دانشگاه پیام نور

دانشکده علوم پایه

پایان نامه

برای دریافت مدرک کارشناسی ارشد

رشته آماررياضی

گروه آمار

عنوان :

مقایسه کاربرد رگرسیون لجستیک با روش پروبیت در تجزیه و تحلیل متغیرهای دوتایی

جهانچهر جواهری

استاد راهنما :

دکتر پرویز نصیری

استاد مشاور :

دکتر مسعود یارمحمدی

دی ماه ۱۳۸۹


بسم الله الرحمن الرحيم

دانشگاه پیام نور

دانشکده علوم پایه

مرکز تهران

پایان نامه

برای دریافت مدرک کارشناسی ارشد

رشته آمارریاضی

گروه آمار

عنوان :

مقایسه کاربرد رگرسیون لجستیک با روش پروبیت در تجزیه و تحلیل متغیرهای دوتایی

جهانچهر جواهری

استاد راهنما :

دکتر پرویز نصیری

استاد مشاور :

دکتر مسعود یارمحمدی

دی 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 .....

صورتجلسه دفاع از پایان نامه کارشناسی ارشد

جلسه دفاع ازپایان نامه کارشناسی ارشد خانم جهانچهر جواهری
دانشجوی رشته آمار ریاضی به شماره دانشجویی: ۸۷ ۸۶۷۱۰۴۶

تحت عنوان:

"مقایسه کاربرد رگرسیون لجستیک با روش پروبیت در تجزیه و تحلیل داده های دوتایی"

جلسه دفاع با حضور داوران نامبرده ذیل در روزچهار شنبه مورخ: ۸۹/۱۰/۸ ساعت : ۸/۳۰-۹/۳۰ درمحل مجتمع علوم پایه و کشاورزی برگزار شد. و پس از بررسی پایان‌نامه مذکور با نمره به عدد ۱۷/۷۰ به حروف هفده و هفت دهم و با درجه ارزشیابی بسیار خوب مورد قبول واقع شد☑ نشد☐

ردیف	نام و نام خانوادگی	هیات داوران	مرتبه دانشگاهي	دانشگاه/ موسسه	امضاء
۱	دکتر پرویز نصیری	استاد راهنما	استادیار	پیام نور	
۲	دکتر مسعود یارمحمدی	استاد مشاور	استادیار	پیام نور	
۳	دکترصادق رضایی	استاد داور	استادیار	دانشگاه شهید چمران اهواز	
٤	دکتر مسعود یارمحمدی	نماینده علمي گروه	استادیار	پیام نور	

تهران، خیابان استاد نجات‌الهی، خیابان شهید فلاح‌پور، پلاک ۲۷
تلفن: ۸۸۸۰۰۲۵۲
دورنگار: ۸۸۳۱۹۴۷۵
www.tpnu.ac.ir
science.agri@tpnu.ac.ir

اینجانب **جهانچهر جواهری** دانشجوی ورودی سال ۱۳۸۶ مقطع کارشناسی ارشد رشته آمار ریاضی گواهی می نمایم چنانچه در پایان نامه خود از فکر، ایده و اندیشه و نوشته دیگری بهره گرفته‌ام با نقل قول مستقیم یا غیر مستقیم منبع و ماخذ آن را نیز در جای مناسب ذکر کرده‌ام. بدیهی است مسئولیت تمامی مطالبی که نقل قول دیگران نباشد بر عهده خویش می‌دانم و جوابگوی آن خواهم بود.

دانشجو تأیید می‌نماید که مطالب مندرج در این پایان نامه (رساله) نتیجه تحقیقات خودش می باشد و در صورت استفاده از نتایج دیگران مرجع آن را ذکر نموده‌است.

