دانشگاه پیام نور

دانشکده علوم پایه

پایان‌نامه

برای دریافت مدرک کارشناسی ارشد

رشته آمار ریاضی

گروه آمار

عنوان پایان‌نامه:

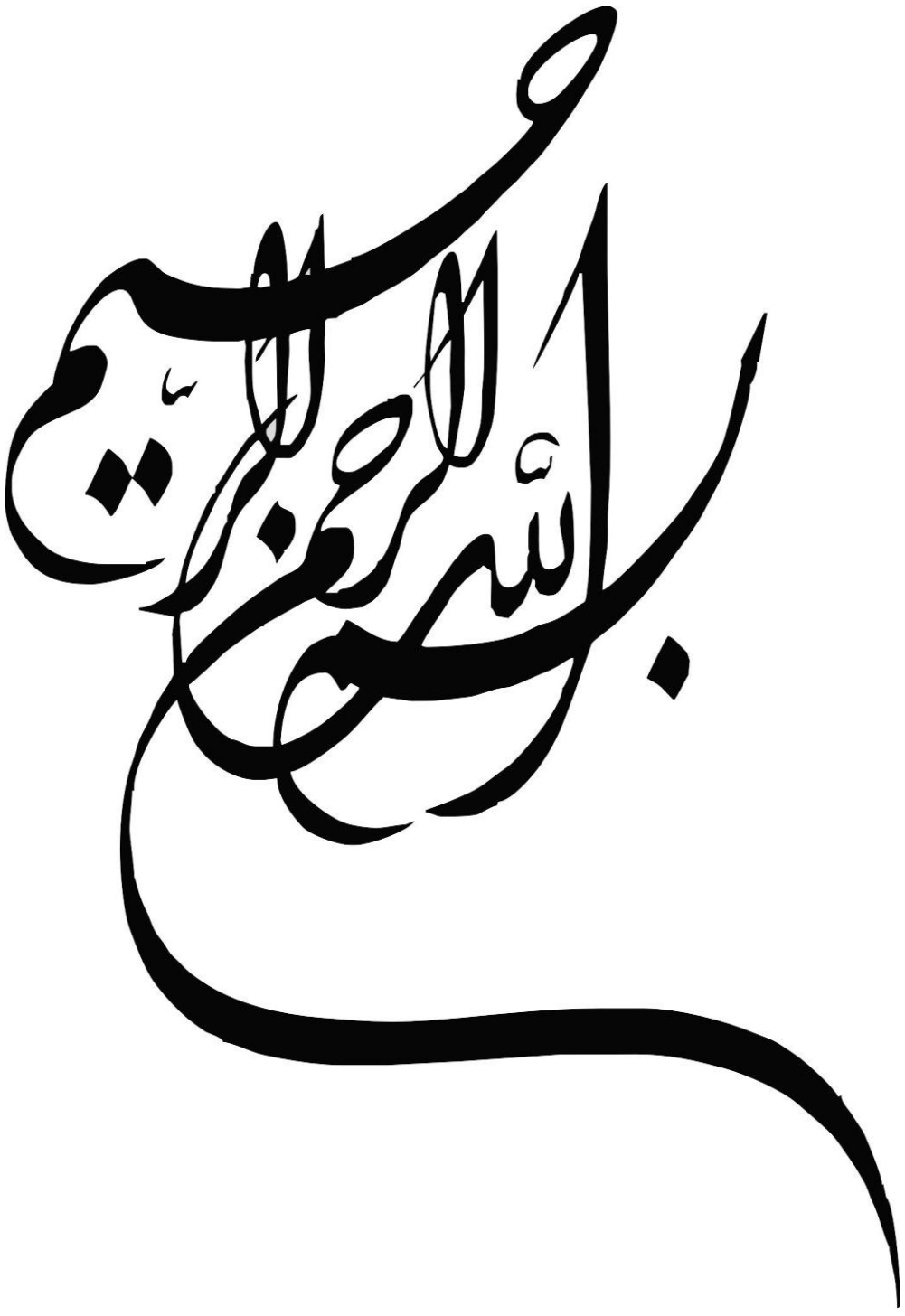
مطالعه‌ای بر برآوردیابی استوار در مدل‌های رگرسیونی

مریم افتخاری

استاد راهنما :

