دانشکده علوم پایه

گروه: آمـــار

عنوان پایان نامه:

رگرسیـــون پواســـــن استـــــوار

پایان نامه برای دریافت عنوان کارشناسی ارشد

رشته: آمـــار ریاضی

نگارش:

اعظم کامیـاب تیموری

استاد راهنما:

دکتر پرویز نصیری

استاد مشاور:

دکتر صادق رضایی

خرداد 1387

چکیده

اغلب داده‌های شمارشی تحت مدل پواسن تحلیل می‌شوند(اگرستی،۱۹۹۶) و چه بسا اگر توزیع مفروض پواسن نباشد، استنباط استخراج شده عموما نادرست است(کاکس،۱۹۸۳). ما در این پایان نامه رهیافت رگرسیون پواسن استوار را برای تحلیل داده‌های شمارشی در نظر گرفته‌ایم. مدل رگرسیون پواسن بطور مجانبی به مقدار استنباط پارامترهای رگرسیونی مطابق می‌شود؛ حتی اگر فرض پواسن برقرار نباشد. روش استوار سازی مطرح شده در این پایان نامه یک تابع درستنمایی معقول در نمونه‌های بزرگ برای پارامترهای رگرسیونی با این فرض که توزیع واقعی، گشتاورهای مرتبه دوم متناهی داشته باشد، ارائه می‌دهد. تعدیل‌هایی که رگرسیون پواسن را استوار[1] می‌سازند، به ترتیب تحت توابع پیوند لگاریتم[2] و پیوند همانی[3] داده خواهند شد. همچنین مطالعات شبیه سازی شده به منظور بررسی کارایی مدل رگرسیون پواسن استوار برای داده‌هایی که بطور واقعی از توزیع‌های غیر پواسن تولید شده‌اند، ارائه می‌گردد.

واژه‌های کلیدی:درستنمایی نیمرخی استوار، رگرسیون پواسن، شبه درستنمایی، رگرسیون پواسن استوار، آزمون نسبت درستنمایی

[1] Robust
[2] Log link
[3] Identity link

فهرست مندرجات

چکیده فارسی..... الف

فصل اول : مقدمه ۱

۱-۱- شرح مسئله ۲

۱-۲- تاریخچه ۳

۱-۳-ساختار ٤

فصل دوم :خانواده مدل‌های خطی تعمیم یافته..... ٦

۲-۱- مقدمه..... ۷

۲-۲- مدل‌های خطی ۸

۲-۳- خانواده نمایی توزیع‌ها ۱۲

۲-٤- خانواده مدل‌های خطی تعمیم یافته ۱۵

۲-٤-۲-مولفه‌های مدل‌های خطی تعمیم یافته ۱٦

۲-۵- ویژگی‌های مدل‌های خطی تعمیم یافته..... ۱۸

۲-٦- ساختار رسمی خانواده مدل‌های خطی تعمیم یافته ۲۰

۲-۷-معادلات درستنمایی برای خانواده مدل‌های خطی تعمیم یافته..... ۲۱

۲-۷-۱- حالت کلی ۲۱

۲-۷-۲- حالت خاص ۲۳

۲-۸- برآورد Ψ..... ۲٦

۲-۹- تابع شبه درستنمایی ۲۷

۲-۹-۱-قضایا ۲۹

۲-۹-۲-تابع شبه درستنمایی در خانواده نمایی ۳۰

فصل سوم : رگرسیون پواسن خطی تعمیم یافته ۳۳

۳-۱- مقدمه.....۳٤

۳-۲- مدل سازی داده های شمارشی برای متغیر پاسخ.....۳۵
۳-۳- بررسی مدل سازی با مثال..... ٤٤
۳-۳-۱- مثال خرچنگ نعلی و انگل‌های آنها..... ٤٤
۳-۳-۲- مثال مربوط به مدل های نرخ ٤۹
۳-٤- بررسی نیکویی برازش مدل پواسن خطی تعمیم یافته ۵۱
۳-۵- بیش پراکنش در رگرسیون پواسن خطی تعمیم یافته ۵۳
۳-٦- معادلات درستنمایی ماکزیمم در رگرسیون پواسن خطی تعمیم یافته.....۵٤
فصل چهارم : رگرسیون پواسن استوار..... ۵٦
٤-۱- مقدمه ۵۷
٤-۲- تعاریف و مفاهیم ۵۸
٤-۳- رد مدل ٦۱
٤-٤- تابع درستنمایی نیمرخی ٦۷
٤-۵- داده های شمارشی مستقل و همتوزیع ۷۲
٤-٦- رگرسیون پواسن استوار – پیوند همانی ۷۳
٤-۷- رگرسیون پواسن استوار – پیوند لگاریتم ۷٦
٤-۸- مطالعات شبیه سازی ۷۸
نتیجه گیری ۸۵
پیوست الف ۸٦
پیوست ب: برنامه‌های شبیه سازی Matlab ۹۵
پیوست ج: جدول داده‌های مثال خرچنگ نعلی و انگل‌های آنها ۱۱٤
فهرست مراجع..... ۱۱۷
چکیده انگلیسی ۱۱۹

