فصل دوازدهم

S-۱

۱- دگرسیون-۱

دگرسیون

در فصلهای قبلی بیشتر تجزیه و تحلیل‌های آماری روی یک صفت از جامعه یا متغیر تصادفی متمرکز بود در این فصل قصد داریم بصورت همزمان دو صفت از جامعه یا دو متغیر تصادفی را مورد بررسی قرار دهیم. این بررسی شامل پیش بینی مقادیر یکی از متغیرۀا از روی مقادیر متغیر دیگراست که به مساله برگشت یا دگرسیون معروف می‌باشد.

به عنوان مثال فرض کنید بخواهیم تاثیر میزان مصرف شیر را در افزایش قد بدست بیاوریم و یا بخواهیم میزان وزن فرزند را از روی وزن پدرش پیش بینی کنیم. ملاحظه می‌کنید که در این گونه مسایل دو متغیر تصادفی مورد مطالعه به نوعی به یکدیگر وابسته می‌باشند. به عبارت دقیقتر یک متغیر تصادفی مثل X را مستقل و متغیر تصادفی Y را وابسته به آن در نظر می‌گیریم و یا برعکس Y را مستقل و X را وابسته به آن در نظر می‌گیریم.

آشکار است که انتخاب هر یک از دو حالت به نوع مساله بستگی دارد.

در مسایل دگرسیون برای یافتن رابطه بین متغیر تصادفی مستقل X و متغیر وابسته Y ابتدا یک نمونه nتایی از متغیر تصادفی x جمع آوری می‌کنیم که نتایج آن بصورت $x_۱, x_۲, \cdots, x_n$ می‌باشند. سپس مقادیر متناظر با هر یک از نمونه‌های بدست آمده (x_iها) را که همان مقادیر معادل متغیر تصادفی وابسته Y می‌باشند بدست می‌آوریم. به این ترتیب برای x_iها مقادیر متناظر $y_۱, y_۲, \cdots, y_n$ بدست می‌آیند. که می‌توانیم نتیجه را بصورت زوج‌های مرتب زیر نشان دهیم.

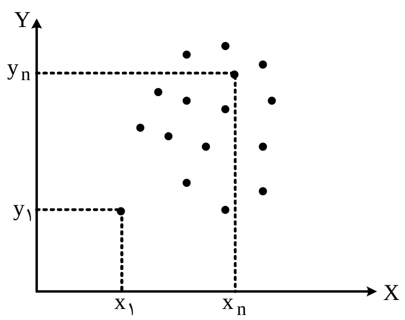
$$(x_۱, y_۱), (x_۲, y_۲), (x_۳, y_۳), \cdots, (x_n, y_n)$$

۲- دگرسیون-۲ به عنوان مثال مقادیر x_iها می‌توانند میزان طول قد افراد یک جامعه و مقادیر y_iها میزان مصرف شیر هر یک از نمونه‌ها باشند. به این ترتیب زوج مرتب (۱۸۰,۲) بیانگر این است که در نمونه‌گیری یکی از افراد جامعه دارای صول قد cm۱۸۰ بوده است و و وی روزانه دو لیوان شیر مصرف کرده است.

پس از بدست آمدن زوج‌های مرتب (x_i, y_i) ملاحظه می‌نید که هر زوج مرتب می‌تواند معادل یک نقطه در صفحه باشد. با رسم نقاط مورد نظر در صفحه یک تصویر کلی از رابطه X و Y بدست می‌آوریم. به شکلهای زیر که برای سه نمونه جداگانه می‌باشند توجه کنید:

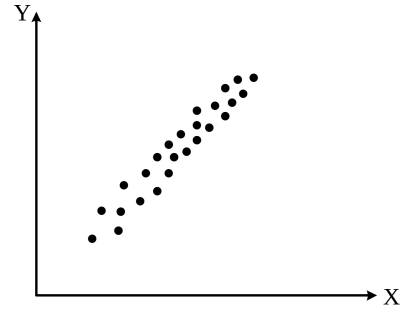
در این نمودار ملاحظه می‌کنید که زوج‌های مرتب (x_i, y_i) بصورت کاملاً پراکنده توزیع شده‌اند و به این ترتیب نتیجه می‌گیریم که رابطه‌ای بین متغیرهای تصادفی X و Y وجود ندارد.

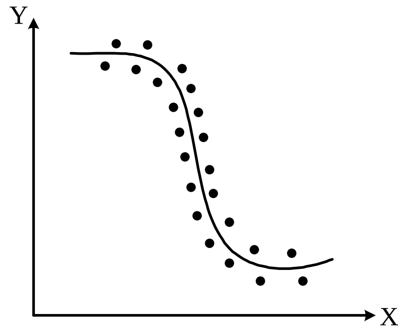
نمودار ۱



ملاحظه می‌کنید که در این نمونه یک رابطه خطی بین مقادیر x_i و y_i وجود دارد و می‌توان نوشت $y_i = a x_i + b$.

نمودار ۲





نمودار ۳

در این نمودار X و Y به یکدیگر وابسته می‌باشند اما این وابستگی از نوع غیر خطی می‌باشد.

