت زیر به دست می آید:

$$P(|X - \mu| \geq k\sigma) \leq \frac{۱}{k^۲} \quad , k > ۰$$

قضیه **۳.۲** (نامساوی کوشی شوارتز[۲])[۴۲]. اگر X و Y دو متغیر تصادفی با تابع توزیع توأم $F(X,Y)$ باشند بطوری که $E(|X|^۲)$ و $E(|Y|^۲)$ متناهی هستند، آن‌گاه:

$$|E(XY)| \leq E(|XY|) \leq \sqrt{E(X^۲)E(Y^۲)}$$

حالت دیگری از نامساوی شوارتز، به فرم زیر است:

$$Cov^۲(X,Y) \leq Var(X)Var(Y).$$

قضیه **۴.۲** (نامساوی جنسن)[۳] [۴۱]. اگر φ یک تابع محدب باشد که روی بازه‌ای شامل برد X تعریف شده است و اگر امید ریاضی X و $\varphi(X)$ هر دو موجود باشند، آن‌گاه

$$\varphi(E(X)) \leq E(\varphi(X))$$

.

تعریف **۲.۲** [۱۳]. دنباله $\{X_n\}$ از متغیرهای تصادفی روی فضای احتمال $(\Omega, \mathcal{A}, P)$ تقریباً همه جا[۴] (a.s) به متغیر تصادفی X همگراست، هر گاه یک مجموعه N با احتمال صفر وجود داشته باشد که به ازای هر $w \in \Omega - N$

$$\lim_{n\to\infty} X_n(w) = X(w)$$

این همگرایی با $X_n \xrightarrow{a.s} X$ نمایش داده می‌شود.

---

[۲] Cauchy Schwartz

[۳] Jensen

[۴] almost surely

تعریف **۳.۲** [۱۳]. دنباله‌ی متغیرهای تصادفی $\{X_n\}$ در احتمال به X همگراست هر گاه به ازای هر $\varepsilon > ۰$

$$\lim_{n\to\infty} P[|X_n - X| > \varepsilon] = ۰$$

این همگرایی با $X_n \xrightarrow{p} X$ نمایش داده می‌شود.

تعریف **۴.۲** [۱۳]. هر گاه $\{F_n\}$ دنباله‌ای از توابع توزیع برای دنباله $\{X_n\}$، و F تابع توزیع متغیر تصادفی X باشد و به ازای هر نقطه پیوستگی F، $\lim_{n\to\infty} F_n(x) = F(x)$، گفته می‌شود که X_n در توزیع به X همگراست و با $X_n \xrightarrow{D} X$ نمایش داده می‌شود. ($X_n \xrightarrow{D} X$ اگر و تنهااگر $\lim_n P[X_n \le x] = P(X \le x)$ به ازای تمام مقادیر که برای آن‌ها $P(X = x) = ۰$)

قضیه **۵.۲** (قانون ضعیف اعداد بزرگ[۵]) [٦]. اگر $X_۱, X_۲, ...$ متغیرهای تصادفی مستقلی (نه لزوماً با توزیع یکسان) هر یک با میانگین متناهی و واریانس متناهی باشند. علاوه بر آن واریانس‌ها به طور یکنواخت کراندار باشند، در این صورت وقتی برای هر n، $S_n = X_۱ + ... + X_n$، آن‌گاه $[S_n - E(S_n)]/n$، در احتمال به صفر همگرا است. یعنی به ازای هر $\epsilon > ۰$

$$\lim_{n\to\infty} P\{|\frac{S_n - E(S_n)}{n}| \ge \epsilon\} = ۰$$

قضیه **٦.۲** (قانون قوی اعداد بزرگ[٦])[٦]. اگر $\{X_n\}$ها یک دنباله‌ای از متغیرهای تصادفی مستقل باشند، به نحوی که برای هر n، $E[X_n] = ۰$، در این صورت هر گاه $\sum_n \frac{var(X_n)}{n^۲} < \infty$، آن‌گاه $n^{-۱}\sum_{k=۱}^{n} X_k \xrightarrow{a.s} ۰$

قانون قوی اعداد بزرگ در حالتی که X_i ها مستقل و هم توزیع باشند، به صورت زیر است:

قضیه **۷.۲** [٦]. اگر متغیرهای تصادفی $X_۱, X_۲, ...$ مستقل و دارای توزیع یکسانی باشند $(i.i.d)$ و $E[X_i] = \mu$ در این صورت $n^{-۱}\sum_{k=۱}^{n} X_k \xrightarrow{a.s} \mu$.