فصـــــل اول

کلیـــات و مقدمه

۱-۱- شرح مسئله و هدف

داده‌های شمارشی در بسیاری از کاربردهای آماری مورد استفاده قرار می‌گیرند، ترخیص روزانه بیماران از بیمارستان، تعداد دزدی‌های ماشین، تعداد زلزله‌های ثبت شده، . . . مثال‌هایی از این قبیل داده‌های شمارشی می‌باشند. توزیع پواسن اکثر اوقات بعنوان مدل نمونه‌گیری برای داده‌های شمارشی فرض می‌شود، علیرغم این که در برخی موارد توزیع واقعی داده‌های شمارشی، پواسن نمی‌باشد(فرام، ۱۹۸۳). در این حالت باید برآورد پارامتر بدست آمده را بطور مجانبی به مقدار واقعی پارامتر منطبق نمود. به همین علت، مفهوم استوارسازی[1] این برآورد مطرح می‌گردد که برای این منظور از رگرسیون پواسن استوار[2] برای برآورد پارامترهای رگرسیونی استفاده می‌کنیم.

بطور کلی رگرسیون پواسن یک مدل خطی تعمیم یافته[3] است که درآن میانگین شمارش‌ها با متغیرهای تبیینی در ارتباط می‌باشد. فرض کنید Y متغیر تصادفی پاسخ با توزیع پواسن باشد بطوری‌که $E(Y)=m$ و X متغیر تبیینی مدل پواسن خطی باشد؛ بنابراین داریم $m=b_0+b\,x$ که در آن bها ضرایب رگرسیونی‌اند. برای مدل لگ خطی پواسن فرم $\log m=b_0+b\,x$ در نظر گرفته می‌شود و بسیاری از نرم افزارهای مرتبط با GLM مانند SAS، مدل‌های رگرسیون پواسن را برای تحلیل داده‌های شمارشی با انتخاب توابع پیوند که پیوندهای لگ یا همانی هستند؛ اجرا می‌کنند و چه بسا سهولت محاسبات مدل لگ خطی پواسن، اعتبار نتایج آن را تضمین نمی‌کند. هدف ما در این پایان نامه استنباط استوار برای داده‌های شمارشی است. مدل رگرسیون پواسن استوار برای استنباط در مورد پارامترهای رگرسیونی در اغلب داده‌های شمارشی مورد استفاده قرار گرفته است و بنابراین در مورد صحیح بودن فرض پواسن نگرانی وجود ندارد. مدل رگرسیون پواسن استوار بطور مجانبی یک تابع درستنمایی معقول برای

[1] Robustness
[2] Robust Poisson Regression(RPR)
[3] General linear model(GLM)

پارامترهای رگرسیونی در نظر می‌گیرد. بنابراین ابزارهای آماری که به مشخص‌سازی و تعیین کامل تابع درستنمایی نیازمندند، بر پایه تابع درستنمایی استوار ارائه شده استخراج می‌شوند.

۱-۲- تاریخچه

انگل[۱] در سال ۱۹۸٤ و لاولس[۲] در سال ۱۹۸۷ انواع ویژه‌ای از مدل‌های رگرسیون دوجمله‌ای برای شمارش‌هایی که بیش پراکنش دارند را مطرح ساختند و نشان دادند که تصحیح روابط میانگین–واریانس برای اعتبار این دیدگاه پارامتری ضروری است. روش‌های دیگری که به مدل وابستگی کمتری دارند مانند روش شبه درستنمایی[۳] توسط ودربرن[٤] در سال ۱۹۷٤ و مک کولاق و نلدر[٥] در سال ۱۹۸۹ معرفی شده‌اند.

عمدتا" بعلت نداشتن یک تابع درستنمایی مناسب، دسترسی به برآورد پارامترهای مورد علاقه درستنمایی ماکزیمم امکان پذیر نمی‌باشد؛ در نتیجه در این پایان نامه استنباط درستنمایی کامل مورد توجه نیست و از درستنمایی نیمرخی[٦] و شبه درستنمایی استفاده می‌کنیم. گولد و لاولس در سال ۱۹۸۸ نشان دادند که برآوردهای ضرایب رگرسیونی در مدل‌های خطی با پارامتر مکانی، حتی اگر توزیع خطا به درستی تعیین نشده باشد، سازگار باقی می‌مانند. برآوردهای واریانس معتبر بوسیله برآوردگر واریانس ارائه شده توسط رویال[۷] در سال ۱۹۸٦ بدست آید.