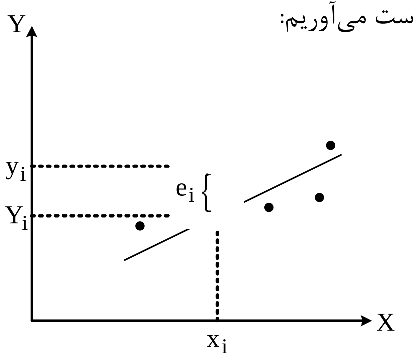
با توجه به دو نمودار ۲ و ۳ این طور به نظر می‌رسد که در حالت کلی دو متغیر تصادفی X و Y اگر از یکدیگر مستقل نباشند یا بصورت خطی و یا بصورت غیر خطی به یکدیگر وابسته می‌باشند.

قبلاً نشان دادیم که برای اندازه‌گیری میزان وابستگی دو متغیر X و Y می‌توان از ضریب همبستگی استفاده نمود. اما در مسایل دگرسیون پیش بینی متغیر Y از روی X و یا بالعکس از اهمیت ویژه‌ای برخوردار است. بنابراین نیازمند روشی هستیم که بتوان در صورت نیاز با ثابت در نظر گرفتن یکی از مقادیر X یا Y مقدار دیگری را بدست بیاوریم. برای این منظور مفهوم بردازش منحنی را مطرح می‌کنیم.

۱- دگرسیون خطی – ۳ ۱۲ . ۱ دگرسیون خطی

اگر بین دو متغیر X و Y یک رابطه خطی وجود داشته باشد می‌توانیم یک خط را طوری رسم کنیم که نقاط (x_i , y_i) کمترین فاصله را با خط مورد نظر داشته باشند. به این عمل برداژش منحنی می‌گویند. معادله خط را بصورت $Y = aX + b$ در نظر می‌گیریم که در آن a و b مقادیر مجهول می‌باشند و در این حالت X متغیر مستقل و Y متغیر وابسته به آن در نظر گرفته می‌شود. مقادیر aو b می‌بایستی طوری محاسبه شوند که مجموع فاصله نقاط (x_i , y_i) از خط $Y = aX + b$ حداقل شود. در این حالت به $Y = aX + b$ معادله دگرسیون Y می‌گویند.

برای حداقل نمودن فاصله نقاط (x_i , y_i) از خط دگرسیون مقدار خطای e_i را مطابق نمودار زیر بدست می‌آوریم:



با توجه به نمودار Y_i مقدار پیش بینی شده توسط خط دگرسیون می‌باشد که با مقدار واقعی y_i به اندازه $e_i = |y_i - Y_i|$ فاصله دارد که این فاصله همان خطای پیش بینی می‌باشد.

برای بدست آوردن بهترین نتیجه، مجموع مربعات خطا را حداقل می‌کنیم که عبارتست از:

$$SSE = E = \sum_{i=1}^{n} (y_i - Y_i)^2$$

که در آن:

$$Y = aX + b \Rightarrow Y_i = aX_i + b \Rightarrow E = \sum_{i=1}^{n} (aX_i + b - Y_i)^2$$

۲- دگرسیون خطی-۴ برای مینیمم نمودن E مشتقات پاره‌ای آنرا نسبت به a و b بدست آورده و برابر صفر قرار می‌دهیم:

$$\frac{\partial E}{\partial a}=\sum_{i=1}^{n} 2x_i(ax_i+b-y_i)=0$$

$$\frac{\partial E}{\partial b}=\sum_{i=1}^{n} 2(ax_i+b-y_i)=0$$

به این ترتیب دستگاه معادلات زیر بدست می‌آید:

$$\begin{cases}\sum_{i=1}^{n} 2x_i(ax_i+b-y_i)=0\\ \sum_{i=1}^{n} 2(ax_i+b-y_i)=0\end{cases}$$

با حل این دستگاه مقادیر مجهول a و b بدست می‌آیند. که عبارتند از:

$$a=\frac{S_{xy}}{S_{xx}} \quad , \quad b=\bar{y}-a\bar{X}$$

که در آن:

$$S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_i y_i-n\bar{x}\bar{y}$$

$$S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2$$

به معادله $y=ax+b$ خط دگرسیون Y روی x می‌گویند.یعنی Y از روی x بدست آمده است.

۱–مثال۱–۵ مثال ۱: در یک بررسی میزان قد پدران و فرزندان آنها بصورت جدول زیر بدست آمده است:

قد پدران X	۱۶۵	۱۶۰	۱۷۰	۱۶۳	۱۷۳	۱۵۷	۱۷۸	۱۶۸	۱۷۳	۱۷۰	۱۷۵	۱۸۰
قد فرزندان Y	۱۷۳	۱۶۸	۱۷۳	۱۶۵	۱۷۵	۱۶۸	۱۷۳	۱۶۵	۱۸۰	۱۷۰	۱۷۳	۱۷۸

مطلوبست:

الف) پردازش خطی داده‌ها

ب) رسم نمودار داده‌ها و خط دگرسیون.

