بسم الله الرحمن الرحيم

دانشگاه شهید باهنر کرمان
دانشکده ریاضی و کامپیوتر
گروه آمار

پایان‌نامه تحصیلی برای دریافت درجه کارشناسی ارشد

رشته آمار گرایش آمار ریاضی

آزمون فرضیه خطی با استفاده از آماره آزمون فازی

مولف:

علیرضا جیریایی

استاد راهنما:

دکتر ماشاالله ماشین‌چی

استاد مشاور:

دکتر علیرضا عرب‌پور

آذرماه ۱۳۹۱

این پایان نامه به عنوان یکی از شرایط درجه کارشناسی ارشد به

گروه آمار

دانشکده ریاضی و کامپیوتر
دانشگاه شهید باهنر کرمان

تسلیم شده و هیچگونه مدرکی به عنوان فراغت از تحصیل دوره مزبور شناخته نمی شود.

دانشجو: علیرضا جیریایی — امضاء:

استاد راهنما: دکتر ماشاالله ماشین‌چی — امضاء:

استاد مشاور: دکتر علیرضا عرب‌پور — امضاء:

داور اول : دکتر وحید امیرزاده — امضاء:

داور دوم: دکتر احد جمالیزاده — امضاء:

نماینده تحصیلات تکمیلی: دکتر محسن مددی — امضاء:

تقدیم به پیشگاه امام عصر (عج)

ای نسیم سحر آرامگه یار کجاست — منزل آن مه عاشق کش عیار کجاست
شب تار است و ره وادی ایمن در پیش — آتش طور کجا موعد دیدار کجاست
هر که آمد به جهان نقش خرابی دارد — در خرابات بگویید که هشیار کجاست
آن کس است اهل بشارت که اشارت داند — نکته‌ها هست بسی محرم اسرار کجاست
هر سر موی مرا با تو هزاران کار است — ما کجاییم و ملامت گر بی‌کار کجاست
بازپرسید ز گیسوی شکن در شکنش — کاین دل غمزده سرگشته گرفتار کجاست
عقل دیوانه شد آن سلسله مشکین کو — دل ز ما گوشه گرفت ابروی دلدار کجاست
ساقی و مطرب و می جمله مهیاست ولی — عیش بی یار مهیا نشود یار کجاست
حافظ از باد خزان در چمن دهر مرنج — فکر معقول بفرما گل بی خار کجاست

حافظ

تشکر و قدردانی

حمد وسپاس پروردگاری را سزاست که به من توفیق تحصیل عنایت فرمود، که اگر در پرتو فضل او نبود، هرگز موفق به انجام این مهم نمی‌شدم.

در راستای نگارش این پایان‌نامه اندیشمندان بزرگی راهنمایم بودند که قلم چون منی عاجز از سپاسگزاری از ایشان است. اما برحسب تکلیف، هر چند با زبانی قاصر تشکر از این عزیزان را وظیفه خود می‌دانم.

تشکر و تقدیر بی‌پایان خود را تقدیم استاد گرامی جناب آقای دکتر ماشین‌چی می‌نمایم که بدون نظارت صبورانه و دلسوزانه این بزرگ‌مرد عرصه علمی کشور تدوین وتحقیق این پایان‌نامه بسیار دشوار می‌نمود.

همچنین از استادان ارجمند جناب آقای دکتر عربپور و جناب آقای پرچمی از صمیم قلب قدردانی می‌نمایم، چه اینکه راهنمایی‌ها و تذکرات ایشان راه‌گشای انجام مراحل این تحقیق بود.

از تمامی اعضای خانواده‌ام خصوصا پدر و مادر بزرگوارم بسیار متشکرم که آن‌چه در توان داشتند برای تحصیل من دریغ نکردند.