[1] Engel , 1984
[2] Lawless , 1987.
[3] Quasi likelihood
[4] Wedderburn , 1974
[5] Mc Cullagh & Nelder , 1989
[6] Profile likelihood
[7] Royall , 1986.

۱-۳- ساختار

چون توزیع پواسن عضو خانواده مدل‌های خطی تعمیم یافته می‌باشد؛ لذا آشنایی با این خانواده بسیار مرتبط و پیش نیاز موضوع این پایان نامه است. در فصل دوم به بررسی خانواده مدل‌های خطی تعمیم یافته که به اختصار از آن به عنوان GLM یاد می‌کنیم می‌پردازیم. نلدر و ودربرن[1] در سال ۱۹۷۲ مطرح کردند که GLMها، مدل‌های خطی عمومی هستند که متعلق به خانواده نمایی می‌باشند و خطاهایی با توزیع نرمال دارند. در این فصل ساختار GLM و مولفه های تشکیل دهنده آن بررسی شده است و معادلات درستنمایی نیز تحت توابع پیوند همانی و پیوند کانونی[2] ارائه گردیده است.

در فصل سوم با رگرسیون پواسن خطی تعمیم یافته آشنا می‌شویم. فرض می‌کنیم که Y متغیر تصادفی پواسن است و $X_0, X_1, .., .., X_K$ (که در آن $X_0 = 1$)، مجموعه‌ای از متغیرهای پیشگو یا رگرسورها، $b_0, b_1, ..., b_k$ مجموعه پارامترهای نامعلوم و متغیر تصادفی خطای e دارای توزیع نرمال با میانگین صفر و واریانس ثابت باشد. از مدل $\ln m = X_t' b = b_0 + \sum_{t=1}^{k} b_t X_t$ بعنوان مدل لگ خطی پواسن یاد می‌کنیم و چون داریم $E(Y) = m = \exp(b_0 + \sum_{t=1}^{k} b_t X_t)$؛ بنابراین تابعی که میانگین را بصورت خطی به پیشگوها ارتباط می‌دهد، تابع لگاریتم است. در نتیجه گوییم تابع پیوند در رگرسیون پواسن تابع لگاریتم می‌باشد. مدل سازی داده‌های شمارشی یا نرخ‌ها برای متغیر پاسخ توسط رگرسیون پواسن خطی تعمیم یافته صورت می‌گیرد و می‌توان نیکویی برازش این مدل به داده‌های شمارشی را نیز آزمون نمود. همچنین در این فصل سه مثال که در درک مطلب کمک شایانی می‌کنند، بررسی کرده‌ایم.

در فصل چهارم به مبحث اصلی این پایان نامه یعنی رگرسیون پواسن استوار می‌پردازیم و تکنیک تابع درستنمایی نیمرخی را معرفی می‌کنیم و از ضریب $\frac{A}{B}$ که آن را در متن توضیح خواهیم داد برای

[1] Nelder & Wedderburn , 1972.
[2] Canonical link

تعدیل کردن[1] تابع درستنمایی نیمرخی استفاده می‌کنیم. همچنین با مفاهیمی چون شواهد گمراه کننده قوی[2] و ضعیف[3] آشنا می‌شویم و شرایطی را که تحت آن تابع درستنمایی نیمرخی، استوار می‌باشد؛ بررسی می‌کنیم. در حقیقت رگرسیون پواسن خطی تعمیم یافته‌ای را که به مدل برازش داده‌ایم، توسط تابع درستنمایی نیمرخی تعدیل شده، استوار می‌سازیم و به این صورت برآورد پارامترها را تحت مدل "بد مشخص سازی شده"[4] بهبود می‌بخشیم و بدین ترتیب برآورد این پارامترها بطور مجانبی به مقدار مورد استنباط منطبق می‌گردد. همچنین در این فصل نتایج برنامه‌های شبیه‌سازی شده از توزیع‌های دوجمله ای منفی و وایبل را برای حجم نمونه های N=50 و N=100 ارائه می‌دهیم.

پیوست الف شامل جزئیات محاسباتی برای مدل‌های خطی تعمیم یافته و الگوریتم نیوتن–رافسون می‌باشد. پیوست ب دستورات برنامهMatlab برای شبیه سازی است و در پیوست ج، داده‌های مربوط به مثال بروکمن (۱۹۹٦) که در فصل سوم به آن پرداخته‌ایم، آمده است. در انتها واژه‌نامه فارسی به انگلیسی و سپس مراجع مورد استفاده در این پایان نامه را قرار داده‌ایم. مقاله‌های رویال و تسو(۲۰۰۵)، رویال و تسو(۲۰۰۳)، رویال و تسو(۲۰۰۰)، بلوم و همکاران (۲۰۰۷) و کتاب مدل‌های خطی تعمیم یافته‌ی مونتگمری (سال۱۹۳۷) و کتاب مقدمه‌ای بر تحلیل داده‌های رسته‌ای اگرستی(سال۱۹۹٦)، اساس تالیف این پایان نامه را تشکیل داده‌اند و گردآورنده تمام تلاش خود را به کار بسته است تا بتواند مفاهیم این مقاله‌ها را به بهترین صورت ممکن مورد بررسی قرار دهد.