ج) با توجه به معادله دگرسیون پیش بینی کنید که اگر قد پدری cm۱۸۵ باشد قد فرزند او چقدر خواهد بود؟

حل: برای بدست آوردن a و b در معادله $y=ax+b$ جدول زیر را تشکیل می‌دهیم:

x	y	x^2	xy	y^2
۱۶۵	۱۷۳	۲۷۲۲۵	۲۸۵۴۵	۲۹۹۲۹
۱۶۰	۱۶۸	۲۵۶۰۰	۲۶۸۸۰	۲۸۲۲۴
۱۷۰	۱۷۳	۲۸۹۰۰	۲۹۴۱۰	۲۹۹۲۹
۱۶۳	۱۶۵	۲۶۵۶۹	۲۶۸۹۵	۲۷۲۲۵
۱۷۳	۱۷۵	۲۹۹۲۹	۳۰۲۷۵	۳۰۶۲۵
۱۵۸	۱۶۸	۲۴۹۶۴	۲۶۵۴۴	۲۸۲۲۴
۱۷۸	۱۷۳	۳۱۶۸۴	۳۰۷۹۴	۲۹۹۲۹
۱۶۸	۱۶۵	۲۸۲۲۴	۲۷۷۲۰	۲۷۲۲۵
۱۷۳	۱۸۰	۲۹۹۲۹	۳۱۱۴۰	۳۲۴۰۰
۱۷۰	۱۷۰	۲۸۹۰۰	۲۸۹۰۰	۲۸۹۰۰
۱۷۵	۱۷۳	۳۰۶۲۵	۳۰۲۷۵	۲۹۹۲۹
۱۸۰	۱۷۸	۳۲۴۰۰	۳۲۰۴۰	۳۱۶۸۴

۲- مثال ۱-۶

$$\overline{X}=\frac{۱}{n}\sum x_i=\frac{۱}{۱۲}\,۲۰۳۳=۱۶۹/۴۱$$

$$\overline{Y}=\frac{۱}{n}\sum y_i=\frac{۱}{۱۲}\,۲۰۶۱=۱۷۱/۷۵$$

$$S_{xy}=\sum_{i=۱}^{n} x_i\,y_i-n\,\overline{x}\,\overline{y}=۳۴۹۴۱۸-۱۲(۱۶۹/۴۱)\,(۱۷۱/۷۵)=۲۵۰/۲۵$$

$$S_{xx}=\sum x_i^{۲}-n\,\overline{x}^{۲}=۳۴۴۹۴۹-۱۲\,(۱۶۹/۴۱)^{۲}=۵۲۴/۹۱$$

$$a=\frac{S_{xy}}{S_{xx}}=\frac{۲۵۰/۲۵}{۵۲۴/۹۱}=۰/۴۷۷$$

$$b=\overline{Y}-a\,\overline{X}=۱۷۱/۷۵-۰/۴۷۷\,(۱۶۹/۴۱)=۹۰/۹$$

بنابراین معادله خط دگرسیون عبارتست از:

$$y=۰/۴۷۷\,x+۹۰/۹$$

ب) نمودار داده‌ها و خط دگرسیون عبارتست از:

ج) با توجه به خط دگرسیون می‌توان قد فرزند یک پدر با قد ۱۸۵ را به فرم زیر بدست آورد.

$$y=۰/۴۷۷\times۱۸۵+۹۰/۹=۱۷۹/۱۴۵$$

۳-S ۱۲. ۲ ضریب همبستگی و دگرسیون

۱-۷

قبلاً ضریب همبستگی دو متغیر X و Y را بصورت زیر تعریف نمودیم:

$$\rho_{XY}=\frac{\sigma_{XY}}{\sigma_X\,\sigma_Y}$$

می‌توان با ساده نمودن معادله دگرسیون مقدار ضریب همبستگی را وارد معادله نمود:

$$y = ax+b \qquad a=\frac{S_{xy}}{S_{xx}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}\left(\frac{\sigma_y}{\sigma_x}\right) = \rho_{xy}\frac{\sigma_y}{\sigma_x}$$

توجه کنید که در اینجا در محاسبه ρ_{xy} و σ_x و σ_y از مقادیر نمونه‌های مشاهده شده استفاده می‌شود.

به این ترتیب معادله خط دگرسیون بصورت زیر بدست می‌آید:

$$y-\bar{y}=\rho_{xy}\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

توجه کنید که همواره برای داده‌های مشاهده شده (x_i, y_i) دو معادله دگرسیون وجود دارد یک معادله بر حسب y نسبت به x می‌باشد و معادله دیگر بر حسب x نسبت به y می‌باشد معادله خط دگرسیون x روی y را می‌توان بصورت زیر بدست آورد:

$$x = \alpha y + \beta$$

$$\alpha = \frac{S_{xy}}{S_{yy}} \qquad \beta = \bar{x} - \alpha \bar{y}$$

$$S_{yy} = \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i - n\bar{y}^2$$

