بسم الله الرحمن الرحیم

دانشکده علوم پایه و کشاورزی

مرکز تهران

پایان‌نامه

برای دریافت مدرک کارشناسی ارشد

رشته آمار ریاضی

گروه آمار

موضوع:

همبستگی متعارف استوار

محمد جعفری

استاد راهنما:

جناب آقای دکتر مسعود یارمحمدی

استاد مشاور:

جناب آقای دکتر پرویز نصیری

مهر ۱۳۹۰

شماره ....
تاریخ ....
پیوست ....

دانشگاه پیام نور
دانشگاه پیام نور استان تهران
اللهم عجل لولیک الفرج والعافیه والنصر

جمهوری اسلامی ایران
وزارت علوم تحقیقات و فناوری

مجتمع علوم پایه و کشاورزی

صورتجلسه دفاع از پایان نامه کارشناسی ارشد

جلسه دفاع ازپایان نامه کارشناسی ارشد آقای محمد جعفری
دانشجوی رشته آمار ریاضی به شماره دانشجویی: ۸۸۰۲۷۰۹۸۲
تحت عنوان:

"همبستگي متعارف استوار"

جلسه دفاع با حضور داوران نامبرده ذیل در روزیکشنبه مورخ: ۹۰/۰۷/۲۴ ساعت : ۱۰/۳۰-۹/۳۰ درمحل مجتمع علوم پایه و کشاورزی برگزار شد. و پس از بررسی پایان‌نامه مذکور با نمره به عدد...۱۸/۷۵... به حروف ...هیجده و هفتاد و پنج صدم... و با درجه ارزشیابی...عالی...مورد قبول واقع شد ☒ نشد ☐

ردیف	نام و نام خانوادگی	هیات داوران	مرتبه دانشگاهي	دانشگاه/ موسسه	امضاء
۱	دکترمسعود یارمحمدي	استاد راهنما	استادیار	پیام نور	
۲	دکترپرویز نصیري	استاد مشاور	"	"	
۳	دکترزهرارضایي قهرودي	استاد داور	"	پژوهشکده آمار	
۴	دکترعلي اصغر شکري	نماینده علمي گروه	استادیار	پیام نور	
۵	دکترعلي اصغر شکري	نماینده تحصیلات تکمیلي	"	"	

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 **محمد جعفری** دانشجوی ورودی سال **۸۹–۸۸** مقطع کارشناسی ارشد رشته **آمارریاضی** گواهی می‌نمایم چنانچه در پایان‌نامه خود از فکر، ایده و نوشته دیگری بهره گرفته‌ام با نقل قول مستقیم یا غیرمستقیم 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 پایان‌نامه نتیجه تحقیقات خودش می‌باشد و در صورت استفاده از نتایج دیگران مرجع آن را ذکر نموده است.

نام و نام خانوادگی دانشجو

تاریخ و امضاء

۱۳۹۰/۷/۲۴

اینجانب **محمد جعفری** دانشجوی ورودی سال **۸۹–۸۸** مقطع کارشناسی ارشد رشته **آمارریاضی** گواهی می‌نمایم چنانچه براساس پایان‌نامه 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 دانشجو

تاریخ و امضاء

۱۳۹۰/۷/۲۴

مهر ۱۳۹۰

تقدیم به

همسر عزیزم که با همراهی و صبر بی‌دریغ در امر تحصیل مرا یاری نمــوده و مشوق من در ادامه تحصیل بودند؛

و تقدیم به

فرزند دلبندم، محمدطه، که با حضورش در زندگی‌ام باعث دلگرمی و امید به آینده شده است.

با تشکر از

پدر و مادر عزیزم که در طی کردن پله‌های ترقی و پیشرفت مرا یاری نموده و در این راه از هیچ کمکی دریغ نفرمودند. با دلی پر از عشق و محبت در برابر وجود گرامیتان زانو برمی‌نهم و بر دستان گرم و پرمحبت‌تان بوسه می‌زنم؛

و با تشکر از

استاد ارجمند و گرامیم، جناب آقای دکتر مسعود یارمحمدی، که با محبـــت و راهنمائی‌شان مرا در تکمیل این پروژه یاری نمودند.

