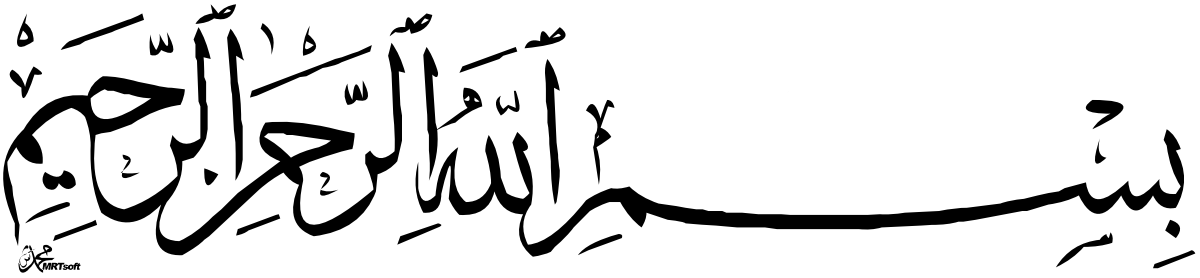
بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ
MRTsoft

دانشگاه پیام نور

دانشکده علوم پایه و کشاورزی

مرکز تهران

پایان نامه

برای دریافت درجه کارشناسی ارشد

در رشته آمار ریاضی

گروه علمی آمار

عنوان پایان نامه:

مقایسه خطای نوع اول برخی آزمونهای ناپارامتری روی ضرائب مدلهای رگرسیونی چندگانه

طاهره محمدی

استاد راهنما: جناب آقای دکتر علی شادرخ

استادمشاور: جناب آقای دکترمسعود یارمحمدی

شهریور ۱۳۹۰

اینجانب طاهره محمدی دانشجوی ورودی سال ۸۶ مقطع کارشناسی ارشد رشته آمار ریاضی گواهی می‌نمایم چنانچه در پایان نامه خود از فکر، ایده 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 ان نباشد بر عهده خویش می دانم و جوابگوی آن خواهم بود.

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

نام و نام خانوادگی: طاهره محمدی

تاریخ و امضاء: ۱۳۹۰/۶/۱۳

اینجانب طاهره محمدی دانشجوی ورودی سال ۸۶ مقطع کارشناسی ارشد رشته آمار ریاضی گواهی می‌نمایم چنانچه براساس 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 : طاهره محمدی

تاریخ و امضاء: ۱۳۹۰/۶/۱۳

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

شهریور ۱۳۹۰

تقدیم به

امید چشمان منتظر

و

همسر عزیزم

تشکر و قدر دانی:

با تشکر از استاد گرانقدرجناب آقای دکتر شادرخ که بنده حقیر را با راهنمایهای ارزنده خود در تحریر این رساله یاری فرمودند وهمچنین جناب آقای دکتر یار محمدی که نکات مفیدی را به بنده یادآوری نمودند. جا دارد از جناب آقای دکتر احتشامی همسرم که با صبر و حوصله فراوان در این امر، مرا یاری نمودند نیز تشکر بنمایم. امید است این رساله چراغ راهی برای محققان شود.

چکیده

روش های ناپارامتری متفاوتی برای انجام آزمون فرض روی ضرائب رگرسیونی ارائه شده است.در این پایان نامه ابتدا مهمترین آنها که از نتیجه مقایسات سایر آماردانان بدست آمده است را معرفی می کنیم. سپس در جستجوی روشی هستیم که از بقیه بهتر عمل کند. در این راستا هو و جون(۲۰۰۱) روش کندی را که در سال ۱۹۹۵ ارائه شده بود اصلاح نموده و با شبیه سازی نشان دادند که روش کندی اصلاح شده که ما آن را به نام روش هو – جون می نامیم، دارای خطای نوع اول کمتری است. از طرف دیگر شادرخ (۲۰۰۵) بطوری تئوری و شبیه سازی نشان داد که روش فریدمن – لان که یک روش نسبتا بهتری نسبت به روش کندی است، دارای عملکرد بهتری می باشد. ما در این پایان نامه اصلاحاتی را که هو–جون روی روش کندی انجام دادند روی روش فریدمن–لان انجام دادیم و این روش اصلاح شده را با خود روش فریدمن– لان و هو–جون مقایسه کردیم و نتیجه گرفتیم که با افزایش حجم نمونه روش فریدمن– لان اصلاح شده دارای خطای نوع اول کمترو توان بالاتری نسبت به روش فریدمن– لان و هو– جون است.