۲–۸ با نوشتن معادله دگرسیون x روی y و استفاده از ضریب همبستگی بدست می‌آوریم:

$$x-\bar{x}=\rho_{xy}\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$

بنابراین در حالت کلی دو خط دگرسیون y روی x و x روی y خواهیم داشت که عبارتند از:

$$y-\bar{y} = \rho_{xy}\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

$$x-\bar{x} = \rho_{xy}\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$

با برابر قرار دادن $x = y$ در معادلات فوق محل تلاقی دو خط دگرسیون نقطه $\begin{cases} x = \bar{x} \\ y = \bar{y} \end{cases}$ بدست می‌آید. همچنین توجه کنید که:

$$a = \rho_{xy}\frac{\sigma_y}{\sigma_x}$$

$$\Rightarrow a\alpha = \rho_{xy}\frac{\sigma_y}{\sigma_x}\,\rho_{xy}\frac{\sigma_x}{\sigma_y} = \rho_{xy}^2$$

$$\alpha = \rho_{xy}\frac{\sigma_x}{\sigma_y}$$

بنابراین $a\alpha = \rho_{xy}^2$ و می‌بایستی مقداری در بازه $[0, 1]$ داشته باشد.

۲–۹ مثال ۲: در مثال ۱ مطلوبست:

الف) محاسبه ضریب همبستگی نمونه با توجه به مقادیر محاسبه شده a و b.

ب) معادله دگرسیون x روی y از روی ضریب همبستگی.

ج) رسم نمودار داده‌ها و دو خط دگرسیون y روی x و x روی y.

د) پیش بینی قد پدر اگر قد فرزند وی ۱۷۹cm باشد.

حل: الف) ضریب همبستگی با توجه به روابط ارایه شده عبارتست از:

$$a=\rho_{xy}\frac{\sigma_y}{\sigma_x}=۰/۴۷۷$$

$$S_{yy}=\sum y_y^{۲}-n\,\bar{y}^{۲}=۳۵۴۲۲۳-۱۲(۱۷۱/۷۵)^{۲}=۲۴۶/۲۵$$

$$\sigma_y=\sqrt{S_{yy}}=\sqrt{۲۴۶/۲۵}=۱۵/۶۹$$

$$\sigma_x=\sqrt{S_{xx}}=\sqrt{۵۲۴/۹۱}=۲۲/۹۱$$

$$\Rightarrow a=۰/۴۷۷=\rho_{xy}\frac{۱۵/۶۹}{۲۲/۹۱}\quad\Rightarrow\quad\rho_{xy}=۰/۶۹۶$$

ب) معادله دگرسیون x روی y با توجه به ρ_{xy} برابر است با:

$$x=\alpha y+\beta=\rho_{xy}\frac{\sigma_x}{\sigma_y}(y-\bar{y})+\bar{x}=۰/۶۹۶\left(\frac{۲۲/۹۱}{۱۵/۶۹}\right)(y-۱۷۱/۷۵)+۱۶۹/۴۱$$

$$\Rightarrow\quad x=۱/۰۱۶\,y-۵/۱۲$$

ج) نمودار داده‌ها و دو خط دگرسیون و روی x و x روی y عبارتست از:

د) با استفاده از خط دگرسیون x روی y مقدار x را برای $y=۱۷۹$ پیش بینی می‌کنیم:

$$x=۱/۰۱۶\,y-۵/۱۲=۱/۰۱۶\,(۱۷۹)-۵/۱۲=۱۷۶/۷۴۴$$

۳ .۱۲ دگرسیون منحنی‌های چند جمله‌ای

با تعمیم روشی که برای برازش داده‌ها بصورت خطی ارایه شد به سادگی می‌توان به داده‌ها یک منحنی چند جمله‌ای بفرم $y=a_n x^n+a_{n-۱}x^{n-۱}+\cdots+a_۱ x+a_۰$ برازش داد. این کار را برای یک منحنی درجه دوم نشان می‌دهیم:

$$Y=a x^{۲}+b x+c$$

$$E=\sum_{i=۱}^{n}(y_i-Y_i)^{۲}=\sum_{i=۱}^{n}(a x_i^{۲}+b x_i+c-y_i)^{۲}$$

حال مجدداً با مشتق‌گرفتن نسبت به مقادیر مجهول a و b و c و برابر صفر قرار دادن معادلات یک دستگاه بدست می‌آوریم که با حل آن a و b و c بدست می‌آیند:

$$\frac{\partial E}{\partial a}=\sum_{i=1}^{n} 2x_i^2\,(a x_i^2+b x_i+c-y_i)=0$$

$$\frac{\partial E}{\partial b}=\sum_{i=1}^{n} 2x_i\,(a x_i^2+b x_i+c-y_i)=0$$

$$\frac{\partial E}{\partial c}=\sum 2\,(a x_i^2+b x_i+c-y_i)=0$$

با داشتن مقادیر نمونه‌های (x_i , y_i) به سادگی می‌توان دستگاه فوق را حل نموده و مقادیر مجهول a و b و c را بدست آورد.

