رگرسیون خطی ساده
- آمار و احتمال مهندسی -

مدرس: مشکانی فراهانی

دانشگاه صنعتی امیرکبیر

۹ دی ۱۳۹۹

از رگرسیون برای به الگو در آوردن رابطه بین متغیرهای آماری استفاده می‌شود.

زیرا هدف بیشتر تحقیق‌ها ارزیابی روابط میان مجموعه‌ای از متغیرهاست.

رگرسیون چیست؟ می‌توان گفت رگرسیون تعیین روابط نادقیق بین متغیرهای آماری و تحلیل این روابط است.

فرض کنید می‌خواهیم بدانیم که آیا مصرف سیگار با متغیرهای اجتماعی مثل سن، تحصیلات، درامد و قیمت سیگار رابطه دارد یا خیر.

رابطه بین مصرف سیگار با متغیرهای ذکر شده به شکل یک معادله یا الگویی است که یک متغیر وابسته (متغیر پاسخ) را به یک یا چند متغیر مستقل (متغیر پیش‌گو) مربوط می‌کند.

در رگرسیون خطی متغیر پاسخ یک متغیر پیوسته است اما متغیرهای پیشگو می‌توانند گسسته یا پیوسته باشند.

متغیر پاسخ را با Y و متغیر پیشگو را با X نشان می‌دهند.

در ساده‌ترین روش‌های رگرسیون خطی فرض می‌شود که متغیرها به‌وسیله یک معادله خط راست در ارتباط هستند.

شکل کلی یک خط راست به صورت $Y = \alpha + \beta X$ است، که در آن α را عرض از مبدأ و β را شیب خط می‌نامند.

الگوی رگرسیون خطی ساده عبارت است از: $$Y_i = \alpha + \beta X_i + E_i, \qquad i = ۱, ۲, \dots, n$$

در معادله بالا

α عرض از مبدأ و β شیب خط است؛ که ثابت اما مجهول هستند.

α و β را ضرایب رگرسیونی می‌نامند.

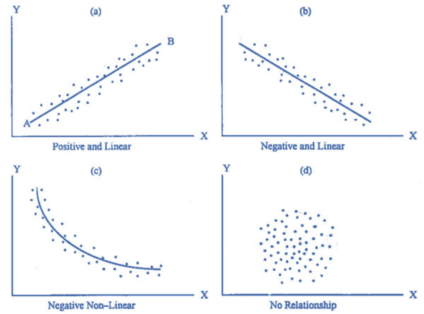
E جمله‌ی خطا است که متغیری تصادفی و غیر قابل مشاهده است: $E(E_i) = ۰ \quad Var(E_i) = \sigma^۲$

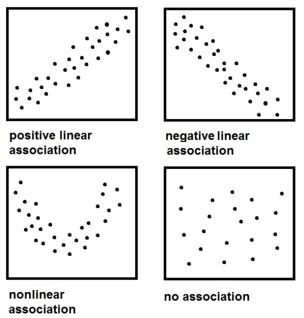
در مدل‌های رگرسیونی رابطه میان متغیر پیشگو و متغیر پاسخ به صورت یک رابطه ریاضی نیست. زیرا به‌ازای هر مقدار از متغیر پیشگو ممکن است چند مقدار برای متغیر پاسخ مشخص شود. بنابراین یک مؤلفه تصادفی در رابطه‌ی آن‌ها وجود دارد E_i.

برای به دست آوردن رابطه‌ی بین متغیر پیشگو x و متغیر پاسخ Y ابتدا یک نمونه تصادفی از جامعه‌ی مد نظر جمع‌آوری می‌کنیم.

یعنی به ازای مقادیر $x_1, x_2, \ldots, x_n$ از متغیر پیشگو مقادیر متغیر پاسخ $Y|x$ یعنی $y_1, y_2, \ldots, y_n$ را اندازه‌گیری می‌کنیم. حال مقادیر مشاهده شده از نمونه‌ی تصادفی یعنی (x_n, y_n) , $\cdots$, (x_1, y_1) را داریم.

برای پی بردن به رابطه بین x و $Y|x$ این مشاهدات را به صورت نقاطی در دستگاه مختصات رسم می‌کنیم. این نمودار را **نمودار پراکندگی یا پراکنش** می‌نامند.





در صورتی که بین x و y رابطه‌ای وجود داشته باشد، می‌توان یک خط یا منحنی را بر این نقاط عبور داد. این خط یا منحنی رابطه میان x و y را مشخص می‌کند.