[1] Adjusting
[2] Strong misleading evidence
[3] Weak misleading evidence
[4] Misspecified

فصـــــل دوم

خانواده مدل‌های خطی تعمیم یافته

۲-۱- مقدمه

تحلیل رگرسیون مجموعه‌ای از تکنیک‌های آماری است که برای مدل‌سازی و بررسی رابطه بین یک متغیر پاسخ مورد علاقه Yو مجموعه‌ای از متغیرهای پیشگو (مستقل) $X_1, X_2, ..., X_K$ می‌باشد. کاربردهای رگرسیون گسترده بوده و تقریبا در هر زمینه کاربردی چون مهندسی و علوم مربوط به شیمی، فیزیک، علوم زیستی، مدیریت، اقتصاد و . . . استفاده می‌شود. مدل‌های رگرسیون خطی بطور گسترده بعنوان مدل‌های تجربی برای تقریب یک رابطه تابعی پیچیده و معمولا نامعلوم بین متغیرهای پاسخ و پیشگو به کار می‌رود.

نلدر و ودربرن در سال ۱۹۷۲، مدل‌های خطی عمومی را به مدل‌های خطی تعمیم یافته توسعه دادند. مدل‌های خطی عمومی بعنوان یک مورد از مدل خطی تعمیم یافته با تابع پیوند همانی، در نظر گرفته می‌شود. در حقیقت مدل‌های خطی تعمیم یافته، مدل‌های خطی عمومی هستند که متعلق به خانواده نمایی می‌باشند و خطاهایی با توزیع نرمال دارند. GLMها به فرض مشخص بودن توزیع کامل[1] نیاز دارند، اما در برخی موقعیت‌ها مخصوصا در مسایل داده‌های گسسته، گاهی اوقات فرض توزیع کامل محدود می‌شود. برای رفع نیاز به وجود توزیع کامل ودربرن در سال ۱۹۷٤ تابعی به نام تابع شبه درستنمایی[2] معرفی نمود که فقط بر پایه فرض وجود گشتاور مرتبه دوم متناهی متغیر پاسخ برقرار می‌باشد(استفورد،۱۹۹٦)[3]. از روش تابع شبه درستنمایی فقط در برآورد کردن میانگین یا ضرایب رگرسیونی استفاده می‌کنیم. خانواده مدل‌های خطی تعمیم یافته این امکان را فراهم می‌سازد تا یک الگوریتم عمومی برای برآورد درستنمایی ماکزیمم در تمام این مدل‌ها به کار ببریم.

از آنجایی که رگرسیون پواسن یک حالت خاص از مدل‌های خطی تعمیم یافته است، لازم می‌باشد تا در ابتدای فصل به معرفی خانواده مدل‌های خطی تعمیم یافته بپردازیم. برای این منظور با خانواده

[1] Full Distribution
[2] Quasi Likelihood(QL)
[3] Srafford,1996.

نمایی آشنا می‌شویم و سپس به معرفی خانواده مدل‌های خطی تعمیم یافته و معادلات درستنمایی در این خانواده می‌پردازیم.

۲-۲- مدل‌های خطی

بطور کلی می‌توان مدل‌ها را به دو دسته مدل‌های قطعی و مدل‌های احتمالی رده‌بندی کرد. در مدل‌های قطعی پاسخ‌ها اغلب بوسیله مجموعه‌ای از معادلات دقیق تعریف می‌شوند که این نوع مدل‌ها بطور گسترده در علوم مهندسی بکار می‌روند. قانون اهم $(E = IR)$ یا انرژی پتانسیل مثال‌هایی از مدل‌های قطعی به شمار می‌روند. در مدل‌های احتمالی پاسخ‌ها تغییرپذیری را نشان می‌دهند؛ زیرا مدل یا عناصر تصادفی را شامل می‌شوند یا این که به نوعی با خطاهای تصادفی سر و کار دارند.

