

فصل نهم

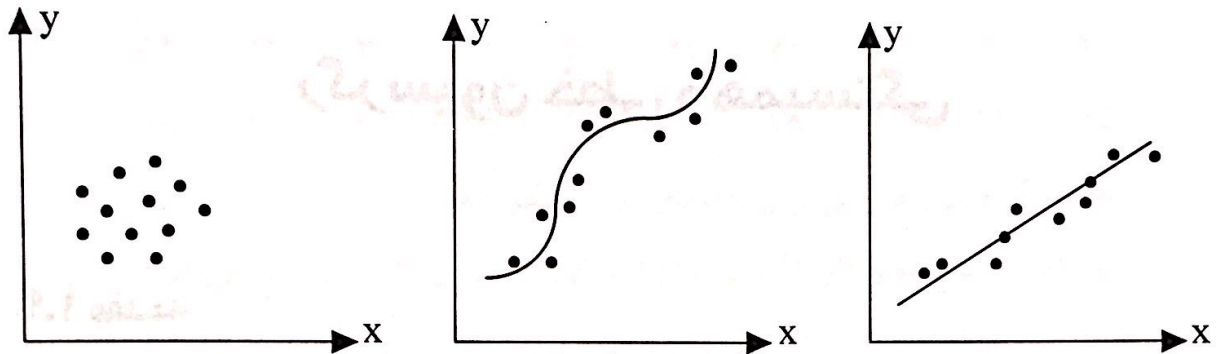
رگرسیون خطی و همبستگی

۱.۹ مقدمه

در اغلب بررسیهای آماری نیاز به پیش بینی مقدار یک متغیر وابسته از روی مقدار یک متغیر مستقل می باشد. برای مثال پیش بینی طول قد فرزند از روی طول قد پدرش، یا پیش بینی معدل کل یک دانشجو در یک نیمسال از روی نمره درس ریاضیات او و یا پیش بینی مقدار مصرف بنزین از روی مسافت طی شده اتومبیل از این نوع مسائل می باشند. چنین مسائلی را مسائل برگشت یا رگرسیون گویند. متغیر مستقل را با x و متغیر وابسته را با $Y | x$ و یا برای راحتی با Y نمایش می دهند. برای مثال در پیش بینی طول قد فرزند از روی طول قد پدرش، طول قد پدر را که یک مقدار ثابت و بخصوص است با x نمایش می دهیم و چون برای یک طول قد پدر x فرزندان او می توانند طول قد های متفاوت داشته باشند بنابراین طول قد فرزند او را با متغیر تصادفی $Y | x$ یا Y نشان می دهیم. به همین ترتیب در پیش بینی مقدار مصرف بنزین، برای مسافت معین طی شده x میزان مصرف بنزین را با Y نمایش می دهیم.

برای یافتن رابطه بین متغیر مستقل x و متغیر وابسته Y ابتدا یک نمونه تصادفی از جمعیت مورد نظر جمع آوری می کنیم. یعنی به ازاء مقادیر x_1, x_2, \dots, x_n و x_n از متغیر مستقل x مقادیر مربوط به متغیر وابسته Y را اندازه گیری می کنیم. فرض کنید این مقادیر اندازه گیری شده y_1, y_2, \dots, y_n و y_n باشند. بنابراین نمونه تصادفی ما به صورت زوجهای $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ است که مقادیر مشاهده شده آن عبارت است از $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. برای پی بردن به رابطه

بین x و Y ابتدا این مشاهدات که به صورت نقاطی در صفحه هستند را در یک دستگاه مختصات رسم می‌کنیم که به آن نمودار پراکندگی گویند. در شکل ۱.۹ نمودار پراکندگی نقاط برای حالت‌های مختلف رسم شده است. توجه کنید که برای یک مقدار x ممکن است چندین مقدار برای Y وجود داشته باشد.



الف - x و Y رابطه خطی دارند ب - x و Y رابطه غیر خطی دارند ج - x و Y رابطه‌ای ندارند

شکل ۱.۹ نمودارهای پراکندگی نقاط

حال برای این نقاط می‌توان یک خط یا منحنی عبور داد و این خط یا منحنی رابطه بین x و Y را مشخص می‌کند. بنابراین در حالت کلی اگر بخواهیم مقدار متغیر Y را از روی مقدار متغیر x پیش بینی کنیم احتیاج به یک رابطه بین x و Y داریم که این رابطه یک معادله پیش بینی کننده است که به آن معادله رگرسیون Y روی x گویند.

۲.۹ رگرسیون ساده خطی

هر گاه بین متغیر مستقل x و متغیر وابسته $Y = Y | x$ یک رابطه خطی برقرار باشد گوئیم یک مدل رگرسیون ساده خطی بین x و Y برقرار است. برای تشکیل این رابطه خطی، فرض کنید یک نمونه تصادفی $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ با مقادیر مشاهده شده $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ داشته باشیم. توجه کنید که $Y_i = Y | x_i, i = 1, 2, \dots, n$. منظور از رگرسیون خطی این است که میانگین $Y | x$ به طور خطی با x در ارتباط باشد یعنی $\mu_{Y|x} = E(Y | x) = \alpha + \beta x$ که به آن خط رگرسیون گویند و در آن α و β پارامترهای نامعلوم هستند که بایستی برآورد شوند. به α و β ضرایب رگرسیون گویند اگر برآورد α و β را با $\hat{\alpha}$ و $\hat{\beta}$ نمایش دهیم در این صورت

مقدار برآورد متغیر وابسته y را با $\hat{y} = \hat{\alpha} + \hat{\beta}x$ نمایش می‌دهند. در زیر روشی برای برآورد ضرایب رگرسیونی α و β از روی نمونه ارائه می‌دهیم و از روی آن بوسیله $\hat{y} = \hat{\alpha} + \hat{\beta}x$ برای مقدار بخصوص x مقدار متغیر وابسته y را پیش بینی می‌کنیم.