این خط یا منحنی را معادله رگرسیونی Y روی x می‌گویند.

منظور از رگرسیون خطی آن است که میانگین $Y|x$ به طور خطی با x در ارتباط است؛ یعنی
$$E(Y|x) = \alpha + \beta x$$

در نمونه تصادفی Y_i برابر با $E(Y)$ نیست؛ اختلاف آن‌ها را مقدار خطای E_i نشان می‌دهند.

مقدار مشاهده شده‌ی E_i را با ε_i نشان می‌دهند: $y_i = \alpha + \beta x_i + \varepsilon_i$

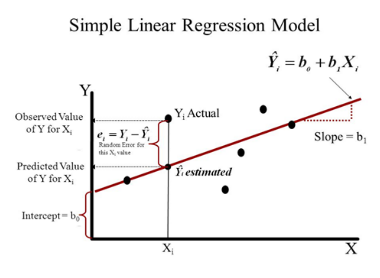
برای پیش‌بینی مقادیر $Y|x_i$ از روی خط رگرسیونی باید پارامترهای α و β را با استفاده از مشاهدات براورد کنیم.

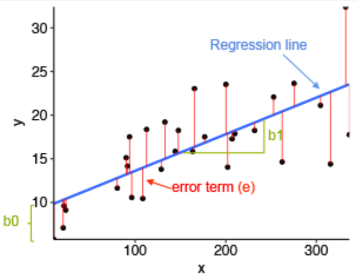
مقدار براورد شده‌ی α و β را با $\hat{\alpha}$ و $\hat{\beta}$ نشان می‌دهیم: $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

α و β را ضرایب رگرسیونی می‌نامند.

تفاوت i-امین مقدار مشاهده شده و i-امین مقدار برازش شده‌ی متناظر با آن را باقی‌مانده i-ام می‌نامند:

$$e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$$





تعیین بهترین خط راست برازش

برای برازش بهترین خط راست از روش کمترین توان‌های دوم خطا استفاده می‌کنیم.

در این روش بهترین خط راست برازش خطی است که مجموع توان دوم خطاها را مینیمم کند.

هرچه انحرافات مقادیر مشاهده شده از این خط کمتر باشد، این خط به داده‌ها نزدیک‌تر است.

پس α و β را طوری براورد می‌کنیم که $SSE = \sum_{i=۱}^{n} e_i^۲$ مینیمم شود.

$$SSE = \sum_{i=۱}^{n} e_i^۲ = \sum_{i=۱}^{n} (y_i - \hat{y}_i)^۲ = \sum_{i=۱}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^۲$$

قضیه: در مدل رگرسیون خطی $Y_i = \alpha + \beta x_i + E_i$ مقادیر $\hat{\alpha}$ و $\hat{\beta}$ که مجموع توان‌های دوم باقی‌مانده‌ها را مینیمم کند، عبارتند از:

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$s_{xx} = \sum (x_i - \bar{x})^۲ = \sum x_i^۲ - n\bar{x}^۲$$

$$s_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$

اثبات:

$$\frac{\partial SSE}{\partial \hat{\alpha}} = ۰ \quad \Rightarrow \quad -۲\left(\sum y_i - n\hat{\alpha} - \hat{\beta}\sum x_i\right) = ۰$$

$$\frac{\partial SSE}{\partial \hat{\beta}} = ۰ \quad \Rightarrow \quad -۲\left(\sum x_i y_i - \hat{\alpha}\sum x_i - \hat{\beta}\sum x_i^۲\right) = ۰$$

اثبات قضیه

$$\sum y_i - n\hat{\alpha} - \hat{\beta}\sum x_i = 0$$
$$n\hat{\alpha} = \sum y_i - \hat{\beta}\sum x_i$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\sum x_i y_i - \hat{\alpha}\sum x_i - \hat{\beta}\sum x_i^2 = 0$$
$$\sum x_i y_i - \left(\bar{y} - \hat{\beta}\bar{x}\right)\sum x_i - \hat{\beta}\sum x_i^2 = 0$$
$$\hat{\beta}\left(\bar{x}\sum x_i - \sum x_i^2\right) = \bar{y}\sum x_i - \sum x_i y_i$$
$$\hat{\beta} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{s_{xy}}{s_{xx}}$$

مثال ۱

جدول زیر مقادیر فسفر موجود در جیره غذایی ۷ گوساله را همراه با افزایش وزن آن‌ها در طول یک هفته نشان می‌دهد. به این داده‌ها خط رگرسیون برازش دهید.