---

[۵] Weak Law of Large Numbers

[٦] Strong Law of Large Numbers

قضیه ۸.۲ (قضیه اسلاتسکی[۷]). اگر $\{X_n, Y_n\}_{n\geq 1}$ دنباله‌ای از جفت متغیرهای تصادفی و c یک مقدار ثابت باشد، $Y_n \xrightarrow{p} c$ و $X_n \xrightarrow{D} X$، آن‌گاه روابط زیر برقرارند:[۲۲]

$$X_n + Y_n \xrightarrow{D} X + c \qquad (۱)$$

$$\begin{cases} X_n Y_n \xrightarrow{L} cX, & c \neq 0 \\ X_n Y_n \xrightarrow{p} 0, & c = 0 \end{cases} \qquad (۲)$$

$$\frac{X_n}{Y_n} \xrightarrow{D} \frac{X}{c}, c \neq 0 \qquad (۳)$$

قضیه ۹.۲ [۱۹]. (قضیه همگرایی غالب[۸]).
اگر $X_n \xrightarrow{a.s} X$، و برای تمام n ها، رابطه زیر برقرار باشد:

$$|X_n| \leq |Y| \quad , \quad E(|Y|) < \infty$$

آن‌گاه X هم انتگرال پذیر است و $E(X) = \lim_{n\to\infty} E(X_n)$

تعریف ۵.۲ دنباله‌ی برآوردگرهای $\{T_n\}$ برای تابعی از پارامتر θ $(\Gamma(\theta))$ سازگار است، هر گاه $T_n \xrightarrow{p} \Gamma(\theta)$ وقتی n به سمت بی‌نهایت میل کند. یعنی برای هر $\varepsilon > 0$ و $\delta > 0$، عدد صحیح مثبتی مانند $n_0 = n_0(\varepsilon, \delta)$ وجود دشته باشد به طوری‌که برای هر $n > n_0$، $P_\theta(|T_n - \Gamma(\theta)| > \varepsilon) < \delta$ یا برای هر $\varepsilon > 0$، $\lim_{n\to\infty} P_\theta(|T_n - \Gamma(\theta)| > \varepsilon) = 0$ [۴۱]

۲—۱ داده پرت

مشاهده دورافتاده یا پرت[۹] داده‌ای است که از توده اکثریت مشاهدات دور باشد. این گونه مشاهدات پس از شناسایی باید به دقت مورد بررسی قرار گیرند تا معلوم شود که آیا دلیلی برای رفتار غیر عادی آن‌هاوجود

[۷] Slutsky
[۸] Dominated Convergence Theorem
[۹] Outlier

دارد یا خیر. وجود داده‌های پرت گاهی غیر عادی، اما قابل توضیح است. اندازه‌گیری غلط، ثبت و یا انتقال نادرست داده‌ها، از کار افتادن وسیله اندازه‌گیری ، مثال‌هایی از این اتفاقات هستند. در چنین مواردی بایستی در صورت امکان، اتفاق مذکور اصلاح و در غیر این صورت آن داده از مجموعه داده‌ها حذف شود. تاکید می‌شود که قبل از خارج کردن داده مذکور بایستی دلایل منطقی کافی برای نامناسب بودن آن داده دور افتاده وجود داشته باشد.
بعضی اوقات داده دور افتاده غیر عادی است، اما آن مشاهده‌ای کاملاً معتبر است. وجود چنین داده‌هایی می‌توانند ناکارآمدی مدل را نشان دهند و اغلب سرنخ‌های با ارزشی جهت ساختن مدل رگرسیونی مناسب‌تر برای تحلیل‌گر فراهم می‌کنند. گاهی نقاط پرت اثر نامتناسبی را بر مدل رگرسیونی اعمال می‌کنند، بدین معنی که برآورد پارامترها تحت تاثیر مجموعه‌ای کوچک از داده‌ها قرار می‌گیرد. این داده‌ها را نقاط تاثیرگذار[۱۰] گویند. با این توضیحات، یک تحلیل‌گر باید داده‌های پرت را شناسایی کند و از تأثیرات آن بر نتایج و تحلیل‌ها آگاه باشد.

تعریف **۶.۲** عدم حساسیت به انحرافات کوچک از فرض‌ها، استواری گفته می‌شود[۳۶].