مهمترین رده از مدل‌های احتمالی مدل‌های خطی می‌باشد که بصورت زیر بیان می‌شوند:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k + e \qquad (۱-۲)$$

که در آن Y متغیر پاسخ، Xها مجموعه‌ای از متغیرهای رگرسور یا پیشگو، $b_0, b_1, ..., b_k$ مجموعه پارامترهای نامعلوم و eبا توزیع $N(0, s^2)$ جمله خطای تصادفی است. معادله شماره (۲-۱) را اغلب مدل رگرسیون خطی می‌نامیم. جمله خطا دارای میانگین صفر است؛ بنابراین میانگین پاسخ در مدل رگرسیون خطی عبارتست از:

$$E(Y) = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k \qquad (۲-۲)$$

معادله (۲-۲) را به این خاطر مدل خطی می‌گویند که در آن میانگین پاسخ یک تابع خطی از پارامترهای مجهول $b_0, b_1, ..., b_k$ می‌باشد.

مدل‌های رگرسیون خطی به دلایل گوناگونی کاربرد وسیعی دارند. دلیل اول این که این مدل‌ها تقریب‌های طبیعی چند جمله‌ای‌ها برای روابط تابعی پیچیده‌ترند. 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 نقطه‌ای مانند x_0 با رابطه (۲-۳) بدست می‌آید.

$$E(Y) \cong f(x_0) + \left.\frac{df(x)}{dx}\right|_{x=x_0} (x - x_0) + R \cong b_0 + b_1(x - x_0) \qquad (۳-۲)$$

که صرفنظر 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 می‌شود. چون اغلب از الگوهای رگرسیون خطی (بطور موفقیت‌آمیزی) بعنوان تقریب چندجمله‌ای‌ها استفاده می‌شود، لذا بعضی اوقات این مدل‌ها را مدل‌های تجربی می‌نامند. دومین دلیل استفاده از مدل‌های رگرسیون خطی این است که از طریق آنها پارامترهای مجهول مدل یعنی $b_0, b_1, ..., b_k$ مستقیما برآورد می‌شوند. در این بخش به دو روش مهم برآورد پارامترها برای مدل رگرسیون خطی که روش‌های کمترین مربعات معمولی[1] و درستنمایی ماکزیمم[2] هستند، اشاره می‌نماییم. همچنین نقش مهم توزیع نرمال در رگرسیون خطی را یادآوری می‌کنیم.

برآورد پارامترها به روش کمترین مربعات

روش کمترین مربعات نوعا برای برآورد ضرایب رگرسیون در یک مدل رگرسیون خطی چندگانه به کار می‌رود. فرض می‌کنیم $n > k$ مشاهده از متغیر پاسخ Y بصورت $y_1, y_2, ..., y_n$ داریم. برای هر پاسخ مشاهده شده y_i یک مشاهده برای هر متغیر پیشگو داریم و فرض می‌کنیم x_{it} مشاهده iام متغیر x_t باشد. جمله خطای e در مدل را دارای میانگین صفر و واریانس ثابت s^2 و $\{e_i\}$ را متغیرهای تصادفی ناهمبسته فرض می‌کنیم. معادله (۲-۱) را بر حسب مشاهدات بصورت معادله (۲-٤) می‌نویسیم.

[1] Ordinary Least Square
[2] Maximum Likelihood

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik} + e_i$$
$$= b_0 + \sum_{t=1}^{k} b_t x_{it} + e_i \qquad (۴-۲)$$

روش کمترین مربعات، bها را در معادله بالا طوری انتخاب می‌کند که مجموع مربعات خطاهای e_i مینیمم شود. برآورد کمترین مربعات bها بصورت زیر است:

$$b = (X'X)^{-1} X'Y \qquad (۵-۲)$$

برآورد پارامترها به روش درستنمایی ماکزیمم

از روش کمترین مربعات برای برآورد پارامترهای یک مدل رگرسیون خطی بدون در نظر گرفتن شکل توزیع متغیر پاسخ Y می‌توان استفاده کرد. اگر شکل توزیع پاسخ معلوم باشد، آنگاه از روش برآورد دیگری به نام درستنمایی ماکزیمم می‌توانیم استفاده کنیم.

مدل رگرسیون خطی $Y = Xb + e$ را در نظر می‌گیریم. فرض می‌کنیم خطاها در این مدل دارای توزیع نرمال مستقل با میانگین صفر و واریانس ثابت s^2 باشند. در این صورت مشاهده y_i در نمونه $(y_i, x_{i1}, ..., x_{ik})$ دارای توزیع نرمال مستقل با میانگین $b_0 + \sum_{t=1}^{k} b_t x_{it}$ و واریانس s^2 خواهد بود. تابع درستنمایی را از توزیع احتمال توام مشاهدات پیدا می‌کنیم. اگر این توزیع توام را با مشاهدات داده شده و پارامترهای مجهول در نظر بگیریم؛ آنگاه تابع درستنمایی را بصورت زیر داریم:

$$L(y, b, s^2) = \frac{1}{(2ps^2)^{n/2}} \exp\left\{\frac{-1}{2s^2}[(y - xb)'(y - xb)]\right\} \qquad (۶-۲)$$