چکیده

برای تحلیل‌های همبستگی متعارف استوار، چندین روش ارائه و بحث شده است. اولین روش براساس تعریف تحلیل‌های همبستگی متعارف می‌باشد که به دنبال یافتن ترکیبات خطی دو مجموعه از متغیرها می‌باشد که دارای ضریب همبستگی ماکسیمم (استوار) می‌باشند. دومین روش، روش تعقیب تصویر می‌باشد. سه روش مختلف که بر اساس تعقیب تصویر می‌باشند، با جزئیات مورد بررسی قرار می‌گیرند. آخرین روش، روش رگرسیون متناوب استوار می‌باشد. یک مطالعه شبیه‌سازی، انجام برآوردگرهای مختلف تحت چند نوع طرح نمونه‌گیری را مقایسه می‌کند. میزان استواری با استفاده از نمودار فروریزش قابل بررسی است.

واژگان کلیدی: همبستگی متعارف، همبستگی متعارف استوار، تعقیب تصویر، برآورد کوواریانس استوار، رگرسیون متناوب، رگرسیون استوار

فهرست مطالب

فصل چهارم

فصل پنجم

فهرست اشکال

فصل اول

همبستگی متعارف

۱–۱ مقدمه

یکی از تعاریف اساسی در علم آمار تعریف همبستگی و رابطه بین دو متغیر می‌باشد. به طور کلی شدت وابستگی دو متغیر به یکدیگر را همبستگی تعریف کرده و ممکن است علاوه بر شدت همبستگی، جهت همبستگی نیز مورد نیاز پژوهشگر باشد.

در آمار انواع زیادی از ضرایب همبستگی متفاوت وجود دارند که هر کدام همبستگی بین دو متغیر را با توجه به نوع داده‌ها و شرایط متغیرها اندازه‌گیری می‌کنند. لذا با توجه به اهمیت این موضوع که چه ضریب همبستگی را در چه زمانی مورد استفاده قرار دهیم، در اینجا به تعریف انواع همبستگی می‌پردازیم. البته بیشتر هدف ما، بررسی همبستگی متعارف بوده و به انواع دیگر همبستگی، به طور کلی اشاره شده است.

نکات قابل توجهی در ضریب همبستگی وجود دارد که ذکر آن‌ها مهم است. ضریب همبستگی یک رابطه متقارن بوده و هر چه مقدار آن به یک نزدیک‌تر باشد، میزان وابستگی دو متغیر بیشتر است. اما توجه داشته باشید که این وابستگی به معنای رابطه علت و معلولی نبوده و ضریب همبستگی حرفی از این که کدام علت و کدام معلول است، به میان نمی‌آورد. اما اگر متغیرهای دیگری نیز بر روی متغیر وابسته تأثیر داشته باشند، آن‌گاه ممکن است هر کوواریانسی که با متغیر مستقل به اشتراک گذاشته‌اند، تأثیر غلطی را بر ضریب همبستگی با متغیر مستقل داشته باشند. همچنین برای تعمیم ضریب همبستگی می‌توان وجود رابطه غیرخطی بین دو متغیر همبسته را در حالی که ضریب همبستگی به غلط آن را نشان ندهد، بررسی نمود. ضریب همبستگی را می‌توان برای سنجش میزان خطای موجود در داده‌ها نیز استفاده کرد.

۱–۲ ضریب همبستگی ساده

بردار تصادفی X دارای ماتریس کوواریانس $\Sigma=[\sigma_{ij}]$ می‌باشد که عناصر روی قطر اصلی واریانس‌های هر مولفه و خارج از قطر اصلی، کوواریانس بین مولفه‌ها را نشان می‌دهد. کمیت

کوواریانس (σ_{ij}) به واحدهای اندازه‌گیری X_i و X_j وابسته است. بنابراین ضریب همبستگی متغیرهای تصادفی X_i و X_j که به واحد اندازه‌گیری وابسته نیست، به صورت زیر تعریف می‌شود.

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{V(X_i)}\sqrt{V(X_j)}}$$

ضریب همبستگی ساده ارتباط خطی بین دو متغیر را اندازه‌گیری می‌کند که دارای خواص زیر می‌باشد:

- $|\rho_{ij}| \le 1$
- $\rho_{ij} = \pm 1 \iff X_i$ و X_j با احتمال یک، رابطه خطی داشته باشند.
- اگر X_i به $aX_i + b$ و X_j به $cX_j + d$ تبدیل شود، ضریب همبستگی بدون تغییر خواهد ماند $a > 0$، $c > 0$.

از روی n مشاهده مستقل از X ، ماتریس کوواریانس نمونه‌ای زیر به دست می‌آید.