واژگان کلیدی: آزمونهای جایگشتی، خطای نوع اول جایگشتی، تعویض پذیری، ضریب رگرسیونی

فهرست **صفحه**

مقدمه ۱

فصل اول :رگرسیون خطی ۲

مقدمه ۳

۱-۱)ضریب همبستگی ٤

۱-۲)معرفی رگرسیون خطی ساده ٥

۱-۳) رگرسیون خطی چندگانه ۹

۱-٤)بررسی مانده ها ۱۳

۱-٥)روشهای ناپارامتری ۱٤

فصل دوم:معرفی آزمون های جایگشتی ۱٦

مقدمه ۱۷

۲-۱)معرفی روش بوت استراپ و جک نایف ۱۷

۲-۲)روشهای انجام آزمون های جایگشتی ۱۹

۲-۳)آزمون های جایگشتی روی ضریب رگرسیون در مدل خطی ساده ۲۱

۲-٤)آزمون روی یک ضریب رگرسیونی در مدل خطی چندگانه ۲۲

۲-٥)روش فریدمن- لان در حالت چندگانه ۲۷

فصل سوم: شبیه سازی ۳٤

مقدمه ۳٥

۳-۱) روش شبیه سازی	۳۵
۳-۲) نتایج شبیه سازی	۱۰۸
نتیجه گیری کلی	۱۱۲
پیوست ۱)برنامه های شبیه سازی	۱۱۳
پیوست ۲) دستور گرام-اشمیت	۱۲۰
پیوست ۳)تجزیه طیفی	۱۲۰
منابع	۱۲۲
واژه نامه	۱۲۳

مقدمه

درحالتی که فرضهای زیربنائی برای انجام آزمون فرض روی ضرائب مدلهای رگرسیونی خطی برقرار نباشد، انجام آزمونهای کلاسیک روی ضرائب رگرسیونی پاسخی مطمئن به ما نخواهد داد و لذا می‌بایست از روشهای ناپارامتری استفاده کرد. یک گروه از این آزمونها، آزمونهای دوباره نمونه‌گیری هستند که آزمونهای جایگشتی جزئی از این گروه می‌باشند که هدف این پایان نامه مقایسه برخی روشهای آزمونهای جایگشتی با روشهای اصلاح شده آنها می‌باشد.

آزمونهای جایگشتی در حدود سال ۱۹۳۰ توسط فیشرمعرفی شد. اما به علت نیاز به محاسبات زیاد چندان مورد توجه قرار نگرفت. تا اینکه کامپیوترهای با سرعت بالا در دسترس قرار گرفت، با توجه به اینکه در اکثر مواقع فرض‌های زیربنائی روی داده‌های حقیقی برقرار نبودند لذا آزمونهای جایگشتی جایگاه ویژه‌ای یافته و مورد توجه محققین رشته‌های اقتصاد، روانشناسی، علوم آزمایشگاهی و...قرار گرفت.

در فصل اول این پایان نامه مدل رگرسیون خطی معرفی شده و در رابطه با خطا و آزمونهای مربوط به آن می پردازد، در فصل دوم آزمونهای متفاوت جایگشتی و نیز روشهای اصلاح شده آنها را معرفی می‌کنیم، در فصل سوم به شیوه شبیه‌سازی روشهای معرفی شده را از دیدگاه خطای نوع اول مورد مقایسه قرار می‌دهیم.