در نمونه تصادفی $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ چون همواره مقدار $Y_i = Y | x_i$ برابر $\mu_{Y|x_i} = E(Y | x_i)$ نیست، بنابراین اختلاف آنها برابر یک مقدار تصادفی E_i می‌باشد یعنی

$$Y_i = \mu_{Y|x_i} + E_i = \alpha + \beta x_i + E_i, \quad i = 1, 2, \dots, n \quad (1.9)$$

این مدل را مدل رگرسیون ساده خطی گویند و E_i را که یک متغیر تصادفی با میانگین صفر و واریانس σ^2 است را مقدار خطا گویند. اگر مقدار مشاهده شده E_i را با e_i نشان دهیم، در این صورت

$$y_i = \alpha + \beta x_i + e_i, \quad i = 1, 2, \dots, n$$

که e_i را مقدار باقیمانده گویند. حال برای یافتن برآوردهای $\hat{\alpha}$ و $\hat{\beta}$ بگونه‌ای عمل می‌کنیم که مجموع مربعات باقیمانده‌ها یعنی $\sum_{i=1}^n e_i^2$ می‌نیمم گردد.

روش حداقل مربعات مجموع مربعات باقیمانده‌ها را معمولاً مجموع مربعات خطاها حول خط رگرسیون گویند و با SSE^(۱) نمایش می‌دهند، یعنی

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \quad (2.9)$$

مقادیری از $\hat{\alpha}$ و $\hat{\beta}$ که SSE را می‌نیمم کند، برآوردهای حداقل مربعات گویند و روش به دست آوردن آنها را روش حداقل مربعات^(۲) می‌نامند که در زیر به ذکر آن می‌پردازیم.

ابتدا کمیت‌های زیر را معرفی می‌کنیم

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

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

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

قضیه ۱.۹ در یک مدل رگرسیون ساده خطی مقادیر $\hat{\alpha}$ و $\hat{\beta}$ که مجموع مربعات خطاها را می‌نیم می‌کنند عبارت اند از

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \hat{\beta} = \frac{S_{xy}}{S_{xx}} \quad (۳.۹)$$

اثبات با مشتق‌گیری از SSE نسبت به α و β و مساوی صفر قرار دادن آنها به معادلات زیر

$$\begin{cases} \sum_{i=1}^n y_i - n\alpha - \beta \sum_{i=1}^n x_i = 0 \\ \sum_{i=1}^n x_i y_i - \alpha \sum_{i=1}^n x_i - \beta \sum_{i=1}^n x_i^2 = 0 \end{cases} \quad \text{می‌رسیم}$$

که از حل این دستگاه جوابهای $\hat{\alpha}$ و $\hat{\beta}$ داده شده در (۳.۹) حاصل می‌شوند و می‌توان نشان داد که این مقادیر SSE را می‌نیم می‌کنند.

بنابراین خط رگرسیون $\mu_{Y|X} = \alpha + \beta x$ بوسیله خط $\hat{y} = \hat{\alpha} + \hat{\beta}x$ برآورد می‌شود، که $\hat{\alpha}$ و $\hat{\beta}$ از رابطه (۳.۹) به دست می‌آیند.

مثال ۱.۲.۹ برای داده‌های جدول زیر برآورد خط رگرسیون را بیابید. سپس نقاط را در صفحه مشخص نموده و خط برآورد شده را رسم کنید.

x_i	۱	۳	۴	۶	۸	۹	۱۱	۱۴
y_i	۱	۲	۴	۴	۵	۷	۸	۹

حل از جدول فوق مقادیر زیر حاصل می‌شوند

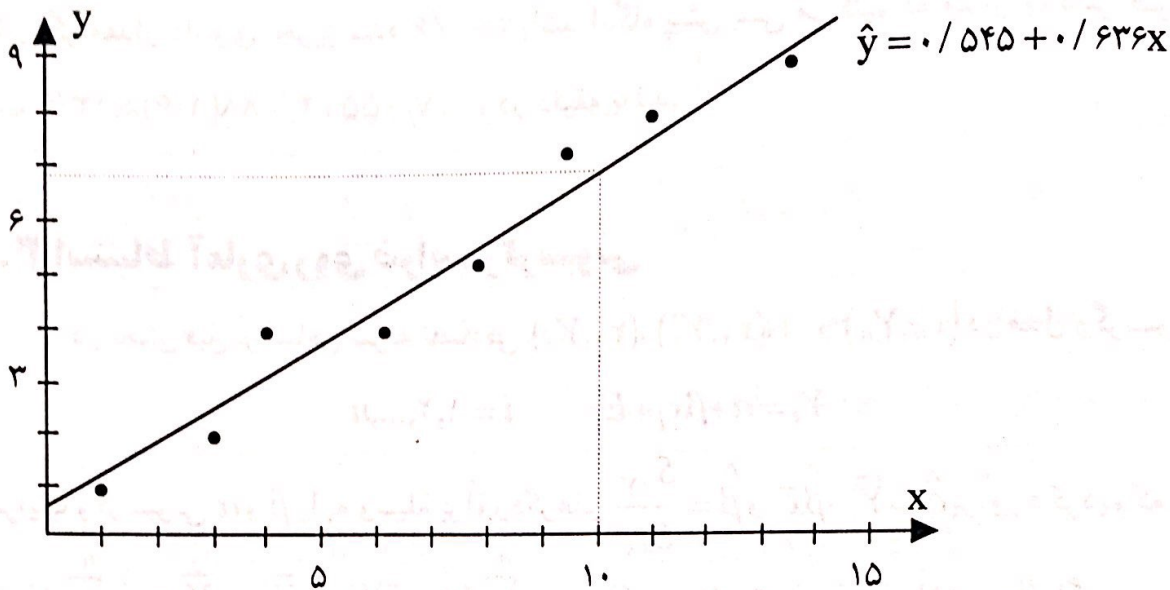
$$\sum x_i = 56, \quad \sum y_i = 40, \quad \sum x_i^2 = 524, \quad \sum y_i^2 = 256, \quad \sum x_i y_i = 364$$

$$S_{xy} = 364 - \frac{(56)(40)}{8} = 84, \quad S_{xx} = 524 - \frac{(56)^2}{8} = 132 \quad \text{بنابراین}$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{84}{132} = 0.636, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{40}{8} - (0.636)\frac{56}{8} = 0.545$$