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})' = [S_{ij}] \qquad (۱-۱)$$

ضریب همبستگی نمونه‌ای بین دو مولفه از بردار تصادفی X مانند X_i و X_j عبارت است از:

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}}$$

۱-۳ ضریب همبستگی جزئی

بردار تصادفی X را به صورت $\left(X_1', X_2'\right)'$ افراز می‌کنیم که در آن X_1 برداری تصادفی با p_1 مولفه و X_2 برداری تصادفی با p_2 مولفه است. توزیع X_1 به شرط معلوم بودن X_2 به صورت زیر است که در آن $\Sigma_{1.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ می‌باشد.

$$N_{p_1}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{1.2}\right)$$

حال اگر تعریف کنیم $\Sigma_{1.2} = [a_{ij}]$ آنگاه ضریب همبستگی جزئی بین X_i و X_j از بردار تصادفی X_1 به شرط معلوم بودن X_2 را به صورت زیر تعریف می‌شود.

$$\rho_{ij.p_1+1,\dots,p_1+p_2} = \frac{a_{ij}}{\sqrt{a_{ii}}\sqrt{a_{jj}}}$$

از روی n مشاهده مستقل از X ، ماتریس کوواریانس نمونه‌ای S که در رابطه ۱-۱ تعریف شد، به دست می‌آید. اگر این ماتریس را به صورت زیر در نظر بگیریم (ابعاد S_{ij} متناظر با Σ_{ij})

$$S=\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

آنگاه ماتریس $S_{1.2}$ برابر است با:

$$S_{1.2}=S_{11}-S_{12}S_{22}^{-1}S_{21}$$

حال اگر تعریف کنیم $S_{1.2}=[b_{ij}]$ آنگاه ضریب همبستگی جزئی نمونه‌ای بین X_i و X_j از بردار تصادفی X_1 به شرط معلوم بودن X_2 را به صورت زیر تعریف می‌شود.

$$r_{ij.p_1+1,\dots,p_1+p_2}=\frac{b_{ij}}{\sqrt{b_{ii}}\sqrt{b_{jj}}}$$

به یکی از خواص مهم ضریب همبستگی جزئی در ذیل اشاره شده است.

- اگر $p=3$ باشد آنگاه فرمول ضریب همبستگی جزئی و همبستگی جزئی نمونه‌ای به صورت زیر خواهد بود.

$$\rho_{12.3}=\frac{\rho_{12}-\rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}}$$

$$r_{12.3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

مثال عددی

فرض کنید $X\sim N_4(0,\Sigma)$ باشد که در آن ماتریس واریانس کوواریانس برابر است با:

$$\Sigma=\begin{bmatrix} 9 & 2 & 2 & 2 \\ 2 & 9 & 2 & 2 \\ 2 & 2 & 9 & 2 \\ 2 & 2 & 2 & 9 \end{bmatrix}$$

ضریب همبستگی جزئی $\rho_{12.34}$ را محاسبه می‌نماییم. ابتدا، به محاسبه Σ_{22}^{-1} می‌پردازیم.

$$\Sigma_{22}^{-1}=\frac{1}{\begin{vmatrix} 9 & -2 \\ -2 & 9 \end{vmatrix}}\begin{bmatrix} 9 & -2 \\ -2 & 9 \end{bmatrix}=\begin{bmatrix} \frac{9}{77} & \frac{-2}{77} \\ \frac{-2}{77} & \frac{9}{77} \end{bmatrix}$$

در ادامه، به محاسبه $\Sigma_{1.2}$ می‌پردازیم.

$$\Sigma_{1.2}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$=\begin{bmatrix}9&2\\2&9\end{bmatrix}-\begin{bmatrix}2&2\\2&2\end{bmatrix}\begin{bmatrix}\frac{9}{77}&\frac{-2}{77}\\\frac{-2}{77}&\frac{9}{77}\end{bmatrix}\begin{bmatrix}2&2\\2&2\end{bmatrix}=\begin{bmatrix}\frac{91}{11}&\frac{14}{11}\\\frac{14}{11}&\frac{91}{11}\end{bmatrix}$$

در نهایت، برای محاسبه $\rho_{12.34}$، داریم:

$$\rho_{12.34}=\frac{\frac{14}{11}}{\sqrt{\frac{91}{11}}\sqrt{\frac{91}{11}}}=\frac{14}{91}=\frac{2}{13}$$

۴–۱ ضریب همبستگی چندگانه

بردار تصادفی X را به صورت زیر افراز می‌کنیم.

$$X=\left(X_1\ \vdots\ X_2,X_3,\ldots,X_p\right)'$$
$$=\left(X_1\ \vdots\ X_2'\right)'$$

