۳۶ استادی یک امتحان ۵۰ نمره‌ای به کلاس ریاضی عمومی داده است. جدول زیر نمره‌های ۵۰ دانشجو می‌باشد

۳۸	۳۴	۲۲	۴۱	۳۵	۴۵	۳۲	۳۷	۳۰	۲۶
۳۶	۲۸	۳۴	۱۶	۳۱	۳۳	۳۸	۳۱	۴۷	۳۶
۲۹	۲۷	۲۵	۴۳	۳۴	۳۶	۲۹	۳۹	۳۳	۳۱
۴۰	۳۵	۳۳	۳۱	۳۷	۴۴	۳۲	۴۱	۱۹	۳۴
۱۵	۴۹	۴۷	۳۸	۳۲	۲۶	۳۹	۳۰	۴۲	۳۵

الف- یک جدول فراوانی برای داده‌های فوق تشکیل دهید.

ب- آیا در سطح معنی‌دار ۰/۱ می‌توان ادعا کرد که نمرات دارای توزیع یکنواخت است؟

ج- آیا در سطح معنی‌دار ۰/۱ می‌توان ادعا کرد که نمرات دارای توزیع نرمال است؟

۳۷ یک سکه را آنقدر پرتاب می‌کنیم تا اینکه یک شیر بیاید. اگر X برابر تعداد پرتاب این سکه باشد، بعد از تکرار این آزمایش در ۲۵۶ بار، نتایج زیر حاصل می‌شود

x	۱	۲	۳	۴	۵	۶	۷	۸
تعداد پرتاب	۱۳۶	۶۰	۳۴	۱۲	۹	۱	۳	۱

آیا در سطح معنی‌دار ۰/۰۵ می‌توان ادعا کرد که توزیع هندسی با پارامتر $\frac{1}{2}$ بر داده‌ها برازنده است؟

۳۸ داده‌های زیر میزان محصول ذرت را در ۱۰۰ مزرعه نشان می‌دهد. اگر در این مزارع $\sum x_i = 9140$۰ و $s = 331/8$ باشد، آیا میزان محصول ذرت این مزارع از توزیع نرمال پیروی می‌کند؟

تعداد مزارع	محصول (برحسب کیلوگرم)
۳	$۹۹/۵ \le x < ۲۹۹/۵$
۷	$۲۹۹/۵ \le x < ۴۹۹/۵$
۱۵	$۴۹۹/۵ \le x < ۶۹۹/۵$
۲۶	$۶۹۹/۵ \le x < ۸۹۹/۵$
۲۲	$۸۹۹/۵ \le x < ۱۰۹۹/۵$
۱۳	$۱۰۹۹/۵ \le x < ۱۲۹۹/۵$
۹	$۱۲۹۹/۵ \le x < ۱۴۹۹/۵$
۵	$۱۴۹۹/۵ \le x < ۱۶۹۹/۵$

فصل نهم

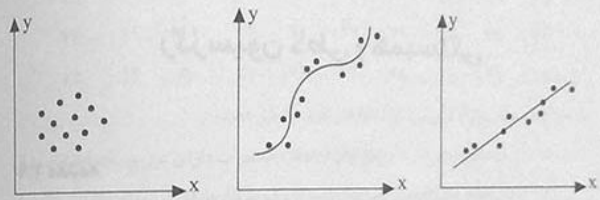
رگرسیون خطی و همبستگی

۱.۹ مقدمه

در اغلب بررسیهای آماری نیاز به پیش بینی مقدار یک متغیر وابسته از روی مقدار یک متغیر مستقل می‌باشد. برای مثال پیش بینی طول قد فرزند از روی طول قد پدرش، یا پیش بینی معدل کل یک دانشجو در یک نیمسال از روی نمرهٔ درس ریاضیات او و یا پیش بینی مقدار مصرف بنزین از روی مسافت طی شدهٔ اتومبیل از این نوع مسائل می‌باشند. چنین مسائلی را مسائل **برگشت یا رگرسیون** گویند. متغیر مستقل را با x و متغیر وابسته را با $Y \mid x$ و یا برای راحتی با Y نمایش می‌دهند. 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 می‌دهیم و چون برای یک طول قد پدر x، فرزندان او می‌توانند طول قدهای متفاوت داشته باشند بنابراین طول قد فرزند او را با متغیر تصادفی $Y \mid x$ یا Y نشان می‌دهیم. به همین ترتیب در پیش بینی مقدار مصرف بنزین، برای مسافت معین طی شدهٔ x میزان مصرف بنزین را با Y نمایش می‌دهیم.