x	۲۵	۳۰	۳۵	۴۰	۴۵	۵۰	۵۵
y	۴/۱	۴/۳	۴/۳	۵/۵	۵/۹	۶/۵	۶/۸
$x_i^۲$	۶۲۵	۹۰۰	۱۲۲۵	۱۶۰۰	۲۰۲۵	۲۵۰۰	۳۰۲۵
$x_i y_i$	۱۰۲/۵	۱۲۹	۱۵۰/۵	۲۲۰	۲۶۵/۵	۳۲۵	۳۷۴

$$\bar{x} = \frac{\sum x_i}{n} = \frac{۲۵ + \cdots + ۵۵}{۷} = ۴۰ \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{۴/۱ + \cdots + ۶/۸}{۷} = ۵/۳۴$$

$$s_{xx} = \sum x_i^۲ - n\bar{x}^۲ = ۱۱۹۰۰ - (۷ \times ۴۰^۲) = ۷۰۰$$

$$s_{xy} = \sum x_i y_i - n\bar{x}\bar{y} = ۱۵۶۶/۵ - (۷ \times ۴۰ \times ۵/۳۴) = ۷۱/۳$$

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{۷۱/۳}{۷۰۰} = ۰/۱۰۲ \qquad \& \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = ۵/۳۴ - (۰/۱۰۲ \times ۴۰) = ۱/۲۶$$

$\Rightarrow$ $\hat{y}_i = ۱/۲۶ + ۰/۱۰۲ x_i$:معادله خط رگرسیونی

مثال ۲

یک کمپانی می‌خواهد تأثیر تبلیغات را در فروش کالاهای تولید شده بررسی کند. بدین منظور داده‌های ثبت شده در جدول زیر را بعد از ۱۰ ماه به دست آورد.

الف- معادله خط رگرسیونی را به دست آورید.

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

ج- اگر هزینه تبلیغات ۱ باشد، حجم فروش را پیش‌بینی کنید.

ماه	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
هزینه تبلیغات	۱/۲	۰/۸	۱	۱/۳	۰/۷	۰/۸	۱	۰/۶	۰/۹	۱/۱
حجم فروش	۱۰۱	۹۲	۱۱۰	۱۲۰	۹۰	۸۲	۹۳	۷۵	۹۱	۱۰۰
$x_i^۲$	۱/۴۴	۰/۶۴	۱	۱/۶۹	۰/۴۹	۰/۶۴	۱	۰/۳۶	۰/۸۱	۱/۲۱
$x_i y_i$	۱۲۱/۲	۷۳/۶	۱۱۰	۱۵۶	۶۳	۶۵/۶	۹۳	۴۵	۸۱/۹	۱۱۰

$$\bar{x} = \frac{۱/۲ + \cdots + ۱/۱}{۱۰} = ۰/۹۴$$

$$\bar{y} = \frac{۱۰۱ + \cdots + ۱۰۰}{۱۰} = ۹۵/۴$$

مثال ۲

ادامه راه‌حل:

$$s_{xx} = \sum x_i^{۲} - n\bar{x}^{۲} = ۹/۲۸ - (۱۰ \times ۰/۹۴^{۲}) = ۰/۴۴۴$$

$$s_{xy} = \sum x_i y_i - n\bar{x}\bar{y} = ۹۲۴/۸ - (۱۰ \times ۰/۹۴ \times ۹۵/۴) = ۲۸/۰۴$$

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{۲۸/۰۴}{۰/۴۴۴} = ۶۳/۱۵$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = ۹۵/۴ - (۶۳/۱۵ \times ۰/۹۴) = ۳۶/۰۴$$

الف- $\hat{y}_i = ۳۶/۰۴ + ۶۳/۱۵\, x_i$

ج- $\hat{y}_i = ۳۶/۰۴ + (۶۳/۱۵ \times ۱) = ۹۹/۱۹$

به ازای هر واحد افزایش در x، مقدار براورد شده y به اندازه $\hat{\beta}$ تغییر می‌کند.
زمانی که $x = ۰$ است، مقدار متوسط y برابر با $\hat{\alpha}$ است.