دکتر نرگس عباسی

شهریور ۱۳۹۲


بسم الله الرحمن الرحيم

دانشگاه پیام نور

دانشکده علوم پایه

مرکز شیراز

پایان‌نامه

برای دریافت مدرک کارشناسی ارشد

رشته آمار ریاضی

گروه آمار

عنوان پایان‌نامه:

مطالعه‌ای بر برآوردیابی استوار در مدل‌های رگرسیونی

مریم افتخاری

استاد راهنما :

دکتر نرگس عباسی

شهریور ۱۳۹۲

جمهوری اسلامی ایران
وزارت علوم، تحقیقات و فناوری
دانشگاه پیام نور استان فارس

دانشگاه پیام نور شیراز
باسمه تعالی

تاریخ : ۹۲/۰۶/۲۸
شماره : ۱۰۴/۱۱۲۷۶
پیوست :

صورتجلسه دفاع از پایان‌نامه دوره کارشناسی ارشد

جلسه دفاع از پایان‌نامه دوره کارشناسی ارشد **خانم مریم افتخاری** دانشجوی رشته **آمار ریاضی** به شماره دانشجویی **۹۰۰۰۱۴۲۲۸** با عنوان:

" مطالعه‌ای بر برآوردیابی استوار در مدل‌های رگرسیونی"

با حضور هیأت داوران در روز **پنج‌شنبه** مورخ **۱۳۹۲/۰۶/۲۸** ساعت **۷:۳۰ صبح** در محل ساختمان **غدیر دانشگاه پیام نور شیراز** برگزار شد و هیأت داوران پس از بررسی، پایان‌نامه مذکور را شایسته نمره به عدد ۱۹ به حروف نوزده با درجه عالی تشخیص داد.

ردیف	نام و نام خانوادگی	هیات داوران	مرتبه دانشگاهی	دانشگاه	امضاء
۱	دکتر نرگس عباسی	راهنما	دانشیار	پیام‌نور شیراز	
۲	عبدالرضا بازرگان‌لاری	داور	استادیار	شیراز	
۳	امیر اکبری	نماینده تحصیلات تکمیلی	مربی	پیام‌نور شیراز	

رئیس اداره تحصیلات تکمیلی

شیراز - شهرک گلستان-بلوار دهخدا
قبل از نمایشگاه بین المللی
تلفن : ۶۲۲۲۲۵۵ - ۰۷۱۱
دورنگار : ۶۲۲۲۲۴۹ - ۰۷۱۱
صندوق پستی: ۱۳۶۸ - ۷۱۹۵۵
www.spnu.ac.ir
Email : admin@spnu.ac.ir

گواهی اصالت نشر و حقوق مادی ومعنوی اثر

اینجانب **مریم افتخاری** دانشجوی ورودی سال **۱۳۹۰** مقطع کارشناسی ارشد رشته **آمار ریاضی** گواهی می نمایم چنانچه در پایان نامه خود از فکر ، ایده و نوشته دیگری بهره گرفته ام با نقل قول مستقیم یا غیر مستقیم منبع و ماخذ آن را نیز در جای مناسب ذکر کرده ام . بدیهی است مسئولیت تمامی مطالبی که نقل قول دیگران نباشد بر عهده خویش می دانم و جوابگوی آن خواهم بود .

دانشجو تائید می نماید که مطالب مندرج در این پایان نامه نتیجه تحقیقات خودش می باشد و در صورت استفاده از نتایج دیگران مرجع آن را ذکر نموده است .

نام و نام خانوادگی دانشجو مریم افتخاری

تاریخ و امضاء ۹۲/۶/۲۸

اینجانب **مریم افتخاری** دانشجوی ورودی سال **۱۳۹۰** مقطع کارشناسی ارشد رشته **آمار ریاضی** گواهی می نمایم چنانچه بر اساس مطالب پایان نامه خود اقدام به انتشار مقاله ، کتاب ، و ... نمایم ضمن مطلع نمودن استاد راهنما ، با نظر ایشان نسبت به نشر مقاله ، کتاب ، و ... و به صورت مشترک و با ذکر نام استاد راهنما مبادرت نمایم .

نام و نام خانوادگی دانشجو مریم افتخاری

تاریخ و امضاء ۹۲/۶/۲۸

کلیه حقوق مادی مترتب از نتایج مطالعات ، آزمایشات و نوآوری ناشی از تحقیق موضوع این پایان نامه مطعلق به دانشگاه پیام نور می باشد .