و در نتیجه

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = 0.545 + 0.636x$$



شکل ۲.۹ نمودار پراکنندگی نقاط و خط رگرسیون برآورد شده

مثال ۲.۲.۹ آزمایشی به منظور مطالعه اثر یک داروی معین در پایین آوردن ضربان قلب در افراد بالغ انجام شده است. مقدار داروی تجویز شده بر حسب میلی گرم و تفاوت ضربان قلب پس از استعمال دارو و قبل از آن برای یک نمونه ۱۳ تایی در جدول زیر آورده شده است

مقدار داروی تجویز شده	۰/۵	۰/۷۵	۱/۰	۱/۲۵	۱/۵	۱/۷۵	۲/۰
واکنش در ضربان قلب	۱۰	۸	۱۲	۱۲	۱۴	۱۲	۱۶
مقدار داروی تجویز شده	۲/۲۵	۲/۵	۲/۷۵	۳	۳/۲۵	۳/۵	
واکنش در ضربان قلب	۱۸	۱۷	۲۰	۱۸	۲۰	۲۱	

برآورد خط رگرسیون را به دست آورید. اگر مقدار داروی تجویز شده ۱/۶ باشد، واکنش در ضربان قلب در دقیقه را به چه میزان پیش بینی می کنید؟

حل از جدول فوق مقادیر زیر حاصل می شوند:

$$\sum x_i = 26, \sum x_i^2 = 63/375, \sum y_i = 198, \sum y_i^2 = 3226, \sum x_i y_i = 442/5$$

$$S_{xy} = 442/5 - \frac{(26)(198)}{13} = 46/5, S_{xx} = 63/375 - \frac{(26)^2}{13} = 11/375$$

بنابراین

$$\hat{\beta} = \frac{46/5}{11/375} = 4/0.88, \hat{\alpha} = \frac{198}{13} - (4/0.88) \frac{26}{13} = 7/0.55$$

در نتیجه

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = 7/0.55 + 4/0.88x$$

حال اگر مقدار داروی تجویز شده $x = 1/6$ باشد آنگاه پیش بینی می کنیم که میزان واکنش ضربان قلب $\hat{y} = 7/0.55 + 4/0.88(1/6) \approx 13/6$ در دقیقه باشد.

۳.۹ استنباط آماری روی ضرایب رگرسیونی

در بخش قبل بر اساس نمونه تصادفی $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ در مدل رگرسیونی

$$Y_i = \alpha + \beta x_i + E_i \quad i = 1, 2, \dots, n$$

ضرایب رگرسیونی α و β را به وسیله برآوردهای $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$ و $\hat{\beta} = \frac{S_{xY}}{S_{xx}}$ برآورد کردیم که در

آن $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ و $S_{xY} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})$ دو متغیر تصادفی می باشند. حال اگر در مدل

رگرسیونی فوق فرض کنیم $E_i \sim N(0, \sigma^2)$ ، $i = 1, 2, \dots, n$ و E_n از یکدیگر مستقل

باشند، در این صورت چون Y_i یک ترکیب خطی از متغیرهای تصادفی نرمال E_i می باشد، پس

$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$ ، $i = 1, 2, \dots, n$ بر اساس توزیع متغیرهای تصادفی Y_i می توان توزیع $\hat{\alpha}$ ، $\hat{\beta}$

و SSE را به دست آورد. توزیع این برآوردها را در قضیه زیر بدون اثبات می آوریم.

قضیه ۲.۹ در مدل رگرسیونی ساده خطی با فرض $E_i \sim N(0, \sigma^2)$ ، $i = 1, 2, \dots, n$ داریم که

$$\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2 \sum x_i^2}{nS_{xx}}\right) \quad \text{الف}$$

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

$$\text{ج- اگر } S^2 = \frac{SSE}{n-2} \text{ آنگاه } S^2 \sim \chi_{(n-2)}^2 \text{ و } \frac{SSE}{\sigma^2} = \frac{(n-2)S^2}{\sigma^2}$$

د- $\hat{\alpha}$ و S^2 و همچنین $\hat{\beta}$ و S^2 از یکدیگر مستقل هستند.

با استفاده از قضیه ۲.۹ می توان نتایج زیر را به دست آورد

$$\frac{\hat{\beta} - \beta}{\frac{S}{\sqrt{S_{xx}}}} \sim t_{(n-2)} \quad (4.9)$$

$$\frac{\hat{\alpha} - \alpha}{S \sqrt{\frac{\sum x_i^2}{n S_{xx}}}} \sim t_{(n-2)}$$

(۵.۹)

که در آن

$$S^2 = \frac{SSE}{n-2} = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n-2} [S_{YY} - \hat{\beta}S_{XY}]$$

(۶.۹)

با استفاده از تابع محورهای (۴.۹) و (۵.۹) می توان فواصل اطمینان $(1-\alpha) \cdot 100\%$ برای α و

β به دست آورد که این فواصل اطمینان عبارتند از

$$\alpha \in \left(\hat{\alpha} - t_{1-\frac{\alpha}{2}}(n-2) S \sqrt{\frac{\sum x_i^2}{n S_{xx}}}, \hat{\alpha} + t_{1-\frac{\alpha}{2}}(n-2) S \sqrt{\frac{\sum x_i^2}{n S_{xx}}} \right)$$

$$\beta \in \left(\hat{\beta} - t_{1-\frac{\alpha}{2}}(n-2) \frac{S}{\sqrt{S_{xx}}}, \hat{\beta} + t_{1-\frac{\alpha}{2}}(n-2) \frac{S}{\sqrt{S_{xx}}} \right)$$

(۷.۹)

با استفاده از روابط (۴.۹) و (۵.۹) می توان آزمونهای آماری را روی α و β انجام داد که

این آزمونها در جدول ۱.۹ آورده شده اند.