فصل اول

رگرسیون خطی

مقدمه

واژه رگرسیون[۱] را نخستین بار فرانسیس گالتن[۲] (۱۹۱۱-۱۸۲۲) دانشمند انگلیسی در کتاب وراثت خود بکار برد (بهبودیان،جواد،۱۳۸۳،تهران،پیام نور).در این فصل بنا داریم به طور مختصر رگرسیون خطی ساده و چندگانه[۳] را بیان و به بررسی مانده‌های[٤] حاصل از برازش خط رگرسیون بپردازیم و توضیح بسیار کوتاهی در مورد آزمونهای ناپارامتری دهیم. آنهایی که در زمینه علوم زیستی یا تجربی یا اجتماعی مطالعه و پژوهش می‌کنند معمولا با دو نوع متغیر برخورد دارند یک نوع آن اغلب قابل کنترل و نوع دیگر تصادفی می‌باشد مثلا سن نوزاد (برحسب روز) یک نوع متغیر قابل کنترل ولی وزن او (برحسب گرم) متغیر تصادفی می‌باشد. کشف ارتباط و مدل سازی میان این گونه متغیرها، به منظور پیش بینی مقدار متغیر تصادفی وابسته و مطالعه‌ی عوامل مؤثر در مقدار آن، برای پژوهشهای علمی ضرورت دارد. البته ممکن است حالتهای پیش بیاید که هر دو متغیر مستقل[٥] و وابسته[٦] متغیرهای تصادفی باشند که ما در این پایان نامه بدان نمی‌پردازیم، همچنین ممکن است متغیر وابسته متغیر کیفیت[۷] باشد که منجر به رگرسیون لوژستیک[۸] می‌گردد که پرداختن به آن خارج از بحث مربوط به این رساله می باشد.

---

[۱] Regression
[۲] Francis Galton
[۳] Multi linear regression
[٤] Residual
[٥] Independent
[٦] Dependent
[۷] Qualitative variable
[۸] Logistic regression

۱-۱) ضریب همبستگی

کوواریانس[۱] بین X و Y معیاری عددی است که برای اندازه گیری تغییرات توأم دو متغیر تصادفی بکار می‌رود و به صورت زیر تعریف می‌شود:

$$Cov(X,Y)=E[(X-\mu_X)(Y-\mu_Y)] =E(XY) - \mu_X\mu_Y \qquad (۱-۱)$$

بطور شهودی می‌توان گفت که اگر احتمال رخ دادن مقادیر بزرگ X همراه با مقادیر بزرگ Y و مقادیر کوچک X همراه با مقادیر کوچک Y زیاد باشد،X و Y در یک جهت تغییر می‌کنند،در چنین وضعیتی هر دو مقدار انحراف $(X-\mu_X)$ و $(Y-\mu_Y)$ با احتمال زیادی مثبت یا منفی خواهند بود، به طوری که حاصلضرب $(X-\mu_X)(Y-\mu_Y)$ به احتمال قوی مثبت است.در نتیجه،امید ریاضی حاصلضرب مثبت است.از سوی دیگر،اگرX و Y گرایش به تغییر در جهت عکس هم داشته باشند،مقادیر مثبت $(X-\mu_X)$ اغلب متناظر با مقادیر منفی $(Y-\mu_Y)$ می‌شوند و بعکس. در نتیجه حاصلضرب به احتمال قوی منفی است و امید ریاضی منفی خواهد بود. با توجه به مطالب مذکور علامت و مقدار کوواریانس نشان دهنده جهت و میزان بستگی بین X و Y است. این معیار به واحدهای اندازه گیری X و Y بستگی دارد برای داشتن معیاری که رابطه بین دو متغیر را اندازه گیری می‌کند که بستگی به واحدهای متغیرها نداشته باشد، کوواریانس را بر انحراف معیار دو متغیر تقسیم می‌کنیم که این معیار را ضریب همبستگی بین X و Y می‌نامیم و بصورت ذیل محاسبه می نمائیم:

$$\boldsymbol{\rho = \rho(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}}=\frac{\sigma_X}{\sigma_X\ \sigma_Y} \qquad (۲-۱)$$

که برآورد کننده گشتاوری آن به صورت ذیل است:

$$R=S_{XY}/S_XS_Y \qquad (۳-۱)$$

$$S_{XY} = 1/n\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) = \overline{XY} - \bar{X}.\bar{Y}$$

$$S_X^2 = 1/n\sum_{i=1}^{n}(X_i-\bar{X})^2 = \overline{X^2} - \bar{X}^2$$

$$S_Y^2 = 1/n\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \overline{Y^2} - \bar{Y}^2$$

[۱] Covariance

شایان ذکر است که محاسبه ضریب همبستگی بین دو متغیر با توجه به نوع متغیرمتفاوت است انواع دیگری از ضریب همبستگی هم موجود است که با توجه به خلاصه بودن این مبحث از آوردن آنها خودداری می‌نمائیم.