نام و نام خانوادگی دانشجو : **جهانچهر جواهری**

تاریخ و امضاء

اینجانب **جهانچهر جواهری** دانشجوی ورودی سال ۱۳۸۶ مقطع کارشناسی ارشد رشته آمار ریاضی گواهی می‌نمایم چنانچه بر اساس مطالب پایان نامه خود اقدام به انتشار مقاله، کتاب و .....نمایم ضمن مطلع نمودن استاد راهنما، با ایشان نسبت به نشر مقاله، کتاب و .....و به صورت مشترک و با نام ذکر نام استاد راهنما مبادرت نمایم.

نام و نام خانوادگی دانشجو : **جهانچهر جواهری**

تاریخ و امضاء

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

دی ماه ۱۳۸۹

چکیده

آنالیز رگرسیون ابزاری بسیار مفید برای بررسی رابطه و همبستگی بین دو و یا چندمتغیر است و به علت قابلیت های خوب کاربردی، امروزه وسیله مفیدی برای رشته های مختلف علوم است. در آنالیز رگرسیون در حالت خاص با متغیر های وابسته ای که فقط ۲ مقدار دارند سروکار داریم. در این حالت می‌توان از رگرسیون لجستیک و رگرسیون پروبیت برای آنالیز روابط بین متغیر وابسته و چند متغیر مستقل استفاده کرد. هدف این آنالیز تنظیم عوامل موثر، بخش بندی کواریت های مهم مربوط به متغیر وابسته و پیشگوئی مقدار متغیر وابسته است. این دو رگرسیون از نظر کیفی نتایج مشابهی دارند، ولی برآورد پارامترهای دو مدل را نباید به طور مستقیم مورد مقایسه قرار داد. لجیت و پروبیت هر دو شکلی از تبدیل داد های دوحالتی (دوتایی) هستند. هر دوی این آنالیزها از مدل‌های احتمال خطی بمنظور برآورد پارامترهای مدل استفاده می‌کنند و حالت خاصی از مدل‌های عمومی خطی هستند. برای محاسبه پارامترهای مدل، تخمین حداکثر درستنمایی از پارامترهای رگرسیون را محاسبه می‌کنند و برای بیشینه کردن و رسیدن به حداکثر درستنمایی روش Fisher's scoring مورد استفاده قرار می گیرد. در نهایت در هر دو روش در صورت رسیدن به حداکثر درستنمایی پارامترهای مدل برآورد می شود. هر دو روش برای رسیدن به حداکثر درستنمایی از رویه مکرر استفاده می‌کنند.

در حالتی که توزیع تجمعی پاسخ‌ها در برابر داده‌ها از توزیع نرمال تبعیت نکند پیشنهاد می‌شود بجای تبدیل پروبیت از تبدیل لجیت استفاده کنیم و در صورتی که توزیع تجمعی داده‌ها از توزیع نرمال تبعیت کنند، آنالیز پروبیت ترجیح داده می‌شود. در حضور تعداد زیادی سطوح متغیر مستقل تبدیل لجیت برازش بهتری را فراهم می‌کند و بر عکس پروبیت مدل‌های با اثرات تصادفی و مجموعه داده های در حد متوسط را بهتر برازش می‌کند. در صورتی که طرح مدونی بکار برده‌نشده باشد خطای برآورد احتمالات در آنالیز پروبیت نسبت به لجستیک بیشتر است و لذا در این حالت آنالیز لجستیک توصیه می‌شود.

از مشکلات مهم همراه با این نوع رگرسیون‌ها می‌توان، نرمال نبودن متغیرها، عدم پوشش کامل نتایج و احتمال واقع شدن $\hat{Y}$ خارج از دامنه (۱و۰)، ناهمگنی واریانس توزیع ها و مشکوک بودن مقدار R^2 به عنوان معیار نکویی برازش را نام برد. مدلهای احتمال خطی از نظر منطقی خیلی جالب نیستند زیرا این مدلها فرض می کنند احتمال همزمان با افزایش X بطور خطی افزایش می یابد و اثر افزایشی X بطور ثابت باقی می ماند. این موضوع گاهی اوقات غیر واقعی است.