ماتریس کوواریانس این افراز به صورت زیر می‌باشد که در آن $\sigma_{21}'=\left(\sigma_{12}\ \ \sigma_{13}\ \ldots\ \sigma_{1p}\right)$ می‌باشد.

$$\Sigma=\begin{pmatrix}\sigma_{11}&\sigma_{21}'\\\sigma_{21}&\Sigma_{22}\end{pmatrix}$$

با استفاده از ماتریس کوواریانس فوق، ضریب همبستگی چندگانه X_1 روی $X_2,X_3,\ldots,X_p$ را به صورت زیر تعریف می‌کنیم.

$$\rho_{1.23\ldots p}=\frac{\sqrt{\sigma_{21}'\Sigma_{22}^{-1}\sigma_{21}}}{\sqrt{\sigma_{11}}}$$

ضریب همبستگی چندگانه دارای خواص زیر می‌باشد.

– اگر $p=2$ باشد، آنگاه ضریب همبستگی چندگانه به صورت ضریب همبستگی ساده درمی‌آید.

$$\rho_{1.2}=\frac{\sqrt{\frac{\sigma_{12}^2}{\sigma_{22}}}}{\sqrt{\sigma_{11}}}=\frac{|\sigma_{12}|}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}=|\rho_{12}|$$

– $0\le\rho_{1.23\ldots p}\le 1$

مثال عددی

فرض کنید $X \sim N_4(0,\Sigma)$ باشد که در آن ماتریس واریانس کوواریانس برابر است با:

$$\Sigma = \begin{bmatrix} 9 & 2 & 2 & 2 \\ 2 & 9 & 2 & 2 \\ 2 & 2 & 9 & 2 \\ 2 & 2 & 2 & 9 \end{bmatrix}$$

ضریب همبستگی چندگانه $\rho_{1.234}$ را محاسبه می‌نماییم. این ماتریس به صورت زیر افراز می‌شود.

$$\sigma_{11} = 9 \quad ، \quad \sigma_{12} = \sigma'_{21} = \begin{pmatrix} 2 & 2 & 2 \end{pmatrix} \quad ، \quad \sigma_{21} = \begin{pmatrix} 2 & 2 & 2 \end{pmatrix}' \quad ، \quad \Sigma_{22} = \begin{bmatrix} 9 & 2 & 2 \\ 2 & 9 & 2 \\ 2 & 2 & 9 \end{bmatrix}$$

حال به محاسبه معکوس ماتریس Σ_{22} می‌پردازیم. برای محاسبه معکوس این ماتریس، ابتدا دترمینان آن را محاسبه می‌نماییم که برابر ٦٣٧ به دست می‌آید. در این مرحله، باید به محاسبه ماترس الحاقی بپردازیم.

$$M = \begin{bmatrix} 77 & -14 & -14 \\ -14 & 77 & -14 \\ -14 & -14 & 77 \end{bmatrix}$$

بنابراین

$$\Sigma_{22}^{-1} = \frac{1}{|\Sigma|} M^T = \frac{1}{637} \begin{bmatrix} 77 & -14 & -14 \\ -14 & 77 & -14 \\ -14 & -14 & 77 \end{bmatrix} = \frac{1}{91} \begin{bmatrix} 7 & -2 & -2 \\ -2 & 7 & -2 \\ -2 & -2 & 7 \end{bmatrix}$$

در نهایت، برای محاسبه $\rho_{1.234}$، داریم:

$$\rho_{1.234} = \frac{\sqrt{\begin{pmatrix} 2 & 2 & 2 \end{pmatrix} \Sigma_{22}^{-1} \begin{pmatrix} 2 & 2 & 2 \end{pmatrix}'}}{\sqrt{9}}$$

$$= \frac{1}{3\sqrt{91}} \left(\begin{pmatrix} 2 & 2 & 2 \end{pmatrix} \begin{bmatrix} 7 & -2 & -2 \\ -2 & 7 & -2 \\ -2 & -2 & 7 \end{bmatrix} \begin{pmatrix} 2 & 2 & 2 \end{pmatrix}' \right)^{\frac{1}{2}} = \frac{1}{3\sqrt{91}} (36)^{\frac{1}{2}} = \frac{2}{\sqrt{91}}$$

۱-۵ همبستگی متعارف

۱-۵-۱ مقدمه

تحلیل همبستگی متعارف[1] یک روش آماری چندمتغیره است که به وسـیله هتلینـگ[2] (۱۹۳۶) معرفـی شده و برای بررسی همبستگی بین یک ترکیب خطی از متغیرهای دو مجموعـه مجـزا مـورد اسـتفاده قرار می‌گیرد.