علیرضا جیریایی

چکیده

مسایل کاربردی مختلفی از آزمون فرض‌ها را می‌توان تحت یک حالت کلی از مدل‌های خطی که شامل آنالیز واریانس هم می‌شود، مورد بررسی قرار داد. در این پایان‌نامه یک روش برای آزمون کردن فرضیه خطی با استفاده از تکنیک‌های فازی ارائه شده است. در واقع در این روش مجموعه‌ای از فواصل اطمینان $۱۰۰(۱-\alpha)\%$، برای همه $۰/۰۱ \leq \alpha \leq ۱$، برای ایجاد کردن مفهوم‌های آماره آزمون فازی و مقدار بحرانی فازی بکار گرفته شده‌اند. سپس بر مبنای این مفاهیم یک قاعده تصمیم ارائه شده‌است. استفاده از همه فواصل اطمینان از ٪۹۹ تا ٪۰ به جای استفاده فقط از یک فاصله اطمینان منجر به بهره‌گیری از بیشترین اطلاعات نهفته در داده‌ها برای استنباط آماری خواهد شد. بعلاوه زمانی که یک نوع تردید در پذیرش یا رد فرضیه‌ها وجود داشته باشد، روش پیشنهاد شده برای اصلاح روند آزمون فرضیه‌ها توصیه شده‌است. به طوری که با استفاده از این روش فرضیه‌ها با اطمینان بیشتری رد یا پذیرفته شوند. این حالت با ارائه چندین مثال در طول پایان‌نامه توضیح داده شده‌است. همه محاسبات انجام شده در این مثال‌ها توسط بسته نرافزاری قدرتمند R انجام شده است و دستورات مربوطه در یک پیوست ارائه شده است.

به عنوان مواردی ساده از آزمون فرضیه خطی آنالیز واریانس یکطرفه و آنالیز کوواریانس یکطرفه مورد بررسی قرار گرفته‌اند. با این وجود روش ارائه شده در این پایان‌نامه هنوز قابل اعمال شدن روی تمام صورت‌های مدل خطی است.

کلمات کلیدی: *آنالیز واریانس، فاصله اطمینان، آماره آزمون فازی، فرضیه خطی، مدل خطی*

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

مروری بر فرضیه خطی و نحوه آزمون آن

۱.۱ مقدمه

در گستره وسیعی از مسائل آماری، آزمایشگر به استنباط درباره برداری از پارامترهای مجهول از یک مدل آماری علاقه‌مند است. برای نمونه او ممکن است بخواهد میانگین یک توزیع نرمال چندمتغیره را برآورد کند و یا برخی از فرضیه‌ها را در مورد بردار میانگین، آزمون کند.

در این فصل مدل خطی به عنوان یک حالت کلی مطرح شده و یک صورت کلی از آزمون فرضیه‌ها در مورد این مدل بحث شده‌است. در بخش‌های ۲.۱ ، ۳.۱ و ۴.۱ این فرضیه‌ها مطرح و آماره آزمون و توزیع آن مشخص شده‌است و برآوردگرهایی برای پارامترهای مدل بدست آمده‌است. در بخش‌های ۵.۱ و ۶.۱ آنالیز واریانس و آنالیز کوواریانس یکطرفه به عنوان حالت های خاص از مدل‌های خطی مورد بررسی قرار گرفته‌است.

در نظر گرفتن حالت کلی باعث می‌شود بسیاری از نتایج رساله قابل تعمیم به مدل‌های خطی باشد و این بسیار مطلوب است. مباحث این فصل نیاز به مقداری دانش در جبر خطی دارد برای این منظور مرجع [۷] یا [۱۸] معرفی می‌گردد. در آخر، در بخش ۷.۱ برخی از مطالعات انجام شده درباره آنالیز واریانس در محیط فازی، به خصوص آنالیز واریانس یکطرفه به طور مختصر ذکر شده‌اند. مطالب

این فصل که در رابطه با مدل خطی مطرح شده، عمدتا از مراجع [۲۴]، [۳۰] و [۳۱] است.

۲.۱ فرضیه‌های خطی

گستره وسیعی از مسئله‌های آزمون فرضیه‌ها و برآوردیابی می‌تواند تحت یک حالت کلی در نظر گرفته‌شود، در این بخش این حالت کلی تحت عنوان مدل خطی و فرضیه‌های خطی مطرح شده است. مثال زیر را در نظر بگیرید.