فهرست مطالب

عنوان مطلب.....صفحه

فصل اول: کلیات و مفاهیم

فصل دوم: رگرسیون لجستیک ورگرسیون پروبیت

فصل سوم: نتایج رگرسیون لجستیگ و رگرسیون پروبیت

فصل چهارم: مقایسه رگرسیون لجستیک و پروبیت

منابع

فصل ۱

کلیات و مفاهیم

۱-۱ ارتباط آماری متغیرها

در اغلب آزمایش های آماری در هر یک از واحدهای آزمایشی، متغیرهای مختلفی اندازه گیری می شوند. ولی در اغلب تکنیک های تجزیه و تحلیلی فقط یک متغیر به صورت جداگانه مورد بررسی قرار می گیرد. ارتباط آماری به سنجش ارتباط بین دو یا چند متغیر کمی مربوط می شود. روش‌های مختلفی نیز جود دارد که می‌توان ارتباط بالقوه بین متغیرها را بررسی کرد. بوسیله روش های یک و چند متغیره می‌توان وابستگی بین متغیرها را بررسی کرد ولی روش ها چند متغیره خیلی پیچیده هستند (اوریت و دان ۱۹۹۱- موریسون ۱۹۹۰). یکی از روشهای ساده بمنظور بررسی ارتباط آماری بین متغیرها استفاده از ضرایب همبستگی است. ضریب همبستگی شاخصی است ریاضی که جهت و مقدار رابطه ی بین دو متغیر را توصیف می‌کند(پلاتا ۲۰۰۶). ضریب همبستگی درمورد توزیع های دویا چند متغیره به کار می رود. اندازه های مختلفی از همبستگی وجود دارد ولی همیشه مقدار آن بین ۱- تا ۱+ قرار دارد. اگر مقادیر دو متغیر شبیه هم تغییر کند یعنی با کم یا زیاد شدن یکی، دیگری هم کم یا زیا د شود، به گونه ای که بتوان رابطه آنها را به صورت یک معادله بیان کرد، گوییم بین این دو متغیرهمبستگی وجود دارد. برای سنجش همبستگی بین متغیرها، ضرایب گوناگونی به کار می رود که مهمترین آنها ضریب همبستگی ساده پیرسون ، ضریب همبستگی اسپیرمن و ضریب همبستگی کندال است (هاک و کساوسکی ۲۰۰۷).

روش دیگری که برای بررسی ارتباط آماری بین متغیرها مورد استفاده قرار می گیرد معادلات رگرسیون است. در اینجا انواع رگرسیون بمنظور سنجش ارتباط آماری بین متغیرها مورد بررسی قرار خواهد گرفت (چان ۲۰۰۳ و لیوایز ۱۹۹۳).

۱-۲ تحلیل رگرسیون

تحلیل رگرسیون ، مجموعه ای از تکنیک های آماری است که برای مدل سازی و بررسی رابطه بین یک متغیر پاسخ Y و مجموعه ای از متغیرهای پیشگو (مستقل) $X_1, X_2, \ldots\ldots X_k$ می باشد. کاربردهای رگرسیون گسترده بوده و تقریباً در هر زمینه کاربردی چون پزشکی، مدیریت، اقتصادو ..... استفاده می شود.مدلهای رگرسیون خطی بطور گسترده بعنوان مدلهای تجربی برای تقریب یک رابطه پیچیده و معمولا نامعلوم بین متغیرهای پاسخ و پیشگو بکار می رود.