۱-۲)معرفی رگرسیون خطی ساده

هر گاه بتوانیم به کمک یک تابع، از روی مقادیر یک متغیر ، مقدار یک متغیر تصادفی دیگر را پیش بینی کنیم می‌گوییم یک تابع رگرسیون یا یک تابع برگشت داریم. به کمک نمودار پراکنش[۱] وجود رابطه (خطی)و پراکندگی داده ها بطور قابل تشخیص است. در صورتی که وجود رابطه خطی بین دو متغیر از روی نمودار پراکنش مشاهده شود می‌توان خط رگرسیون زیر را به داده ها برازش داد.

$$Y|x_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (۱-٤)$$

که درفرمول فوق متغیر قابل کنترل را با مقدار ثابت x نشان می‌دهیم و آنرا متغیر مستقل[۲] می‌نامیم. متغیر دوم را که به متغیر x بستگی دارد و مقدار آن تصادفی می‌باشد با متغیر تصادفی $(Y|x)$ نشان می‌دهیم و آنرا متغیر وابسته[۳] می‌گوئیم. برای اینکه کلمه مستقل با مفهوم احتمالی آن اشتباه نشود، گاهی x را متغیر کنترل شده یا پیش بینی کننده یا بازگشت دهنده[٤] می‌گویند، متغیر دوم $(Y|x)$ راهم اغلب متغیر پاسخ[۵] می‌نامند.

یک متغیر تصادفی دیگر که با ε نشان داده می‌شود و آنرا خطای تصادفی یا به سادگی خطا[٦] می‌نامیم. فرضهای زیر بنایی مدل رگرسیونی ذکر شده به صورت ذیل می‌باشد:

۱) ε_i ها ناهمبسته باشند.

۲) ε_i ها دارای میانگین صفر و واریانس ثابت σ^2 هستند.

در صورتی که ε_i ها دارای توزیع نرمال با میانگین صفر و واریانس ثابت σ^2 باشند مدل (۱-٤) را یک مدل رگرسیونی نرمال (گاوس[۱]) می‌نامند.

---

[۱] Scatter plot
[۲] Independent
[۳] Dependent
[٤] Controlled or predictor or reqressor
[۵] Response
[٦] Error

۱-۲-۱)برآورد پارامترهای رگرسیون خطی ساده

برای برآورد کردن خط رگرسیونی لازم است پارامترهای آنرا با استفاده از داده‌های نمونه تصادفی $(x_1 و y_1),(x_2,y_2),\ldots,(x_n و y_n)$ برآورد کنیم.برای انجام این کار دو روش وجود دارد.یکی روش حداقل مربعات[۲] (LS) و دیگری روش ماکسیمم درستنمائی[۳] است.روش حداقل مربعات را نخستین بار ریاضیدان برجسته آلمانی گاوس در محاسبات خود استفاده کرده است. به روایتی هم لژندار[۴] ریاضیدان فرانسوی درباره این روش سه سال قبل از گاوس مقاله‌ای منتشر کرده است. به طور کلی در روش LS سعی می‌کنیم با مینییم کردن مجموع توانهای دوم خطا، پارامترهای مجهول را برآورد کنیم. دلیل مینیمم کردن مجموع یاد شده این است که می‌خواهیم یک خط مستقیم مناسب بر داده های دوتائی برازش دهیم. در اینجا از ذکر محاسبات در بدست آوردن پارامترها با این روش صرف نظر کرده و مستقیما برآوردها را بصورت ذیل بیان می‌کنیم.