مثال ۱.۲.۱. فرض کنید $Y_1, Y_2, \ldots, Y_k$ متغیرهای مستقل با $E(Y_i) = \mu_i$ و $Var(Y_i) = \sigma^2$ برای $i = 1, 2, \ldots, k$ باشند. همچنین n_i مشاهده از هر Y_i گرفته شده‌است و $\sum_{i=1}^{a} n_i = n$. لازم است که فرضیه $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ آزمون شود. مسئله‌هایی از این دست به طور کاملاً طبیعی ایجاد می‌شود، برای نمونه در آزمایش‌های کشاورزی که آزمایشگر به مقایسه میانگین محصول ایجاد شده، می‌پردازد. وقتی k نوع کود متفاوت برای تولید محصولات بکار رفته است.

در تعریف ۱.۲.۱ منظور از مدل خطی و فرضیه خطی، بیان شده‌است.

تعریف ۱.۲.۱. فرض کنید $\boldsymbol{Y} = (Y_1\ Y_2\ \ldots\ Y_n)'$ یک بردار تصادفی ستونی و $\boldsymbol{X}$ یک ماتریس $n \times k$، $k < n$، از ثابت‌های معلوم حقیقی x_{ij}، $i = 1, 2, \ldots, k$ و $j = 1, 2, \ldots, n_i$ باشند. گفته می‌شود که توزیع $\boldsymbol{Y}$ در یک مدل خطی صدق می‌کند اگر رابطه زیر برقرار باشد.

$$E(\boldsymbol{Y}) = \boldsymbol{X\beta} \tag{۱.۱}$$

در بالا $\boldsymbol{\beta} = (\beta_1\ \beta_2\ \ldots\ \beta_k)'$ یک بردار از پارامترهای مجهول $\beta_1, \beta_2, \ldots, \beta_k$ است. معمول و رایج است که نوشته شود،

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{۲.۱}$$

که $\boldsymbol{\epsilon} = (\epsilon_1\ \epsilon_2\ \ldots\ \epsilon_n)'$ یک بردار از متغیرهای تصادفی غیر قابل مشاهده است و $E(\epsilon_i)$ برای هر i، $i = 1, 2, \ldots, n$ برابر با صفر است. رابطه ۲.۱ به نام مدل خطی شناخته می‌شود.

یک فرضیه کلی خطی، درباره بردار $\boldsymbol{\beta}$ در نظر گرفته می‌شود. برای نمونه اینکه $\boldsymbol{\beta}$ در فرضیه $H_{\circ} : \boldsymbol{H\beta} = \circ$ صدق کند، که در آن $\boldsymbol{H}$ یک ماتریس با درایه‌های معلوم حقیقی با بعد $r \times k$، $r \leq k$، است.

در ادامه فرض شده است که $\epsilon_{۱}, \epsilon_{۲}, \ldots, \epsilon_n$ متغیرهای تصادفی مستقل با توزیع نرمال هستند و برای $i = ۱, ۲, \ldots, n$، $E(\epsilon_i) = \circ$ و $Var(\epsilon_i) = \sigma^{۲}$ هستند. از مدل ۲.۱ نتیجه می‌شود که $Y_{۱}, Y_{۲}, \ldots, Y_n$ متغیرهای تصادفی مستقل نرمال با واریانس مشترک $\sigma^{۲}$ و $E(Y_i) = \sum_{j=۱}^{k} x_{ij}\beta_j$ می باشند. تابع چگالی احتمال بردار تصادفی $\boldsymbol{Y}$ به صورت زیر است.

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{۱}{(۲\pi\sigma^{۲})^{\frac{n}{۲}}} \exp\left\{-\frac{۱}{۲\sigma^{۲}}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})\right\} \tag{۳.۱}$$

$$\boldsymbol{y} \in \mathbb{R}^n,$$

$$\sigma^{۲} > \circ,$$

$$\boldsymbol{\beta} \in \mathbb{R}^k.$$

همچنین فرض می‌شود که $\boldsymbol{H}_{r\times k}$ یک ماتریس پر رتبه سطری با رتبه r و $\boldsymbol{X}_{n\times k}$ یک ماتریس پر رتبه ستونی با رتبه k باشند.