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

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 براین نقاط می‌توان یک خط یا منحنی عبور داد و این خط یا منحنی رابطهٔ 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 معادلهٔ پیش بینی کننده است که به آن معادلهٔ رگرسیون Y روی x گویند.

۲.۹ رگرسیون سادهٔ خطی

هر گاه بین متغیر مستقل x و متغیر وابستهٔ $Y=Y\mid x$ یک رابطهٔ خطی برقرار باشد گوئیم یک مدل رگرسیون سادهٔ خطی بین x و Y برقرار است. برای تشکیل این رابطهٔ خطی، فرض کنید یک نمونهٔ تصادفی (x_1,Y_1)، (x_2,Y_2)، ... و (x_n,Y_n) یا مقادیر مشاهده شدهٔ (x_1,y_1)، (x_2,y_2)، ... و (x_n,y_n) داشته باشیم. توجه کنید که $Y_i=Y\mid x_i$ ، $i=1,2,...,n$. منظور از رگرسیون خطی این است که میانگین $Y\mid x$ به طور خطی با x در ارتباط باشد یعنی $\mu_{Y\mid x}=E(Y\mid x)=\alpha+\beta x$ 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 وابستهٔ 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 بوسیلهٔ $\hat{y}=\hat{\alpha}+\hat{\beta}x$ برای مقدار بخصوص x، مقدار متغیر وابستهٔ y را پیش بینی می‌کنیم.

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

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

این مدل را **مدل رگرسیون سادهٔ 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

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

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[1] نمایش می‌دهند، یعنی

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

مقادیری از $\hat{\alpha}$ و $\hat{\beta}$ که SSE را می‌نیمم کند، برآوردهای حداقل مربعات گویند و روش بدست آوردن آنها را روش **حداقل مربعات**[2] می‌نامند که در زیر به ذکر آن می‌پردازیم.

ابتدا کمیتهای زیر را معرفی می‌کنیم

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

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

1- Sum of Square Errors

2- Least Squares Method

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

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

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

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

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

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 می‌نیمم می‌کنند.

بنابراین خط رگرسیون $\mu_{Y|x}=\alpha+\beta x$ بوسیلهٔ خط $\hat{y}=\hat{\alpha}+\hat{\beta}x$ برآورد می‌شود، که $\hat{\alpha}$ و $\hat{\beta}$ از رابطهٔ (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

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

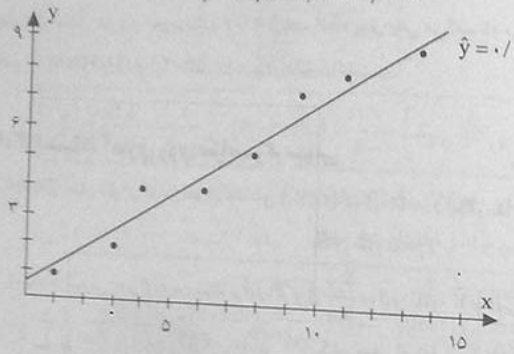
بنابراین

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

و در نتیجه

$$\hat{\beta}=\frac{S_{xy}}{S_{xx}}=\frac{84}{132}=0/636\quad,\quad \hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}=\frac{40}{8}-(0/636)\frac{56}{8}=0/545$$

$$\hat{y}=\hat{\alpha}+\hat{\beta}x=0/545+0/636x$$



شکل ۲.۹ نمودار پراکندگی 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 میلی‌گرم 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 می‌کنید؟

حل از جدول فوق مقادیر زیر حاصل می‌شوند:

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

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

بنابراین

$$\hat{\beta}=\frac{46/5}{11/375}=4/088\quad,\quad \hat{\alpha}=\frac{198}{13}-(4/088)\frac{26}{13}=7/055$$

در نتیجه

$$\hat{y}=\hat{\alpha}+\hat{\beta}x=۷/۰۵۵+۴/۰۸۸x$$

حال اگر مقدار داروی تجویز شده $x=۱/۶$ باشد آنگاه پیش بینی می‌کنیم که میزان واکنش ضربان قلب $\hat{y}=۷/۰۵۵+۴/۰۸۸(۱/۶)\approx ۱۳/۶$ در دقیقه باشد.

