
بسم الله الرحمن الرحيم

دانشگاه پیام نور

دانشکده علوم پایه

پایان نامه

برای دریافت مدرک کارشناسی ارشد

رشته آمار ریاضی

گروه آمار

عنوان پایان نامه:

استوار سازی معادلات برآوردگر تعمیم یافته

سعید معدنی

استاد راهنما:

دکتر مسعود یار محمدی

استاد مشا ور:

دکتر پرویز نصیری

شهریور۹۰

جمهوری اسلامی ایران
وزارت علوم تحقیقات و فناوری

مجتمع علوم پایه و کشاورزی

شماره .....
تاریخ .....
پیوست .....

دانشگاه پیام نور
دانشگاه پیام نور استان تهران
العلم عجل لولیک الفرج و العافیه و النصر

صورتجلسه دفاع از پایان نامه کارشناسی ارشد

جلسه دفاع ازپایان نامه کارشناسی ارشد آقای سعید معدنی
دانشجوی رشته آمار ریاضی به شماره دانشجویی: ۸۶۷۱۰۵۲۶۶
تحت عنوان:

"استوار سازي معادلات برآوردگر تعميم يافته "

جلسه دفاع با حضور داوران نامبرده ذیل در روز دوشنبه مورخ: ۹۰/۰۶/۰۷ ساعت : ۱۰-۹ درمحل مجتمع علوم پایه و کشاورزی برگزار شد. و پس از بررسی پایان‌نامه مذکور با نمره به عدد ۱۸/۵ به حروف هجده و پنج دهم و با درجه ارزشیابی عالی مورد قبول واقع شد ☑ نشد ☐

ردیف	نام و نام خانوادگی	هیات داوران	مرتبه دانشگاهی	دانشگاه/ موسسه	امضاء
۱	دکترمسعود یارمحمدي	استاد راهنما	دانشیار	پیام نور	
۲	دکترپرویز نصیري	استاد مشاور	دانشیار	پیام نور	
۳	دکتر صادق رضایي	استاد داور	دانشیار	امیرکبیر	
٤	دکترسیما نصري	نماینده علمي گروه	استادیار	پیام نور	
٥	دکترسیما نصري	نماینده تحصیلات تکمیلي	استادیار	پیام نور	

تهران، خیابان استاد نجات اللهی
خیابان شهید فلاح پور، پلاک ۲۷
تلفن: ۸۸۸۰۰۲۵۲
دورنگار: ۸۸۳۱۹۴۷۵
WWW.TPNU.AC.IR
science.agri@tpnu.ac.ir

اینجانب سعید معدنی دانشجوی ورودی سال ۱۳۸٦مقطع 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 میدانم و جوابگوی آن خواهم بود.

سعید معدنی

اینجانب سعید معدنی دانشجوی ورودی سال ۱۳۸٦مقطع 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 ، بانظر ایشان نسبت به نشر مقاله، کتاب، و...وبه صورت مشترک و با ذکر نام استاد راهنما مبادرت نمایم .
سعید معدنی

کلیه حقوق مادی مترتب از نتایج مطالعات ، آزمایشات و نوآوری ناشی از تحقیق این پایان نامه متعلق به دانشگاه پیام نور می باشد
شهریور ۱۳۹۰

به پاس گرمای امید بخش وجودشان

این مجموعه را به همسر و دخترم تقدیم می کنم.

تقدیر و تشکر

در آغاز بر خود واجب میدانم که از زحمات بی دریغ استاد بزرگوار جناب آقای دکتر مسعود یار محمدی که باراهنمایی های راهگشای خود مرا در تهیه و تدوین این اثر همراهی کردند سپاسگزاری کنم.

همچنین از استاد بزرگوار جناب آقای دکتر پرویز نصیری که در مسیر تدوین پایان نامه از مشاوره ارزنده ایشان بهره گرفته ام تشکر می کنم.