۳.۱ برآورد پارامترهای مجهول در مدل خطی

پارامترهای مجهول در مدل خطی معرفی شده در تعریف ۱.۲.۱ بردار $\boldsymbol{\beta}$ و پارامتر پراکندگی $\sigma^{۲}$ می باشند. در واقع دو روش عمده برای برآورد بردار $\boldsymbol{\beta}$ وجود دارد که هر چند باهم تفاوت دارند ولی منجر به برآورد یکسان برای $\boldsymbol{\beta}$ خواهند شد. در اینجا از ذکر جزئیات پرهیز شده است و خواننده علاقه‌مند می‌تواند برای مشاهده جزئیات بیشتر به مراجع [۲۴] و [۳۱] مراجعه نماید.

۱.۳.۰۱ برآوردگر پارامترهای مجهول با استفاده از اصل کمترین مربعات

یکی از روش‌های معمول برای برآوردکردن پارامترها در مدل خطی استفاده از اصل کمترین مربعات است. در این روش فرض نرمال بودن بردار $\boldsymbol{\epsilon}$ لازم نیست و برآوردگر از مینیمم کردن مجموع مربعات زیر بر حسب $\boldsymbol{\beta}$ بدست می‌آید.

$$\begin{aligned}\sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} [Y_i - E(Y_i)]^2 \\ &= (\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta}) \\ &= \boldsymbol{Y'Y} - \boldsymbol{Y'X\beta} - \boldsymbol{\beta' X' Y} + \boldsymbol{\beta' X' X \beta} \qquad (4.1)\end{aligned}$$

با مشتق‌گیری از مجموع مربعات ۴.۱ نسبت به $\boldsymbol{\beta}$ و برابر صفر قراردادن آن برآوردگر $\boldsymbol{\beta}$ یعنی $\hat{\boldsymbol{\beta}}$، عبارت است از جواب‌های معادله $\boldsymbol{X'X\hat{\beta}} = \boldsymbol{X'Y}$.

اکنون اگر $\boldsymbol{X}$ دارای رتبه k $(k < n)$ باشد بنابراین $\boldsymbol{X'X}$ که دارای همان رتبه است، یک ماتریس نامنفرد خواهد بود و یک برآوردگر منحصربه‌فرد برای $\boldsymbol{\beta}$ به صورت $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$ است.

اگر $\boldsymbol{X}$ دارای رتبه‌ای کمتر از k باشد آنگاه $\boldsymbol{X'X}$ یک ماتریس منفرد خواهد بود و حل معادله $\boldsymbol{X'X\hat{\beta}} = \boldsymbol{X'Y}$ منجر به یک جواب منحصربه‌فرد نمی‌شود.

۲.۳.۰۱ برآوردگر ماکزیمم درستنمایی پارامترهای مجهول

روش دیگری که عمدتا برای برآورد پارامترها در مدل خطی بکار برده می‌شود، استفاده از تابع درستنمایی است. اگر فرض شود که $\boldsymbol{\epsilon} \sim N_n(0, I_n\sigma^2)$ آنگاه برآوردگر $\boldsymbol{\beta}$ و σ^2 از ماکزیمم کردن تابع درستنمایی زیر برحسب $\boldsymbol{\beta}$ و σ^2 حاصل خواهدشد.

$$L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})\right\}$$

یا معادلا ماکزیمم کردن تابع زیر

$$\begin{aligned}\ell(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) &= LnL(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) \\ &= -\frac{n}{2} Ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})\end{aligned}$$

که پس از مشتق‌گیری و برابر صفر قرار دادن، برآوردگر $\boldsymbol{\beta}$ از حل معادله $\boldsymbol{X'X\hat{\beta}} = \boldsymbol{X'Y}$ حاصل می‌شود که مشابه با روش برآوردیابی با استفاده از اصل کمترین مربعات است. برآوردگر اریب σ^2 به صورت $\widehat{\sigma^2} = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}})'(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}})$ خواهد بود.

۴.۱ آماره آزمون و توزیع آماری آن

برای آزمون کردن فرضیه خطی $H_0 : \boldsymbol{H\beta} = 0$ درباره مدل خطی معرفی شده در تعریف ۱.۲.۱، یک آماره آزمون براساس روش نسبت درستنمایی تعمیم یافته در قضیه ۱.۴.۱ مطرح و توزیع این آماره آزمون بدست آمده است.