$$\widehat{\beta_0}=\bar{Y}-\hat{\beta}_1\bar{x} \qquad \hat{\beta}_1=\frac{S_{xY}}{S_x^2}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad (۵-۱)$$

در نتیجه برآورد خط رگرسیون ساده به صورت زیر می شود:

$$\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i \qquad (۶-۱)$$

می‌دانیم که برآوردهای فوق برآوردهای نااریبی برای پارامترهای خط رگرسیونی هستند. به منظور برآورد واریانس‌ها از مشاهده مانده‌ها استفاده می‌کنیم که تفاضل بین دو مقدار مشاهده شده y_I وپیش بینی شده $\hat{y}_i$ بر اساس خط رگرسیونی می‌باشد این برآورد به صورت زیر است:

$$\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2=\frac{n}{n-2}(S_Y^2-\hat{\beta}_1^2 s_x^2) \qquad (۷-۱)$$

که در آن $S_Y^2=\sum(y_i-\bar{y})^2$

درواقع تعبیر هندسی روش LS بدین صورت است که خط برازش داده شده به داده‌ها به گونه‌ای از میان آنها عبور می‌کند که خطاها دارای کمترین مقدار شود.

---

۱ Gauss
۲ Least sqares Method
۳ Maximm Liklihood Estimator
۴ Legendr

۱-۲-۲) توزیع برآوردها و آماره های آزمون

با توجه به فرض‌های که درباره مدل ε_i ها داشتیم و Y_i ها هم مستقل و دارای توزیع نرمال هستند داریم:

$$Y_i \sim N(\beta_0+\beta_1 x_i,\sigma^2) \qquad (۸-۱)$$

اگر تابع درستنمائی[۱] ،یعنی چگالی توأم را برای $Y_1,Y_2,\ldots,Y_n$ را تشکیل دهیم و ماکزیمم کنیم، برآوردهای بدست آمده در رابطه مذکور برآوردهای درستنمائی ماکزیمم هستند که هر سه با برآوردهای که به روش حداقل مربعات خطا بدست آمده یکی هستند ونیز این برآوردها دو به دو مستقل‌اند و با توجه به اینکه $\widehat{\beta_0}$ و $\widehat{\beta_1}$ ترکیب خطی از Y_iهاکه نرمال هستند پس آنها دارای توزیع نرمال می‌باشند و داریم:

$$\begin{cases} \hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{n s_x^2}) \\ \hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{n}(1+\frac{\bar{x}^2}{s_x^2})) \\ \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim x^2_{(n-2)} \end{cases} \qquad (۹-۱)$$

با توجه به توزیع‌های فوق و تعریف توزیع t–استودنت[۲] آماره‌های ذیل دارای توزیع t–استودنت با n-2 درجه آزادی[۳] می‌باشند.

$$T_1=\frac{\sqrt{n}(\widehat{\beta_1}-\beta_1)}{\frac{\hat{\sigma}}{s_x}}=T_{(n-2)} \qquad (۱۰-۱)$$

$$T_2=\frac{\sqrt{n}(\widehat{\beta_0}-\beta_0)}{\hat{\sigma}\sqrt{1+\frac{\bar{x}^2}{s_x^2}}}=T_{(n-2)} \qquad (۱۱-۱)$$

بنابراین با توجه به نکات فوق می‌توان به راحتی در مورد پارامترهای مدل خطی ساده آزمونهای آماری ذیل را انجام دهیم و یا برای آنها فواصل اطمینان بدست آوریم.

---

[۱] Likelihood function
[۲] Students t distribution
[۳] Degrees of freedom

$$\begin{cases} H_0: \beta_0=0 \\ H_1: \beta_0 \neq 0 \end{cases} \qquad \begin{cases} H_0: \beta_1 =0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

۱-۲-۳)تحلیل واریانس

در این بخش به اختصار تحلیل واریانس برای رگرسیون خطی ساده را بیان می‌کنیم.برای اینکه مدل رگرسیون خطی ساده را با پارامترهای کامل و با دو مدل که هر کدام از حذف یکی از پارامترهای β_0, β_1 بدست آمده‌اند ،مقایسه کنیم از تحلیل واریانس استفاده می‌نمائیم، با فرض اینکه مدل کامل است جمله زیر را به دو قسمت تفکیک می‌نمائیم.