شهریور ۹۲

تقدیم به:

پدر، مادر گرامی‌ام

و به تمام آزاد مردانی که نیک می‌اندیشند و عقل و منطق را پیشه خود نموده و جز رضای الهی و پیشرفت و سعادت جامعه، هدفی ندارند.دانشمندان، بزرگان، و جوانمردانی که جان و مال خود را در حفظ و اعتلای این مرز و بوم فدا نموده و می‌نمایند.

با تشکر و قدردانی فراوان

از استاد گرامی‌ام سرکار خانم دکتر نرگس عباسی به دلیل یاری‌ها و راهنمایی‌های بی‌چشمداشت ایشان که بسیاری از سختی‌ها را برایم آسان‌تر نمودند، چرا که بدون راهنمایی‌های ایشان تأمین این پایان‌نامه بسیار مشکل می‌نمود .

چکیده:

رگرسیون استوار یکی از مهمترین روش‌های آنالیز داده‌هایی می‌باشد که آلوده هستند و نقاط پرت دارند. این روش می‌تواند جهت تعیین نقاط پرت استفاده شود و نتایج پایداری را در حضور داده‌های پرت فراهم آورد. در این پایان‌نامه روش رگرسیون استوار معرفی و در نرم افزار SAS/STAT- Version 9 اجرا شده است. این متن روش‌های مهم رگرسیون استوار را معرفی می‌کند. این روش‌ها شامل برآورد M، برآورد LTS، برآورد S و برآورد MM می‌باشد. روش استوار در آزمون مدل‌های جزئی خطی تعمیم‌یافته بررسی می‌شود به این صورت که یک خانواده از آماره‌های استوار معرفی می‌شود که امکان انتخاب بین یک مدل پارامتری و نیمه‌پارامتری را فراهم می‌آورد. به طور دقیق‌تر تحت یک مدل جزئی خطی تعمیم‌یافته یعنی زمانی که (y_i, x_i, t_i) مشاهدات مستقل هستند بطوری که $y_i|(\mathrm{x}_i, t_i) \sim F(\circ, \mu_i)$ با $Var\left(y_i|(\mathrm{x}_i, t_i)\right) = V(\mu_i)$ و $\mu_i = H(\eta(t_i) + \mathrm{x}_i^T\beta)$ که درآن $H = g^{-1}$ یک تابع پیوندی شناخته شده و $\beta \in \mathbb{R}^p$ یک پارامتر ناشناخته و $\eta(t_i)$ یک تابع هموار ناشناخته می‌باشد. در این روش برآوردگرهای حاصل از β دارای توزیع نرمال مجانبی و سازگاری ریشه n خواهد بود.

واژگان کلیدی:

مدل‌های جزئی خطی تعمیم‌یافته، آزمون‌سازی استوار، برآوردهای استوار، نقاط نفوذی، داده پرت

فهرست

مقدمه

در این پایان‌نامه روش‌های آماری استوار را معرفی می‌شود، روش‌های استوار به روش‌هایی اتلاق می‌شود که حساسیت کمی نسبت به داده‌های پرت داشته و به سادگی تحت تاثیر داده‌های کوچک و بزرگ قرار نگیرند. اولین بار باکس[۱] (۱۹۵۳) از واژه‌ی استوار استفاده کرد. اما چندان مورد توجه قرار نگرفت. آماردانان تا مدت‌ها از بررسی چنین موضوعی گریزان بودند. هوبر[۲] (۱۹۶۴) برای اولین بار برآوردگرهای استوار مکان را مورد بررسی قرار داد.

آمار استوار به مفهوم روش‌هایی برای حل مسائل ناشی از اثر داده‌های پرت است. داده‌های پرت اثرات بدی بر روی برآورد پارامتر و همچنین فرضیات مدل دارند. به طور کلی آمار استوار وقتی بکار می‌رود که فرضیات ریاضی داده‌های برقرار نباشد، ولی عموماً داده‌های پرت باعث از بین بردن فرضیات مدل می‌شود که در این حالت از آمار استوار استفاده می‌شود. تقریباً همیشه وقتی صحبت از آمار استوار به میان می‌آید داده‌های پرت به فکر ما خطور می‌کند.