۳.۹ استنباط آماری روی ضرایب رگرسیونی

در بخش قبل بر اساس نمونهٔ تصادفی $(x_۱,Y_۱)$، $(x_۲,Y_۲)$، ... و (x_n,Y_n) در مدل رگرسیونی

$$Y_i=\alpha+\beta x_i+E_i \qquad i=۱,۲,...,n$$

ضرایب رگرسیونی α و β را به وسیلهٔ برآوردگرهای $\hat{\beta}=\frac{S_{xY}}{S_{xx}}$ و $\hat{\alpha}=\overline{Y}-\hat{\beta}\overline{x}$ برآورد کردیم که در آن $\overline{Y}=\frac{۱}{n}\sum_{i=۱}^{n}Y_i$ ، $S_{xY}=\sum_{i=۱}^{n}(x_i-\overline{x})(Y_i-\overline{Y})$ و S_{xY} متغیر تصادفی می‌باشند. حال اگر در مدل رگرسیونی فوق فرض کنیم $E_i\sim N(۰,\sigma^۲)$ ، $i=۱,۲,...,n$ و $E_۱$، $E_۲$، ... و E_n از یکدیگر مستقل باشند، در این صورت چون Y_i یک ترکیب خطی از متغیرهای تصادفی نرمال E_i می‌باشد، پس $Y_i\sim N(\alpha+\beta x_i,\sigma^۲)$ ، $i=۱,۲,...,n$. بر اساس توزیع متغیرهای تصادفی Y_i می‌توان توزیع $\hat{\alpha}$ $\hat{\beta}$ و SSE را به دست آورد. توزیع این برآوردگرها را در قضیهٔ زیر بدون اثبات می آوریم.

قضیهٔ ۲.۹ در مدل رگرسیونی سادهٔ خطی با فرض $E_i\sim N(۰,\sigma^۲)$ ، $i=۱,۲,...,n$ داریم که

الف- $\hat{\alpha}\sim N(\alpha,\frac{\sigma^۲\sum x_i^۲}{nS_{xx}})$

ب- $\hat{\beta}\sim N(\beta,\frac{\sigma^۲}{S_{xx}})$

ج- اگر $S^۲=\frac{SSE}{n-۲}$ آنگاه $\frac{SSE}{\sigma^۲}=\frac{(n-۲)S^۲}{\sigma^۲}\sim\chi^۲_{(n-۲)}$

د- $\hat{\alpha}$ و $S^۲$ و همچنین $\hat{\beta}$ و $S^۲$ از یکدیگر مستقل هستند.

با استفاده از قضیهٔ ۲.۹ می‌توان نتایج زیر را به دست آورد

$$\frac{\hat{\beta}-\beta}{\frac{S}{\sqrt{S_{xx}}}}\sim t_{(n-۲)} \qquad (۴.۹)$$

$$\frac{\hat{\alpha}-\alpha}{S\sqrt{\frac{\sum x_i^۲}{nS_{xx}}}}\sim t_{(n-۲)} \qquad (۵.۹)$$

که در آن

$$S^۲=\frac{SSE}{n-۲}=\frac{۱}{n-۲}\sum_{i=۱}^{n}(Y_i-\hat{\alpha}-\hat{\beta}x_i)^۲=\frac{۱}{n-۲}\left[S_{YY}-\hat{\beta}S_{xY}\right] \qquad (۶.۹)$$

با استفاده از تابع محورهای (۴.۹) و (۵.۹) می‌توان فواصل اطمینان $(۱-\alpha)۱۰۰\%$ برای α و β به دست آورد که این فواصل اطمینان عبارتند از

$$\alpha\in(\hat{\alpha}-t_{۱-\frac{\alpha}{۲}}(n-۲)S\sqrt{\frac{\sum x_i^۲}{nS_{xx}}}\ ,\ \hat{\alpha}+t_{۱-\frac{\alpha}{۲}}(n-۲)S\sqrt{\frac{\sum x_i^۲}{nS_{xx}}})$$

$$\beta\in(\hat{\beta}-t_{۱-\frac{\alpha}{۲}}(n-۲)\frac{S}{\sqrt{S_{xx}}}\ ,\ \hat{\beta}+t_{۱-\frac{\alpha}{۲}}(n-۲)\frac{S}{\sqrt{S_{xx}}}) \qquad (۷.۹)$$

با استفاده از روابط (۴.۹) و (۵.۹) می‌توان آزمونهای آماری را روی α و β انجام داد که این آزمونها در جدول ۱.۹ آورده شده‌اند.

