٤٠٨١٨

۳۳۸۱ /۴/ ۳۶

مرکز اطلاعات و مدارک علمی ایران
تحویل مدارک

دانشگاه صنعتی اصفهان

دانشکده علوم ریاضی

برآورد در رگرسیون خطی ساده زمانی که ناهمگونی خطای مدل به فرم نامعلومی باشد

پایان نامه کارشناسی ارشد آمار

صدیقه میرزائی صالح آبادی

۴۰ ۸۱۸

استاد راهنما

دکتر احمد پارسیان

دکتر علی زینل همدانی

۱۳۸۰

۴۰ ۸۱۸

دانشگاه صنعتی اصفهان

دانشکده علوم ریاضی

پایان نامه کارشناسی ارشد رشته آمار خانم صدیقه میرزائی صالح آبادی

تحت عنـــوان

برآورد در رگرسیون خطی ساده زمانی که ناهمگونی خطای مدل به فرم نامعلومی باشد

در تاریخ ۱۱/۷/۸۰ توسط کمیته تخصصی زیر مورد بررسی و تصویب نهائی قرار گرفت.

۱ـ استاد راهنمای پایان‌نامه — دکتر احمد پارسیان و دکتر علی زینل همدانی

۲ـ استاد مشاور پایان‌نامه — دکتر علی رجالی

۳ـ استاد داور ۱ — دکتر عبدالرسول برهانی حقیقی

۴ـ استاد داور ۲ — دکتر سروش علیمرادی

سرپرست تحصیلات تکمیلی دانشکده — دکتر امیر نادری

الحمدلله رب العالمین

...رحمت واسعه پروردگار فرصتی مغتنم داد تا به اقتضای توان خود از محضر اساتیدی گرانقدر بهره جسته و ره‌توشه‌ای برگیرم، بی‌پیرایه بهترین ایام زندگیم را در محضر این عزیزان که نه فقط رهنمون به دانشم بودند، بلکه درس اخلاق وتواضع را آموختند سپری کرده‌ام. تمام مباهات من در طول تحصیل نه دست‌یازیدن به درجه‌ای از دانش، بلکه فراسوی آن شاگردی نزد معلمانی بوده‌است‌که خود دریایی از معرفت هستند و سهم من پرتویی از تشعشع معرفت ایشان بر اندیشه بوده‌است، انسانهای بزرگواری که حق قدردانی بر عهدهٔ اینجانب دارند:

معلمان ارجمندم، دکتر احمد پارسیان و دکتر علی زینل همدانی

مشاور پرتلاشم، دکتر علی رجالی

ناظرین محترم پایان‌نامه‌ام، دکتر عبدالرسول برهانی حقیقی و دکتر سروش علیمرادی

اساتید همراهم، دکتر ایوب ساعی، دکتر محمد صالحی

سرپرست محترم تحصیلات تکمیلی دانشکده، دکتر امیر نادری

اساتید گرانقدر، کارشناسان، مسؤولین و کارکنان محترم دانشکدهٔ علوم ریاضی

دوستان و همکلاسیهای خوبم در دوران کارشناسی و کارشناسی‌ارشد به‌ویژه آقای مسعود ماهوش که زحمت تایپ این پایان‌نامه را به عهده داشتند و خانمها مهتاب بقائی،فرخنده سجادی و مریم قائمی.

و خانوادهٔ بزرگوارم

پدر ارجمند و مادر مهربانم - دوگوهر پر تلالؤ زندگی‌ام -که وجودم برایشان همه رنج بود و وجودشان برایم همه مهر، مویشان سپیدی گرفت تا رویم سپید بماند، آنان که فروغ نگاهشان و گرمی کلامشان سرمایه‌های جاودانی زندگیم هستند. آنان که راستی قامتم در شکستگی قامتشان تجلی یافت. دربرابر وجود گرامیشان زانوی ادب بر زمین می‌نهم و با دلی مملو از عشق و محبت بر دستانشان بوسه می‌زنم...

و خواهران عزیز و برادران خوبم.

امید است که خداوند توفیق و رسیدن به کمال را جبران محبت‌هایشان قرار دهد.

صدیقه‌میرزائی صالح‌آبادی

تابستان ۱۳۸۰

کلیه حقوق مادی مترتب بر نتایج مطالعات، ابتکارات و نوآوریهای ناشی از تحقیق موضوع پایان‌نامه متعلق به دانشگاه صنعتی اصفهان است.