H_0	آماره آزمون	H_1	ناحیه بحرانی
$\alpha = \alpha_0$	$T = \frac{\hat{\alpha} - \alpha_0}{S \sqrt{\frac{\sum x_i^2}{n S_{xx}}}}$	$\alpha > \alpha_0$	$T > t_{1-\alpha}(n-2)$
		$\alpha < \alpha_0$	$T < -t_{1-\alpha}(n-2)$
		$\alpha \neq \alpha_0$	$ T > t_{1-\frac{\alpha}{2}}(n-2)$
$\beta = \beta_0$	$T = \frac{\hat{\beta} - \beta_0}{\frac{s}{\sqrt{S_{xx}}}}$	$\beta > \beta_0$	$T > t_{1-\alpha}(n-2)$
		$\beta < \beta_0$	$T < -t_{1-\alpha}(n-2)$
		$\beta \neq \beta_0$	$ T > t_{1-\frac{\alpha}{2}}(n-2)$

جدول ۱.۹ آزمونهای آماری روی ضرایب رگرسیونی α و β

مثال ۱.۳.۹ مواد اولیه‌ای که برای ساختن الیاف مصنوعی به کار می‌رود در انبار مرطوبی نگهداری می‌شود. نتایج حاصل از اندازه‌گیری‌های رطوبت نسبی در انبار و میزان رطوبت در یک نمونه مواد اولیه (هر دو بر حسب درصد) در ۱۲ روز در جدول زیر ثبت شده است

رطوبت انبار x	۴۲	۳۵	۵۰	۴۳	۴۸	۶۲
رطوبت مواد اولیه y	۱۲	۸	۱۴	۹	۱۱	۱۶
رطوبت انبار x	۳۱	۳۶	۴۴	۳۹	۵۵	۴۸
رطوبت مواد اولیه y	۷	۹	۱۲	۱۰	۱۳	۱۱

الف- برآورد خط رگرسیون را به دست آورید.

ب- یک فاصله اطمینان ۹۵ درصدی برای β بسازید.

ج- آیا در سطح معنی‌دار ۰/۰۱ می‌توان ادعا کرد که $\alpha = -۱$ است؟

حل الف- از جدول فوق مقادیر زیر حاصل می‌شوند:

$$\sum x_i = 533, \quad \sum x_i^2 = 24529, \quad \sum y_i = 132, \quad \sum y_i^2 = 1526, \quad \sum x_i y_i = 6093$$

بنابراین

$$S_{xy} = 6093 - \frac{(533)(132)}{12} = 230, \quad S_{xx} = 24529 - \frac{(533)^2}{12} = 854/917$$

$$S_{yy} = 1526 - \frac{(132)^2}{12} = 74, \quad \hat{\beta} = \frac{230}{854/917} = 0/269$$

$$\hat{\alpha} = \frac{132}{12} - (0/269) \frac{533}{12} = -0/948, \quad s^2 = \frac{1}{10} [74 - (0/269)(230)] = 1/213$$

در نتیجه

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = -0/948 + 0/269x$$

ب- با استفاده از فواصل اطمینان (۷.۹) داریم که

$$t_{1-\frac{\alpha}{2}}(n-2) = t_{0/975}(10) = 2/23$$

$$\beta \in (0/269 - 2/23 \sqrt{\frac{1/213}{854/917}}, 0/269 + 2/23 \sqrt{\frac{1/213}{854/917}}) = (0/185, 0/353)$$

بنابراین ۹۵ درصد اطمینان داریم که β در فاصله فوق قرار دارد.

ج- در این مثال با آزمون $\begin{cases} H_0 : \alpha = -1 \\ H_1 : \alpha \neq -1 \end{cases}$ مواجه هستیم و با استفاده از جدول ۱.۹ فرض

H_0 رد می شود اگر و فقط اگر $|T| > t_{1-\frac{\alpha}{2}}(n-2)$ که در آن

$$t_{1-\frac{\alpha}{2}}(n-2) = t_{0.995}(10) = 3/17$$

$$T = \frac{\hat{\alpha} - \alpha_0}{s \sqrt{\frac{\sum x_i^2}{nS_{xx}}}} = \frac{-0.948 + 1}{\sqrt{\frac{(1/213)(24529)}{(12)(854/917)}}} = 0.305$$

چون $|T| = 0.305 < t_{1-\frac{\alpha}{2}}(n-2) = 3/17$ پس فرض H_0 رد نمی شود یعنی $\alpha = -1$ می باشد.

مثال ۲.۳.۹ نمره های امتحان میان ترم و پایان ترم یک کلاس ۹ نفره از دانشجویان به صورت زیر است

میان ترم	۶	۵	۷	۷	۴	۶	۴	۵	۳
پایان ترم	۱۰	۸	۱۱	۱۲	۱۱	۹	۱۰	۹	۶

الف- برآورد خط رگرسیونی را برای پیش بینی نمره پایان ترم از روی نمره میان ترم به دست آورید.

ب- یک فاصله اطمینان ۹۰ درصدی برای α پیدا کنید.