به طور مثال در اکثر برنامه‌های کاربردی، پارامترهای مدل آمیخته بوسیله برآوردکننده‌ی درستنمایی ماکسیمم[۳] از طریق الگوریتم امید ماکسیمم‌سازی[۴] برآورد شده است. در حالی که برآورد ماکسیمم درستنمایی در برخورد با داده‌های پرت بسیار حساس می‌باشد. هوبر (۱۹۸۱) همپل[۵] و همکاران (۱۹۸۶)روسو[۶] و لروی[۷] (۱۹۸۷). برازش مدل‌های آمیخته را می‌توان مستقیماً برای داده‌ها توسط برآوردکننده‌های استوار محدود استفاده کردند. دلیل آن این است که این برآوردکننده‌های استوار به تناسب مدل پارامتری برای اکثریت داده‌ها طراحی شده‌اند. در حالی که داده‌های باقی مانده که از مدل تبعیت نمی‌کنند به عنوان داده‌های پرت در نظرگرفته می‌شوند. با این حال، در عمل، داده‌ها می‌توانند کاملا بدون داشتن بخش همگن متشکل از حداقل ۵۰٪ از داده‌ها ناهمگن باشند.

در بسیاری موقعیت‌ها مدل خطی برای بیان رابطه‌ی بین متغیر پاسخ و متغیر کمکی متناظر نارسا است.

---

[1] Box
[2] Huber
[3] MLE: Maximum Likelihood Estimator
[4] EM: Expectation-Maximum
[5] Hampel
[6] Rousseeuw
[7] Leroy

آزمون فرض برای مدل‌های جزئی خطی تعمیم‌یافته، عمدتاً مبنی بر مقایسه‌ی برآوردگرهای پارامتری هموار شده، متمرکز است. برای نمونه هاردل[1] و همکارانش(۱۹۹۸) یک آزمون آماری برای انتخاب بین یک مدل خطی و یک مدل نیمه‌پارامتری را مد نظر قرار دادند. طرح پیشنهادی آن‌ها مبنی بر شیوه‌ی تعدیل‌یافته برآورد مطرح شده توسط سورینی[2] و استنیس‌ولس[3] (۱۹۹۴) جهت مقابله با درستنمایی هموارنشده و هموارشده می‌باشد. یک مطالعه‌ی تطبیقی از روش‌های مختلف توسط مولر[4](۲۰۰۱) ارائه گردید در حالی‌که یک رویکرد متفاوت توسط رودریگزکامپاس[5] و همکاران (۱۹۹۸) بررسی شده بود.

معروف است چنین برآوردهایی در مواجه با نقاط پرت ناکارآمد هستند و در پی آن آزمون‌های آماری که از روی آن‌ها ساخته می‌شود منجر به نتیجه‌گیری نادرست می‌گردد. روش‌های استوار برای مدل‌های خطی تعمیم‌یافته توسط استفانسکی[6] و همکاران(۱۹۸۶)، کانش[7] و همکاران(۱۹۸۹)،بیانکو[8] و یوهای[9] (۱۹۹۵)، کانتونی[10] و رانچتی[11] (۲۰۰۱)، کراکس[12] و هاسبروک[13] (۲۰۰۲) و بیانکو و همکاران (۲۰۰۵) مورد توجه قرار گرفته است.

علاوه بر این آزمون‌های استوار برای فرض مقابل، تحت مدل رگرسیون جزئی خطی توسط بیانکو و همکاران (۲۰۰۶) مورد مطالعه قرار گرفت.

در این پایان‌نامه روش‌های آماری استوار برای انواع مدل‌های رگرسیونی معرفی شده و تمرکز بیشتر بر روی مدل‌های جزئی خطی تعمیم‌یافته است. جنبه جدید بودن این پایان‌نامه بررسی روش‌های آماری استوار در مدل‌های جزئی خطی تعمیم‌یافته می‌باشد.

در فصل اول به تعاریف و مفاهیم مربوط به مدل‌های رگرسیون، مدل‌های جزئی خطی تعمیم‌یافته و مدل‌های خطی تعمیم‌یافته پرداخته شده است.در فصل دوم رگرسیون استوار و تشخیص نقاط پرت، در

[1]Hardli
[2]Severini
[3]Staniswalis
[4]Müller
[5]RodrIguez Campos
[6]Stefanski
[7]Künsch
[8]Bianco
[9]Yohai
[10]Cantoni
[11]Ronchetti
[12]Croux
[13]Haesbroeck

فصل سوم آزمون‌سازی درمدل‌های جزئی خطی تعمیم‌یافته با یک رویکرد استوار و در فصل چهارم شبیه‌سازی و مقایسه روش استوار با روش کلاسیک و نتیجه‌ی مطالعات مونت‌کارلو[1] ارائه شده است. و بالاخره در قسمت ضمیمه خطای اندازه‌گیری و اثبات‌ها مورد بررسی قرار گرفته است.