برآوردگرهای درستنمایی ماکزیمم مقادیری از پارامترهای b و s^2 هستند که تابع درستنمایی را ماکزیمم می‌کنند. ماکزیمم کردن تابع درستنمایی L معادل ماکزیمم کردن تابع لگاریتم درستنمایی یعنی $\ln L$ می‌باشد. تابع لگاریتم درستنمایی بصورت زیر است:

$$\ln\left[L(y, b, s^2)\right] = \frac{-n}{2}\ln(2p) - \frac{n}{2}\ln(s^2) - \frac{1}{2s^2}[(y - xb)'(y - xb)] \qquad (۷-۲)$$

مشتق تابع لگاریتم درستنمایی را "تابع امتیاز" می‌نامند. اگر از تابع لگاریتم درستنمایی نسبت به پارامترهای b و s^2 مشتقات جزیی گرفته و مساوی صفر قرار دهیم خواهیم داشت:

$$\frac{1}{s^2} X'(Y - Xb) = 0 \qquad (۸-۲)$$

دستگاه معادلات $k \times k$ بالا را "معادلات امتیاز درستنمایی ماکزیمم"[1] می‌نامند. برآوردگر درستنمایی ماکزیمم، پاسخ دستگاه معادلات امتیاز است.

$$b = (X'X)^{-1} X'Y \qquad (۹-۲)$$

اگر نسبت به s^2 مشتق بگیریم و مساوی صفر قرار دهیم، برآوردگر s^2 بصورت زیر بدست می‌آید:

$$\hat{s}^2 = \frac{1}{n}(Y - Xb)'(Y - Xb) \qquad (۱۰-۲)$$

بطور کلی برآوردگرهای درستنمایی ماکزیمم نسبت به برآوردگرهای کمترین مربعات دارای خواص آماری بهتری هستند که بخاطر فرض‌های اضافی می‌باشد که برای آنها در نظر گرفته می‌شود. مثلا برآوردگرهای درستنمایی ماکزیمم نیاز به نرمال بودن توزیع مشاهدات دارند ولی در روش برآورد کمترین مربعات برقراری این فرض لازم نیست. برآوردگرهای درستنمایی ماکزیمم نااریب یا بطور مجانبی نااریب هستند که با افزایش حجم نمونه نااریب می‌شوند و نیز مجموعه‌ای از آماره‌های بسنده را تشکیل می‌دهند و سازگارند.

۲-۳- خانواده نمایی توزیع‌ها

تعاریف

چون خانواده نمایی توزیع‌ها یک مفهوم مهم در GLM می‌باشد، لذا ضروری است که در ابتدا با این خانواده بیشتر آشنا شویم.

[1] Score Equations Maximum Likelihood

اعضای خانواده نمایی همگی دارای تابع چگالی احتمال برای پاسخ مشاهده شده Y هستند که آن را به فرم ارائه شده در (۲-۱۱) نشان می‌دهیم:

$$f(y;q,\Psi)=\exp\{\frac{yq-b(q)}{a(\Psi)}+c(y,\Psi)\} \qquad (۲-۱۱)$$

که در آن a(.) و b(.) و c(.) توابعی مشخص هستند. θ پارامتر مکان و Ψ پارامتر پراکندگی نام دارد. بطور مثال توزیع‌های دوجمله‌ای، پواسن، نرمال و . . . متعلق به این خانواده می‌باشند.

بعنوان مثال تابع چگالی احتمال متغیر تصادفی نرمال Y با پارامترهای μ و σ^2 را در نظر بگیرید که:

$$f(y;m,s)=\frac{1}{s\sqrt{2p}}\exp\left\{\frac{-[y-m]^2}{2s^2}\right\}$$

$$=\exp\{\frac{1}{s^2}(ym-\frac{m^2}{2})-\frac{1}{2}[\frac{y^2}{s^2}+\ln(2ps^2)]\}$$

اگر تابع چگالی فوق را به صورت ارائه شده در (۲-۱۱) بنویسیم، $q=m$، $b(q)=m^2/2$، $a(\Psi)=\Psi$، $\Psi=s^2$ و $c(y,\Psi)=-1/2[\frac{y^2}{s^2}+\ln(2ps^2)]$. نتیجه می‌شود که در آن $q=m$ پارامتر مکان و $\Psi=s^2$ پارامتر پراکندگی می‌باشند.

برای توزیع گسسته، متغیر تصادفی Y با توزیع پواسن و پارامتر μ را در نظر می‌گیریم:

$$f(y;m)=\frac{e^{-m}m^y}{y!}=\exp\{y\ln m-m-\ln y!\}$$

در نتیجه داریم:

$q=\ln m$ ، $b(q)=e^q$ و $c(y,\Psi)=-\ln(y!)$. پارامتر مکانی m و پارامتر پراکندگی $\Psi=1$ است.