چکیده

معادلات برآوردگرتعمیم یافته ارایه شده توسط لیانگ وزیگر (۱۹۸٦) ، روشی برای برآورد پارامترها در مدلهای خطی تعمیم یافته است ، هنگامی که همبستگی نامشخصی در میان مشاهدات موجود باشد. این روش در برابر داده های غیر معمول و نقاط دورافتاده تحت تاثیر قرار دارد . روشهای تشخیصی و روشهای استوارسازی روشهائی هستند که به ترتیب جهت شناسائی نقاط دورافتاده و نیز کاهش اثرات این نقاط به کار می روند.

در این تحقیق دو روش شناسائی داده های دورافتاده و استوارسازی را مورد بررسی قرار داده و آنها را در مورد معادلات برآوردگر تعمیم یافته به کار می بریم . روشهای تشخیصی به کاررفته تعمیم روشهائی می باشند که توسط کوک (۱۹۷۷) و بلسلی و همکاران (۱۹۸۰) در رابطه با رگرسیون خطی مطرح شده اند. روش استوارسازی به کار رفته که در اینجا آنرا معادلات برآوردگر تعمیم یافته استوار می نامیم تعمیمی از روش معادلات برآوردگر تعمیم یافته می باشد که در برابر داده هائی با قدرت نفوذ بالا استوار است . همچنین با ارائه یک مثال کاربردی روشهای تشخیصی و روش های استوارسازی را بررسی کرده و با استفاده از روش شبیه سازی روشهای استوار را با روش معادلات برآوردگر تعمیم یافته مقایسه خواهیم کرد.

واژه های کلیدی :معادلات برآوردگر تعمیم یافته – معادلات برآوردگر تعمیم یافته استوار

فهرست مطالب

مقدمه

معادلات برآوردگر تعمیم یافته[۱] (GEE) که توسط لیانگ[۲] وزیگر[۳] (۱۹۸٦) ارایه شده است امروزه به طور وسیعی در بیولوژی و پزشکی به کار رفته و کاربرد آن در رشته‌های علمی دیگر مانند کشاورزی و صنایعی که در تحلیل داده های آنها همبستگی بین مشاهدات وجود دارد دیده می‌شود. به عنوان مثال در کشاورزی ممکن است داده های جمع آوری شده روی یک کرت از زمین دارای همبستگی باشند و یا در تحقیقات صنعتی مشاهدات جمع آوری شده مربوط به یک کوره مجموعه ای از داده های همبسته را بوجود آورند.

هرگاه از هر عضو نمونه در چند موقعیت مکانی یا زمانی مشاهداتی جمع آوری شود ، این داده ها را اندازه گیری‌های مکرر[٤] می نامیم. داده های مربوط به هر عضو مجموعه ای از مشاهدات همبسته را تشکیل می دهند که آنها را خوشه نامیده و داده‌های مکرر را که در آن موقعیت تکرار مشاهدات ، نقاط زمانی هستند ، داده‌های طولی و مطالعات از این نوع را مطالعات طولی می نامند . در مطالعات طولی اغلب بین متغیرهای درون هر خوشه همبستگی وجود داشته و به همین دلیل برای تحلیل این نوع از داده ها باید تمهیدات خاصی را به کار برد که روش معادلات براوردگر تعمیم یافته یکی از آنها است.

به علت کاربرد زیاد روش (GEE) در برآورد پارامترهای مدل های خطی تعمیم یافته بررسی تأثیر مشاهدات جمع آوری شده روی برآورد پارامترها مهم بوده و اینکه مشاهدات دور افتاده کدامند و تأثیر آنها روی برآورد پارامترها چیست حائز اهمیت است.

---

۱ .generalized estimating equations
۲ .liang
۳ .zeger
٤ Repeated measurements

به منظور بررسی و تشخیص اینکه در اثر حذف مشاهدات یا خوشه ها چه تغییراتی در برازش مدل اتفاق می‌افتد، روش های تشخیصی رگرسیون را بکاربرده ایم. استفاده از روش‌های تشخیصی برای اثرات داده ها، خصوصاً در مدلهای چند متغیره کارایی بالائی نداشته بعلاوه شناسائی نقاط دور افتاده و تشخیص آن در رگرسیون دودویی که در آنها متغیر پاسخ به صورت دو حالتی ۰ یا ۱ مطرح می شود از تشخیص هویت آنها در داده های پیوسته دشوارتر است.