$H_۰$	آمارهٔ آزمون	$H_۱$	ناحیه بحرانی
$\alpha=\alpha_۰$	$T=\frac{\hat{\alpha}-\alpha_۰}{S\sqrt{\frac{\sum x_i^۲}{nS_{xx}}}}$	$\alpha>\alpha_۰$	$T>t_{۱-\alpha}(n-۲)$
		$\alpha<\alpha_۰$	$T<-t_{۱-\alpha}(n-۲)$
		$\alpha\neq\alpha_۰$	$\lvert T\rvert>t_{۱-\frac{\alpha}{۲}}(n-۲)$
$\beta=\beta_۰$	$T=\frac{\hat{\beta}-\beta_۰}{\frac{S}{\sqrt{S_{xx}}}}$	$\beta>\beta_۰$	$T>t_{۱-\alpha}(n-۲)$
		$\beta<\beta_۰$	$T<-t_{۱-\alpha}(n-۲)$
		$\beta\neq\beta_۰$	$\lvert T\rvert>t_{۱-\frac{\alpha}{۲}}(n-۲)$

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 فاصلهٔ اطمینان ۹۵ درصدی برای β بسازید.

ج- آیا در سطح معنی‌دار ۰/۰۱ می‌توان ادعا کرد که $\alpha=-1$ است؟

حل الف- از جدول فوق مقادیر زیر حاصل می‌شوند:

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

بنابراین

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

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

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

در نتیجه

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

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

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

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

بنابراین ۹۵ درصد اطمینان داریم که β در فاصلهٔ 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 می‌شود 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 نمی‌شود یعنی $\alpha=-1$ می‌باشد.

مثال ۲.۳.۹ نمره‌های 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 نمرهٔ پایان ترم از روی نمرهٔ میان ترم به دست آورید.

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

ج- آیا در سطح معنی‌دار ۰/۰۵ می‌توان ادعا کرد که $\beta<2$ است؟

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

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

بنابراین

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

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

$$\hat{\alpha}=\frac{86}{9}-(0/828)\frac{47}{9}=5/23\ ,\ s^2=\frac{1}{7}\left[26/22-(0/828)(12/89)\right]=2/22$$

در نتیجه

$$\hat{y}=۵/۲۳+۰/۸۲۸x$$

ب- چون $t_{۱-\frac{\alpha}{۲}}(n-۲)=t_{۰/۹۵}(۷)=۱/۹$ بنابراین از (۷.۹) داریم که

$$\alpha\in(۵/۲۳-۱/۹\sqrt{\frac{(۲/۲۲)(۲۶۱)}{(۹)(۱۵/۵۶)}},۵/۲۳+۱/۹\sqrt{\frac{(۲/۲۲)(۲۶۱)}{(۹)(۱۵/۵۶)}})=(۱/۳۶۵,۹/۰۹۵)$$

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

ج- در این مثال با آزمون $\begin{cases}H_۰:\beta=۲\\H_۱:\beta<۲\end{cases}$ مواجه هستیم و با استفاده از جدول ۱.۹ فرض $H_۰$ رد می‌شود اگر و فقط اگر $T<-t_{۱-\alpha}(n-۲)$ که در آن

$$t_{۱-\alpha}(n-۲)=t_{۰/۹۵}(۷)=۱/۹$$

$$T=\frac{\hat{\beta}-\beta_۰}{\frac{s}{\sqrt{s_{xx}}}}=\frac{۰/۸۲۸-۲}{\sqrt{\frac{۲/۲۲}{۱۵/۵۶}}}=-۳/۱۰۳$$

چون $-۱/۹=-t_{۱-\alpha}(n-۲)>T=-۳/۱۰۳$ پس فرض $H_۰$ رد می‌شود یعنی $\beta<۲$ است.