قضیه ۱.۴.۱. با در نظر گرفتن مدل خطی $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ که در آن $\boldsymbol{X}$ یک ماتریس با بعد $n \times k$ از ثابت‌های معلوم x_{ij}، $i = 1, 2, \ldots, n$ و $j = 1, 2, \ldots, k$، با رتبه k $(k < n)$ است و $\boldsymbol{\beta}$ یک بردار از پارامترهای مجهول $\beta_1, \beta_2, \ldots, \beta_k$ است و $\boldsymbol{\epsilon} = (\epsilon_1\ \epsilon_2 \ldots \epsilon_n)'$ یک بردار از متغیرهای تصادفی مستقل نرمال با میانگین مشترک صفر و واریانس مشترک σ^2 است. آزمون نسبت درستنمایی تعمیم یافته برای آزمون کردن فرضیه خطی $H_0 : \boldsymbol{H\beta} = 0$، که در آن $\boldsymbol{H}$ یک ماتریس از بعد $r \times k$ و با رتبه r، $(r \le k)$ است، فرض H_0 را در سطح معنی‌داری γ رد خواهد کرد اگر $F \ge F_{1-\gamma}$، که در آن $P(F \le F_{1-\gamma}|H_0) = 1 - \gamma$ و F یک متغیر تصادفی به صورت زیر است.

$$F = \frac{SS^*/r}{SS/(n-k)} \tag{۵.۱}$$

در رابطه ۵.۱،

$$SS^* = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\circ)'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\circ) - (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

$$SS = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

و $\hat{\boldsymbol{\beta}}$ برآوردگر ماکزیمم درستنمایی $\boldsymbol{\beta}$ و همچنین $\hat{\boldsymbol{\beta}}_\circ$ برآوردگر ماکزیمم درستنمایی $\boldsymbol{\beta}$ تحت فرض $H_\circ$ هستند، به علاوه متغیر تصادفی F تحت فرض $H_\circ$ دارای توزیع آماری F با r درجه آزادی در صورت و $n-k$ درجه آزادی در مخرج خواهد بود.

برهان. آزمون نسبت درستنمایی تعمیم یافته برای فرض $\boldsymbol{H\beta} = \circ$، فرض $H_\circ$ را رد می‌کند اگر و تنها اگر $\lambda(\boldsymbol{Y}) < c$، که $\lambda(\boldsymbol{Y})$ به صورت زیر است.

$$\lambda(\boldsymbol{Y}) = \frac{\sup_{\boldsymbol{\theta}\in\Theta_\circ} f_{\boldsymbol{\beta},\sigma^۲}(\boldsymbol{y})}{\sup_{\boldsymbol{\theta}\in\Theta} f_{\boldsymbol{\beta},\sigma^۲}(\boldsymbol{y})},$$

در برابری بالا $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^۲)'$ و $\Theta_\circ = \left\{(\boldsymbol{\beta}', \sigma^۲)' : \boldsymbol{H\beta} = \circ\right\}$ هستند.
فرض کنید $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \widehat{\sigma^۲})'$ برآوردگر ماکزیمم درستنمایی $\boldsymbol{\theta} \in \Theta$ باشد و $\hat{\boldsymbol{\theta}}_\circ = (\hat{\boldsymbol{\beta}}'_\circ, \widehat{\sigma^۲_\circ})'$ برآوردگر ماکزیمم درستنمایی $\boldsymbol{\theta}$ تحت فرض $H_\circ$ باشد. همان‌طور که قبلا اشاره شد با داشتن تابع درستنمایی، $\hat{\boldsymbol{\beta}}$ مقداری از $\boldsymbol{\beta}$ است که عبارت $(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta})$ را مینیمم می‌کند و برآوردگر ماکزیمم درستنمایی $\sigma^۲$ برابر است با $\widehat{\sigma^۲} = \frac{۱}{n}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$. به طور مشابه $\hat{\boldsymbol{\beta}}_\circ$ مقداری از $\boldsymbol{\beta}$ است که عبارت $(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta})$ را تحت شرط $\boldsymbol{H\beta} = \circ$ مینیمم می‌کند (این کار با استفاده از ضرایب لاگرانژ همیشه امکان پذیر است.) و برآوردگر ماکزیمم درستنمایی $\sigma^۲$ تحت فرض صفر عبارت است از $\widehat{\sigma^۲_\circ} = \frac{۱}{n}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\circ)'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\circ)$. در نتیجه $\lambda(\boldsymbol{Y}) = \left(\frac{\widehat{\sigma^۲}}{\widehat{\sigma^۲_\circ}}\right)^{\frac{n}{۲}}$ است و ناحیه بحرانی برابر با $\lambda(\boldsymbol{Y}) < c$ و معادل با ناحیه $\left(\lambda(\boldsymbol{Y})\right)^{-\frac{۲}{n}} > (c)^{-\frac{۲}{n}}$ است، که به شکل زیر