از آنجا که هدف از روشهای استوار سازی شناسایی و کاهش اثر داده های دورافتاده و پر نفوذ است، در این تحقیق ابتدا روشهای شناسائی نقاط دورافتاده و پرنفوذ را مطرح کرده و اثر داده های غیرمتجانس را روی پارامترهای مدلهای خطی تعمیم یافته که به روشGEE تحلیل می شوند مورد بررسی قرار می دهیم. سپس روشهای استوارسازی را به منظور کاهش اثر نقاط دورافتاده و پر نفوذ مورد توجه قرار داده و آنرا معادلات برآوردگرتعمیم یافته استوار و به اختصار[۱] (REGEE) می نامیم. این روشها بر اساس وزن دهی به مشاهدات برمبنای میزان تأثیر آنها روی برآورد پارامترهای مدل طراحی شده اند. روشهای برآورد پارامتر در (REGEE) توسعه یافته ی روشهای استوارسازی مدلهای خطی تعمیم یافته اند که توسط پرگیبون (۱۹۸۲)[۲] و کارول و پدرسن (۱۹۹۳)[۳] در استوارسازی رگرسیون لوژستیک و پرایسرو کاکیش (۱۹۹۵)[٤] در استوارسازی مدلهایی با پاسخ همبسته مطرح شده اند ، در این روش ها کاهش وزن مشاهدات با استفاده از الگوریتم معینی انجام شده و برای هر مشاهده وزنی بین ۰ و۱ در نظر گرفته می شود که این وزن متناسب با اثر آن مشاهده روی برآورد پارامترهای مدل است.

[۱] .Resistant generalized estimating equations
[۲] Pergibon
[۳] Carroll and Pederson
[٤] Preisser and Qaqish

درفصل اول راجع به مدلهای خطی تعمیم یافته و روش معادلات برآوردگر تعمیم یافته بحث می شود. روش‌های تشخیصی و اصول استوارسازی در فصل دوم مورد بررسی قرار می گیرد. در فصل سوم روشهای تشخیصی را برای معادلات برآوردگر تعمیم یافته مورد بحث قرار داده و مثالی کاربردی از داده های پزشکی را تحلیل خواهیم کرد، روشهای استوارسازی را برای معادلات برآوردگر تعمیم یافته در فصل چهارم مورد بررسی قرار داده و با استفاده از شبیه سازی و یک مثال کاربردی روشهای استوار معادلات برآوردگر تعمیم یافته را با روش معمول آن مقایسه خواهیم کرد.

فصل اول

معادلات برآوردگر تعمیم یافته

فصل اول: معادلات برآوردگر تعمیم یافته

۱-۱ مقدمه

با توجه به کاربرد فراوان الگوهای خطی تعمیم یافته وضعیت های بسیاری وجود دارد که پاسخ های تکراری از یک واحد، مشاهداتی همبسته را به وجود می آورد. مثالی از این مورد را می توان در مطالعات طولی که در آن پاسخها روی یک آزمودنی در طول زمان به تکرار اندازه گیری می شوند در نظر گرفت. بدیهی است که مشاهدات روی یک آزمودنی همبسته اند و در نظر گرفتن همبستگی میان مشاهدات مشکلاتی را بوجود می آورد. در واقع وجود همبستگی استفاده از روشهای حداکثر درستنمایی معمول را مشکل کرده و لازم است از روش های شبه درستنمایی و یا تعمیمی از آن که روش معادلات برآوردگر تعمیم یافته است استفاده شود. در این فصل ابتدا مدلهای خطی تعمیم یافته ، روش شبه درستنمایی و در پایان روش معادلات برآورد گر تعمیم یافته را معرفی می کنیم .