[1]Monte Carlo

فصل اول

مقدمه‌ای بر مدل‌های آماری

۱–۱ مقدمه

در آمار رگرسیون یعنی یک نوع رابطه یا تابع ریاضی که بین متغیر وابسته از یک سو ومتغیرهای مستقل از سوی دیگر برقرار می‌باشد. تحلیل رگرسیونی یک ابزار آماری مفید برای دیدن رابطه‌ی بین متغیرها به کار می‌رود.

تحلیل رگرسیون فن و تکنیکی آماری برای بررسی و به مدل در آوردن ارتباط بین متغیرهاست. کاربردهای رگرسیون متعدد است وتقریباً در هر زمینه‌ای از جمله مهندسی، فیزیک، اقتصاد، مدیریت، علوم زیستی و بیولوژی و علوم اجتماعی صورت می‌پذیرد. در حقیقت تحلیل رگرسیونی ممکن است فن و تکنیک آماری با بیشترین و وسیعترین کاربرد بین تکنیک‌های آماری باشد.

در این فصل مدل‌های رگرسیونی،مدل‌های جزئی خطی تعمیم‌یافته ومدل‌های خطی تعمیم‌یافته مورد بررسی قرارگرفته است. برای مدل‌های رگرسیون روش کمترین مربعات و برای مدل‌های خطی تعمیم‌یافته روش برآورد درستنمایی ماکسیمم آورده شده است که هر دوی این روش‌های برآورد جزء روش‌های کلاسیک می‌باشد. همان‌طور که اشاره شد روش‌های کلاسیک در مقابل داده‌های پرت به نتایج اشتباه منجر می‌شوند. در فصل بعدی روش‌های رگرسیون استوار برای حل این مشکل ارائه شده است.

۱–۲رگرسیون خطی ساده

مدل رگرسیون خطی ساده را برای n مشاهده می‌توان به صورت زیر نوشت:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,, \qquad i = 1,2,3,\dots,n. \qquad (۱–۱)$$

یک تخصیص ساده نشان می‌دهد که فقط یک x برای پیش بینی y وجود دارد، و مدل خطی یعنی مدل (۱–۱) بر حسب β_0 و β_1 خطی است. که در واقع داریم $\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i$ که خطی است. برای مثال،یک مدل مانند $y_i = \beta_0 + e^{\beta_1 x_i} + \varepsilon_i$ نسبت به β_0 و β_1 خطی است، ولی مدل $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$ خطی نیست.

برای تحلیل مدل (۱–۱)، فرض‌های زیر را در نظر می‌گیریم:

(۱) برای تمام $\mathbb{E}(\varepsilon_i) = 0\ ,\ i = 1,2,\dots,n$،یا به عبارت دیگر معادل $\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i$.

(۲) برای تمام $Var(\varepsilon_i) = \sigma^2 \ , \ i = 1,2,\dots,n$ یا به عبارت دیگر $Var(y_i) = \sigma^2$.

(۳) برای تمام $\text{cov}(\varepsilon_i, \varepsilon_j) = ۰ \ , \ i = 1,2,\dots,n$ یا به عبارت دیگر $\text{cov}(y_i, y_j) = ۰$.

فرض (۱) بیان می‌کند که مدل (۱-۱) صحیح است، و نتیجه می‌دهد که y_i فقط به x_i بستگی دارد و تمام تغییرات دیگر y_i تصادفی است.

۱-۲-۱ روش کمترین مربعات برآورد پارامترهای مدل

با استفاده از یک نمونه تصادفی n مشاهده $y_1, y_2, \dots, y_n$ به همراه مقادیر ثابت $\text{x}_1, \text{x}_2, \dots, \text{x}_n$، می‌توانیم پارامتر های $\beta_۰$ و β_1 و σ^2 را برآورد کنیم. برای به دست آوردن $\hat{\beta}_۰$ و $\hat{\beta}_1$ از روش کمترین مربعات استفاده می‌کنیم، که به هیچ گونه مفروضات توزیعی نیاز ندارد.