۱۱ – estenbate amari

استنباط آماری بر روی ضرایب دگرسیونی

در بخش قبل بر اسا نمونه $(x_1,y_1)\,,\,(x_2,y_2)\,,\cdots,(x_n,y_n)$ مدل ساده دگرسیونی را معرفی نمودیم و به کمک روش کمترین مربعات $y_i = a + b x_i + E_i$ برای a و b برآوردگرهای $\hat{a}=\bar{y}-\hat{b}\,\bar{x}\quad,\quad \hat{b}=\frac{S_{xy}}{S_{xx}}$ را معرفی نمودیم.

در این بخش می‌خواهیم مدل احتمالی دگرسیون را معرفی کرده و بر اساس آن برخی از فرمهای آماری را بر روی a و b آزمون کنیم.

فرض کنید در مدل ساده دگرسیونی خطا یعنی E_i متغیر تصادفی با توزیع نرمال به فرم زیر باشد:

$$E_i \sim N(0,\sigma^2)$$

همچنین متغیرهای تصادفی E_i به ازای $i=1,2,\cdots,n$ را مستقل از هم در نظر می‌گیریم در این صورت چون Y_i یک ترکیب خطی از متغیر تصادفی E_i می‌باشد بنابراین:

$$Y_i \sim N(a+b x_i,\sigma^2)$$

۱۲ – estenbate amari

محاسبه برآوردگرهای a و b و σ^2

به ازای هر مقدار ثابت X_i تابع چگالی متغیر تصادفی Y_i برابر سات با:

$$f_{Y_i|X_i}(y_i|x_i)=\frac{1}{\sigma\sqrt{2\pi}}\;e^{-\frac{1}{2}\left(\frac{Y_i-(a+b x_i)}{\sigma}\right)^2}\quad -\infty < y_i < \infty$$

می‌توان به کمک روش ماکزیمم درستنمایی که روش دیگری است ۶ برآوردگرهای a و b و σ بذای بدست آوردن برآوردگرهای ماکزیمم درستنمایی پارامترهای a و b و σ به کمک نمونه تصادفی $(X_1,Y_1)\,,\,(X_2,Y_2)\,,\cdots,(X_n,Y_n)$ ابتدا از تابع درستنمایی (یا لگاریتم آن که که ساده‌تر است) نسبت به a و b و σ مشتق می‌گیریم.

عبارتهای حاصل را برابر صفر قرار می‌دهیم، و سپس دستگاه معادلات حاصل را حل می‌کنیم.

بنابراین با مشتق گیری جزیی از:

$$Ln(L)=-n\,Ln(G)-\frac{n}{2}\,Ln(2\pi)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i-(a+b x_i)\right]^2$$

نسبت به a و b و σ و برابر گذاشتن عبارت‌های حاصل ۶ به دست می‌آوریم:

$$\frac{\partial Ln(L)}{\partial a}=\frac{1}{\sigma^{2}}+\sum_{i=1}^{n}\left[Y_i-(a+bx_i)\right]=0$$

$$\frac{\partial Ln(L)}{\partial b}=\frac{1}{\sigma^{2}}+\sum_{i=1}^{n}X_i\left[Y_i-(a+bx_i)\right]=0$$

$$\frac{\partial Ln(L)}{\partial \sigma}=\frac{-n}{\sigma}+\frac{1}{\sigma^{3}}+\sum_{i=1}^{n}\left[Y_i-(a+bx_i)\right]^{2}=0$$

از دو معادله اول $\hat{a}$ و $\hat{b}$ مشابه با برآوردگرهای به روش کمترین مربعات به صورت زیر به دست می‌آید.

$$\hat{b}=\frac{S_{xy}}{S_{xx}}$$

$$\hat{a}=\overline{Y}-\hat{b}\,\overline{X}$$

با قرار دادن $\hat{a}$ و $\hat{b}$ در معادله سوم برآوردگر $\hat{\sigma}^{2}$ برابر می‌شود با:

$$\hat{\sigma}^{2}=\frac{S_{yy}-\hat{b}\,S_{xy}}{n}=\frac{SSE}{n}$$

اگر در براوردگر ماکزیمم درستنمایی $\hat{\sigma}^{2}$، n را به $n-2$ تبدیل کنیم آنگاه برآوردگر $S^{2}=\frac{SSE}{n-2}$ نااریب خواهد شد یعنی در این حالت:

$E\left[S^{2}\right]=E\left[\frac{SSE}{n-3}\right]=\sigma^{2}$ می‌شود.

۱۳ mesal

مثال ۱۳: فرض کنیم یک کمپانی می‌خواهد تأثیر تبلیغات را در فروش کالاهای تولید شده بررسی کند بدین منظور داده‌های زیر را بعد از ۱۰ ماه بدست می‌آورد. این داده‌ها در جدول زیر مرتب شده است.

به کمک داده‌های بدست آمده برآورد a و b و σ^{2} را بدست اورید.