۱-۲ مدلهای خطی تعمیم یافته

در الگوهای رگرسیون خطی و غیر خطی توزیع نرمال نقش محوری داشته و در برخی از روشهای استنباطی مربوط به رگرسیون خطی و غیر خطی فرض آن است که متغیر پاسخ Y از توزیع نرمال پیروی می کند. اما مثال های کاربردی زیادی وجود دارد که این فرض حتی به طور تقریبی هم برقرار نیست. مثلاً زمانی که متغیر پاسخ یک متغیر گسسته و شمارشی است این فرض نقض می گردد. به عنوان مثال در مدلهائی با پاسخ های دوتایی نیز که فقط دو حالت موفقیت (۱) و شکست (۰) مورد نظر است فرض نرمال بودن نقض شده است. مواردی وجود دارد که متغیر پاسخ پیوسته بوده ولی فرض نرمال بودن برقرار نیست به عنوان مثال می توان به توزیع فشار در اجزاء مکانیکی و یا زمان زوال اجزای الکترونیکی اشاره کرد.

الگوهای خطی تعمیم یافته [۱](GLM) برای برازش الگوهای رگرسیون به داده هایی هستند که متغیر پاسخ در آنها از توزیع خانواده نمایی تبعیت می کنند. خانواده توزیع های نرمال، دوجمله‌ای، پواسون، هندسی، دو جمله‌ای منفی، نمایی، گاما، از این نوع هستند. شکل کلی مدل خطی تعمیم یافته به صورت زیر است :

$$g(\mu) = g[E(y)] = X'\beta \qquad (۱-۱)$$

X بردار متغیرهای کمکی و β بردار پارامترها یا ضرایب رگرسیون است. g تابع پیوند یا ربط می باشد می توان گفت هر الگوی خطی تعمیم یافته سه جزء اصلی دارد:

۱. توزیع متغیر پاسخ.

۲. پیشگوی خطی که متغیرهای برگشت یا کمکی را شامل می شود.

[۱] .Generalized Linear model.

۳. تابع پیوند g که پیشگوی خطی را به میانگین متغیر پاسخ مربوط می کند.

۱-۲-۱ توزیع متغیر پاسخ

همانطورکه گفته شد متغیر پاسخ y با میانگین $E(y)=\mu$ در مدلهای خطی تعمیم یافته دارای تابع چگالی از خانواده نمایی می باشد.این تابع چگالی به صورت کلی زیر مطرح می شود.

$$f(y,\theta \text{و } \phi)=\exp\left\{\frac{y\theta - b(\theta)}{a(\phi)}+c(y,\phi)\right\} \qquad (۲-۱)$$

اگرتابع لگاریتم درستنمایی را $L(\theta)$ بنامیم ، داریم :

$$L(\theta)=\sum\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)}+c(y_i,\phi)\right\}$$

و از اصول استنباط آماری خواهیم داشت :

$$E\left(\frac{\partial L(\theta)}{\partial \theta}\right)=0 \qquad \Rightarrow \qquad \mu=E(y)=\frac{db(\theta)}{d\theta}=b'(\theta)$$

$$var(y)=\frac{d^2b(\theta)}{d\theta^2}a(\phi)=b''(\theta)\,a(\phi)=\frac{d\mu}{d\theta}a(\phi)$$

خانواده تابع چگالی نمایی روش یکسانی را برای برآورد پارامترها در مدلهای خطی تعمیم یافته فراهم می کند . مثلاً توزیع نرمال که برجسته ترین عضو خانواده نمایی است دارای تابع چگالی به صورت زیر است :

$$f(y,\mu,\sigma)=\exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}.\frac{1}{\sqrt{2\pi\sigma^2}}$$

$$=\exp\left\{\frac{\left[y\mu-\frac{\mu^2}{2}\right]}{\sigma^2}-\frac{1}{2}\left[\frac{y^2}{\sigma^2}+L_n(2\pi\sigma^2)\right]\right\}$$

با توجه به صورت کلی توزیع نمایی در (۱-۲) داریم:

$$\theta=\mu \qquad b(\theta)=\frac{\mu^2}{2} \qquad a(\phi)=\phi=\sigma^2$$

$$c(y,\phi) = -\frac{1}{2}\left[\frac{y^2}{\sigma^2} + L_n(2\pi\sigma^2)\right]$$