واریانس $\hat{\beta}$

$$\begin{aligned}
Var(\hat{\beta}) = Var\left(\frac{s_{xy}}{s_{xx}}\right) &= \frac{۱}{s_{xx}^{۲}} Var\left[\sum_{i=۱}^{n}(x_i - \bar{x})(Y_i - \bar{Y})\right] \\
&= \frac{۱}{s_{xx}^{۲}} \sum_{i=۱}^{n}(x_i - \bar{x})^{۲} Var(Y_i - \bar{Y}) \\
&= \frac{۱}{s_{xx}^{۲}} \sum_{i=۱}^{n}(x_i - \bar{x})^{۲} Var(Y_i) \\
&= \frac{۱}{s_{xx}^{۲}} \sum_{i=۱}^{n}(x_i - \bar{x})^{۲}\sigma^{۲} \\
&= \frac{\sigma^{۲}}{s_{xx}}
\end{aligned}$$

برآورد $\sigma^۲$

برای برآورد $\sigma^۲$ از مجموع انحراف‌های Y_iها حول برآورد آن‌ها در هر سطح از x_i یعنی $\hat{Y}_i$ استفاده می‌کنیم:

$$\hat{\sigma^۲} = S^۲ = \frac{SSE}{n-۲} = \frac{\sum(Y_i - \hat{Y}_i)^۲}{n-۲} = \frac{\sum(Y_i - \hat{\alpha} - \hat{\beta}x_i)^۲}{n-۲} = \frac{s_{yy} - \hat{\beta}s_{xy}}{n-۲}$$

اثبات:

$$\begin{aligned}
SSE &= \sum_{i=۱}^{n}\left[Y_i - \hat{\alpha} - \hat{\beta}x_i\right]^۲ = \sum_{i=۱}^{n}\left[Y_i - \bar{Y} + \hat{\beta}\bar{x} - \hat{\beta}x_i\right]^۲ \\
&= \sum_{i=۱}^{n}\left[(Y_i - \bar{Y}) - \hat{\beta}(x_i - \bar{x})\right]^۲ \\
&= \sum_{i=۱}^{n}\left(Y_i - \bar{Y}\right)^۲ - ۲\hat{\beta}\sum_{i=۱}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) + \hat{\beta}^۲\sum_{i=۱}^{n}(x_i - \bar{x})^۲ \\
&= s_{yy} - ۲\hat{\beta}s_{xy} + \hat{\beta}^۲ s_{xx} = s_{yy} - \hat{\beta}s_{xx}
\end{aligned}$$

فاصله اطمینان برای $\hat{\beta}$

برای ساختن فاصله اطمینان نیاز به داشتن یک کمیت همراه با تابع توزیع آن داریم.

تاکنون هیچ فرضی روی توزیع متغیرهای تصادفی E_i و Y_i نگذاشتیم.

در مدل رگرسیونی $Y_i = \alpha + \beta x_i + E_i$ چون α، β و x_i ثابت هستند، با تعیین توزیع E_i می‌توان توزیع Y_i را تعیین کرد.

فرض کنید $E_1, \dots, E_n$ از یکدیگر مستقل بوده و از توزیع نرمال با میانگین صفر و واریانس σ^2 تبعیت کنند:
$$E_i \sim N(0, \sigma^2)$$

پس $$Y_i = \alpha + \beta x_i + E_i \sim N(\alpha + \beta x_i, \sigma^2)$$

از طرفی چون $$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{1}{s_{xx}} \sum (x_i - \bar{x})(Y_i - \bar{Y}) = \frac{1}{s_{xx}} \sum (x_i - \bar{x}) Y_i$$
ترکیب خطی از متغیرهای تصادفی مستقل نرمال است، پس $$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{s_{xx}})$$

فاصله اطمینان برای $\hat{\beta}$

با توجه به تعریف توزیع t داریم:

$$\frac{\hat{\beta}-\beta}{S/\sqrt{s_{xx}}}=\frac{\frac{\hat{\beta}-\beta}{\sigma/\sqrt{s_{xx}}}}{\sqrt{\frac{(n-۲)S^۲}{\sigma^۲(n-۲)}}}=\frac{N(۰,۱)}{\sqrt{\frac{\chi^۲_{(n-۲)}}{n-۲}}}\sim t_{(n-۲)}$$

پس یک فاصله اطمینان $(۱-\alpha)$ ٪۱۰۰ برای β به صورت زیر است:

$$\beta\in\left(\hat{\beta}-t_{۱-\frac{\alpha}{۲},(n-۲)}\frac{s}{\sqrt{s_{xx}}}\ ,\ \hat{\beta}+t_{۱-\frac{\alpha}{۲},(n-۲)}\frac{s}{\sqrt{s_{xx}}}\right)$$

که در آن $\hat{\beta}=\frac{s_{xy}}{s_{xx}}$ و $s^۲=\frac{SSE}{n-۲}=\frac{s_{yy}-\hat{\beta}s_{xy}}{n-۲}$ برقرار است.