در روش کمترین مربعات،برآوردگرهای $\hat{\beta}_۰$ و $\hat{\beta}_1$ را به قسمی پیدا می‌کنیم که مجموع مربعات $y_i - \hat{y}_i$، برای n مشاهده y_i را از مقادیر پیش‌بینی آنها $\hat{y}_i = \hat{\beta}_۰ + \hat{\beta}_1 \text{x}_i$ مینیمم سازد.

$$\hat{\varepsilon}'\hat{\varepsilon} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_۰ - \hat{\beta}_1 \text{x}_i) \quad (۲-۱)$$

توجه کنید که مقدار پیش‌بینی شده $\hat{y}_i$ مقدار $\mathbb{E}(y_i)$ را برآورد می‌کند نه y_i را، یعنی $\hat{\beta}_۰ + \hat{\beta}_1 \text{x}_i$ مقدار $\beta_۰ + \beta_1 \text{x}_i$ را برآورد می‌کند نه $\beta_۰ + \beta_1 \text{x}_i + \varepsilon_i$ یک نماد بهتر باید $\mathbb{E}(\hat{y}_i)$ باشد،ولی $\hat{y}_i$ اغلب به‌کار برده می‌شود.

برای یافتن مقادیر $\hat{\beta}_۰$ و $\hat{\beta}_1$ که $\hat{\varepsilon}'\hat{\varepsilon}$ را در (۱-۲) مینیمم سازد، نسبت به $\hat{\beta}_۰$ و $\hat{\beta}_1$ مشتق می‌گیریم و حاصل را برابر صفر قرار می‌دهیم:

$$\frac{\partial \hat{\varepsilon}'\hat{\varepsilon}}{\partial \hat{\beta}_۰} = -2 \sum_{\text{i}=1}^{\text{n}} \left(y_i - \hat{\beta}_۰ - \hat{\beta}_1 \text{x}_i\right)^2 = ۰ \quad (۳-۱)$$

$$\frac{\partial \hat{\varepsilon}'\hat{\varepsilon}}{\partial \hat{\beta}_1} = -2 \sum_{\text{i}=1}^{\text{n}} \left(y_i - \hat{\beta}_۰ - \hat{\beta}_1 \text{x}_i\right)^2 \text{x}_i = ۰ \quad (۴-۱)$$

جواب (۱-۳) و (۱-۴) به صورت زیر داده می‌شود:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \text{x}_i y_i - n\bar{\text{x}}\bar{y}}{\sum_{i=1}^{n} \text{x}_i^2 - n\bar{\text{x}}^2} = \frac{\sum_{i=1}^{n} (\text{x}_i - \bar{\text{x}})(y_i - \bar{y})}{\sum_{i=1}^{n} (\text{x}_i - \bar{\text{x}})^2} \quad (۵-۱)$$

$$\hat{\beta}_۰ = \bar{y} - \hat{\beta}_1 \bar{\text{x}} \quad (۶-۱)$$

برای تحقیق این که $\hat{\beta}_0$ و $\hat{\beta}_1$ در (5.1) و (6.1) مقدار $\hat{\varepsilon}'\hat{\varepsilon}$ را در (2.1) مینیمم می‌سازد، می‌توانیم از مشتق دوم استفاده کنیم یا به طور ساده‌تر ملاحظه کنیم که $\hat{\varepsilon}'\hat{\varepsilon}$ دارای اکسترمم نیست، در نتیجه مشتق اول مینیمم آن را نشان می‌دهد.

۱–۳ الگوی رگرسیون خطی چندگانه

پاسخ y اغلب تحت تاثیر بیش از یک متغییر مستقل است.به عنوان مثال، محصول یک بذر ممکن است به مقدار نیتروژن، پتاسیم و فسفات در کود مورد استفاده بستگی داشته باشد. این متغییرها توسط آزمایشگر کنترل می‌شوند، ولی محصول ممکن است به متغییرهای کنترل نشده‌ای مانند آن‌ها که با آب و هوا در ارتباط‌اند نیز وابسته باشد.

یک الگوی خطی متغییر پاسخ y را به چندین متغیر مستقل به صورت زیر مربوط می‌سازد.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots + \beta_k x_k + \varepsilon_0 \qquad (۷–۱)$$

پارامترهای β_0،β_1،... و β_k را ضرایب رگرسیون می‌نامند. ε_0 تغییرات تصادفی y را که به وسیله xها بیان نمی‌شوند نشان می‌دهد. این متغیر تصادفی ممکن است قسمتی به علت سایر متغیرهایی باشد که بر y اثر می‌گذارند ولی مجهولند یا در نظر گرفته نشده‌اند. الگوی (۱–۷) بر حسب βها خطی است؛ لازم نیست بر حسب xها خطی باشد.پس الگوهایی به صورت زیر:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2 + \varepsilon$$

نیز جزء الگوهای خطی به حساب می‌آیند.