نلدر و دربرن در سال ۱۹۷۲ مدلهای خطی عمومی را به مدلهای خطی تعمیم یافته توسعه دادند.در حقیقت مدلهای خطی تعمیم یافته ، مدلهای خطی عمومی هستد که متعلق به خانواده نمائی می باشند و خطاهایی با توزیع نرمال دارند.GLM ها به فرض مشخص بودن توزیع کامل نیاز دارد ، اما در برخی موقعیت ها مخصوصا در بررسی داده های گسسته ، گاهی اوقات فرض توزیع کامل محدود می شود. برای رفع نیاز به وجود توزیع کامل ودربرن در سال ۱۹۷۴ تابعی به نام تابع شبه درستنمائی معرفی نمود که فقط بر اساس فرض وجود گشتاور مرتبه دوم متناهی متغیرپاسخ برقرار می باشد.(استفورد،۱۹۹۶) .از روش تابع شبه درستنمائی فقط در برآورد کردن میانگین یا ضرایب رگرسیونی استفاده می شود.

۱-۲-۱ مدلهای خطی

مدلهای خطی بصورت زیر بیان می شود:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon \qquad (۱-۱)$$

که در آن Y متغیر پاسخ ، X ها متغیر پیشگو ، $\beta_0, \beta_1, ..., \beta_k$ مجموعه پارامترهای نا معلوم و ε جمله خطای تصادفی است. مدل بالا را اغلب مدل رگرسیون خطی می نامند. جمله خطا دارای میانگین صفر است. پس میانگین پاسخ در مدل رگرسیون خطی عبارت است از:

$$E(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \qquad (۱-۲)$$

در این مدل میانگین پاسخ یک تابع خطی از پارامترهای مجهول β_0 و β_1 و....و β_k است.
مدلهای رگرسیون خطی به دلایل گوناگونی کاربرد وسیعی دارند. دلیل اول اینکه این مدلها تقریب طبیعی چندجمله ای برای روابط تابعی پیچیده ترند. یعنی اگر $E(Y)=f(X)$ رابطه دقیق یا قطعی بین متغیر پاسخ و متغیر پیشگو باشد، آنگاه تقریب اول سری تیلور این رابطه در نقطه ای مانند x_0 با رابطه زیر بدست می آید.

$$\begin{aligned} E(Y) &\cong f(x_0) + \left.\frac{df(x)}{dx}\right|_{x=x_0} (x - x_0) + R \\ &\cong \beta_0 + \beta_1 (x - x_0) \end{aligned} \qquad (۱-۳)$$

که صرف نظر از باقیمانده R (بجز جمله خطا) یک الگوی رگرسیون خطی یک متغیری است. وقتی K متغیر پیشگو داشته باشیم ، تقریب مرتبه اول تیلور مستقیما به یک مدل رگرسیونی خطی K متغیری منجر می شود. چون اغلب از الگوهای رگرسیون خطی (به طور موفقیت آمیزی) بعنوان تقریب چند جمله ای ها استفاده می شود، لذا بعضی اوقات این مدلها را مدل تجربی می نامند. دومین دلیل استفاده از مدلهای رگرسیون خطی اینست که از طریق آنها پارامترهای مجهول مدل یعنی β_0 و β_1 و....و β_k مستقیماً برآورد شوند.

در ادامه به دو روش مهم برآورد پارامترهای مدل رگرسیون خطی که روشهای کمترین مربعات معمولی و درستنمائی ماکزیمم هستند ، اشاره می نمائیم.

۱-۲-۲ برآورد پارامترها

۱-۲-۲-۱ برآورد پارامترها به روش کمترین مربعات[1]

این روش نوعاً برای برآورد ضرایب رگرسیون دریک مدل رگرسیون خطی چندگانه بکار می رود. فرض می کنیم $n \succ k$ مشاهده از متغیر پاسخ Y بصورت $y_1, y_2, \ldots\ldots y_n$ داریم. برای هر پاسخ مشاهده شده y_i یک مشاهده برای هر متغیر پیشگو داریم و فرض می کنیم x_{it}مشاهده iام متغیر x_t باشد. جمله خطای ε در مدل را دارای میانگین صفر و واریانس ثابت σ^2 و $\{\varepsilon_i\}$ را متغیرهای تصادفی ناهمبسته فرض می کنیم. معادله ۳ را بر حسب مشاهدات بصورت معادله زیر می نویسیم:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ldots + \beta_k x_{ik} + \varepsilon_i \\ &= \beta_0 + \sum_{t=1}^{k} \beta_t x_{it} + \varepsilon_i \end{aligned} \qquad (۴-۱)$$