و یا در مورد توزیع دو جمله ای داریم:

$$f(y,n,p) = \binom{n}{y} p^y (1-p)^{n-y}$$

$$= \exp\left\{L_n\binom{n}{y} + yL_n p + (n-y)L_n(1-p)\right\}$$

$$\theta = L_n\left(\frac{p}{1-p}\right) \qquad b(\theta) = nL_n(1+e^\theta) \qquad \phi = 1 \qquad c(\cdot) = L_n\binom{n}{y} \quad a(\theta) = 1$$

۱-۲-۲ پیشگوی خطی

پیشگوی خطی یا همان مؤلفه سیستماتیک متغیرهایی مانند x را نشان می دهد که نقش متغیرهای تبیینی را در رگرسیون به عهده دارند. فرض کنیم $\mu = E(y)$ میانگین احتمال باشد مقدار μ در GLM با تغییر متغیرهای تبیینی تغییر می کند. به طور کلی مؤلفه سیستماتیک به صورت زیر مطرح می شود.

$$X'\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \qquad (۳-۱)$$

این ترکیب خطی، متغیرهای پیشگو نامیده می شود.

۱-۲-۳ تابع پیوند (تابع ربط)

یک تابع پیوند چگونگی وابستگی $E(y) = \mu$ را به $X'\beta$ شرح می دهد و معمولاً به صورت $g(\mu)$ نوشته می شود که یک تابع غیر خطی بوده و به صورت زیر مطرح می شود.

$$g(\mu)=\beta_{\circ}+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n=X'\beta \qquad (۱-٤)$$

ساده ترین تابع پیوند دارای شکل $g(\mu)=\mu$ است که پیوند همانی نامیده می شود. این تابع میانگین را به طور مستقیم مدلسازی کرده و مدلی شبیه مدل رگرسیون معمولی پدید می آورد که معمولا برای پاسخ های پیوسته مورد استفاده قرار می گیرد .

$$\mu=\beta_{\circ}+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n=X'\beta \qquad (۱-۵)$$

از جمله توابع پیوند دیگر $g(\mu)=\log(\mu)$ است که مدل را بر حسب لگاریتم میانگین بیان می کند. انتخاب توابع پیوند بستگی به ماهیت متغیر پیشامد کنترل دارد. انواع توابع ربط (پیوند) برای توزیع های مختلف درجدول (۱-۱) آمده است .

جدول ۱-۱

پیوند متعارف	توزیع
(پیوند همانی) $\eta_i=\mu_i$	نرمال
(پیوند لوجیت) $\eta_i=L_n\left(\frac{p_i}{۱-p_i}\right)$	دو جمله ای
(پیوند لگاریتمی) $\eta_i=L_n(\mu_i)$	پواسون
(پیوند وارون) $\eta_i=\left(\frac{۱}{\mu_i}\right)$	نمایی
(پیوند وارون) $\eta_i=\left(\frac{۱}{\mu_i}\right)$	گاما

۱-۳ معادلات درستنمایی برای الگوهای خطی تعمیم یافته

روش برآورد پارامترها در GLM، روش حداکثر درستنمایی می باشد ولی تغییرات واقعی حداکثر درستنمایی به الگوریتمی می رسد که برپایه روش تکراری کمترین مربعات وزنی قرار دارد . در

صورتی که از پیوند متعارف $\eta_i = g[\mu_i] = X_i'\beta$ استفاده کنیم، تابع لگاریتم درستنمایی عبارت است از:

$$\mathrm{L} = \log L(\mathrm{y},\beta) = \sum_{i=\backslash}^{n}\left\{\frac{(y_i\theta_i - b(\theta_i))}{a(\phi)} + c(y_i,\phi)\right\}$$

$$\frac{\partial L}{\partial \beta} = \frac{\partial L}{\partial \theta_i}\times\frac{\partial \theta_i}{\partial \beta} = \sum_{i=1}^{n}\frac{1}{a(\phi)}\left[y_i - \frac{db(\theta_i)}{d\theta_i}\right]X_i = \sum_{i=1}^{n}\frac{1}{a(\phi)}(y_i-\mu_i)X_i$$