۲—۲ اندازه استواری یک برآورد

دو معیار سنجش استواری به صورت زیر است:

۱) نقطه فروریزش بیانگر آن است که چگونه برآوردگر توسط کسری از داده‌ها تحت تأثیر قرار می‌گیرد.

۲) منحنی تأثیر که با استفاده از آن، می‌توان اطلاعاتی از تأثیر داده پرت بر برآوردگر را به دست آورد.

۲—۲—۱ نقطه فروریزش

ابتدا مفهوم اریبی تعریف می‌شود و سپس با استفاده از آن مفهوم نقطه فروریزش ارائه می‌گردد.

تعریف **۷.۲** اگر $Z = \{(\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{ip}, y_i),\ i = 1, ..., n\}$ یک نمونه n تایی و T یک برآوردگر رگرسیونی باشد(یعنی با اعمال T بر نمونه Z، برآورد ضرایب معادله رگرسیون $T(Z) = \hat{\beta}$ حاصل شود) و تمام نمونه‌های ممکن Z' که توسط جایگزینی هر m تا از نقاط داده‌های اصلی توسط مقادیر دلخواه (حتی مشاهدات خیلی پرت) با $\mathbb{Z}$ نمایش داده می‌شود، آن‌گاه

[۱۰] Influential points

$$bias(m;T,Z) = Sup_{Z' \in \mathbb{Z}} ||T(Z') - T(Z)||$$

که در آن ||.|| نرم اقلیدسی است، ماکزیمم اریبی نامیده می‌شود[۳۶].

اگر این اریبی نامتناهی باشد، بدین معنی است که m داده پرت می‌توانند اثر زیادی بر T داشته باشند.در این صورت می‌توان این مسأله را با جمله «برآوردگر به سمت پایین فرو می‌ریزد» توصیف کرد.

تعریف **۸.۲** [۳۶]. نقطه فروریزش برآوردگر T براساس نمونه n تایی Z به صورت زیر تعریف می‌شود

$$\epsilon_n^*(T,Z) = min\{\frac{m}{n}; bias(m;T,Z) = \infty\}$$

به عبارت دیگر، این معیار کوچکترین کسر عدم خلوص است که باعث می‌شود برآوردگر T از $T(Z)$ دور شود. هر چه این کسر بزرگ‌تر باشد، نشان دهنده استوارتر بودن برآوردگر T است.

۲–۲–۲ منحنی تأثیر

اگر F یک تابع توزیع k بعدی و θ پارامتری از جامعه باشد که F به آن وابسته است، آن‌گاه می‌توان نوشت $\theta = T(F)$ که در این صورت T را تابعک آماری نامند. ساده‌ترین مثال از تابعک آماری میانگین متغیر تصادفی X، $E_F[X]$ است.

$$E_F[X] = T[F] = \int x dF(x)$$

می‌توان مقادیر کوچک ناخالصی تابع توزیع F در $z_\circ = (x_\circ, y_\circ)$ را با در نظرگرفتن یک توزیع آمیخته به صورت زیر مدل‌سازی کرد:

$$F_t = (1-t)F + t\delta_{z_\circ}$$

که در آن $\delta_{z_\circ}$ تابع توزیع مقادیر ناخالصی $z_\circ$ است (صفحه ۸۲ از کتاب سیبر [۳۹]).
منحنی تأثیر به صورت زیر تعریف می‌شود.

تعریف **۹.۲** منحنی تأثیر یک تابعک آماری T، مشتق $T(F_t)$ نسبت به t در نقطه $t = \circ$ است که میزان واکنش T نسبت به مقادیر ناخالصی کوچک $z_\circ$ را نشان می‌دهد[۳۹].

$$IC(F, z_\circ) = \frac{dT(F_t)}{dt}|_{t=\circ}$$

نکته ۱.۲ اگر IC نسبت به $x(y)$ کران‌دار باشد نتیجه می‌شود که نسبت به x (y) استوار است[۳۹].

تعریف ۱۰.۲ برآوردگر استوار برآوردگری است که با ورود داده پرت به نمونه فروریزش نمی‌کند[۳۹].

تعریف ۱۱.۲ روش رگرسیون استوار به روشی گفته می‌شود که نه تنها در صورت نرمال بودن توزیع خطاها خوب عمل می‌کند، بلکه نسبت به انحراف‌های کوچک از فرض نرمال بودن نیز حساسیت اندکی دارد[۲۷].