حل:

$$S_{xx} = \sum_{i=1}^{n} x_i^{۲} - \frac{(\sum_{i=1}^{n} x_i)^{۲}}{n} = ۹/۲۸ - \frac{(۹/۴)^{۲}}{۱۰} = ۰/۴۴۴$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} = ۹۲۴/۸ - \frac{(۹/۴)(۹۵۹)}{۱۰} = ۲۳/۳۴$$

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{۹۵۹}{۱۰} = ۹۵/۹ \quad , \quad \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{۹/۴}{۱۰} = ۰/۹۴$$

پس:

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{۲۳/۳۴}{۰/۴۴۴} = ۵۲/۵۶۷۶ \approx ۵۲/۵۷$$

$$\hat{a} = \overline{Y} - \hat{b}\,\overline{X} = ۹۵/۹ - (۵۲/۵۶۷۶)(۰/۹۴) \approx ۴۶/۴۹$$

بنابراین معادله خط دگرسیون به صورت: $y = ۴۶/۴۹ + ۵۲/۵۷\ x$ می‌شود.

۱۴Mesal

برای محاسبه برآوردگر نااریب $\sigma^{۲}$ یعنی $S^{۲}$ به صورت زیر عمل می‌کنیم:

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^{۲} = \sum_{i=1}^{n} y_i^{۲} - \frac{(\sum_{i=1}^{n} y_i)^{۲}}{n} = ۹۳/۵۶۹ - \frac{(۹۵۹)^{۲}}{۱۰} = ۱۶۰۰/۹$$

با جایگذاری S_{yy} و $\hat{b}$ و S_{xy} در فرمول SSE خواهیم داشت:

$$SSE = S_{xy} - \hat{b}\, S_{xy} = ۱۶۰۰/۹ - (۵۲/۵۶۷۶)(۲۳/۳۴) = ۳۷۰۰$$

پس:

$$S^{۲} = \frac{SSE}{n-۲} = \frac{۳۷۳/۹۷}{۸} = ۴۶/۷۵۰$$

حال از برآوردگرهای ماکزیمم ددرستنمایی ضرایب دگرسیون ساده در آزمون فرضهایی درباره a و b و در ساختن فاصله‌های اطمینان برای این پارامترها استفاده می‌کنیم. بدین منظور نتایج زیر را بدون اثبات می‌پذیریم:

(۱) $$\hat{a} \sim N\left(a\ ,\ \frac{\sigma^{۲} \sum_{i=1}^{n} X_i^{۲}}{n\, S_{xx}}\right)$$

(۲) $$\hat{b} \sim N\left(b\ ,\ \frac{\sigma^{۲}}{S_{xx}}\right)$$

(۳) اگر $S^{۲} = \frac{SSE}{n-۲}$ آنگاه $\frac{SSE}{\sigma^{۲}} = \frac{(n-۲)\, S^{۲}}{\sigma^{۲}} \sim X^{۲}_{(n-۲)}$

(۴) $S^{۲}, \hat{a}$ مستقل از هم هستند.

(۵) $S^{۲}, \hat{b}$ مستقل از هم هستند.

۱۵ – estenbate amari

استباط آماری بر روی b

رابطه خطی بین x و y اولین موضوعی است که مورد بررسی قرار می‌گیرد. بنابراین علاقمندیم بدانیم که آیا داده‌ها به اندازه کافی اطلاعات در رابطه با ارتباط خطی بودن y با x می‌دهد یا نه؟ آزمون فرضهای آماری که بر روی پارامتر b ساخته می‌شود ارتباط بین y و x را روشن می‌کند بعنوان مثال اگر y با افزایش یا کاهش x تغییر می‌کند می‌توان فرض $b = ۰$ را در مقابل فرض $b \neq ۰$ ازمون کرد. با توجه به اینکه برآوردگر $\hat{b}$ دارای توزیع $\hat{b} \sim N(b, \frac{\sigma^۲}{S_{xx}})$ می‌باشد می‌توان از تست آماری $T = \frac{\hat{b}-b_۰}{\frac{\sigma}{S_{xx}}} \sim t(n-۲)$.

برای آزمون فرض‌های

$$(۱)\quad \begin{cases} H_۰ : b = b_۰ \\ H_۱ : b \neq b_۰ \end{cases} \qquad (۲)\quad \begin{cases} H_۰ : b = b_۰ \\ H_۱ : b > b_۰ \end{cases} \qquad (۳)\quad \begin{cases} H_۰ : b = b_۰ \\ H_۱ : b < b_۰ \end{cases}$$

استفاده نمود.

نواحی بحرانی برای آزمون فرض‌های بالا در سطح معنی‌دار α را می‌توان در جدول زیر خلاصه نمود.

هم‌چنین می‌توان با استفاده از تست آماری T با ضریب اطمینان $۱ - \alpha$ فاصله اطمینان زیر را برای پارامتر b معرفی نمود.

$$P(\hat{b} - t^{(n-۲)}_{(۱-\frac{\alpha}{۲})} + \frac{S}{\sqrt{S_{xx}}} < b < \hat{b} + t^{(n-۲)}_{(۱-\frac{\alpha}{۲})} \frac{S}{\sqrt{S_{xx}}}) = ۱-\alpha$$

۱۶–مثال ۱-۴

مثال ۴: با توجه به داده‌های مثال ۱ تعیین کنید که آیا پارامتر b اختلاف معنی‌داری از مقدار ۰ دارد یا نه (بوسیله استفاده از یک مدل خطی بین فروش ماهانه و مقدار تبلیغات)