هدف تحلیل همبستگی متعارف، شناسایی و کمی نمودن رابطه بین دو مجموعه از متغیرها مـی‌باشـد. این روش، همبستگی بین یک ترکیب خطـی از متغیرهـای یـک مجموعـه بـا یـک ترکیـب خطـی از متغیرهای یک مجموعه دیگر را مورد بررسی و مطالعه قرار می‌دهد. ایده این روش این است که ابتـدا زوج ترکیب خطی که بیشترین همبستگی دارند را معین نموده، سپس زوج ترکیب خطـی بعـدی کـه بزرگترین همبستگی را بین همه زوج‌ها داشته به طوری که با زوج انتخاب شده قبلی ناهمبسته باشـد، انتخاب می‌شود. این فرایند به همین صورت ادامه می‌یابد. در عمل ماتریس کوواریـانس (همبـستگی) نمونه از همبستگی بین زوج‌های ترکیب‌های خطی به دست می‌آید و هدف، محقق می‌گردد (جانسون و ویچرن[3]، ۲۰۰۰).

این روش زمانی مناسب است که متغیرهای مستقل و وابسته دارای مقیاس فاصله‌ای باشند. همبستگی متعارف رابطه چندین متغیر مستقل را با چندین متغیر وابسته انجام مـی‌دهـد. یکـی از مـشکلاتی کـه پژوهشگر در بسیاری از مطالعات رگرسیون و تجزیه و تحلیل ممیز بـا آنهـا روبـرو اسـت همبـستگی متقابل بالا میان متغیرهای توصیف‌کننده است، که با توجه به اثرات ایـن متغیرهـا بـر روی متغیرهـای وابسته به برآوردهای اریب‌دار منجر می‌شود.

از آنجایی که ماتریس کوواریانس (همبستگی) به شـدت بـه داده‌هـای دورافتـاده حـساس مـی‌باشـد، نتیجه‌های به دست آمده استوار نیست. برای استوار نمـودن روش‌هـای کلاسـیک، کرنـل[4] (۱۹۹۱)

[1] Canonical correlation analysis
[2] Hotelling
[3] Johnson & Wichern
[4] karnel

M- برآوردگرها را پیشنهاد نمود. پس از وی کراکس و دهون[1] (۲۰۰۲) برآوردگرهای MCD را پیشنهاد نموده و تابع نفوذ همبستگی‌های متعارف و بردارهای متعارف را مورد مطالعه و بررسی قرار دادند. استواری و ویژگی‌های مجانبی تحلیل همبستگی متعارف استوار توسط تاسکینن و همکاران[2] (۲۰۰۴) ارائه شده است. البته در حال حاضر نیز مطالعات ارزیابی در زمینه تحلیل همبستگی متعارف در حال انجام است.

۱-۵-۲ تحلیل همبستگی متعارف

فرض می‌کنیم اولین گروه از متغیرها دارای p متغیر باشد که با بردار تصادفی $X^{(1)}_{p\times1}$ نشان داده می‌شود و دومین گروه از متغیرها دارای q متغیر باشد که با بردار تصادفی $X^{(2)}_{q\times1}$ نشان داده می‌شود. ضمناً فرض می‌کنیم که رابطه $p \le q$ همواره برقرار باشد.

در صورتی که دو بردار تصادفی $X^{(1)}$ و $X^{(2)}$ را توأماً در نظر بگیریم، به صورت زیر خواهد بود:

$$X_{((p+q)\times1)} = \begin{bmatrix} X^{(1)} \\ -- \\ X^{(2)} \end{bmatrix} = \begin{bmatrix} X^{(1)}_1 \\ X^{(1)}_2 \\ \vdots \\ X^{(1)}_p \\ -- \\ X^{(2)}_1 \\ X^{(2)}_2 \\ \vdots \\ X^{(2)}_q \end{bmatrix}$$

بردار امید ریاضی و ماتریس واریانس کوواریانس بین دو بردار تصادفی $X^{(1)}$ و $X^{(2)}$ به صورت زیر خواهد بود:

$$\mu_{((p+q)\times1)} = E(X) = \begin{bmatrix} E\left(X^{(1)}\right) \\ ---- \\ E\left(X^{(2)}\right) \end{bmatrix} = \begin{bmatrix} \mu^{(1)} \\ -- \\ \mu^{(2)} \end{bmatrix}$$

---
[1] Croux & Dehon
[2] Taskinen et al.