۲—۳ مدل رگرسیونی

یکی از روش‌های کاربردی برای تحلیل داده‌ها در بین ابزارهای آماری، تحلیل رگرسیونی است. تحلیل رگرسیونی، روشی کارآمد برای بررسی و مدل سازی ارتباط بین متغیرها است که از آن در توصیف داده‌ها، برآورد پارامترهای مجهول، پیش‌گویی و کنترل استفاده می‌شود. به منظور دست‌یابی به یک مدل رگرسیونی به مشاهداتی از متغیرهای مستقل (توضیحی) $\mathbf{x} = (x_1, x_2, ..., x_p)$ و نیز مشاهداتی از متغیر وابسته (پاسخ) y نیاز است. ساده‌ترین حالت که در آن تنها یک متغیر مستقل x وجود داشته باشد، عبارت است از وقتی که رابطه بین این دو متغیر یک رابطه خطی به صورت زیر بین مشاهدات وجود داشته باشد:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, 2, ...n \qquad (۱—۲)$$

که در آن β_0 عرض از مبدا و β_1 شیب خط است و مولفه‌های ϵ_i، به عنوان خطای تصادفی تلقی می‌شوند. بدین معنی که به عنوان مقادیر مشاهده شده متغیری تصادفی، اندازه ناتوانی مدل در برازش دقیق داده‌ها را اندازه‌گیری می‌کنند. این خطا ممکن است به دلیل عدم حضور برخی از متغیرهای موثر، خطاهای تصادفی مربوط به مشاهدات و اندازه‌گیری‌ها و موارد دیگر باشد [۳۶].

روش‌های مختلفی برای برآورد پارامترهای یک مدل رگرسیونی وجود دارد که در زیر تعدادی از آن‌ها معرفی می‌شوند.

۲—۴ برآوردگر حداقل مربعات

معمول‌ترین روش در برآورد پارامترهای یک مدل خطی استفاده از روش کمترین مربعات است. اگر توزیع خطاها نرمال باشد برآوردگر حداقل مربعات[۱۱] β به گونه‌ای انتخاب می‌شود که مجموع توان دوم انحراف‌ها یعنی

$$\sum_{i=1}^{n} \varepsilon_i^2 \qquad (۲—۲)$$

را کمینه کند. منحنی تأثیر LSE عبارت است از $IC(z_0, F) = \Sigma_F^{-1} x_0 [y_0 - x_0' T(F)]$ که در آن $T(F) = \{\Sigma_F\}^{-1} \gamma_F$، $\Sigma_F = E_F[XX']$، $\gamma_F = E_F[XY]$، وقتی F تابع چگالی Y، X یک بردار تصادفی و $z_0 = (x_0, y_0)$ مقدار توأم مشاهده شده برای (X, Y) است. منحنی تأثیر نسبت به هر دو مقدار X_0 و Y_0 کران دار نیست و این بدین معنی است که LSE استوار نیست[۳۶] و [۳۹].