روش کمترین مربعات، ها را در معادله بالا طوری انتخاب می کند که مجموع مربعات خطاهای ε_iمینیمم شود. برآورد کمترین مربعات βها بصورت زیر است:

$$b = (X'X)^{-1} X'Y$$

1-ordinary least square

۱-۲-۲-۲ برآورد پارامترها به روش درستنمائی ماکزیمم[1]

از روش کمترین مربعات برای برآورد پارامترهای یک مدل رگرسیون خطی بدون در نظرگرفتن شکل توزیع متغیر پاسخ Y می توان استفاده کرد. اگر شکل توزیع اسخ معلوم باشد، آنگاه از روش برآورد دیگری به نام درستنمائی ماکزیمم می توانیم استفاده کنیم.

مدل رگرسیون خطی $Y = X\beta + \varepsilon$ را در نظر می گیریم. فرض می کنیم خطاها در این مدل دارای توزیع نرمال مستقل با میانگین صفر و واریانس ثابت σ^2 باشند. در این صورت مشاهده در نمونه $(y_i, x_{i1}, \ldots\ldots, x_{ik})$ دارای توزیع نرمال مستقل با میانگین $\beta_0 + \sum_{t=1}^{k} \beta_t x_{it}$ و واریانس σ^2 خواهد بود. تابع درستنمائی را از توزیع احتمال توام مشاهدات پیدا می کنیم . اگر این توزیع توأم را با مشاهدات داده شده و پارامترهای مجهول در نظر بگیریم، آنگاه تابع درستنمائی را به صورت زیر داریم:

$$L(y,\beta,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{\frac{-1}{2\sigma^2}\left[(y-x\beta)'(y-x\beta)\right]\right\} \qquad (۵-۱)$$

برآوردگرهای درستنمائی ماکزیمم مقادیری از پارامترهای β و σ^2 هستند که تابع درستنمائی را ماکزیمم می کنند. ماکزیمم کردن تابع درستنمائی L معادل ماکزیمم کردن تابع لگاریتم درستنمائی یعنی lnL می باشد.

تابع لگاریتم درستنمائی بصورت زیر است :

$$\ln\left[L(y,\beta,\sigma^2)\right] = \frac{-n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\left[(y-x\beta)'(y-x\beta)\right] \qquad (۶-۱)$$

مشتق تابع لگاریتم درستنمائی را تابع امتیاز می نامند. اگر از تابع لگاریتم درستنمائی نسبت به پارامترهای β و σ^2 مشتقات جزئی گرفته و مساوی صفر قرار دهیم خواهیم داشت:

$$\frac{1}{\sigma^2} X'(Y - X\beta) = 0 \qquad (۷-۱)$$

[1]-Maximom Liklihood

دستگاه معادلات $k \times k$ بالا را معادلات امتیاز درستنمائی ماکزیمم[1] مـی نامنـد. بـرآورد درسـتنمائی ماکزیمم، پاسخ دستگاه معادلات امتیاز است.

$$\beta = (X'X)^{-1} X'Y$$

اگر نسبت به σ^2 مشتق بگیریم و مساوی صفر قرار دهیم، برآوردگر σ^2 بصورت زیر بدسـت مـی آید:

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\beta)'(y - X\beta)$$

بطور کلی برآوردگرهای درستنمائی نسبت به برآوردگرهای کمترین مربعـات دارای خـواص آمـاری بهتری هستند که بخاطر فرض های اضافی می باشد کـه بـرای آنهـا در نظـر گرفتـه مـی شـود. مـثلا برآوردگرهای درسـتنمائی مـاکزیمم نیـاز بـه نرمـال بـودن توزیـع مـشاهدات دارنـد ولـی در روش برآوردکمترین مربعات

برقراری این فرض لازم نیست. برآوردگرهای درسـتنمائی مـاکزیمم نااریـب بطـور مجـانبی نااریـب هستند و با افزایش حجم نمونه اریب می شوند و نیزمجموعه ای از آماره های بسنده را تـشکیل مـی دهند و سازگارند.

