

## فصل دوازدهم

S-1

### 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 متغیرها از روی مقادیر متغیر دیگر است که به مساله برگشت یا دگرسیون معروف می‌باشد.

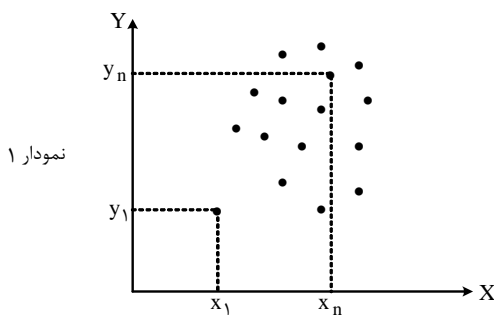
به عنوان مثال فرض کنید بخواهیم تاثیر میزان مصرف شیر را در افزایش قد بدست بیاوریم و یا بخواهیم میزان وزن فرزند را از روی وزن پدرش پیش بینی کنیم. ملاحظه می‌کنید که در این گونه مسایل دو متغیر تصادفی مورد مطالعه به نوعی به یکدیگر وابسته می‌باشند. به عبارت دقیقتر یک متغیر تصادفی مثل  $X$  را مستقل و متغیر 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  $n$  تایی از متغیر تصادفی  $X$  جمع آوری می‌کنیم که نتایج آن بصورت  $X_1, X_2, \dots, X_n$  می‌باشند. سپس مقادیر متناظر با هر یک از نمونه‌های بدست آمده ( $X_i$  ها) را که همان مقادیر معادل متغیر تصادفی وابسته  $Y$  می‌باشند بدست می‌آوریم. به این ترتیب برای  $X_i$  ها مقادیر متناظر  $Y_1, Y_2, \dots, Y_n$  بدست می‌آیند. که می‌توانیم نتیجه را بصورت زوج‌های مرتب زیر نشان دهیم.

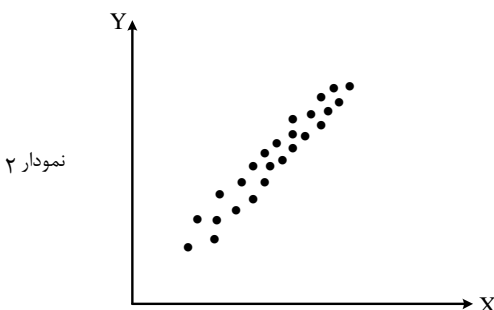
$$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots, (X_n, Y_n)$$

**۲- دگرسیون-۲** به عنوان مثال مقادیر  $X_i$  ها می‌توانند میزان طول قد افراد یک جامعه و مقادیر  $Y_i$  ها میزان مصرف شیر هر یک از نمونه‌ها باشند. به این ترتیب زوج مرتب  $(180, 2)$  بیانگر این است که در نمونه‌گیری یکی از افراد جامعه دارای صول قد  $180$  cm بوده است و وی روزانه دو لیوان شیر مصرف کرده است.

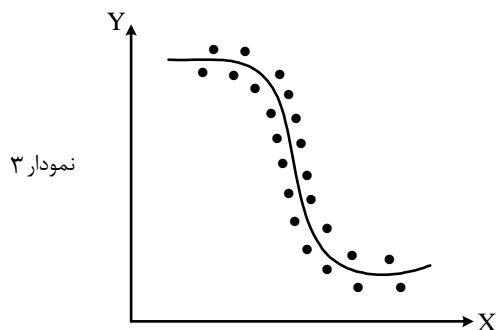
پس از بدست آمدن زوج‌های مرتب  $(X_i, Y_i)$  ملاحظه می‌نماید که هر زوج مرتب می‌تواند معادل یک نقطه در صفحه باشد. با رسم نقاط مورد نظر در صفحه یک تصویر کلی از رابطه  $X$  و  $Y$  بدست می‌آوریم. به شکل‌های زیر که برای سه نمونه جداگانه می‌باشند توجه کنید:



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.



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



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

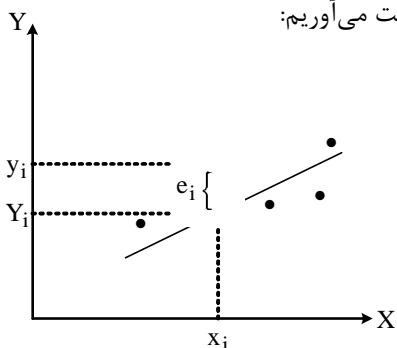
با توجه به دو نمودار ۲ و ۳ این طور به نظر می‌رسد که در حالت کلی دو متغیر تصادفی  $X$  و  $Y$  اگر از یکدیگر مستقل نباشند یا بصورت خطی و یا بصورت غیر خطی به یکدیگر وابسته می‌باشند.