۴.۹ ضریب همبستگی خطی

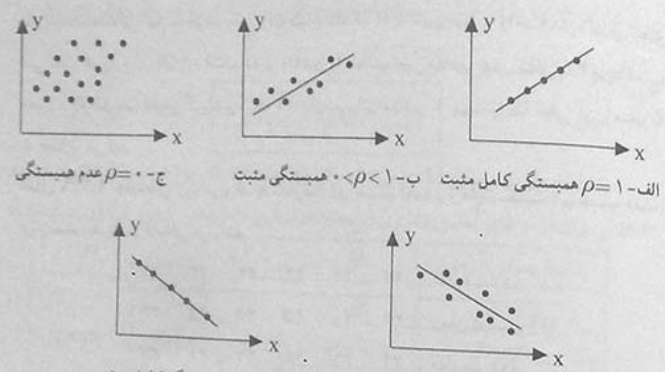
تاکنون فرض کردیم که متغیر مستقل x یک متغیر کنترل شده است و یک متغیر تصادفی نیست. حال فرض کنید که هم متغیر X و هم Y هر دو متغیر تصادفی باشند.

روشهای رگرسیونی موقعی مناسب هستند که متغیر تصادفی Y به متغیر تصادفی X که اغلب به وسیلهٔ پژوهشگر کنترل می‌شود بستگی داشته باشد. برای سنجش میزان وابستگی دو متغیر تصادفی X و Y از معیاری بنام ضریب همبستگی خطی استفاده می‌شود 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 اندازه‌گیری داده‌ها بستگی ندارد و همواره $-۱\leq\rho\leq۱$ می‌باشد. در شکل ۳.۹ حالتهای مختلف از همبستگی خطی X و

Y نشان داده شده است.



الف- $\rho=۱$ همبستگی کامل مثبت ب- $۰<\rho<۱$ همبستگی مثبت ج- $\rho=۰$ عدم همبستگی

د- $-۱<\rho<۰$ همبستگی منفی ه- $\rho=-۱$ همبستگی کامل منفی

شکل ۳.۹ حالتهای مختلف همبستگی خطی بین متغیرهای X و Y

برای برآورد ضریب همبستگی یک نمونهٔ تصادفی $(X_۱,Y_۱),(X_۲,Y_۲),\ldots$ و (X_n,Y_n) از (X,Y) را انتخاب می‌کنیم و از روی این نمونهٔ تصادفی کواریانس X و Y، واریانس X و واریانس Y را به صورت زیر برآورد می‌کنیم

$$\hat{\sigma}_X^۲=\frac{۱}{n-۱}\sum_{i=۱}^{n}(X_i-\bar{X})^۲=\frac{۱}{n-۱}S_{XX}$$

$$\hat{\sigma}_Y^۲=\frac{۱}{n-۱}\sum_{i=۱}^{n}(Y_i-\bar{Y})^۲=\frac{۱}{n-۱}S_{YY}$$

$$\hat{\sigma}_{XY}=\frac{۱}{n-۱}\sum_{i=۱}^{n}(X_i-\bar{X})(Y_i-\bar{Y})=\frac{۱}{n-۱}S_{XY}$$

حال با قرار دادن این برآوردگرها به جای پارامترهای $\sigma_X^۲,\sigma_Y^۲,\sigma_{XY}$ در فرمول ضریب همبستگی خطی، برآوردگر ضریب همبستگی خطی به صورت زیر به دست می‌آید:

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

اگر مقدار مشاهدهٔ شدهٔ این نمونهٔ تصادفی $(x_۱,y_۱),(x_۲,y_۲),\ldots$ و (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. می‌توان نشان داد که R با تغییر مبدأ و واحد اندازه‌گیری تغییر نمی‌کند و همواره $-1\le R\le 1$ است. تعبیر مقادیر R همانند تعبیر مقادیر ρ در شکل ۳.۹ می‌باشد. به مقدار R^2 **ضریب تعیین** گویند و $R^2\times 100\%$ از تغییرات مقادیر Y جهت رابطهٔ خطی آن با متغیر X به حساب می‌آید.

مثال ۱.۴.۹ داده‌های زیر مربوط به مقاومت (بر حسب اهم) و زمان شکست (بر حسب دقیقه) ترانزیستورها با بار اضافی می‌باشد

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 می‌شوند

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

بنابراین

$$S_{xx}=18075-\frac{(459)^2}{12}=518/25\quad ,\quad S_{xy}=15907-\frac{(459)(401)}{12}=568/75$$