۱-۳ رگرسیون خطی ساده

۱-۳-۱ تاریخچه رگرسیون خطی

رگرسیون شاخه ای از علم آمار است که به بررسی رابطه بـین متغیرهـا مـی پـردازد (کـاتنر ۲۰۰۵) موضوع رگرسیون ابتدا در قرن هیجدهم با استفاده از نجوم با هدف هدایت کشتی ها مورد توجه قرار گرفت . لژاندر[2] در سال ۱۸۰۵ روش کمترین توان های دوم را بسط و توسعه داد. گـوس[3] ادعـا مـی کند که او این روش را چند سال قبل توسعه داده و در سال ۱۸۰۹ نشان داد که وقتـی خطاهـا دارای

1-Score Equations Maximum Likelihood
1-catner
2- Legendre
3-Guss-۱

توزیع نرمال اند، روش کمترین توانهای دوم به جواب بهینه منجر می شود. تا قرن نوزدهم این روش شناسی تقریبا در علوم فیزیک مورد استفاده قرار می گرفت. فرانسیس گالتن [۱] در سال ۱۸۷۵ اصطلاح برگشت به مقدار متوسط را در مورد یک معادلـه رگرسـیونی سـاده در نظـر گرفـت . گـالتن از ایـن اصطلاح برای بیان این پدیده که پسران پدران قد بلند،بلند قد بوده ولی نه به بلندی قد پدرشان ، در حالی که پسران پدران کوتاه قد میل به کوتاه قد بودن دارند ولی این تمایل به اندازه کوتاه قـد بـودن پدرانشان نیست ، استفاده کرد . برای اطلاع بیشتر در مورد تاریخچه استفاده از رگرسیون می توانید به (استیگر[۲] ۱۹۸۶) مراجعه کنید. کاربرد های رگرسیون متعدد است و همانند سـایر شـاخه هـای علـم آمار ، تقریبا در هر زمینه ای از جمله مهندسی ، فیزیک ، اقتصاد، مـدیریت ، علـوم زیـستی و علـوم اجتماعی صورت می پذیرد. تحلیل رگرسیونی فن و تکنیکی آماری برای بررسی و به مدل در آوردن ارتباط بین متغیرهاست.در حقیقت این تحلیل ممکن است فن آماری با بیشترین و وسیعترین کـاربرد بین فنون آماری باشد . (مونت گمری و پک[۱]، ۱۹۹۲). هدف در این تحلیل یافتن یـک مـدل مناسـب برای نمایش ارتباط بین متغیرهای وابسته و مستقل می باشد.

ضریب کورولاسیون فقط یک معیار از ارتباط خطی را می دهد و هیچ نشانی از این کـه کـدام خط راست را ارائه نمی دهد (اگر یک خط راست مناسب باشد) و نشان دهنده بهترین ارتباط اسـت. آنالیز رگرسیون ساده خطی به معنی تعیین این ارتباط است. ساده در اینجا اشاره به این حقیقـت دارد که یک متغیر مستقل در آنالیز قرار داده شده است و ربطی به پیچیده یا ساده بودن آنالیز ندارد. بوسیله نمودار داده ها نیز می‌توان نشان داد که مدل خط راست مناسب است یا نه.