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2=\sum_{i=1}^{n}\left[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})\right]^2=\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2+\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \qquad (۱-۱۲)$$

که جملات را با نمادهای زیر نشان می‌دهیم:

$$\text{SST}=\sum_{i=1}^{n}(Y_i - \bar{Y})^2= \text{n}S_Y^2 \qquad (۱-۱۳)$$

تغییر پذیری نسبت به معدل را می‌سنجد درجه آزادی آن (n-1) در مدل شامل هر دو پارامتر تغییرپذیری خطا با (n-2) درجه آزادی است.

$$\text{SSE}=\sum_{i=1}^{n}\hat{E}_i^2 = (\text{n}-2)\,\hat{\sigma}^2 \qquad (۱-۱۴)$$

و جمله دیگر را تغییر پذیری رگرسیون با ۱ درجه آزادی می‌نامیم هر چقدر نسبت تغییر پذیری رگرسیون نسبت به خطا بیشتر باشد برازش داده‌ها با این خط رگرسیونی موفقیت آمیزتر بوده است.

$$\text{SSR}=\text{n}\hat{\beta}_1^2 s_x^2 \qquad (۱-۱۵)$$

نسبت SSE به واریانس و نسبت SSR به واریانس دارای توزیع کی دو است و با توجه به رابطه توزیع F–فیشر[۱] و کی دو[۲]، به منظور آزمون اینکه $\beta_1=0$ از آماره زیر استفاده می‌نمائیم:

$$\text{MSE}=\frac{SSE}{df_\varepsilon} \qquad \text{MSR}=\frac{SSR}{df_R} \qquad \text{F}=\frac{MSR}{M\,SE} \qquad (۱-۱۶)$$

[۱] F-Fisher distribution
[۲] Chi-square distribution

ناحیه رد(بحرانی[۱])H_0 به صورت ذیل می‌باشد.

$$C.R_{H_0}:F>F_{(1-\alpha)}(1,n-2) \qquad (۱۷-۱)$$

محاسبه نسبت SSR به SST نیز برای درک این مطلب که Y را می‌توان به کمک خط رگرسیون خطی پیش بینی کرد لازم است، این نسبت را ضریب تعیین[۲] می‌نامند.

$$r^2 = \frac{SSR}{SST}=1-\frac{SSE}{SST} \qquad (۱۸-۱)$$

در واقع استراتژی کلی به صورت ذیل است که:

۱.اجرای مدل کامل و بدست آوردن SSE

۲.اجرای مدل کاهش یافته بر اساسH_0 وبدست آوردن SSE

۳.محاسبه آماره آزمون

٤.تصمیم گیری

(بزرگ نیا،آذرنوش،۱۳۸۲،۳۹۳الی٤۳۲)

۱-۳) رگرسیون خطی چندگانه[۳]

در دو بخش قبل در مورد اینکه یک متغیر وابسته تنها به یک متغیر کنترل شده به صورت خطی ارتباط دارد مطالبی را ارائه کرده‌ایم،اینک می‌خواهیم متغیر وابسته‌ای را مورد بررسی قرار دهیم که به صورت خطی به چندین متغیر کنترل شده ارتباط دارد. به عبارتی دیگر می‌خواهیم در مورد رگرسیون خطی چند متغیره یا رگرسیون چند گانه به اختصار صحبت کنیم.