معلمان بزرگوارم:

نهال دانشی را که دیروز در وجودم نهادید به لطف حق و صبر و تلاش مستمر شما عزیزان امروز به درختی جوان بدل گشته و اولین شاخه‌اش به گل نشسته است، اینک این اولین شکوفه‌اش را هرچند ناچیز به شما تقدیم می‌کنم.

فهرست مطالب

چکیده:

عدم کارائی روش حداقل مربعات برای برآورد پارامترها، وقتی که خطای مدل دارای توزیعی با دم سنگین باشد و یا خطای مدل دارای واریانس ثابت نباشد، مسأله‌ای آشنا در رگرسیون خطی است.

در چنین حالتهایی فاصله اطمینان استاندارد پارامترهای مدل، احتمال پوششی کاملاً متفاوت از آنچه راکه باید باشد خواهد داشت. در این پایان‌نامه ابتدابه معرفی سه روش پیشنهادی برای حل این مشکل، با استفاده از تکنیک شبیه‌سازی می‌پردازیم وکارائی آنها را با هم مقایسه می‌نمائیم. نشان داده می‌شود که روش برآورد کراندار اثر گذار M با وزن شوئیپ، دارای بالاترین کارائی است، اگرچه در بعضی موارد ممکن است دو روش دیگر هم نتایج مناسبی را ارائه دهند.

در پایان، برای شیب خط رگرسیونی، یک فاصله اطمینان، با استفاده از تکنیک بوت استرپ و هرکدام از روشهای معرفی شده، به دست می‌آوریم.

فصل اول

مقدمه:

مطالعهٔ روابط بین متغیرها در زمینه‌های زیادی از فعالیتهای علمی متداول شده است. تجزیه و تحلیل رگرسیونی یکی از گسترده‌ترین روشهای آماری است که در بررسی روابط بین متغیرها مورد استفاده قرار می‌گیرد. این روش به علت ظرافت مباحث نظری آن و به دلیل فراهم آوردن یک روش سادهٔ مفهومی برای بررسی رابطهٔ بین متغیرها در علوم مختلف قابل توجه است.

تحلیل رگرسیونی و استفاده از مدلهای رگرسیونی ممکن است برای اهداف مختلفی از جمله توصیف داده‌ها، برآورد پارامترها، کنترل متغیرها و پیش‌بینی به‌کار رود. شاید مهمترین کاربرد رگرسیون در این است که بر اساس مشاهدات، متغیر پاسخ را در محدودهٔ داده‌ها پیش‌بینی نماید که این امر در انجام برنامه‌ریزی‌ها می‌تواند بسیار مفید باشد. از طرف دیگر، از لحاظ نظری بخش عمده‌ای از تحلیلهای رگرسیونی در ارتباط با مدلهای ریاضی است، مدلهایی که تعیین‌کنندهٔ شکل و نوع ارتباط بین متغیرهاهستند.

برای تجزیه و تحلیل رگرسیونی لازم است فرضهایی درنظر گرفته می‌شوند. اعتبار نتایج حاصل از رگرسیون، مشروط به صحت این فرضیات است. فرض اساسی که در تجزیه و تحلیل رگرسیون خطی مبنای کار قرار می‌گیرد این است که ارتباط بین متغیر پاسخ Y و متغیرهای مستقل X_1، X_2، ... X_p به صورت زیر است:

$$Y=\beta_0+\beta_1X_1+\ldots+\beta_pX_p+\varepsilon \qquad (1-1)$$

که در آن ε عامل خطای تصادفی مدل است . در اینجا معمولاً پارامترهای مدل به روش حداقل مربعات برآورد می‌شوند.

اگر علاوه بر برآورد پارامترها و برازش مدل نیاز به استنباطهای آماری دربارهٔ پارامترهای مدل باشد، برای سهولت درکار به جز فرض بالا فرضهای زیر نیز باید برقرار باشند:

۱) خطاهای تصادفی ε مستقل از هم باشند.

۲) خطاهای تصادفی ε دارای توزیع نرمال با میانگین صفر و واریانس ثابت σ^2 باشند.

امّا در بسیاری از مطالعات عملی مشاهده می‌شود که برخی از فرضیات فوق قابل حصول نیستند. در آن صورت برای رسیدن به فرضهای مذکور از روشهای متفاوتی از جمله انتخاب تبدیلات مناسب (دوجمله‌ای، پواسون، دوجمله‌ای منفی و...) یا انتخاب وزن مناسب برای مشاهدات و در نظرگرفتن رگرسیون وزنی و..... استفاده می‌شود[۱]. در بسیاری موارد از جمله در بررسی‌های اقتصادی یا مطالعات روانشناسی با عدم تثبیت واریانس خطاهای تصادفی ε مواجه می‌شویم که در این صورت می‌توان با درنظرگرفتن وزن برای مشاهدات و استفاده از رگرسیون وزنی واریانس خطا را ثابت نمود.