دو متغیر موجود در رگرسیون ساده یکسان نیستند. تغیر در یکی (متغیر مستقل) باعـث تغییـر در دیگری (متغیر وابسته) می شـود و رگرسـیون y برحـسب x هماننـد رگرسـیون x برحـسب y نیست و باید متوجه این اختلاف بود. به عنوان مثال چون تغییر در سـطوح بیمـاری مـی توانـد روی

4-Galton-۲
5-Stigler- [2]
6-gomeri

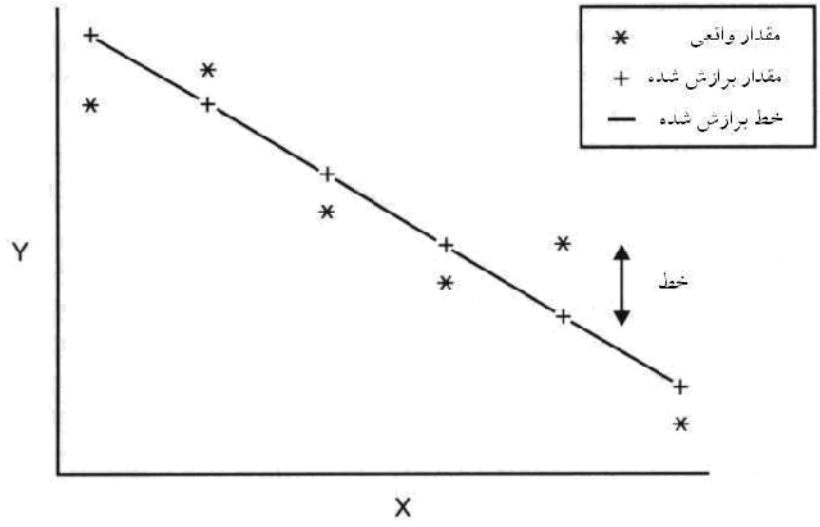
عملکرد تأثیر بگذارد، امکان پیدا کردن رگرسیون عملکرد محصول (به عنوان متغیر وابسته y) بر حسب سطوح بیماری در آن محصول (به عنوان متغیر مستقل x) وجود دارد. ولی این احتمال وجود ندارد که عملکرد روی سطح بیماری تأثیر بگذارد. لذا رگرسیون بیماری بر حسب عملکرد غیر ضروری خواهد بود. اگر x و y همبستگی کامل داشته باشند، در این صورت مدل رگرسیون y بر حسب x و x بر حسب y یکسان خواهد بود ولی تفسیر آنها هنوز متفاوت خواهد بود.

اگر متغیر مستقل به وسیله x و متغیر وابسته را به وسیله y نشان دهیم، آنالیز رگرسیون ساده یک مدل رگرسیون خطی ساده را به شکل زیر برازش خواهد نمود.

$$y = a + bx$$

که فقط یک x برای پیش بینی yوجود دارد.و خطی یعنی اینکه مدل برحسب β_0و β_1خطی است. این مدل کاملاً مناسب داده ها نخواهدبود ولی همان طوری که در شکل نشان داده شده است، مشاهده ها به صورت تصادفی از خط راست انحراف دارند. مقادیر y محاسبه شده به وسیله مدل در واقع مقادیر برازش شده هستند و تشکیل یک خط راست را می دهند. اختلاف بین مقادیر برازش شده و مقادیر y مشاهده شده خطاها هستند. آنالیز رگرسیون مقادیر پارامترهای b,a را برای مدل تخمین می زند و بنابراین مدل خط راست را برازش می کند. روش های مختلفی برای تخمین پارامترهای b,a وجود دارد. متداول ترین روش «روش حداقل مربعات» است که در آن، مقادیر b,a را به گونه ای پیدا می کنند که حداقل مجموع مربعات خطا را داشته باشیم. مفروضات مورد نیاز برای آنالیز رگرسون خطی همان

مفروضات آنالیز واریانس، یعنی ثابت بودن واریانس، توزیع تصادفی و استقلال خطاها است. همانند آنالیز واریانس، یک روش خوب برای ارزیابی اعتبار مدل برازش شده، ترسیم انحراف های مدل در برابر مقادیر برازش شده است. مجدداً توزیع تصادفی، به صورت یک نوار افقی در اطراف خطای صفر نشان دهنده مناسب بودن مدل مورد استفاده است. نباید الگو یا تمایل خاصی در نمودار وجود داشته باشد.