۲—۵ برآوردگر L_1

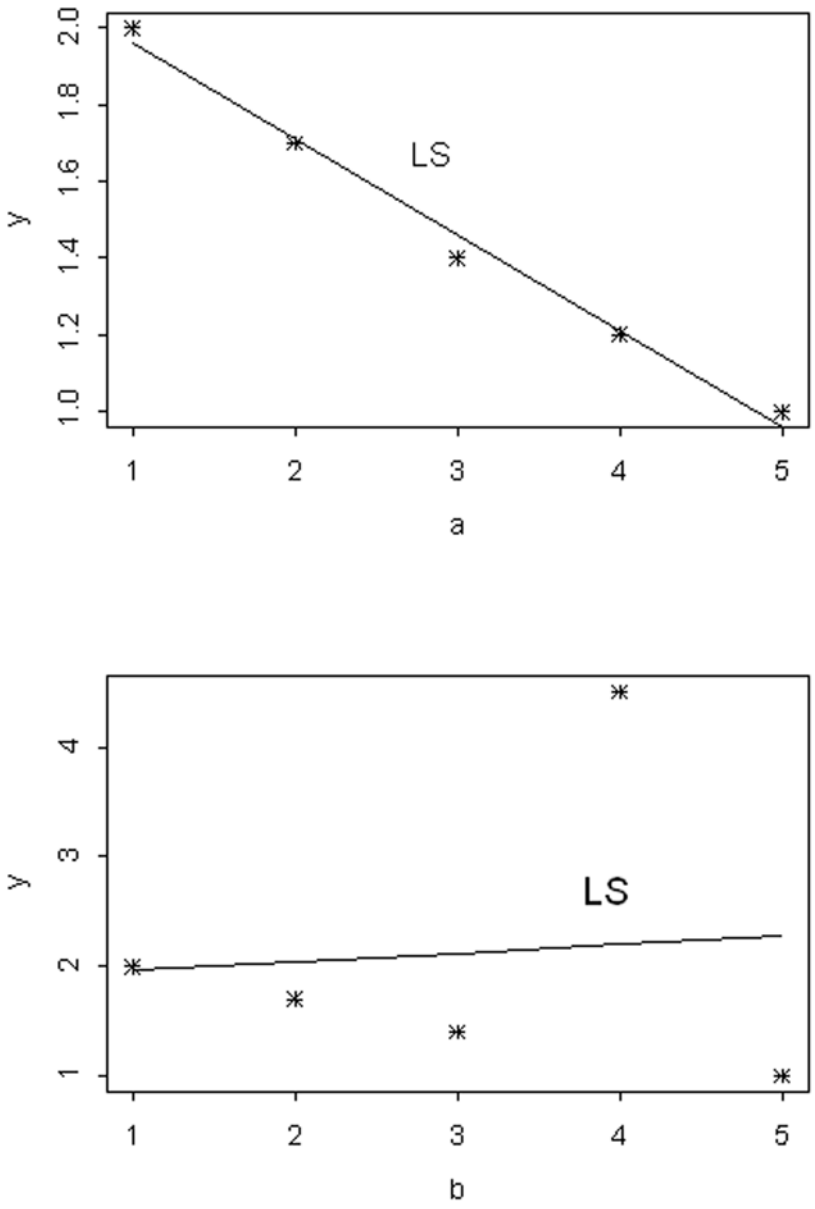
با قرار دادن $|\varepsilon_i|$ به‌جای ε_i^2 در رابطه‌ی (۲—۲) برآوردگر L_1 به دست می آید. در واقع ساده‌ترین روش در برآورد پارامترهای یک مدل رگرسیونی، استفاده از کمترین قدرمطلق انحرافات است که در مقایسه با روش حداقل مربعات دارای حساسیت کمتری نسبت به داده‌های پرت است، با این حال نمی‌توان از برآوردگر L_1 به عنوان جایگزین قوی برای برآوردگر حداقل مربعات استفاده کرد، زیرا به شدت توسط تک مشاهده پرت در جهت x (داده بانفوذ) تحت تأثیر قرار می‌گیرد،[۳۶] که این موضوع در بخش بعد بررسی می‌شود.

۲—۶ بررسی استواری روش LS و L_1 نسبت به داده‌های پرت

برای بررسی اثر داده‌های پرت، به مدل رگرسیونی ساده می‌توان توجه کرد. نمودار وضعیت داده‌ها نسبت به یکدیگر را به سادگی می‌توان در یک نمودار پراکنش نمایش داد. شکل ۱.۲a نمودار پراکنش[۱۲] پنج نقطه است که تقریباً روی یک خط مستقیم قرار می‌گیرند و خط رگرسیونی کمترین مربعات (LS) به خوبی به آن‌ها برازش می‌یابد. حال اگر مشاهده چهارم اشتباه گزارش داده‌شده باشد و یا نوشته شود. در این صورت نقطه (x_4, y_4) از بقیه مشاهدات و نیز از خط ایده‌آل دور می‌افتد. شکل ۱.۲b چنین موقعیتی را نشان می‌دهد. این نقطه یک « داده پرت در مسیر y» نامیده می‌شود و اثر نسبتا زیادی بر خط LS دارد، چرا که