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 برداشش منحنی را مطرح می‌کنیم.

### ۱- دگرسیون خطی - ۱.۱۲.۳

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 برداشش 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  $a$  و  $b$  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:

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

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

به این ترتیب دستگاه معادلات زیر بدست می آید:

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

با حل این دستگاه مقادیر مجهول  $a$  و  $b$  بدست می آیند. که عبارتند از:

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

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  $X$  می گویند. یعنی  $Y$  از روی  $X$  بدست آمده است.

**۱-مثال 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) بردارش خطی داده ها

ب) رسم نمودار داده ها و خط دگرسیون.

ج) با توجه به معادله دگرسیون پیش بینی کنید که اگر قد پدری ۱۸۵cm باشد قد فرزند او چقدر خواهد بود؟

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

x	y	x <sup>2</sup>	xy	y <sup>2</sup>
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

$$\bar{X} = \frac{1}{n} \sum x_i = \frac{1}{12} 2033 = 169/41$$

$$\bar{Y} = \frac{1}{n} \sum y_i = \frac{1}{12} 2061 = 171/75$$

$$S_{xy} = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} = 349418 - 12(169/41)(171/75) = 250/25$$

$$S_{xx} = \sum x_i^2 - n \bar{x}^2 = 344949 - 12(169/41)^2 = 524/91$$

$$a = \frac{S_{xy}}{S_{xx}} = \frac{250/25}{524/91} = 0/477$$

$$b = \bar{Y} - a \bar{X} = 171/75 - 0/477(169/41) = 90/9$$

بنابراین معادله خط دگرسیون عبارتست از:

$$y = 0/477 x + 90/9$$

(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.

$$y = 0/477 \times 185 + 90/9 = 179/145$$

## ۳-۱۲. ضریب همبستگی و دگرسیون

۷-۱

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:

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

توجه کنید که در اینجا در محاسبه  $\rho_{xy}$  و  $\sigma_x$  و  $\sigma_y$  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:

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

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

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

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 ۱۷۹ cm باشد.

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

$$a = \rho_{xy} \frac{\sigma_y}{\sigma_x} = 0.477$$

$$S_{yy} = \sum y_y^2 - n\bar{y}^2 = 354223 - 12(171/75)^2 = 246/25$$

$$\sigma_y = \sqrt{S_{yy}} = \sqrt{246/25} = 15/69$$

$$\sigma_x = \sqrt{S_{xx}} = \sqrt{524/91} = 22/91$$

$$\Rightarrow a = 0.477 = \rho_{xy} \frac{15/69}{22/91} \Rightarrow \rho_{xy} = 0.696$$

ب) معادله دگرسیون X روی Y با توجه به  $\rho_{xy}$  برابر است با:

$$x = \alpha y + \beta = \rho_{xy} \frac{\sigma_x}{\sigma_y} (y - \bar{y}) + \bar{x} = 0.696 \left( \frac{22/91}{15/69} \right) (y - 171/75) + 169/41$$

$$\Rightarrow x = 1/0.16 y - 5/12$$

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  $y = 179$  پیش بینی می‌کنیم:

$$x = 1/0.16 y - 5/12 = 1/0.16 (179) - 5/12 = 176/744$$

### 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

$$y = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$$

برازش داد. این کار را برای یک منحنی درجه دوم نشان می‌دهیم:

$$Y = a x^2 + b x + c$$

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

حال مجدداً با مشتق گرفتن 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:

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

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

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

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 اساس نمونه  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  مدل ساده دگرسیونی را معرفی نمودیم و به کمک روش کمترین مربعات

$$y_i = a + b x_i + E_i \quad \text{برای } a \text{ و } b \text{ برآوردگرهای } \hat{a} = \bar{y} - \hat{b} \bar{x} \quad , \quad \hat{b} = \frac{S_{xy}}{S_{xx}} \text{ را معرفی نمودیم.}$$

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  $i = 1, 2, \dots, n$  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 $\sigma^2$

به ازای هر مقدار ثابت  $X_i$  تابع چگالی متغیر تصادفی  $Y_i$  برابر است 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  $\sigma$  بذای بدست آوردن برآوردگرهای ماکزیمم درستنمایی پارامترهای  $a$  و  $b$  و  $\sigma$  به کمک نمونه تصادفی  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$  ابتدا از تابع درستنمایی (یا لگاریتم آن که که ساده‌تر است) نسبت به  $a$  و  $b$  و  $\sigma$  مشتق می‌گیریم.