حل: می‌خواهیم آزمون فرض زیر را انجام دهیم:

$$H_۰ : b = b_۰$$
$$H_۱ : b \neq b_۰$$

از تست آماری:

$$T = \frac{\hat{b} - ۰}{\frac{S}{\sqrt{S_{xx}}}} \sim t^{(n-۲)}_{(۱-\frac{\alpha}{۲})} = t^{(۸-۲)}_{(۱-\frac{\alpha}{۲})} = t^{(۶)}_{(۱-\frac{\alpha}{۲})}$$

استفاده می‌کنیم. با استفاده از جدول توزیع T به ازای $\alpha = ۰/۵$ $t^{(۶)}_{۰/۹۷۰} = ۲/۳۰۶$ بنابراین چون:

$$T = \frac{\hat{b}}{\frac{S}{\sqrt{S_{xx}}}} = \frac{۵۲/۵۷}{\frac{۶/۸۴}{\sqrt{۰/۴۴۴}}} = ۵/۱۲ > ۲/۳۰۶$$

باشد پس فرض $H_\circ$ را رد می‌کنیم. بنابراین بر اساس این مشاهدات می‌توان گفت که هزینه‌های مربوط به تبلیغات تأثیر در پیش بینی فروش ماهانه کالا دارد. با توجه به اطلاعات داده شده می‌توان یک فاصله اطمینان با ضریب اطمینان $1-\alpha=۰/۹۵$ نیز برای یک پارامتر b بدست آورد.

جایگذاری مقادیر $\hat{b}$ و S و S_{xx} و $t^{(n-۲)}_{(۱-\frac{\alpha}{۲})}$ در $\hat{b} \pm t^{(n-۲)}_{(۱-\frac{\alpha}{۲})} \times \frac{S}{\sqrt{S_{xx}}}$ اطمینان بدست می‌آید.

$$۵۲/۵۷ \pm ۲/۳۰۶ \frac{۶/۸۴}{\sqrt{۰/۴۴۴}} \quad \Rightarrow \quad ۵۲/۵۷ \pm ۲۳/۶۷ \quad \Rightarrow \quad [۲۸, ۹۰, ۷۶]$$

این فاصله نشان می‌دهد که با احتمال ۰/۹۵ مقدار پارامتر b که بوسیله برآوردگر $\hat{b}$ برآورد می‌شود را دارا می‌باشد.

۱۷-مثال ۲-۴

بطور مشابه می‌توان برای پارامتر a فاصله اطمینان و آزمون فرض ساخت. با توجه به اینکه برآوردگر $\hat{a}$ دارای توزیع $\hat{a} \sim N(a, \frac{\sigma^۲ \sum x_i^۲}{n S_{xx}})$

می‌باشد می‌توان از تست آماری $T = \frac{\hat{a}-a}{S\sqrt{\frac{\sum x_i^۲}{n S_{xx}}}} \sim t(n-۲)$ استفاده نمود و خواص بحرانی را برای فرض‌های آماری مختلف بر روی پارامتر b

در سطح معنی‌دار α در جدول زیر ساخت. همچنین یک فاصله اطمینان با ضریب اطمینان $(۱-\alpha)$ ۱۰۰% برای a بصورت:

$$P(\hat{a} - t^{(n-۲)}_{(۱-\frac{\alpha}{۲})} S\sqrt{\frac{\sum x_i^۲}{n S_{xx}}} < a < \hat{a} + t^{(n-۲)}_{(۱-\frac{\alpha}{۲})} S\sqrt{\frac{\sum x_i^۲}{n S_{xx}}}) = ۱-\alpha = \circ$$

برآورد $E[Y|X]$ مقدار میانگین Y به شرط مقدار داده شده X.

برآورد میانگین Y به شرط مقدار داده شده X یکی از مسائل مهم است که می‌بایست مورد بررسی قرار گیرد. بعنوان مثال ممکن است مسئول حفاظت جان کارگران کارخانه علاقمند باشد که متوسط حوادثی که برای کارگران اتفاق می‌افتد، به ازای ساعات خاصی از آموزش که به کارگران داده می‌شود را برآورد کند.

فرض کنید که x و y یک رابطه خطی بر اساس مدل احتمالی دگرسیونی داشته باشد بطوریکه:

$$E[Y|X] = a + bx$$

دیدیم که تابع چگالی شرطی متغیر تصادفی Y به ازای مقدار داده شده X برابر است با:

$$f_{Y|X}(y|x) = \frac{۱}{\sigma\sqrt{۲\pi}} e^{-\frac{۱}{۲\sigma^۲}(y-(a+bx))^۲} \qquad -\infty < y < \infty$$

۱۸-مثال ۳-۴

حال می‌خواهیم با فرض خطی بودن رابطه بین x و y فاصله اطمینان و آزمون فرض برای پارامتر $E[Y|X] = a + bx$ به ازای مقدار داده شده x بدست آوریم. همانطور که قبلاً اشاره شد $a+bx_p$ میانگین شرطی Y به شرط $X = x_p$ است و از $\hat{Y} = \hat{a} + \hat{b}x_p$ به عنوان یک برآورد برای $a+bx_p$ یاد کردیم. حال می‌خواهیم فرض $H_\circ: E[Y|X=x_p] = E_\circ$ را در مقابل فرض $H_۱: E[Y|X=x_p] \neq E_\circ$ ازمون می‌کنیم