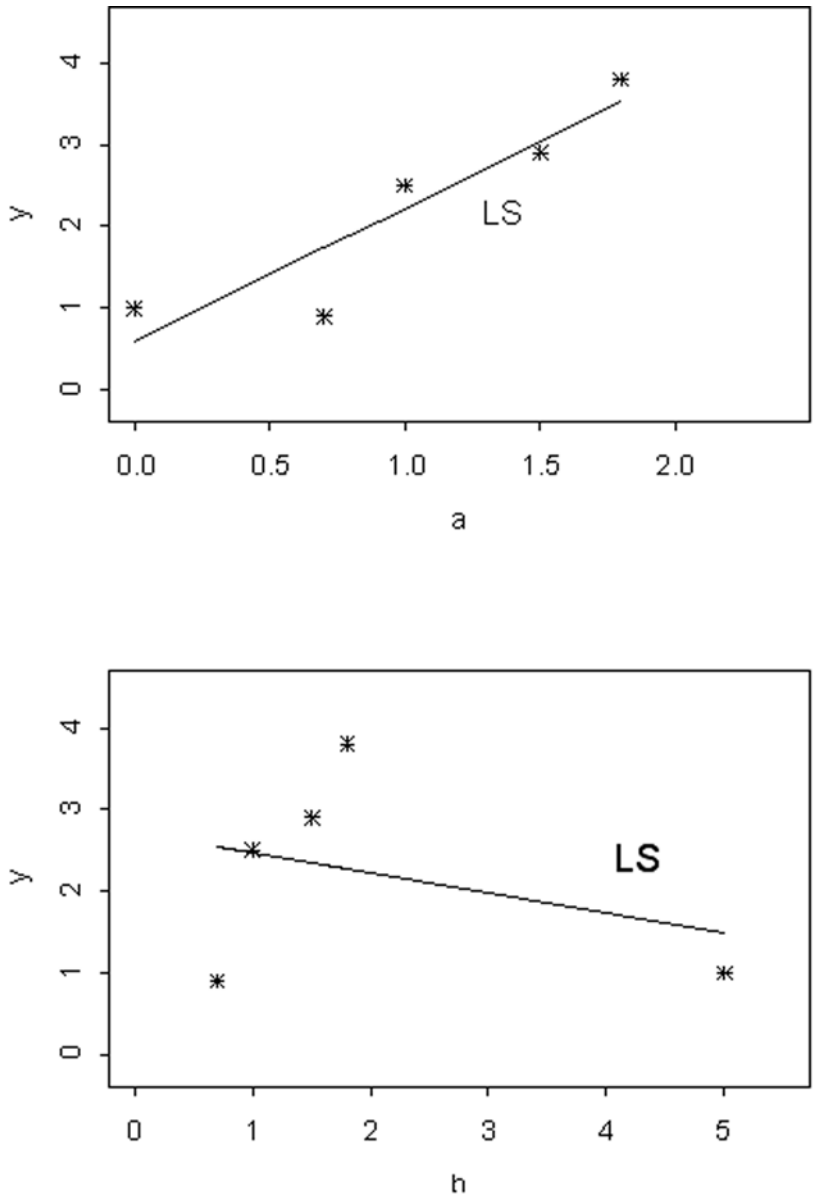
[۱۱] Least Squares Estimator(LSE)
[۱۲] Scatter Plot

خط برازش شده در این حالت با خط ایده‌آل در شکل ۱.۲a تفاوت دارد [۳۶]. همچنین این داده دارای باقیمانده بزرگی است. در رگرسیون چندگانه (بیش از یک متغیر توضیحی) طریق مقادیر باقیمانده‌ها شناسایی می‌شوند [۴۴].



شکل ۱.۲. تاثیر داده پرت در جهت متغیر پاسخ بر برآورد کمترین مربعات

امکان وجود داده پرت در متغیرهای توضیحی وجود دارد. پس موقعیتی در نظر گرفته می‌شود که متغیر توضیحی شامل یک داده پرت باشد. شکل ۲.۲a حاوی پنج نقطه با برازش خوب LS است. اکنون فرض می‌شود که در مقدار x_1 تغییری رخ دهد. شکل ۲.۲b نشان دهنده چنین وضعیتی است. نقطه (x_1, y_1) « داده پرت در مسیر x» نامیده می‌شود و اثر آن روی خط رگرسیون بسیار زیاد است [۳۶].



شکل ۲.۲. تاثیر داده اهرمی بر برآورد کمترین مربعات

چون وضعیت قرار گرفتن نقاط در فضای x، در تعیین خواص مدل اهمیت دارد، به همین دلیل، معمولا نقاط پرت فضای x را نقاط بانفوذ[۱۳] می‌نامند. چنین نقاطی معمولا باقی‌مانده کمی دارند چرا که خط رگرسیونی را به سمت خود کشیده‌اند. بنابراین از طریق بررسی باقی‌مانده‌ها شناسایی نمی‌شوند. تاثیرگذاری نقاط پرت را می‌توان با کنار گذاشتن این نقاط و برازش مجدد معادله رگرسیونی و مقایسه مقادیر به دست آمده جدید و قبلی، مشخص کرد [۴۴].

زمانی که یک نقطه پرت در نزدیکی خط رگرسیونی که توسط اکثریت داده‌ها تعیین می‌شود قرار گیرد،

[۱۳] Leverage Point