ج- آیا در سطح معنی دار ۰/۰۵ می توان ادعا کرد که $\beta < 2$ است؟

حل الف- اگر x نمره میان ترم و y نمره پایان ترم باشند آنگاه از جدول فوق مقادیر زیر حاصل

می شوند

$$\sum x_i = 47, \quad \sum x_i^2 = 261, \quad \sum y_i = 86, \quad \sum y_i^2 = 848, \quad \sum x_i y_i = 462$$

بنابراین

$$S_{xy} = 462 - \frac{(47)(86)}{9} = 12/89, \quad S_{xx} = 261 - \frac{(47)^2}{9} = 15/56$$

$$S_{yy} = 848 - \frac{(86)^2}{9} = 26/22, \quad \hat{\beta} = \frac{12/89}{15/56} = 0/828$$

$$\hat{\alpha} = \frac{86}{9} - (0/828) \frac{47}{9} = 5/23, \quad s^2 = \frac{1}{V} [26/22 - (0/828)(12/89)] = 2/22$$

در نتیجه

$$\hat{y} = 5/23 + 0/828x$$

ب- چون $t_{1-\alpha/2}(n-2) = t_{0.95}(7) = 1/9$ بنابراین از (۷.۹) داریم که

$$\alpha \in \left(5/23 - 1/9 \sqrt{\frac{(2/22)(261)}{(9)(15/56)}}, 5/23 + 1/9 \sqrt{\frac{(2/22)(261)}{(9)(15/56)}} \right) = (1/365, 9/0.95)$$

بنابراین ۹۰ درصد اطمینان داریم که α در فاصله فوق قرار دارد.

ج- در این مثال با آزمون $\begin{cases} H_0 : \beta = 2 \\ H_1 : \beta < 2 \end{cases}$ مواجه هستیم و با استفاده از جدول ۱.۹ فرض

H_0 رد می شود اگر و فقط اگر $T < -t_{1-\alpha}(n-2)$ که در آن

$$t_{1-\alpha}(n-2) = t_{0.95}(7) = 1/9$$

$$T = \frac{\hat{\beta} - \beta_0}{\frac{s}{\sqrt{S_{xx}}}} = \frac{0/828 - 2}{\sqrt{\frac{2/22}{15/56}}} = -3/10.3$$

چون $-3/10.3 = T < -t_{1-\alpha}(n-2) = -1/9$ پس فرض H_0 رد می شود یعنی $\beta < 2$ است.

۴.۹ ضریب همبستگی خطی

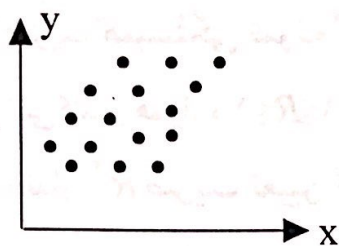
تاکنون فرض کردیم که متغیر مستقل X یک متغیر کنترل شده است و یک متغیر تصادفی نیست. حال فرض کنید که هم متغیر X و هم Y هر دو متغیر تصادفی باشند.

روشهای رگرسیونی موقعی مناسب هستند که متغیر تصادفی Y به متغیر تصادفی X که اغلب به وسیله پژوهشگر کنترل می شود بستگی داشته باشد. برای سنجش میزان وابستگی دو متغیر تصادفی X و Y از معیاری بنام ضریب همبستگی خطی استفاده می شود که در بخش ۴.۴ آن را در جمعیت به صورت زیر تعریف کردیم

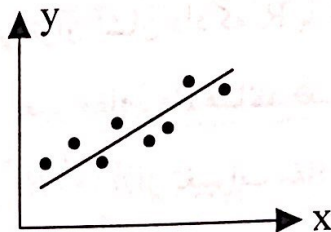
$$\rho = \rho(X, Y) = \frac{COV(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

و در بخش ۴.۴ مشاهده کردیم که ضریب همبستگی خطی به مبداء و واحد اندازه گیری داده ها بستگی ندارد و همواره $-1 \leq \rho \leq 1$ می باشد. در شکل ۳.۹ حالت های مختلف از همبستگی خطی X و

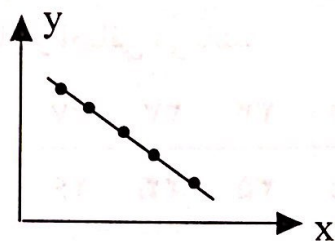
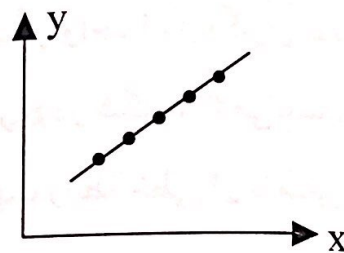
Y نشان داده شده است.



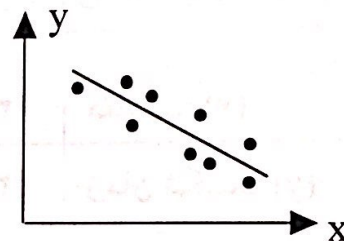
ج - $\rho = 0$ عدم همبستگی



ب - $0 < \rho < 1$ همبستگی مثبت



ه - $\rho = -1$ همبستگی کامل منفی



د - $-1 < \rho < 0$ همبستگی منفی

شکل ۳.۹ حالت‌های مختلف همبستگی خطی بین متغیرهای X و Y

برای برآورد ضریب همبستگی یک نمونه تصادفی $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ از (X, Y) را انتخاب می‌کنیم و از روی این نمونه تصادفی کواریانس X و Y، واریانس X و واریانس Y را به صورت زیر برآورد می‌کنیم

$$\hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} S_{XX}$$

$$\hat{\sigma}_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n-1} S_{YY}$$

$$\hat{\sigma}_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1} S_{XY}$$

حال با قرار دادن این برآوردگرها به جای پارامترهای $\sigma_X^2, \sigma_Y^2, \sigma_{XY}$ در فرمول ضریب همبستگی خطی، برآوردگر ضریب همبستگی خطی به صورت زیر به دست می‌آید:

$$\hat{\rho} = R = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} \quad (۸.۹)$$

اگر مقدار مشاهده شده این نمونه تصادفی $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ باشد آنگاه برآورد

ضریب همبستگی خطی یعنی مقدار مشاهده شده R عبارت از $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ خواهد بود که به آن ضریب همبستگی نمونه گویند. می توان نشان داد که R با تغییر مبدأ و واحد اندازه گیری تغییر نمی کند و همواره $-1 \leq R \leq 1$ است. تعبیر مقادیر R همانند تعبیر مقادیر ρ در شکل ۳.۹ می باشد. به مقدار R^2 ضریب تعیین گویند و R^2 از تغییرات مقادیر Y جهت رابطه خطی آن با متغیر X به حساب می آید.