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

بنابراین با مشتق گیری جزئی از:

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

نسبت به  $a$  و  $b$  و  $\sigma$  و برابر گذاشتن عبارت‌های حاصل ۶ به دست می‌آوریم:

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

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

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

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

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

$$\hat{a} = \bar{Y} - \hat{b} \bar{X}$$

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

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

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

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

### 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  $\sigma^2$  را بدست آورید.

حل:



$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} = 9/28 - \frac{(9/4)^2}{10} = 0.444$$

$$S_{xy} = \sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n} = 924/8 - \frac{(9/4)(959)}{10} = 23/34$$

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} = \frac{959}{10} = 95.9, \quad \bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{9/4}{10} = 0.94$$

پس:

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{23/34}{0.444} = 52/5676 \approx 52/57$$

$$\hat{a} = \bar{Y} - \hat{b}\bar{X} = 95.9 - (52/5676)(0.94) \approx 46/49$$

بنابراین معادله خط دگرسیون به صورت:  $y = 46/49 + 52/57 x$  می‌شود.

#### ۱۴ Mesal

برای محاسبه برآوردگر نارایب  $\sigma^2$  یعنی  $S^2$  به صورت زیر عمل می‌کنیم:

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} = 93/569 - \frac{(959)^2}{10} = 1600/9$$

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

$$SSE = S_{xy} - \hat{b} S_{xy} = 1600/9 - (52/5676)(23/34) = 3700$$

پس:

$$S^2 = \frac{SSE}{n-2} = \frac{373/97}{8} = 46/750$$

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

$$\hat{a} \sim N\left(a, \frac{\sigma^2 \sum_{i=1}^n X_i^2}{n S_{xx}}\right) \quad (1)$$

$$\hat{b} \sim N\left(b, \frac{\sigma^2}{S_{xx}}\right) \quad (2)$$

$$\frac{SSE}{\sigma^2} = \frac{(n-2) S^2}{\sigma^2} \sim X_{(n-2)}^2 \quad \text{اگر } S^2 = \frac{SSE}{n-2} \quad (3)$$

(4)  $\hat{a}$ ,  $S^2$  مستقل از هم هستند.

(5)  $\hat{b}$ ,  $S^2$  مستقل از هم هستند.

## ۱۵ – estenbate amari

### استنباط آماری بر روی $b$

رابطه خطی بین  $X$  و  $Y$  اولین موضوعی است که مورد بررسی قرار می‌گیرد. بنابراین علاقمندییم 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  $b = 0$  را در مقابل فرض  $b \neq 0$  آزمون کرد. با توجه به اینکه برآوردگر  $\hat{b}$  دارای توزیع

$$\hat{b} \sim N\left(b, \frac{\sigma^2}{S_{XX}}\right) \quad T = \frac{\hat{b} - b_0}{\frac{S}{\sqrt{S_{XX}}}} \sim t(n-2)$$

برای آزمون فرضهای

$$\begin{array}{lll} H_0: b = b_0 & H_0: b = b_0 & H_0: b = b_0 \\ (1) & (2) & (3) \\ H_1: b \neq b_0 & H_1: b > b_0 & H_1: b < b_0 \end{array}$$

استفاده نمود.

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

هم‌چنین می‌توان با استفاده از تست آماری  $T$  با ضریب اطمینان  $1 - \alpha$  فاصله اطمینان زیر را برای پارامتر  $b$  معرفی نمود.

$$P\left(\hat{b} - t_{\frac{1-\alpha}{2}}^{(n-2)} \frac{S}{\sqrt{S_{XX}}} < b < \hat{b} + t_{\frac{1-\alpha}{2}}^{(n-2)} \frac{S}{\sqrt{S_{XX}}}\right) = 1 - \alpha$$

### 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:

$$H_0: b = b_0$$

$$H_1: b \neq b_0$$

از تست آماری:

$$T = \frac{\hat{b} - b_0}{\frac{S}{\sqrt{S_{XX}}}} \sim t_{\frac{1-\alpha}{2}}^{(n-2)} = t_{\frac{1-\alpha}{2}}^{(8-2)} = t_{\frac{1-\alpha}{2}}^{(6)}$$

استفاده می‌کنیم. با استفاده از جدول توزیع  $T$  به ازای  $\alpha = 0.05$   $t_{0.025}^{(6)} = 2.306$  بنابراین چون:

$$T = \frac{\hat{b}}{S} = \frac{52/57}{\frac{6/84}{\sqrt{0.444}}} = 5/12 > 2/3.06$$

باشد پس فرض  $H_0$  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  $1-\alpha=0.95$  نیز برای یک پارامتر  $b$  بدست آورد.