بازنویسی شده است.

$$\frac{\widehat{\sigma_\circ^{\text{۲}}}}{\widehat{\sigma^{\text{۲}}}} > c_{\text{۱}}$$

ناحیه بالا می‌تواند به شکل $\frac{(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_\circ)'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_\circ)}{(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})} > c_{\text{۱}}$ نوشته شود و یا معادلا به صورت زیر:

$$\frac{(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_\circ)'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_\circ)-(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})} > c_{\text{۱}} - \text{۱}$$

اکنون تنها چیزی که از برهان باقی می‌ماند بدست آوردن توزیع آماره آزمون است. برای این منظور فرض کنید V_n یک فضای برداری باشد که شامل تمام مشاهدات $\boldsymbol{y}$ باشد و V_k یک زیر فضای برداری از V_n باشد که توسط بردارهای ستونی مستقل خطی $\boldsymbol{x}_{\text{۱}}, \boldsymbol{x}_{\text{۲}}, \ldots, \boldsymbol{x}_k$ از ماتریس $\boldsymbol{X}$ تولید شده است. و V_{k-r} یک زیر فضا از V_k باشد که $E(\boldsymbol{Y})$ تحت فرض $H_\circ$ در آن قرار خواهد گرفت[۱]. پایه متعامد یکه‌ای شامل $(k-r)$ بردار ستونی $\boldsymbol{\alpha}_i$ برای زیر فضای V_{k-r} مثل $(\boldsymbol{\alpha}_{r+\text{۱}}\ \boldsymbol{\alpha}_{r+\text{۲}}\ \ldots\ \boldsymbol{\alpha}_k)$ را در نظر بگیرید با تعمیم این پایه به پایه متعامد یکه $(\boldsymbol{\alpha}_{\text{۱}}\ \boldsymbol{\alpha}_{\text{۲}}\ \ldots\ \boldsymbol{\alpha}_{r+\text{۱}}\ \boldsymbol{\alpha}_{r+\text{۲}}\ \ldots\ \boldsymbol{\alpha}_k)$ برای زیر فضای V_k و تعمیم دوباره این پایه به پایه متعامد یکه $(\boldsymbol{\alpha}_{\text{۱}}\ \boldsymbol{\alpha}_{\text{۲}}\ \ldots\ \boldsymbol{\alpha}_k\ \boldsymbol{\alpha}_{k+\text{۱}}\ \ldots\ \boldsymbol{\alpha}_n)$ برای V_n، یک پایه متعامد یکه برای فضای برداری V_n خواهیم داشت. فرآیند بالا همیشه امکان پذیر است.

[۱] نکته تکمیلی در مورد مطالب مطرح شده این است که، به وضوح $\boldsymbol{Y}$ در یک مدل خطی صدق می‌کند اگر بردار $E(\boldsymbol{Y}) = (E(Y_{\text{۱}})\, E(Y_{\text{۲}}) \ldots E(Y_n))'$ در یک فضای برداری k بعدی تولید شده توسط بردارهای ستونی مستقل خطی $\boldsymbol{x}_{\text{۱}}, \boldsymbol{x}_{\text{۲}}, \ldots, \boldsymbol{x}_k$ از ماتریس $\boldsymbol{X}$ قرار داشته باشد. در واقع (۱.۱) باعث می‌شود که $E(\boldsymbol{Y})$ ترکیبی خطی از بردارهای معلوم $\boldsymbol{x}_{\text{۱}}, \boldsymbol{x}_{\text{۲}}, \ldots, \boldsymbol{x}_k$ باشد. فرضیه خطی $H_\circ : \boldsymbol{H\beta} = \circ$ ایجاب می‌کند که پارامترهای $\beta_{\text{۱}}, \beta_{\text{۲}}, \ldots, \beta_k$ در r معادله همگن خطی صدق کنند.