مثال ۱.۴.۹ داده های زیر مربوط به مقاومت (بر حسب اهم) و زمان شکست (بر حسب دقیقه) ترانزیستورها با بار اضافی می باشد

مقاومت (x)	۴۳	۲۹	۴۴	۳۳	۳۳	۴۷
زمان شکست (y)	۳۲	۲۰	۴۵	۳۵	۲۲	۴۶
مقاومت (x)	۳۴	۳۱	۴۸	۳۴	۴۶	۳۷
زمان شکست (y)	۲۸	۲۶	۳۷	۳۳	۴۷	۳۰

ضریب همبستگی نمونه را به دست آورید و آن را تعبیر کنید.

حل از جدول فوق مقادیر زیر حاصل می شوند

$$\sum x_i = 459, \quad \sum x_i^2 = 18075, \quad \sum y_i = 401, \quad \sum y_i^2 = 14301, \quad \sum x_i y_i = 15907$$

بنابراین

$$S_{xy} = 15907 - \frac{(459)(401)}{12} = 561/75, \quad S_{xx} = 18075 - \frac{(459)^2}{12} = 511/25$$

$$S_{yy} = 14301 - \frac{(401)^2}{12} = 900/917$$

$$r = \frac{561/75}{\sqrt{(511/25)(900/917)}} = 0/832$$

در نتیجه

چون مقدار r به یک نزدیک است پس یک رابطه خطی نسبتاً خوبی در جهت مثبت بین X و Y برقرار است. همچنین چون $r^2 = 0/693$ پس $r^2 = 69/3\%$ از تغییرات Y جهت رابطه خطی آن با X به حساب می آید.

استنباط آماری روی ρ

برای استنباط آماری روی ρ نیاز به داشتن توزیع احتمال R ضریب همبستگی نمونه

تصادفی داریم. فیشر^(۱) آماردان انگلیسی ثابت کرده است که آماره $W = \frac{1}{2} \ln \frac{1+R}{1-R}$ برای اندازه نمونه بزرگ تقریباً دارای توزیع نرمال با میانگین تقریبی $\mu_w \approx \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$ و واریانس تقریبی $\sigma_w^2 \approx \frac{1}{n-3}$ می باشد. یعنی به طور تقریبی داریم که:

$$Z = \frac{W - \mu_w}{\sigma_w} \sim N(0, 1) \quad (9.9)$$

با استفاده از تابع محور (۹.۹) می توان برای ρ فاصله اطمینان پیدا کرد و با استفاده از توزیع (۹.۹) می توان آزمونهای آماری را روی ρ انجام داد که این آزمونها در جدول ۲.۹ آورده شده اند.

H_0	آماره آزمون	H_1	ناحیه بحرانی آزمون
$\rho = \rho_0$	$Z = \frac{W - \mu_w}{\sigma_w}$	$\rho > \rho_0$	$Z > z_{1-\alpha}$
		$\rho < \rho_0$	$Z < -z_{1-\alpha}$
		$\rho \neq \rho_0$	$ Z > z_{1-\frac{\alpha}{2}}$

جدول ۲.۹ آزمونهای آماری روی ρ که $\mu_w = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0}$ می باشد.

مثال ۲.۴.۹ در مثال ۱.۴.۹ آیا در سطح معنی دار ۰/۰۵ می توان ادعا کرد که $\rho > ۰/۵$ است.

حل در این مثال با آزمون $\begin{cases} H_0 : \rho = ۰/۵ \\ H_1 : \rho > ۰/۵ \end{cases}$ مواجه هستیم و با استفاده از جدول ۲.۹ فرض H_0 رد می شود اگر و فقط اگر $Z > z_{1-\alpha}$ که در آن

$$r = ۰/۸۳۲, \quad w = \frac{1}{2} \ln \frac{1+r}{1-r} = \frac{1}{2} \ln \frac{1/۸۳۲}{۰/۱۶۸} = ۱/۱۹۵$$

$$\mu_w = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} = \frac{1}{2} \ln \frac{1/۵}{۰/۵} = ۰/۵۴۹$$

$$\sigma_w = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{9}} = ۰/۳۳۳, \quad Z = \frac{w - \mu_w}{\sigma_w} = \frac{1/۱۹۵ - ۰/۵۴۹}{۰/۳۳۳} = ۱/۹۴$$

$$z_{1-\alpha} = z_{۰/۹۵} = ۱/۶۴۵$$

چون $۱/۹۴ = Z > z_{1-\alpha} = ۱/۶۴۵$ پس فرض H_0 رد می شود یعنی $\rho > ۰/۵$ است.