این تحقیق تلاشی برای شناخت و مقایسه چند روش ابداع شده برای نحوهٔ انتخاب وزن مناسب برای مشاهدات در رگرسیون است به طوری که برآورد پارامترها و استنباطهای آماری دربارهٔ آنها از اعتبار لازم برخوردار باشند.

مرکز اطلاعات و مدارک علمی ایران
تحویل سیستم مرکزی

۱.۱) تاریخچه

استفاده از روشهای رگرسیونی از همان موقعی که لژاندر[1] (۱۸۰۵) و گوس[2] (۱۸۰۹)[۲] داده‌های «مسیر حرکت اشیاء نجومی» را بررسی می‌کردند مورد علاقه محققان بوده است. در دهه‌های بعد، با گسترش روشهای رگرسیونی در علوم اجتماعی متوجه شدند که بعضی از فرضیات مدل (حتی به طور تقریبی) برقرار نیستند. یول[3] (۱۸۹۹)[۱] یکی از اولین کسانی بود که در مقالهٔ خود تحت عنوان «یک بررسی در علل تغییرات در گدائی در انگلستان، مخصوصاً در دودهه اخیر» با این مشکل مواجه گردید او برای رفع این اشکال در داده‌های موجود تغییراتی ایجاد نمود که سبب تعدیل در فرضیات مدل شد. پس از آن فعالیتهای زیادی در این زمینه انجام شده و هنوز هم ادامه دارد. امّا سابقهٔ تثبیت واریانس خطای تصادفی به سال ۱۹۶۴ برمی‌گردد. در این سال هوبر[4] با تعریف تابع Ψ-هوبر و درنظر گرفتن وزن مناسبی برای مشاهدات، رگرسیون وزنی را مورد مطالعه قرار داد.[۳]

از سایر مطالعات انجام شده در زمینهٔ ثابت‌کردن واریانس خطای تصادفی در مدل رگرسیونی می‌توان به مطالعات هوبر(۱۹۷۳)[۴]، باکس[5] و هیل[6] (۱۹۷۴)[۵]، شوئیپ[7] (۱۹۷۵)[۶]، موستلر[8] و توکی[9] (۱۹۷۷)[۷]، هیل و هلند[10] (۱۹۷۷)[۸]، جابسون[11] و فولر[12] (۱۹۸۰)[۹] که با معرفی وزنهای متفاوت، برآوردی برای واریانس خطا درنظر گرفته‌اند و نیز مطالعات اندروز[13] (۱۹۷۴)[۱۰]، مالوز[14] (۱۹۷۵)[۱۱]، بیکل[15] و داکسام[16] (۱۹۸۱)[۱۲]، هوبر(۱۹۸۱)[۳]، کارول[17] و راپرت[18] (۱۹۸۲)[۱۳]، همپل[19] (۱۹۸۶)[۱۴]، مارونا[20] و مورجنتالر[21] (۱۹۸۶)[۱۵]، رابینسون[22] (۱۹۸۷)[۱۶]، وینز[23] (۱۹۹۲)[۱۷]، کوهن[24]، دالال[25] و توکی(۱۹۹۳)[۱۸] و ویلکاکس[26] (۱۹۹۵)[۱۹] که با معرفی وزنهای مختلف، برآورد پایداری برای واریانس خطا ارائه داده‌اند، اشاره نمود.

هدف اصلی این تحقیق مطالعهٔ اهمیت فرض ثابت بودن واریانس خطا در مباحث رگرسیونی و بررسیِ چند روش مهمِ معرفی وزن و در نظر گرفتن رگرسیون وزنی در برقراری این فرض، در فصل سوم و نیز شبیه سازی و مقایسه این روشها در فصل چهارم است .

1- Legender
2- Gauss
3- Yule
4- Huber
5- Box
6- Hill
7- Schweppe
8- Mostteller
9- Tukey
10- Holland
11- Jobson
12- Fuller
13- Andrews
14- Mallows
15- Bickel
16- Doksom
17- Carroll
18- Ruppert
19- Hample
20- Marrona
21- Morgenthaler
22- Robinson
23- Wiens
24- Cohen
25- Dalal
26- Wilcox

۲.۱) روشهای تجزیه و تحلیل رگرسیون

در این پایان‌نامه از واژه‌هائی مانند رگرسیون خطی ساده[1]، روش حداقل مربعات[2]، رگرسیون وزنی[3]، برآوردگر رگرسیونی M[4] و..... استفاده شده است. برای سهولت در مطالعهٔ این پایان‌نامه، در این بخش به اختصار از آنها یاد می‌کنیم.