تعریف مدل خطی:فرض کنید $x_1,x_2,\ldots,x_k$ متغیرهای کنترل شده و متغیر تصادفی شرطی $(Y|x_1,x_2,\ldots,x_k)$ متغیر وابسته باشد همچنین متغیرهای کنترل شده بدون خطای اندازه گیری و متغیر وابسته دارای خطای اندازه گیری و خطای تصادفی که روی هم با ε نشان می‌دهیم باشند رابطه

[۱] Critical level
[۲] Cofficient of determination
[۳] Multiple linear regression

زیر یک مدل خطی kگانه با پارامترهای $\beta_1,\beta_2,\ldots,\beta_k$ را نشـان مـی‌دهـد. ε دارای میـانگین صـفر و واریانس ثابت σ^2 می‌باشدکه اگر نرمال باشد یک مدل نرمال داریم.

$$\{Y|X_1,X_2,\ldots,X_n\}=\beta_1X_1+\beta_2X_2+\ldots+\beta_kX_k+\varepsilon \qquad (۱۹-۱)$$

اگر n بار آزمایش را انجام دهیم داده‌ها را می‌توان صورت ماتریسی ذیل نمایش داد با این فـرض کـه آزمایشها مستقلا انجام می‌گیرد ε_iها مستقل و همتوزیع‌اند .

$$Y=\begin{pmatrix}Y_1\\ \vdots\\ Y_n\end{pmatrix}\quad \beta=\begin{pmatrix}\beta_1\\ \vdots\\ \beta_n\end{pmatrix}\quad \varepsilon=\begin{pmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_k\end{pmatrix} \qquad \mathrm{X}=\begin{pmatrix}x_{11} & \ldots & x_{1k}\\ \vdots & \ldots & \vdots\\ x_{n1} & \ldots & x_{nk}\end{pmatrix}$$

اینک می‌توان n رابطه ی(۱-۱۹) را که بصورت ماتریسی بیان شد، بطور خلاصه ذیل نشان دهیم:

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \qquad \mathrm{E}(\boldsymbol{\varepsilon})=0\,,\ \sum_{\varepsilon}=\sigma^2I \qquad (۲۱-۱)$$

جهت برآورد پارامترهای مدل با روش مجموع توانهای دوم خطـا و اسـتفاده از معـادلات نرمـال[۱] بـه شرط اینکه ماتریس $\acute{X}X$ وارون پذیر باشد(عدم وجود هم خطی) داریم:

$$\left(\acute{X}X\right)^{-1}\acute{X}Y=\ \hat{\beta}=\begin{pmatrix}\hat{\beta}_1\\ \vdots\\ \hat{\beta}_k\end{pmatrix} \qquad (۲۲-۱)$$

با توجه به فرضیات مدل، ماتریس کوواریانس $\hat{\beta}$ به صورت زیر است:

$$\sum_{\hat{\beta}}=\sigma^2(\acute{X}X)^{-1} \qquad (۲۳-۱)$$

اما در رابطه (۱-۲۳) مقدارσ^2 مجهول است یک برآورد نااریب برای آن به صورت زیر است:

$$\hat{\sigma}^2=\frac{||Y-X\hat{\beta}||^2}{n-k} \qquad (۲٤-۱)$$

و بردار میانگین و برآورد آنرا بصورت زیر نشان میدهیم:

$$\mu=E(Y)=X\beta \qquad \hat{\mu}=X\hat{\beta}=X(\acute{X}X)^{-1}\acute{X}Y \qquad (۲۵-۱)$$

اگر رابطه (۱-۲۱)را، مدل (۱) در نظر بگیریم. مدل زیر را با عنوان(۲) معرفی می‌کنیم.

[۱] Normal equations

$$Y=X\beta+\varepsilon \quad , \quad \varepsilon\sim N(0,\sigma^2 I) \qquad (۲۶-۱)$$

برآورد پارامترها در مدل(۲) همانند مدل (۱) می‌باشد با ذکر این نکته که برآوردها دارای توزیع‌های به صورت زیر می‌باشند:

$$\begin{cases} \hat{\beta}\sim N(\beta,\sigma^2(\acute{X}X)^{-1}) \\ \widehat{\mu}\sim N(\mu,\sigma^2 X(\acute{X}X)^{-1}\acute{X}) \\ \frac{(n-k)\hat{\sigma}^2}{\sigma^2}\overset{d}{\Rightarrow}x^2_{(n-k)} \end{cases} \qquad (۲۷-۱)$$

(بزرگ نیا،آذرنوش،۱۳۸۲،۱۶۹الی۱۷٤)