می‌توان $\boldsymbol{X}$ را با عملیات سطری و ستونی به ماتریس هم‌ارز $\boldsymbol{X}^*$ تبدیل کرد، و $\boldsymbol{H}$ را با فقط عملیات سطری به ماتریس هم ارز سطری $\boldsymbol{H}^*$ طوری تبدیل کرد که بتوان $\boldsymbol{X}^*$ را به زیر ماتریس‌هایی افراز کرد که یکی از آنها $\boldsymbol{H}^*$ باشد. از آنجایی که $\boldsymbol{H}^*$ هم ارز سطری $\boldsymbol{H}$ است رابطه $\boldsymbol{H}^*\boldsymbol{\beta} = \circ$ برقرار است و از این نتیجه گرفت که تحت $H_\circ$، $E(\boldsymbol{Y})$ در یک زیر فضای $(k-r)$ بعدی از فضای k بعدی تولید شده توسط $\boldsymbol{x}_{\text{۱}}, \boldsymbol{x}_{\text{۲}}, \ldots, \boldsymbol{x}_k$ قرار دارد.

فرض کنید $\boldsymbol{Z} = (Z_۱\, Z_۲\, \ldots\, Z_n)'$ مختصات مربوط به $\boldsymbol{Y}$ وابسته به پایه $(\boldsymbol{\alpha}_۱\, \boldsymbol{\alpha}_۲\, \ldots\, \boldsymbol{\alpha}_n)$ باشد، بنابراین رابطه $\boldsymbol{Y} = \boldsymbol{\alpha}_۱ Z_۱ + \boldsymbol{\alpha}_۲ Z_۲ + \ldots + \boldsymbol{\alpha}_n Z_n$ و یا $\boldsymbol{Y} = (\boldsymbol{\alpha}_۱\, \boldsymbol{\alpha}_۲\, \ldots\, \boldsymbol{\alpha}_n)\boldsymbol{Z}$ برقرار است. اگر $\boldsymbol{P} = (\boldsymbol{\alpha}_۱\, \boldsymbol{\alpha}_۲\, \ldots\, \boldsymbol{\alpha}_n)$ باشد بنابراین از آنجایی که $\boldsymbol{P}\boldsymbol{P}' = I_n$ است، $\boldsymbol{P}$ یک ماتریس متعامد است. می‌توان نوشت $\boldsymbol{P}'\boldsymbol{Y} = \boldsymbol{Z}$ یعنی $Z_i = \boldsymbol{\alpha}_i'\boldsymbol{Y}$ خواهد بود. بنابراین داریم $E(\boldsymbol{Z}) = \boldsymbol{P}'E(\boldsymbol{Y}) = \boldsymbol{P}'\boldsymbol{X}\boldsymbol{\beta}$ یا $E(Z_i) = \boldsymbol{\alpha}_i'\boldsymbol{X}\boldsymbol{\beta}$. از آنجایی‌که $\boldsymbol{X}\boldsymbol{\beta} \in V_k$ است، در نتیجه برای هر $i > k$، $\boldsymbol{\alpha}_i'\boldsymbol{X}\boldsymbol{\beta} = ۰$ است. به طور مشابه تحت فرض $H_۰ : \boldsymbol{H}\boldsymbol{\beta} = ۰$، $\boldsymbol{X}\boldsymbol{\beta} \in V_{k-r} \subset V_k$ است. پس برای هر $i \leq r$، $\boldsymbol{\alpha}_i'\boldsymbol{X}\boldsymbol{\beta} = ۰$ است. بنابراین اگر فرض کنیم $\boldsymbol{w} = \boldsymbol{P}'\boldsymbol{X}\boldsymbol{\beta}$ باشد آنگاه خواهیم داشت $w_{k+۱} = w_{k+۲} = \ldots = w_n = ۰$ و همچنین تحت فرض $H_۰$، خواهیم داشت $w_۱ = w_۲ = \ldots = w_r = ۰$.