برای برآورد βها نمونه‌ای به حجم n مشاهده روی yها و xهای مربوطه را به کار خواهیم برد. این مدل برای مشاهده iام عبارت است از:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ \ldots + \beta_k x_{ik} + \varepsilon_i\,, \qquad i = 1,2,\ldots,n. \qquad (۸–۱)$$

فرض‌های مربوط به ε_i یا y_i مشابه آن‌هایی است که برای رگرسیون ساده داده شد.

(۱) برای $\mathbb{E}(\varepsilon_i) = 0\ , i = 1,2,\ldots,n$ یا به عبارت معادل

$$\mathbb{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}.$$

(۲) برای $i = 1,2, \dots, n$,$Var(\varepsilon_i) = \sigma^2$ یا به عبارت معادل $Var(y_i) = \sigma^2$

(۳) برای تمام $i \neq j$,$\text{cov}(\varepsilon_i, \varepsilon_j) = ۰$ یا به عبارت معادل $\text{cov}(y_i, y_j) = ۰$

فرض (۱) بیان می‌کند که به طور متوسط مدل صحیح است یعنی تمام xهای مربوطه منظور شده‌اند و مدل خطی است. (۲) بیان می‌کند که واریانس y ثابت است و در نتیجه به x ها بستگی ندارد. فرض (۳) بیان می‌کند که yها با یکدیگر ناهم‌بسته‌اند، معمولاً در یک نمونه تصادفی برقرار است (در یک سری زمانی یا وقتی اندازه‌های مکرر از یک طرح یا حیوان گرفته می‌شود، مشاهدات نوعاً وابسته می‌شوند). بعداً فرض نرمال بودن را اضافه خواهیم کرد که تحت آن yها علاوه بر ناهم‌بسته بودن مستقل نیز خواهند بود.

وقتی تمام سه فرض برقرار است، برآوردگر کمترین مربعات βها دارای چندین خاصیت خوب است. اگر یک فرض یا بیشتر برقرار نباشد،برآوردگرها ممکن است ضعیف باشند. تحت فرض نرمال بودن برآوردگرهای درستنمایی ماکسیمم دارای خواص عالی هستند.

هر یک از سه شرط ممکن است برای داده‌های حقیقی برقرار نباشد. چندین راه حل برای امتحان فرض‌ها پیشنهاد شده است.

۱-۳-۱ برآوردگر کمترین مربعات برای بردار پارامترها

در این بخش از روش کمترین مربعات در برآورد βها در مدل x ثابت بحث می‌کنیم. هیچ فرضی درباره توزیع y برای به دست آوردن برآوردگرها ضروری نیست.برای پارامتر های $\beta_۰$،β_1،... و β_k بر آوردگرهایی را پیدا می‌کنیم که مجموع مربعات انحراف‌های n مقدار مشاهده شده y از مقادیر پیش‌بینی شده $\hat{y}$ را مینیمم سازد. می‌دانیم که $\hat{\beta}_۰$،$\hat{\beta}_1$،... و $\hat{\beta}_k$ عبارت (۱-۹) را می‌نیمم می‌سازد.

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_۰ - \hat{\beta}_1 x_{i1} - \beta_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik})^2 \quad (۱-۹)$$

توجه کنید که مقدار پیش‌بینی شده $\hat{y}_i = \hat{\beta}_۰ + \hat{\beta}_1 \text{x}_{i1} + \beta_2 \text{x}_{i2} + \cdots + \hat{\beta}_k \text{x}_{ik}$ برآورد $\mathbb{E}(y_i)$ است؛ نه y_i؛ یک نماد بهتر $\mathbb{E}(\hat{y}_i)$ است، ولی عموماً از $\hat{y}_i$ استفاده می‌شود.

برای به دست آوردن برآوردگرهای کمترین مربعات،لازم نیست که معادله‌ی پیش‌بینی

$$\hat{y}_i = \hat{\beta}_۰ + \hat{\beta}_1 \text{x}_{i1} + \beta_2 \text{x}_{i2} + \cdots + \hat{\beta}_k \text{x}_{ik}$$