مثال ۳

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

الف- براورد خط رگرسیونی را بنویسید.

ب- یک فاصله اطمینان ۹۵ درصدی برای β بسازید.

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

$$s_{xx} = \sum x_i^{۲} - n\bar{x}^{۲} = ۸۵۴/۹۱۷$$

$$s_{xy} = \sum x_i y_i - n\bar{x}\bar{y} = ۲۳۰$$

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = ۰/۲۶۹ \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = -۰/۹۴۸$$

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = -۰/۹۴۸ + ۰/۲۶۹x$$

مثال ۳

$$s_{yy} = \sum y_i^۲ - n\bar{y}^۲ = ۷۴$$

$$s^۲ = \frac{s_{yy} - \hat{\beta} s_{xy}}{n-۲} = ۱/۲۱۳$$

$$۱ - \alpha = ۰/۹۵ \quad \Rightarrow \quad ۱ - \frac{\alpha}{۲} = ۱ - \frac{۰/۰۵}{۲} = ۰/۹۷۵ \quad \Rightarrow \quad t_{۰/۹۷۵,(۱۲-۲)} = ۲/۲۳$$

$$\beta \in \left(\hat{\beta} - t_{۱-\frac{\alpha}{۲},(n-۲)} \frac{s}{\sqrt{s_{xx}}} \ , \ \hat{\beta} + t_{۱-\frac{\alpha}{۲},(n-۲)} \frac{s}{\sqrt{s_{xx}}} \right)$$

$$\beta \in \left(۰/۲۶۹ - ۲/۲۳ \times \sqrt{\frac{۱/۲۱۳}{۸۵۴/۹۱۷}} \ , \ ۰/۲۶۹ + ۲/۲۳ \times \sqrt{\frac{۱/۲۱۳}{۸۵۴/۹۱۷}} \right)$$

$$\beta \in (۰/۱۸۵ \ , \ ۰/۳۵۳)$$

مراحل انجام آزمون فرضیه‌های مربوط به شیب خط رگرسیونی به صورت زیر است:

۱- نوشتن فرضیه‌های آزمون (شبیه به مرحله ۳)

۲- محاسبه آماره آزمون $T_0 = \frac{\hat{\beta}-\beta_0}{S/\sqrt{s_{xx}}}$

۳- تعیین ناحیه بحرانی

$$\begin{cases} H_0 : \beta = \beta_0 \\ H_1 : \beta \neq \beta_0 \end{cases} \qquad \begin{cases} H_0 : \beta \leq \beta_0 \\ H_1 : \beta > \beta_0 \end{cases} \qquad \begin{cases} H_0 : \beta \geq \beta_0 \\ H_1 : \beta < \beta_0 \end{cases}$$

$$C : |T_0| > t_{1-\frac{\alpha}{2},(n-2)} \qquad C : T_0 > t_{1-\alpha,(n-2)} \qquad C : T_0 < -t_{1-\alpha,(n-2)}$$

۴- نتیجه‌گیری بر مبنای مرحله ۳

مثال ۴

یک سازنده‌ی مواد افزودنی به بنزین ادعا می‌کند که ماده‌ی افزودنی A در کاهش اکسید نیتروژن خارج شده از خودرو مؤثر است. برای این منظور ۱۰ خودرو از یک مدل تحت آزمایش قرار می‌گیرند. ابتدا میزان اکسید نیتروژن خارج شده از هر ماشین بدون اضافه کردن ماده‌ی A اندازه گرفته می‌شود. سپس به هر یک میزان معینی از ماده A را به باک پر بنزین اضافه کرده و میزان نیتروژن خارج شده از آن خودروها را اندازه‌گیری کرده‌اند. تقلیل در میزان اکسید نیتروژن به عنوان متغیر پاسخ ثبت شده است.

الف- براورد خط رگرسیونی را به دست آورید.