۵.۹ تمرینات

۱ برای تعیین رابطه بین هزینه حمل یک نوع کالا و فاصله فروشگاه از محل توزیع کالا، یک نمونه تصادفی شامل ۸ فروشگاه که این کالا را عرضه می‌کنند انتخاب و فاصله فروشگاه تا محل توزیع کالا و هزینه حمل ۱۰۰ واحد از این کالا در جدول زیر ثبت شده است

x فاصله به کیلومتر	۶	۱۳	۲۷	۱۵	۹	۱۱	۲۱	۱۴
y هزینه حمل بر حسب صد تومان	۲۹	۹۳	۱۵۹	۱۱۵	۶۶	۹۰	۱۳۹	۹۸

برآورد خط رگرسیون هزینه حمل کالا بر حسب فاصله را به دست آورید. اگر فاصله یک فروشگاه تا محل توزیع ۱۰ کیلومتر باشد، هزینه حمل ۱۰۰ واحد کالا تا این فروشگاه را به چه میزان پیش بینی می‌کنید؟

۲ می‌خواهیم به داده‌های حاصل از قدرت کشش y ، ده قطعه پلاستیک که هر یک x دقیقه پخته شده‌اند، یک خط راست برازش کنیم. داده‌ها در جدول زیر ثبت شده‌اند. برآورد خط رگرسیون مورد نظر را به دست آورید.

x	۲۳	۳۵	۴۵	۶۵	۷۵	۹۵	۱۰۵	۱۲۵	۱۵۵	۱۸۵
y	۲	۹/۸	۹/۲	۲۶/۲	۱۷/۱	۲۴/۸	۴۳	۵۵/۳	۳۸/۴	۶۳/۳

۳ از یک نمونه تصادفی از ۱۰ مرد ۳۰ ساله اطلاعات زیر در مورد حقوق سالیانه کنونی (بر حسب صد هزار تومان) و تعداد سالهای تحصیلی رسمی که داشته‌اند به دست آمده است.

x سالهای تحصیل	۱۰	۱۲	۱۳	۱۴	۱۶	۱۶	۱۶	۱۶	۱۸	۲۰
y حقوق سالیانه	۷/۰	۶/۲	۸/۱	۷/۵	۶/۵	۱۰/۵	۸/۰	۱۳/۲	۱۲/۸	۱۶/۵

برآورد خط رگرسیون حقوق سالیانه بر حسب تعداد سالهای تحصیل را به دست آورید. اگر مردی ۱۵ سال تحصیل کرده باشد، حقوق سالیانه او را به چه میزان پیش بینی می‌کنید؟

۴ یک نوع ماده شیمیایی داریم که در x درجه حرارت y گرم آن تجزیه می‌شود. در پنج آزمایش داده‌های زیر به دست آمده‌اند:

x	-۲	-۱	۰	۱	۲
y	۵	۴	۳	۲	۱

الف- برآورد خط رگرسیون مقدار ماده تجزیه شده بر حسب درجه حرارت را به دست

آورید.

ب- نقاط داده شده و خط برآورد شده را در یک دستگاه مختصات رسم کنید.

ج- اگر درجه حرارت ۳ باشد، مقدار ماده‌ای را که تجزیه می‌شود پیش بینی کنید.

۵ در جدول زیر نیروی کشش به کار رفته برای یک نمونه فولاد، بر حسب هزار پوند و طول حاصل از کشش بر حسب یک هزارم اینچ می‌باشد

x نیروی کشش	۱	۲	۳	۴	۵	۶
y طول	۱۴	۳۳	۴۰	۶۳	۷۶	۸۵

الف- با رسم نمودار پراکندگی داده‌ها، تأیید کنید که فرض خطی بودن رگرسیون y روی x معقول است.

ب- برآورد خط رگرسیون را به دست آورده و با استفاده از آن وقتی نیروی کشش $3/5$ هزار پوند باشد، طول حاصل را پیش بینی کنید.

ج- برای ضرایب رگرسیونی α و β فواصل اطمینان ۹۵ درصدی را تشکیل دهید.

۶ یک مطالعه توسط یک خرده فروش در ارتباط با رابطه بین مخارج تبلیغ و میزان فروش (هر دو بر حسب هزار تومان) به طور هفتگی صورت پذیرفته است و داده‌های زیر به دست آمده است

مخارج تبلیغ	۴۰	۲۰	۲۵	۲۰	۳۰	۵۰	۴۰	۲۰	۵۰	۴۰	۲۵	۵۰
فروش	۳۸۵	۴۰۰	۳۹۵	۳۶۵	۴۷۵	۴۴۰	۴۹۰	۴۲۰	۵۶۰	۵۲۵	۴۸۰	۵۱۰

الف- معادله خط رگرسیون را جهت پیش بینی فروش هفتگی از روی هزینه تبلیغات پیدا کنید.

ب- فاصله اطمینان ۹۰ درصدی را برای β به دست آورید.

۷ در تمرین ۳، آیا می‌توان در سطح معنی‌دار 0.01 ادعا کرد که $\beta > 0$ است؟

۸ داده‌های زیر مربوط به تعداد کارها بر حسب روز و زمان لازم برای پردازش مرکزی (CPU) می‌باشد.

x تعداد کارها	۱	۲	۳	۴	۵
y زمان (CPU)	۲	۵	۴	۹	۱۰

الف- معادله خط رگرسیون را جهت پیش بینی زمان (CPU) بر حسب تعداد کارها پیدا

کنید.

ب- یک فاصله اطمینان ۹۵ درصدی برای ضریب رگرسیونی α به دست آورید.

ج- فرض صفر $\beta=2$ در مقابل $\beta>2$ را در سطح معنی دار ۰/۰۵ آزمون کنید.

۹ اثر دما بر رنگ محصولی بررسی شده است. داده‌ها در جدول زیر آورده شده‌اند

دما x	۱۰۰	۱۱۰	۱۲۰	۱۴۰	۱۵۰	۱۷۰	۱۸۰	۲۰۰	۲۳۰	۱۶۰
رنگ y	۳/۵	۷/۴	۷/۱	۱۵/۶	۱۱/۱	۱۴/۹	۲۳/۵	۲۷/۱	۲۲/۱	۳۲/۹

الف- معادله خط رگرسیون را جهت پیش بینی رنگ بر حسب دما پیدا کنید.