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

در نتیجه

$$r=\frac{568/75}{\sqrt{(518/25)(900/917)}}=0/832$$

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

استنباط آماری روی ρ

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

تصادفی داریم. فیشر[1] آماردان انگلیسی ثابت کرده است که آماره $W=\frac{1}{2}Ln\frac{1+R}{1-R}$ برای اندازهٔ نمونهٔ بزرگ تقریباً دارای توزیع نرمال با میانگین تقریبی $\mu_W\approx\frac{1}{2}Ln\frac{1+\rho}{1-\rho}$ و واریانس تقریبی $\sigma_W^2\approx\frac{1}{n-3}$ می‌باشد. یعنی به طور تقریبی داریم که:

$$Z=\frac{W-\mu_W}{\sigma_W}\sim N(0,1) \qquad (9.9)$$

با استفاده از تابع محور (۹.۹) می‌توان برای ρ فاصلهٔ اطمینان پیدا کرد و با استفاده از توزیع (۹.۹) می‌توان آزمونهای آماری را روی ρ انجام داد که این آزمونها در جدول ۲.۹ آورده شده‌اند.

ناحیهٔ بحرانی آزمون	H_1	آمارهٔ آزمون	H_0
$Z>z_{1-\alpha}$	$\rho>\rho_0$	$Z=\frac{W-\mu_{W_0}}{\sigma_W}$	$\rho=\rho_0$
$Z<-z_{1-\alpha}$	$\rho<\rho_0$		
$\lvert Z\rvert>z_{1-\frac{\alpha}{2}}$	$\rho\neq\rho_0$		

جدول ۲.۹ آزمونهای آماری روی ρ که $\mu_{W_0}=\frac{1}{2}Ln\frac{1+\rho_0}{1-\rho_0}$ می‌باشد.

مثال ۲.۴.۹ در مثال ۱.۴.۹ آیا در سطح معنی‌دار ۰/۰۵ می‌توان ادعا کرد که $\rho>0/5$ است.

حل در این مثال با آزمون $\begin{cases}H_0:\rho=0/5\\ H_1:\rho>0/5\end{cases}$ مواجه هستیم و با استفاده از جدول ۲.۹ فرض H_0 رد می‌شود اگر و فقط اگر $Z>z_{1-\alpha}$ که در آن

$$r=0/832\quad ,\quad w=\frac{1}{2}Ln\frac{1+r}{1-r}=\frac{1}{2}Ln\frac{1/832}{0/168}=1/195$$

$$\mu_{w_0}=\frac{1}{2}Ln\frac{1+\rho_0}{1-\rho_0}=\frac{1}{2}Ln\frac{1/5}{0/5}=0/549$$

$$\sigma_w=\sqrt{\frac{1}{n-3}}=\sqrt{\frac{1}{9}}=0/333\quad ,\quad Z=\frac{w-\mu_{w_0}}{\sigma_w}=\frac{1/195-0/549}{0/333}=1/94$$

$$z_{1-\alpha}=z_{0/95}=1/645$$

چون $Z=1/94>z_{1-\alpha}=1/645$ پس فرض H_0 رد می‌شود یعنی $\rho>0/5$ است.

1- Fisher

۵.۹ تمرینات

۱ برای تعیین رابطه بین هزینه حمل یک نوع کالا و فاصلهٔ فروشگاه از محل توزیع کالا، یک نمونهٔ تصادفی شامل ۸ فروشگاه که این کالا را عرضه می‌کنند انتخاب و فاصلهٔ فروشگاه تا محل توزیع کالا و هزینهٔ 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 هزینهٔ حمل بر حسب صد تومان	۲۹	۹۳	۱۵۹	۱۱۵	۶۶	۹۰	۱۳۹	۹۸

برآورد خط رگرسیون هزینهٔ حمل کالا بر حسب فاصله را به دست آورید. اگر فاصلهٔ یک فروشگاه تا محل توزیع ۱۰ کیلومتر باشد، هزینهٔ 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	۱۸۵	۱۵۵	۱۲۵	۱۰۵	۹۵	۷۵	۶۵	۴۵	۳۵	۲۳
y	۶۳/۳	۳۸/۴	۵۵/۳	۴۳	۲۴/۸	۱۷/۱	۲۶/۲	۹/۲	۹/۸	۲

۳ از یک نمونهٔ 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 سالیانهٔ او را به چه میزان پیش بینی می‌کنید؟

۴ یک نوع مادهٔ 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 مادهٔ 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 ۳/۵ 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 رابطهٔ 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 هزینهٔ تبلیغات پیدا کنید.

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

۷ در تمرین ۳، آیا می‌توان در سطح معنی‌دار ۰/۰۱ 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- معادلهٔ خط رگرسیون را جهت پیش بینی زمان (CPU) بر حسب تعداد کارها پیدا