تابع مولد گشتاور $\boldsymbol{Z}$ به صورت زیر خواهد بود.

$$\begin{aligned} M_{\boldsymbol{Z}}(\boldsymbol{t}) &= E\left[\exp(t_۱ Z_۱ + t_۲ Z_۲ + \ldots + t_n Z_n)\right] \\ &= E\left[\exp(\boldsymbol{t}'\boldsymbol{Z})\right] \\ &= E\left[\exp(\boldsymbol{t}'\boldsymbol{P}'\boldsymbol{Y})\right] \end{aligned}$$

و از آنجایی‌که $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^۲ I_n)$ رابطه زیر برقرار است.

$$\begin{aligned} M_{\boldsymbol{Z}}(\boldsymbol{t}) &= M_{\boldsymbol{Y}}(\boldsymbol{t}'\boldsymbol{P}') \\ &= \exp\left\{\boldsymbol{t}'\boldsymbol{P}'E(\boldsymbol{Y}) + \frac{۱}{۲}\boldsymbol{t}'\boldsymbol{P}' I_n \sigma^۲ \boldsymbol{P}\boldsymbol{t}\right\} \\ &= \exp\left\{\boldsymbol{t}'\boldsymbol{P}'\boldsymbol{X}\boldsymbol{\beta} + \frac{۱}{۲}\boldsymbol{t}' I_n \sigma^۲ \boldsymbol{t}\right\} \end{aligned}$$

در نتیجه $\boldsymbol{Z} \sim N_n(\boldsymbol{P}'\boldsymbol{X}\boldsymbol{\beta}, \sigma^۲ I_n)$ خواهد بود.

اکنون درایم

$$\begin{aligned}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) &= (\boldsymbol{P}\boldsymbol{Z}-\boldsymbol{P}\boldsymbol{w})'(\boldsymbol{P}\boldsymbol{Z}-\boldsymbol{P}\boldsymbol{w})\\ &= (\boldsymbol{Z}-\boldsymbol{w})'(\boldsymbol{Z}-\boldsymbol{w})\\ &= \sum_{i=۱}^{k}(Z_i-w_i)^۲+\sum_{i=k+۱}^{n} Z_i^۲.\end{aligned}$$

وکمیت $(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})$ مینیمم می‌شود اگر برای هر $i=۱,۲,\ldots,k$، $\hat{w}_i=Z_i$ باشد. پس خواهیم داشت

$$(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})=\sum_{i=k+۱}^{n} Z_i^۲.$$

از آنجای‌که تحت فرض $H_۰$، رابطه $w_۱=w_۲=\ldots=w_r=۰$ برقرار است. بنابراین مینیمم کردن $(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})$ تحت فرض $H_۰$ معادل با انتخاب $\widehat{w_۰}_i=Z_i$ برای هر $i=r+۱,r+۲,\ldots,k$ خواهد بود. در نتیجه رابطه زیر را خواهیم داشت:

$$(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_۰)'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_۰)=\sum_{i=۱}^{r} Z_i^۲+\sum_{i=k+۱}^{n} Z_i^۲.$$

و همچنین آماره F تحت فرض صفر برابر است با:

$$F=\frac{\sum_{i=۱}^{r} Z_i^۲/r}{\sum_{i=k+۱}^{n} Z_i^۲/(n-k)}.$$

اکنون واضح است که $\sum_{i=k+۱}^{n}\frac{Z_i^۲}{\sigma^۲}$ دارای توزیع $\chi^۲$ با $(n-k)$ درجه آزادی است و تحت فرض صفر $\sum_{i=۱}^{r}\frac{Z_i^۲}{\sigma^۲}$ دارای توزیع $\chi^۲$ با r درجه آزادی است و از آنجایی‌که $\sum_{i=k+۱}^{n}\frac{Z_i^۲}{\sigma^۲}$ و $\sum_{i=۱}^{r}\frac{Z_i^۲}{\sigma^۲}$ از یکدیگر مستقل هستند، تحت فرض $H_۰$، متغیر تصادفی F دارای توزیع F با r درجه آزادی در