ب- آیا دلیل قانع‌کننده‌ای برای این وجود دارد که افزایش ماده A باعث کاهش اکسید نیتروژن می‌شود؟ $(\alpha = ۰/۰۵)$

میزان ماده A	۱	۱	۲	۳	۴	۴	۵	۶	۶	۷
تقلیل در میزان NO	۲/۱	۲/۵	۳/۱	۳	۳/۸	۳/۲	۴/۳	۳/۹	۴/۴	۴/۸

$$s_{xx} = \sum x_i^۲ - n\bar{x}^۲ = ۱۹۳ - (۱۰ \times ۳/۹^۲) = ۴۰/۹$$

$$s_{yy} = \sum y_i^۲ - n\bar{y}^۲ = ۱۳۰/۰۵ - (۱۰ \times ۳/۵۱^۲) = ۶/۸۴۹$$

$$s_{xy} = \sum x_i y_i - n\bar{x}\bar{y} = ۱۵۲/۷ - (۱۰ \times ۳/۹۳/۵۱) = ۱۵/۸۱$$

مثال ۴

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = ۰/۳۸۷ \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = ۲$$

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = ۲ + ۰/۳۸۷\, x$$

ب-

$$\begin{cases} H_۰ : \beta = ۰ \\ H_۱ : \beta > ۰ \end{cases}$$

$$\hat{\sigma^۲} = \frac{SSE}{n-۲} = \frac{s_{yy} - \hat{\beta}s_{xy}}{n-۲} = \frac{۶/۸۴۹ - (۰/۳۸۷ \times ۱۵/۸۱)}{۱۰ - ۲} = ۰/۰۹۲$$

$$T_۰ = \frac{\hat{\beta} - \beta_۰}{S/\sqrt{s_{xx}}} = \frac{۰/۳۸۷ - ۰}{\sqrt{\frac{۰/۰۹۲}{۴۰/۹}}} = ۸/۱۴۱$$

$$C : T_۰ > t_{۱-\alpha,(n-۲)} \qquad \Rightarrow \qquad ۸/۱۴۱ > ۱/۸۶$$

$$\alpha = ۰/۰۵ \qquad \Rightarrow \qquad t_{۰/۹۵,(۱۰-۲)} = ۱/۸۶$$

فرض صفر رد می‌شود؛ یعنی افزایش ماده افزودنی A باعث افزایش تقلیل در میزان اکسید نیتروژن می‌شود.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.945[a]	.892	.879	.30365

a. Predictors: (Constant), Material

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.111	1	6.111	66.284	.000[b]
	Residual	.738	8	.092		
	Total	6.849	9			

a. Dependent Variable: NitrogenOxides

b. Predictors: (Constant), Material

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.002	.209		9.600	.000
	Material	.387	.047	.945	8.141	.000

a. Dependent Variable: NitrogenOxides

ضریب همبستگی خطّی

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

برای سنجش میزان وابستگی دو متغیر تصادفی X و Y از معیاری به نام ضریب همبستگی خطی استفاده می‌شود:

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

برای براورد ضریب همبستگی، یک نمونه تصادفی $(X_1, Y_1), \ldots, (X_n, Y_n)$ را انتخاب می‌کنیم و از روی این نمونه کوواریانس X و Y، واریانس X و واریانس Y را به صورت زیر براورد می‌کنیم:

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

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

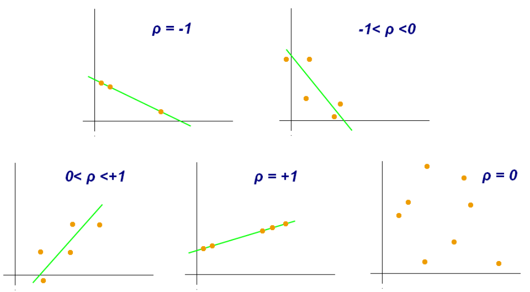
$$\sigma_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{S_{XY}}{n-1}$$

ضریب همبستگی خطی

در نتیجه براوردگر ضریب همبستگی خطی به صورت زیر است:

$$\hat{\rho} = R = \frac{S_{XY}}{\sqrt{S_{XX}\, S_{YY}}}$$

- توجه کنید که همواره $-۱ \leq R \leq ۱$



مثال ۵

در مثال ۴ ضریب همبستگی نمونه را به دست آورده و آن را تحلیل کنید.

راه‌حل:

$$\hat{\rho} = R = \frac{s_{xy}}{\sqrt{s_{xx}\, s_{yy}}} = \frac{۱۵/۸۱}{\sqrt{۴۰/۹ \times ۶/۸۴۹}} = ۰/۹۴۵$$

چون مقدار R به یک نزدیک است پس یک رابطه خطی قوی در جهت مثبت بین X و Y برقرار است.