ب- آیا در سطح معنی دار ۰/۰۵ می توان ادعا کرد که $\alpha=0$ است؟

۱۰ داده‌های جدول زیر را در نظر بگیرید

x	۱	۱/۱	۱/۲	۱/۳	۱/۴	۱/۵	۱/۶	۱/۷	۱/۸	۱/۹	۲
y	۸/۱	۷/۸	۸/۵	۹/۸	۹/۵	۸/۹	۸/۶	۱۰/۲	۹/۳	۹/۲	۱۰/۵

الف- معادله خط رگرسیون را پیدا کنید.

ب- برای ضریب رگرسیونی α و β فواصل اطمینان ۹۰ درصدی را تشکیل دهید.

ج- آیا در سطح معنی دار ۰/۰۱ می توان ادعا کرد که $\alpha>0$ است؟

۱۱ جدول زیر نمره‌های ۷ دانشجو در امتحان ریاضی و زبان می باشد.

نمره ریاضی	۱۲	۱۳	۱۰	۱۳	۱۵	۱۱	۱۲
نمره زبان	۱۲	۱۴	۱۳	۱۵	۱۱	۱۴	۱۱

ضریب همبستگی نمونه را محاسبه و آن را تعبیر کنید.

۱۲ در تمرین ۱ ضریب همبستگی نمونه را به دست آورید و آن را تعبیر کنید.

۱۳ برای تعیین رابطه خطی بین دو متغیر x و y، نمونه تصادفی به اندازه $n=20$ از یک جمعیت استخراج و مقادیر x و y را ثبت کرده‌ایم. نتایج به شرح زیر می باشند

$$\sum x_i = 35, \quad \sum x_i^2 = 680, \quad \sum y_i = 48, \quad \sum y_i^2 = 1348, \quad \sum x_i y_i = 960$$

ضریب همبستگی نمونه را به دست آورید و آن را تعبیر کنید.

۱۴ در جدول زیر نمرات امتحان پایان ترم درسهای جبر و فیزیک ۱۰ دانشجو که به طور تصادفی انتخاب شده‌اند، درج گردیده است

	۷۷	۸۴	۶۸	۹۸	۷۱	۸۷	۶۵	۹۳	۸۰	۷۵
جبر										
فیزیک	۷۴	۸۹	۷۲	۹۵	۸۰	۹۱	۷۲	۸۶	۷۸	۸۲

ضریب همبستگی نمونه را به دست آورید و آن را تعبیر کنید.

۱۵ نتایج زیر از اندازه گیری حجم عبور و مرور و غلظت مونو اکسید کربن در ۱۵ محل نمونه برداری حاصل شده است

حجم عبور و مرور x	۱۰۰	۱۱۰	۱۲۵	۱۵۰	۱۷۵	۱۹۰			
غلظت مونو اکسید کربن y	۸/۸	۹/۰	۹/۵	۱۰/۰	۱۰/۵	۱۰/۵			
x	۲۰۰	۲۲۵	۲۵۰	۲۷۵	۳۰۰	۳۲۵	۳۵۰	۳۷۵	۴۰۰
y	۱۰/۵	۱۰/۶	۱۱/۰	۱۲/۱	۱۲/۱	۱۲/۵	۱۳/۰	۱۳/۲	۱۴/۵

الف - ضریب همبستگی نمونه را محاسبه و آن را تعبیر کنید.

ب - فرضیه ناهمبسته بودن داده‌ها، یعنی $\rho = 0$ را در سطح معنی دار ۰/۰۵ آزمون کنید.

۱۶ با استفاده از تابع محور (۹.۹) یک فاصله اطمینان $(1-\alpha) \cdot 100\%$ برای ضریب همبستگی خطی ρ پیدا کنید.

(راهنمایی: توجه کنید که $f(\rho) = Ln \frac{1+\rho}{1-\rho}$ یک تابع اکیداً صعودی از ρ است.)

۱۷ ضریب همبستگی میان قد و وزن ۱۰۰ نوزاد برابر ۰/۳ است. آیا در سطح معنی دار ۰/۱ می‌توان ادعا کرد که $\rho > 0/25$ است.

۱۸ با استفاده از نتیجه تمرین ۱۶ یک فاصله اطمینان ۹۰ درصدی برای ضریب همبستگی خطی در تمرین ۱۱ به دست آورید.

۱۹ داده‌های زیر اندازه‌های میزان کربن و شاخص نفوذپذیری ۲۲ ترکیب خاکستر می‌باشند.

میزان کربن (درصد)	۴/۴	۵/۵	۴/۲	۳/۰	۴/۵	۴/۹	۴/۶	۵/۰	۴/۷	۵/۱	۴/۴
شاخص نفوذپذیری	۱۲	۱۴	۱۸	۳۵	۲۳	۲۹	۱۶	۱۲	۱۸	۲۱	۲۷
میزان کربن (درصد)	۴/۱	۴/۹	۴/۷	۵/۰	۴/۶	۳/۶	۴/۹	۵/۱	۴/۸	۵/۲	۵/۲
شاخص نفوذپذیری	۱۳	۱۹	۲۲	۲۰	۱۶	۲۷	۲۱	۱۳	۱۸	۱۷	۱۱

الف - ضریب همبستگی نمونه را به دست آورید و آن را تعبیر کنید.

ب - یک فاصله اطمینان ۹۵ درصدی برای ρ تشکیل دهید.

۲۰ داده‌های زیر را در نظر بگیرید:

x	۴	۵	۹	۱۴	۱۸	۲۲	۲۴
y	۱۶	۲۲	۱۱	۱۶	۷	۳	۱۷

الف- ضریب همبستگی نمونه را به دست آورید و آن را تعبیر کنید.

ب- فرض $\rho = -0.8$ را در سطح معنی دار 0.05 آزمون کنید.