اگر L(θ) بصورت $L(q)=\prod_{i=1}^{n} f(y_i;q)$، تابع درستنمایی برای تابع چگالی f(y) باشد؛ به سادگی می‌توان نشان داد که:

$$E(\frac{\partial \ln L(q)}{\partial q})=0 \qquad (۲-۱۲)$$

$$E\left(\frac{\partial^2 \ln L(q)}{\partial q^2}\right) + E\left(\frac{\partial \ln L(q)}{\partial q}\right)^2 = 0 \qquad (۱۳-۲)$$

این نتایج خواص خوبی از تابع درستنمایی هستند که بوسیله دیفرانسیل‌گیری زیر بدست می‌آیند:

$$\int f(y_i; q_i, \Psi) dy_i \equiv 1$$

در آن پارامتر پراکندگی Ψ ثابت در نظر گرفته می‌شود.

تابع لگ درستنمایی با قرار دادن $l(q_i) = \ln L(q_i)$ به صورت زیر بدست می‌آید

$$l(q_i; y_i) = \frac{y_i q_i - b(q_i)}{a(\Psi)} + c(y_i, \Psi)$$

اگر از تابع فوق نسبت به q_i مشتق بگیریم، داریم:

$$\frac{\partial l}{\partial q_i} = \frac{1}{a(\Psi)}\{y_i - b'(q_i)\}$$

$$\frac{\partial^2 l}{\partial q_i^2} = \frac{-b''(q_i)}{a(\Psi)}$$

با توجه به (۲-۱۲) و (۲-۱۳) داریم

$$E\left(\frac{\partial l}{\partial q_i}\right) = \frac{1}{a(\Psi)}\{E(Y_i) - b'(q_i)\} = 0$$

که نتیجه می‌دهد

$$E(Y_i) = b'(q_i) \qquad (۱۴-۲)$$

بنابراین در خانواده نمایی به فرم (۲-۱۱)، $E(Y_i)$ از رابطه (۲-۱۴) بدست می‌آید.

همچنین:

$$0 = E\left(\frac{\partial^2 l}{\partial q_i^2}\right) + E\left(\frac{\partial l}{\partial q_i}\right)^2$$

$$= \frac{-b''(q_i)}{a(\Psi)} + E\left(\frac{Y_i^2 + b'^2(q_i) - 2b'(q_i)Y_i}{a^2(\Psi)}\right)$$

$$= \frac{-b''(q_i)a(\Psi) + E(Y_i^2) + b'^2(q_i) - 2b'(q_i)E(Y_i)}{a^2(\Psi)}$$

با توجه به (۲-۱۴) و این که $b'^2(q_i) = E^2(Y_i)$

$$0 = -b''(q_i)a(\Psi) + E(Y_i^2) + E^2(Y_i) - 2E^2(Y_i)$$

$$0 = -b''(q_i)a(\Psi) + E(Y_i^2) - E^2(Y_i)$$

$$b''(q_i)a(\Psi) = E(Y_i^2) - E^2(Y_i) = \text{var}(Y_i)$$

در نتیجه فرمول زیر را برای پیدا کردن واریانس در خانواده نمایی به کار می‌بریم.

$$\text{var}(Y_i) = b''(q_i)a(\Psi) \qquad (۱۵-۲)$$

و نیز با توجه به روابط اخیر خواهیم داشت:

$$m = E(Y) = \frac{db(q)}{dq} = b'(q)$$
$$\text{var}(Y) = \frac{d^2b(q)}{dq^2}a(\Psi) = b''(q)a(\Psi) = \frac{dm}{dq}a(\Psi)$$

فرض کنید بدون در نظر گرفتن $a(\Psi)$، var_m واریانس پاسخ Y باشد که اندیس m وابسته بودن واریانس پاسخ به میانگینش را نشان می‌دهد. به این ترتیب:

$$\text{var}_m = \frac{\text{var}(Y)}{a(\Psi)} = \frac{dm}{dq} = b''(q) = \frac{\partial^2 b(q)}{\partial q^2}$$

درنتیجه $$\frac{dq}{dm} = \frac{1}{\text{var}_m}$$

این نتایج را می‌توان به آسانی برای توزیع‌های نرمال یا پواسن بررسی کرد.