۱-۳-۱) استنباط آماری در مدلهای خطی چند متغیره

مدل خطی(۲) و ماتریس M راکه h×k و پر رتبه سطری[۱] (سطرها مستقل باشند) می‌باشد در نظر می‌گیریم تحت فرض آماری H_0:Mβ=0 که آنرا مدل (۳) در نظر می‌گیریم. برآورد پارامترهای مدل با استفاده از روش نسبت درستنمائی تعمیم یافته[۲] به صورت ذیل است:

$$\hat{\hat{\sigma}}^2=\frac{||Y-\widehat{\widehat{\mu}}||^2}{n-k-h} \quad ,\widehat{\widehat{\mu}}=proj^{Y}_{W_\circ} \qquad (۲۸-۱)$$

که در آن برآورد میانگین تصویر بردار Y روی فضای ایجاد شده توسط سطرهای ماتریس M است و طبق قضیه تصویری بردار میانگین یکتاست.

۱-۳-۲) تحلیل وراینس

تحلیل واریانس در مدل چندمتغیره به همان صورت مدل خطی ساده صورت می‌گیرد ، فقط به جای اسکالرها از معادل برداری یا ماتریسی آنها استفاده می‌شود که بصورت زیر بیان می‌نمائیم.

الف)مجموع توانهای دوم خطا در مدل(۲)

$$SSE=||Y-\widehat{\mu}||^2=||Y||^2-\widehat{||\mu||^2} \qquad (۲۹-۱)$$

[۱] Full row rank
[۲] Generalized likelihood ratio

ب)مجموع توانهای دوم خطا در مدل (۳)تحت فرض H_0

$$SST=||Y-\hat{\mu}||^2=||Y||^2-||\hat{\mu}||^2 \qquad (۳۰-۱)$$

ج)مجموع توانهای دوم رگرسیونی

$$SSR=||\hat{\mu}-\hat{\mu}||^2=||\hat{\mu}||^2-||\hat{\mu}||^2 \qquad (۳۱-۱)$$

با استفاده از فرم کانونی[۱] Y،تحت فرض صفر داریم:

$$SSE=\sum_{i=k+1}^{n} Z_i^2 \overset{d}{\Rightarrow} \sigma^2 x^2_{(n-k)} \qquad (۳۲-۱)$$

$$SST=\sum_{i=k-h+1}^{n} Z_i^2 \overset{d}{\Rightarrow} \sigma^2 x^2_{(n-k+h)} \qquad (۳۳-۱)$$

$$SSR=\sum_{i=k-h+1}^{k} Z_i^2 \overset{d}{\Rightarrow} \sigma^2 x^2_{(h)} \qquad (۳۴-۱)$$

ملاحظه می‌شود با تعاریف فوق SSRوSSE همواره مستقل هستند. بدین ترتیب برآورد ناااریب σ^2 در دو مدل (2)و (3) تحت فرض صفر به صورت زیر است:

I)در مدل (2): MSE=$\hat{\sigma}^2=\frac{SSE}{n-k}$ دارای توزیع x^2 با n-k درجه آزادی

II) در مدل (3): MST=$\hat{\sigma}^2=\frac{SST}{n-k+h}$ دارای توزیع x^2 با n-k+h درجه آزادی

باز هم مانند بخش (۱-۲-۳) جدول تحلیل واریانس را تشکیل داده و با توجه به آماره F تعریف شده در رابطه(۱-۱۶) در مورد فرض صفر تصمیم می‌گیریم. در مورد زمانی که بردار X پررتبه ستونی[۲] نباشد معادلات نرمال باز هم دارای پاسخ هستند،زیرا به هرحال تصویر y روی فضای ستونی[۳] X همواره وجود دارد .اما پاسخ آن یکتا نیست.این موضوع منجر به بحث وارون گسترده[۴] می‌گردد که از حوصله این بخش خارج است.

(بهبودیان،جواد،۱۳۸۳،رگرسیون،تهران،پیام نور)

[۱] Canonical form
[۲] Full column rank
[۳] column space
[۴] Generalized inverse