۱.۲.۱) رگرسیون خطی ساده

ساده‌ترین مدل رگرسیونی، رگرسیون خطی ساده شامل یک متغیر وابسته(پاسخ) و یک متغیر مستقل است . هدف پیداکردن یک رابطه خطی بین Y و X است . اگر برای یافتن این رابطه، ضریب X با استفاده از حداقل نمودن مربعات خطا، برآورد شود، در این صورت این روش، روش حداقل مربعات در رگرسیون خطی ساده نامیده می‌شود.

فرضهای آماری در رگرسیون خطی ساده عبارتند از:

الف) شرط وجودی: به این معنی که $\mu_{Y|X}$ و $\sigma^2_{Y|X}$ مقادیر متناهی داشته باشند.

ب) شرط استقلال: به این معنی که $Y_1, Y_2, \ldots, Y_n$ مستقل از یکدیگر باشند.

ج) شرط خطی بودن: به این معنی که رابطهٔ بین X و Y خطی باشد.

د) شرط همگونی واریانس: به این معنی که واریانس خطای مدل ثابت باشد.

ه) شرط نرمال بودن: به این معنی که Y_iها دارای توزیع نرمال باشند.

روش کار به صورت زیر است :

اگر زوج مرتب (x_i, y_i) $\quad i=1,2,\ldots,n$ از رابطهٔ خطی

$$\underset{\sim}{y}=X\underset{\sim}{\beta}+\underset{\sim}{\varepsilon} \qquad (۲-۱)$$

که در آن $\underset{\sim}{\beta}=\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}$ و $X=\begin{bmatrix}1 & X_1\\ \vdots & \\ 1 & X_n\end{bmatrix}$ و $\underset{\sim}{\varepsilon}=\begin{bmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_n\end{bmatrix}$ و $\underset{\sim}{y}=\begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix}$

آنگاه با می‌نیمم کردن تابع

$$Q(\beta)=\underset{\sim}{\varepsilon}^T\underset{\sim}{\varepsilon}=\left(\underset{\sim}{y}-X\underset{\sim}{\beta}\right)^T\left(\underset{\sim}{y}-X\underset{\sim}{\beta}\right) \qquad (۳-۱)$$

نسبت به $\underset{\sim}{\beta}$، برآورد حداقل مربعات برای $\underset{\sim}{\beta}$ به صورت زیر بدست می‌آید:

1- Simple Linear Regression 2- Least Squares(LS) Method 3- Weighted Linear Regression
4- M-Estimator

$$\hat{\beta}= \left(X^T X\right)^{-1} X^T Y$$

این برآورد دارای توزیع نرمال با میانگین β و ماتریس واریانس-کواریانس $\sigma^2 \left(X^T X\right)^{-1}$ است.

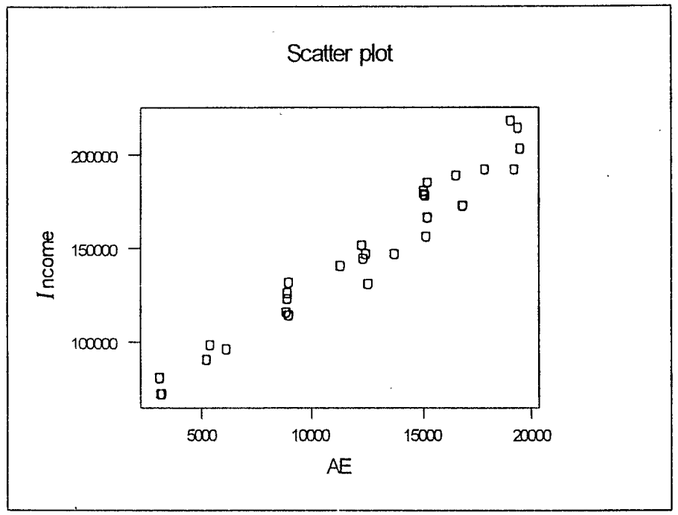
مثال ۱-۱: مدیر رستورانی معتقد است که درآمد ماهانه از فروش غذا و هزینهٔ تبلیغات برای معرفی آنها، رابطه‌ای مستقیم دارند. برای بررسی این مسأله اطلاعات مربوط به درآمد ماهانه حاصل از فروش غذای ۳۰ رستوران و نیز هزینه آگهی‌های سالانه را جمع آوری کرده است[۲۰]. داده‌های جمع آوری شده در جدول ۱-۱ آورده شده‌اند:

جدول (۱-۱) داده‌های مربوط به درآمد ماهانه و هزینه آگهی‌های سالانه مثال (۱-۱)[۲۰]

مشاهده	درآمد	هزینه‌های تبلیغاتی
۱	۸۱۴۶۴	۳۰۰۰
۲	۷۲۶۶۱	۳۱۵۰
۳	۷۲۳۳۴	۳۰۸۵
۴	۹۰۷۴۳	۵۲۲۵
۵	۹۸۵۸۸	۵۳۵۰
۶	۹۶۵۰۷	۶۰۹۰
۷	۱۲۶۵۷۴	۸۹۲۵
۸	۱۱۴۱۳۳	۹۰۱۵
۹	۱۱۵۸۱۴	۸۸۸۵
۱۰	۱۲۳۱۸۱	۸۹۵۰
۱۱	۱۳۱۴۳۴	۹۰۰۰
۱۲	۱۴۰۵۶۴	۱۱۳۴۵
۱۳	۱۵۱۳۵۲	۱۲۲۷۵
۱۴	۱۴۶۹۲۶	۱۲۴۰۰
۱۵	۱۳۰۹۶۳	۱۲۵۲۵

مشاهده	درآمد	هزینه‌های تبلیغاتی
۱۶	۱۴۴۶۳۰	۱۲۳۱۰
۱۷	۱۴۷۰۴۱	۱۳۷۰۰
۱۸	۱۷۹۰۲۱	۱۵۰۰۰
۱۹	۱۶۶۲۰۰	۱۵۱۷۵
۲۰	۱۸۰۷۳۲	۱۴۹۹۵
۲۱	۱۷۸۱۸۷	۱۵۰۵۰
۲۲	۱۸۵۳۰۴	۱۵۲۰۰
۲۳	۱۵۵۹۳۱	۱۵۱۵۰
۲۴	۱۷۲۵۷۹	۱۶۸۰۰
۲۵	۱۸۸۸۵۱	۱۶۵۰۰
۲۶	۱۹۲۴۲۴	۱۷۸۳۰
۲۷	۲۰۳۱۱۲	۱۹۵۰۰
۲۸	۱۹۲۴۸۲	۱۹۲۰۰
۲۹	۲۱۸۷۱۵	۱۹۰۰۰
۳۰	۲۱۴۳۱۷	۱۹۳۵۰

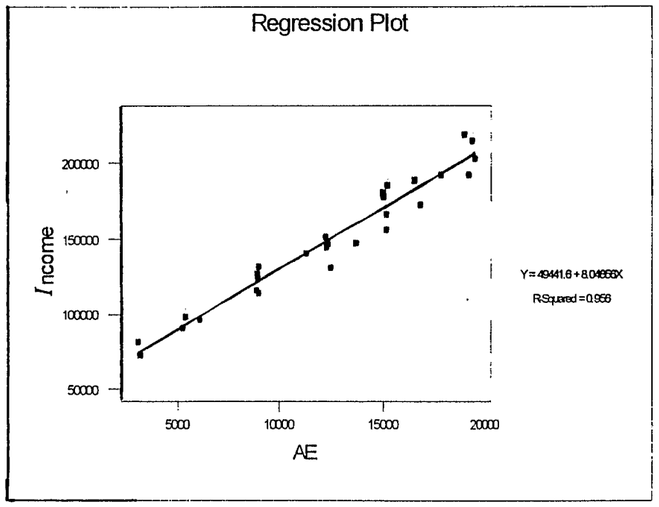
در این مثال درآمد ماهانه، متغیر پاسخ و هزینهٔ آگهی‌ها، متغیر مستقل است. قبل از هر چیز نمودار پراکنش این دو متغیر را رسم نموده و مشاهده می‌کنیم که رابطه‌ای مستقیم بین این دو وجود دارد.(شکل ۱ - ۱)



شکل ۱ - ۱) نمودار پراکنش درآمد در مقابل هزینهٔ آگهی، مثال (۱ - ۱)

با توجه به داده‌ها، مدل برازش شده عبارت است از: $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1 X \equiv Income = ۴۹۴۴۱/۶ + ۸/۰۵\, AE$

که خط برازش شده به داده‌ها در شکل (۱ - ۲)نمایش داده‌شده‌است.



شکل ۱ - ۲) نمودار خط برازش شده به داده‌ها، مثال (۱ - ۱)