در توزیع نرمال به فرم خانواده نمایی داشتیم که:

$$q = m \qquad b(q) = \frac{m^2}{2} \qquad a(\Psi) = s^2$$

و همچنین با استفاده از فرمول‌های ارائه شده، داریم:

$$\text{var}(Y) = \frac{d^2b(q)}{dq^2}a(\Psi) = s^2 \qquad E(Y) = \frac{db(q)}{dq} = m$$

برای توزیع پواسن در فرم خانواده نمایی نشان دادیم که:

$$q = \ln m \quad \Rightarrow \quad m = \exp(q) \quad ، \quad b(q) = m \quad ، \quad a(\Psi) = 1 \quad ، \quad c(y, \Psi) = -\ln y!$$

$$E(Y)=\frac{db(q)}{dq}=\frac{db(q)}{dm}\frac{dm}{dq}$$

چون $\frac{dm}{dq}=\exp(q)=m$ است، لذا واریانس توزیع پواسن عبارت است از $Var(Y)=m\times 1=m$ که در واقع همان میانگین توزیع پواسن است.

۲-٤- خانواده مدل‌های خطی تعمیم یافته

۲-٤-۱- تعاریف

واضح است که وقتی با الگوهای رگرسیون خطی و غیر خطی سر و کار داریم، توزیع نرمال نقش محوری ایفا می‌کند. در حقیقت در روش‌های استنباطی مربوط به مدل‌های رگرسیون خطی و غیر خطی فرض بر این است که متغیر پاسخ Y از توزیع نرمال تبعیت می‌کند؛ اما موارد زیادی وجود دارند که در آنها این فرض حتی بصورت تقریبی هم برقرار نیست. مانند زمانی که متغیر پاسخ یک متغیر گسسته نظیر یک شمارش باشد. ما اغلب با شمارش عیب‌ها یا "پیشامدهای نادر" چون آسیب‌ها، بیمارانی با امراض خاص، زلزله و . . . مواجه می‌شویم.

مدل‌های خطی تعمیم یافته یک تکنیک برای مدل‌سازی رابطه بین k+1 متغیر پیشگو و تابعی از میانگین متغیر پاسخ گسسته یا پیوسته Y است که بیان می‌کند مقدار مورد انتظار متغیر پاسخ Y ، $m=E(Y)$، به متغیرهای پیشگو از طریق رابطه $E(Y)=b_0+b_1X_1+b_2X_2+\ldots+b_kX_k$ وابسته است. الگوهای خطی تعمیم یافته که از توزیع بسیار جامع خانواده نمایی پیروی می‌کنند، برای برازش الگوهای رگرسیون به داده‌های پاسخ یک متغیری توسعه داده شده‌اند. توزیع‌هایی چون نرمال، دوجمله‌ای، پواسن، هندسی نمایی، گاما ،. . . متعلق به خانواده نمایی هستند.

۲-٤-۲- مولفه‌های مدل‌های خطی تعمیم یافته

مدل‌های خطی تعمیم یافته از سه مولفه تشکیل شده‌اند:

مولفه تصادفی

مولفه تصادفی متغیر پاسخ را تعریف می‌کند و توزیع آن را مشخص می‌سازد.

مشاهدات $y_1, y_2, ., ., ., y_n$ که یافته‌های متغیرهای پاسخ مستقل و همتوزیع $Y_1, Y_2, ., ., ., Y_n$ هستند را در نظر می‌گیریم. منظور از مولفه تصادفی GLM، تعیین متغیر پاسخ Y و انتخاب یک توزیع آماری برای $Y_1, Y_2, ., ., ., Y_n$ است. توزیع مولفه تصادفی ممکن است گسسته یا پیوسته باشد. برای مثال در حالتی که برآمدهای Y_i دودویی (برای مثال شکست یا پیروزی) و یا تعداد موفقیت‌ها در تعدادی آزمایش باشند، توزیع مولفه تصادفی دوجمله‌ای در نظر گرفته می‌شود. اگر مشاهدات پیوسته باشند، بهتر است از توزیع نرمال بعنوان مولفه تصادفی استفاده کنیم. بنابراین با توجه به ساختار متغیر پاسخ می‌توان توزیع مولفه تصادفی را انتخاب کرد. برای مثال در الگوی رگرسیون خطی $Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k + e$ تابع پاسخ نرمال بوده و پیشگوی خطی عبارت است از:

$$X_t' = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k . \qquad (۱۶-۲)$$

مولفه سیستماتیک

مولفه سیستماتیک در مدل خطی تعمیم یافته یک پیشگوی خطی است و متغیرهای استفاده شده در مدل را که بعنوان متغیرهای پیشگو هستند شامل می‌شود.

در GLMها مقدار $E(Y) = m$ با تغییر سطوح متغیرهای تبیینی تغییر می‌کند. در هنگام ارائه معادله، متغیرهای تبیینی بصورت خطی تحت عنوان پیشگو در طرف راست مدل قرار می‌گیرند. به عبارت دیگر مولفه سیستماتیک $X_t' = b_0 + b_1 X_1 + ... + b_k X_k$، متغیرهایی که نقش $\{X_j\}$در مدل را به عهده دارند، تعیین می‌کند.