می‌توانیم با توجه به تابع توزیع $\hat{Y}$ تست آماری زیر را برای آزمون فرض فوق معرفی نماییم.

$$T = \frac{\hat{Y} - E_\circ}{S_{\hat{Y}}} = \frac{\hat{Y} - E_\circ}{\sqrt{S^۲\left[\frac{۱}{n} + \frac{(x_p - \bar{x})^۲}{S_{xx}}\right]}} \sim t(n-۲)$$

متغیر تصادفی T دارای توزیع t با $(n-۲)$ درجه آزادی می‌باشد. بنابراین در سطح معنی‌دار α فرض $H_\circ$ زمانی رو می‌شود که $|T| > t^{(n-۲)}_{(۱-\frac{\alpha}{۲})}$ باشد. بطور مشابه می‌توان فرض $H_\circ$ داده شده را در مقابل فرض‌های $H_۱: E\left[Y|X=x_p\right] > E_\circ$ یا $H_۱: E\left[Y|X=x_p\right] < E_\circ$ آزمون نمود که در این صورت فرض $H_\circ$ به ترتیب زمانی رد خواهد شد که داشته باشیم:

$$T < -t^{(n-۲)}_{(۱-\frac{\alpha}{۲})} \quad , \quad T > -t^{(n-۲)}_{(۱-\frac{\alpha}{۲})}$$

از تست آماری فرض شده می‌توان برای پیدا کردن فاصله اطمینان برای $E\left[Y|X=x_p\right]$ نیز استفاده نمود.

$$P\left(\hat{Y} - t^{(n-۲)}_{(۱-\frac{\alpha}{۲})}\sqrt{S^۲\left[\frac{۱}{n} + \frac{(x_p - \bar{x})^۲}{S_{xx}}\right]} < a + b\,x_p\,(\hat{Y} + t^{(n-۲)}_{(۱-\frac{\alpha}{۲})})\sqrt{S^۲\left[\frac{۱}{n} + \frac{(x_p - \bar{x})^۲}{S_{xx}}\right]}\right.$$

۱۹-مثال

مثال: از داده‌های مثال ۲ استفاده کرده و یک فاصله اطمینان با ضریب اطمینان ٪۹۵ برای $X = ۱/۰$ پیدا کنید.

حل:

ابتدا $\hat{Y}$ را در نقطه $X_p = ۱/۰$ تخمین می‌زنیم:

$$\hat{Y} = \hat{a} + \hat{b}\,x_p$$

در مثالهای قبل $\hat{a} = ۴۶/۴۹$ و $\hat{b} = ۵۲/۵۷$ را بدست آوردیم.

پس:

$$\hat{Y} = ۴۶/۴۹ + (۵۲/۵۷)(۱/۰) = ۹۹/۰۶$$

فرمول مناسب برای فاصله اطمینان برابر بود با:

$$\hat{Y} \pm t^{(n-۲)}_{(۱-\frac{\alpha}{۲})}\sqrt{S^۲\left[\frac{۱}{n} + \frac{(x_p - \bar{x})^۲}{S_{xx}}\right]}$$

با جایگذاری مقادیر مناسب در عبارت فوق که قبلاض محاسبه شده است خواهیم داشت:

$$۹۹/۰۶ \pm (۲/۳۰۶)\sqrt{۴۶/۷۵\left[\frac{(۱/۰ - ۰/۹۴)}{۰/۴۴}\right]}$$

بعد از محاسبات انجام شده حاصل می‌شود:

$$۹۹/۰۶ \pm ۵/۱۸ \qquad \Rightarrow \qquad [۹۳/۸۸ \,,\, ۱۰/۲۴]$$

می‌توان از محاسبات فوق نتیجه گرفت که برای هر واحد هزینه تبلیغات $(\$\ ۱۰/۰۰۰)$ متوسط فروش ماهانه با احتمال ٪۹۵ در فاصله $ ۹۳۸۸۰۰ و $ ۱۰۴۲۴۰۰ خواهد بود. در شکل زیر برای $E\left[Y|X\right]$ رسم شده است.

۲۰- Zarebe

ضریب همبستگی

در بسیاری از اوقات نیاز به شاخصی داریم که چگونگی ارتباط بین دو متغیر x و y را اندازه بگیرد. این شاخص ضریب همبستگی خطی بین دو متغیر x و y نامیده می‌شود.

ضریب همبستگی خطی نمونه‌ای برآورد نامناسب برای ضریب همبستگی خطی تعریف شده در فصل‌های قبل است این معیار میزان قوت ارتباط خطی بین متغیرهای x و y را اندازه می‌گیرد و اولین بار دانشمند معروفی انگلیسی به نام کارل پیرسون آنرا معرفی نمود از این جهت به آن ضریب همبستگی خطی پیرسون نیز می‌گویطم و به فرم زیر تعریف می‌شود:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}}$$