جایگذاری مقادیر  $\hat{b}$  و  $S$  و  $S_{xx}$  و  $t_{(1-\frac{\alpha}{2})}^{(n-2)}$  در  $\hat{b} \pm t_{(1-\frac{\alpha}{2})}^{(n-2)} \times \frac{S}{\sqrt{S_{xx}}}$  اطمینان بدست می‌آید.

$$52/57 \pm 2/3.06 \frac{6/84}{\sqrt{0.444}} \Rightarrow 52/57 \pm 23/67 \Rightarrow [28, 90, 76]$$

این فاصله نشان می‌دهد که با احتمال  $0.95$  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  $\hat{a} \sim N(a, \frac{\sigma^2 \sum x_i^2}{nS_{xx}})$

می‌باشد می‌توان از تست آماری  $T = \frac{\hat{a}-a}{S \sqrt{\frac{\sum x_i^2}{nS_{xx}}}} \sim t_{(n-2)}$  استفاده نمود و خواص بحرانی را برای فرض‌های آماری مختلف بر روی پارامتر  $b$

در سطح معنی‌دار  $\alpha$  در جدول زیر ساخت. همچنین یک فاصله اطمینان با ضریب اطمینان  $(1-\alpha) 100\%$  برای  $a$  بصورت:

$$P\left(\hat{a} - t_{(1-\frac{\alpha}{2})}^{(n-2)} S \sqrt{\frac{\sum x_i^2}{nS_{xx}}} < a < \hat{a} + t_{(1-\frac{\alpha}{2})}^{(n-2)} S \sqrt{\frac{\sum x_i^2}{nS_{xx}}}\right) = 1-\alpha = 0$$

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:

$$f_{Y|X}(y|x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-(a+bx))^2} \quad -\infty < y < \infty$$

### 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  $H_0: E[Y|X=x_p] = E_0$  را در مقابل فرض  $H_1: E[Y|X=x_p] \neq E_0$  آزمون می‌کنیم

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

$$T = \frac{\hat{Y} - E_0}{S_{\hat{Y}}} = \frac{\hat{Y} - E_0}{\sqrt{S^2 \left[ \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{XX}} \right]}} \sim t(n-2)$$

متغیر تصادفی  $T$  دارای توزیع  $t$  با  $(n-2)$  درجه آزادی می‌باشد. بنابراین در سطح معنی‌دار  $\alpha$  فرض  $H_0$  زمانی رو می‌شود که  $|T| > t_{\frac{\alpha}{2}}^{(n-2)}$

باشد. بطور مشابه می‌توان فرض  $H_0$  داده شده را در مقابل فرض‌های  $H_1: E[Y|X=x_p] < E_0$  یا  $H_1: E[Y|X=x_p] > E_0$  که در این صورت فرض  $H_0$  به ترتیب زمانی رد خواهد شد که داسته باشیم:

$$T < -t_{\frac{\alpha}{2}}^{(n-2)}, \quad T > t_{\frac{\alpha}{2}}^{(n-2)}$$

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

$$P\left(\hat{Y} - t_{\frac{\alpha}{2}}^{(n-2)} \sqrt{S^2 \left[ \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{XX}} \right]} < a + b x_p < \hat{Y} + t_{\frac{\alpha}{2}}^{(n-2)} \sqrt{S^2 \left[ \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{XX}} \right]}\right)$$

### ۱۹-مثال

**مثال:** از داده‌های مثال ۲ استفاده کرده و یک فاصله اطمینان با ضریب اطمینان ۹۵٪ برای  $X=1/0$  پیدا کنید.

حل:

ابتدا  $\hat{Y}$  را در نقطه  $X_p=1/0$  تخمین می‌زنیم:

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

در مثالهای قبل  $\hat{a}=46/49$  و  $\hat{b}=52/57$  را بدست آوردیم.

$$\hat{Y} = 46/49 + (52/57)(1/0) = 99/06$$

پس:

فرمول مناسب برای فاصله اطمینان برابر بود با:

$$\hat{Y} \pm t_{\frac{\alpha}{2}}^{(n-2)} \sqrt{S^2 \left[ \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{XX}} \right]}$$

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

$$99/06 \pm (2/306) \sqrt{46/75 \left[ \frac{(1/0 - 0/94)}{0/44} \right]}$$

بعد از محاسبات انجام شده حاصل می‌شود:

$$99/06 \pm 5/18 \Rightarrow [93/88, 10/24]$$

می‌توان از محاسبات فوق نتیجه گرفت که برای هر واحد هزینه تبلیغات ( $\$ 10/000$ ) متوسط فروش ماهانه با احتمال ۹۵٪ در فاصله  $\$ 938800$  و  $\$ 1042400$  خواهد بود. در شکل زیر برای  $E[Y|X]$  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 می‌گوییم و به فرم زیر تعریف می‌شود:

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