یک مثال از مفهوم مقادیر برازش شده، مقادیر واقعی و مدل برازش شده (در این حالت یک مدل خط راست). برگرفته از دراپر و اسمیت ۱۹۸۱

اگ

ر نمودار داده های خام اولیه یک عکس العمل منحنی خط را به متغیر مستقل نشان دهد، ممکن است هنوز رگرسیون خطی مناسب باشد. به وسیله یک تابع ریاضی مناسب یا تبدیل کردن، می توان با استفاده از رگرسیون خطی یک مدل برای داده ها برازش نمود (کولین ۱۹۹۹). اصطلاح رگرسیون خطی را می توان برای همه مدل هایی که از نظر ریاضی پارامترهای خطی دارند به کار برد ولی منعکس کننده شکل تابعیت نیست. هر دو متغیر وابسته و مستقل را می توان به منظور برازش مدل تبدیل کرد. به عنوان مثال در مدل های زیر می توان متغیر مستقل x را تبدیل نمود :

$$y = a + bx + cx^2$$

$$y = a + bx + cx^2 dx^3$$

$$y = a + b \ln x$$

معادلات خط راست، درجه دو و درجه سه که در بالا ذکر شد، اولین سه معادله چند جمله ای مناسبی هستند که وابسته به مقایسه های چند جمله ای می باشند.

اگر یک مجموعه داده را برای آنالیز رگرسیون مورد استفاده قرار می گیرد، باید یک دامنه مناسب و کافی از مقادیر x موجود باشد. برای دامنه مقادیر x موجود فقط می توان یک خط یا منحنی خط برازش نمود و هر نتیجه بدست آمده درباره این داده ها، فقط برای محدوده همین دامنه است. مدلی که برای مقادیر از یک دامنه داده شده برازش می شود را نباید برای پیش بینی یا تخمین مقادیری که خارج از این دامنه واقع شده اند به کار برد چون به دلیل موجود نبودن هیچ اطلاعاتی در زمینه این مقادیر نمی دانیم که این مدل برای خارج از این دامنه نیز مناسب است یا خیر. خط برازش شده با اطمینان بالاتری در نواحی که داده های جفتی بیشتری وجود دارد، برآورد می گردد.

بنابراین اگر این امکان وجود دارد که مقادیر y را برای مقادیر x انتخاب شده اندازه گیری کنیم، انتخاب مقادیر x، بسته به هدف تحقیق تغییر خواهد کرد. اگر همه نقاط از اهمیت یکسانی برخوردار هستند که این امر اغلب در مورد خط راست صدق می کند، یک توزیع یکنواخت مقادیر x در کل دامنه x مطلوب خواهدبود.اگر هدف آزمایش برازش یک مدل ریاضی برای یک مجموعه داده است،و دانش اولیه در باره شکل مدل وجود ندارد، یک توزیع یکنواخت از مقادیر مورد نیاز است. اگر مقادیر خاصی به عنوان مثال یک مقدار مطلوب مورد نظر است و تا حدودی می دانیم که ممکن است این مقدار کجای محور x ها واقع شود. اگر بیشتر مقادیر x در اطراف محل مورد انتظار باشند و تعداد کمی از مقادیر در دو منتهی الیه واقع شوند، محل واقعی این مقدار با قطعیت بیشتری یافت خواهد شد.

برای حصول اطمینان از برازش مدل صحیح ، تعداد کافی از داده ها در دامنه مقادیر x مورد نیاز است. تعداد مقادیر مختلف موجود برای x پیچیدگی مدلی را که می توان برای داده ها برازش نمود تعیین می کند (دراپر و اسمیت۱۹۸۱). بین دو نقطه داده همیشه فقط یک خط راست عبور می کند ولی می توان تعداد نامحدودی منحنی های پیچیده تر ، بین همان دو نقطه کشید. برای برازش