$$= \sum_{i=1}^{n}\frac{1}{a(\phi)}(y_i-\mu_i)X_i \qquad (٦\text{-}١)$$

می توانیم برآوردهای حداکثر درستنمایی پارامترها را با حل دستگاه معادلات زیر برای β پیدا کنیم ازآنجا که در بیشتر موارد $a(\phi)$ یک عدد ثابت است، این معادلات به صورت زیر در می آیند:

$$\sum_{i=1}^{n}(y_i-\mu_i)X_i = ٠ \qquad (٧\text{–}١)$$

این دستگاه دارای $p = k+1$ معادله بوده و هر معادله برای یک پارامتر الگو می باشد. شکل ماتریسی این معادله به صورت زیر است :

$$X'(y-\mu) = ٠ \qquad (٨\text{–}١)$$

که در آن $\mu' = [\mu_1,\mu_2,...\mu_n]$ می باشد. این معادلات را معادلات امتیاز حداکثر درستنمایی می نامند.

١–٤ روش حداقل مربعات تعمیم یافته و موزون

در مدل رگرسیون خطی $\mathrm{y} = \mathrm{X}\beta + \varepsilon$ مفروضاتی مانند $\mathrm{E}(\varepsilon) = ٠$ و $\mathrm{V}(\varepsilon) = \sigma^2\mathrm{I}$ در نظر گرفته می شود گاهی این مفروضات قابل اجرا نیستند . به عنوان مثال اگر $V(\varepsilon) = \sigma^2 V$ که V یک ماتریس $n\times n$ شناخته شده باشد اصلاحاتی باید روی روش حداقل مربعات معمولی در نظر گرفته

شود. می دانیم که اگر V یک ماتریس قطری باشد اما اعضای قطر مساوی نباشد در اینصورت متغیرهای y ناهمبسته اند اما واریانسهای مساوی ندارند. اگر بعضی از اعضای غیر قطر اصلی صفر نباشد مشاهدات همبسته بوده و برآورد حداقل مربعات معمولی $\hat{\beta} = (X'X)^{-1}X'y$ بهینه نیست. چون $\sigma^2 V$ ماتریس کوواریانس خطاها می باشد، V بایستی ناویژه و معین مثبت باشد بنابراین ماتریس $n \times n$ ناویژه متقارن k موجود است به طوری که $k'k = kk = V$. ماتریس k ریشه دوم V نامیده می شود. متغیرهای زیر را در نظر می گیریم:

$$Z = k^{-1}y \qquad B = k^{-1}X \qquad g = k^{-1}\varepsilon$$

به عبارت دیگر مدل رگرسیون $y = X\beta + \varepsilon$ به صورت زیر خواهدشد

$$\mathrm{k}^{-1}\mathrm{y} = \mathrm{k}^{-1}\mathrm{X}\beta + \mathrm{k}^{-1}\varepsilon$$

$$Z = B\beta + g \qquad (۹-۱)$$

خطاهای تبدیل یافته در مدل جدید دارای امید ریاضی صفر بوده و ماتریس کوواریانس g هم به صورت زیر خواهد بود.

$$V(g) = E\left\{[g - E(g)][g - E(g)]'\right\} = E(gg')$$

$$= E\left(k^{-1}\varepsilon\varepsilon' k^{-1}\right) = k^{-1}E(\varepsilon\varepsilon')k^{-1} = \sigma^2 k^{-1}Vk^{-1} = \sigma^2 k^{-1}kkk^{-1} = \sigma^2 I$$

بنابراین اعضای g دارای میانگین صفر و واریانس ثابت بوده و ناهمبسته می باشند. چون خطاهای g در مدل $Z = B\beta + g$ و در مفروضات معمول صدق می کند می توانیم روش حداقل مربعات معمول را برای آن به کار ببریم. تابع حداقل مربعات و درنتیجه معادلات نرمال به صورت زیر خواهد بود .

$$S(\beta) = g'g = \varepsilon' V^{-1}\varepsilon = (y - X\beta)' V^{-1}(y - X\beta) \qquad (۱۰-۱)$$