فصل ۱۱ تحلیل رگرسیون خطی

۱۱-۱ مقدمه

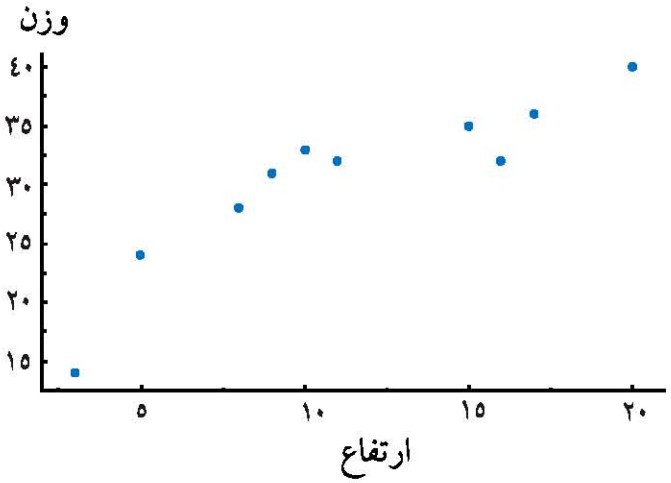
یکی از اهداف اغلب تحقیقات علمی بررسی روابط بین متغیرهای مستقل و پاسخ است. در فصل قبل نحوه بررسی اثر یک یا چند عامل بر متغیر پاسخ در قالب تحلیل واریانس، مورد بحث قرار گرفت. در اینگونه مطالعات محقق می‌توانست سطوح مختلف متغیر مستقل را بطور تصادفی به واحدهای آزمایشی نسبت دهد. مثلاً برای بررسی اثر رژیم غذایی بر روی دام‌ها، به هر دام به طور تصادفی یک نوع رژیم غذایی داده می‌شود. این در حالی است که در مورد بعضی متغیرها مانند اثر وزن بیماران بر روی فشار خون، محقق در تعیین وزن بیماران نقشی ندارد. در اینجا متغیر وزن یک متغیر مشاهده‌ای است زیرا مقادیر آن به طور اتفاقی وجود دارد. در این موارد برای توصیف رابطه متغیرهای مشاهده‌ای با متغیر پاسخ از مدل رگرسیونی استفاده می‌شود. مدل رگرسیونی می‌تواند رابطه متغیرهای کمکی و پاسخ را به صورت خطی یا غیر خطی بیان کند. از مدل‌های رگرسیونی برای توصیف داده‌ها، پیش‌بینی و کنترل متغیر پاسخ استفاده می‌شود.

قبل از مطالعه تحلیل رگرسیون خطی، نمودار پراکنش و معیارهای کواریانس و ضریب همبستگی که اندازه و جهت رابطه بین دو متغیر را نشان می‌دهند، شرح داده می‌شوند.

نمودار پراکنش

هرگاه مشاهدات برای هر عنصر شامل مقادیر دو متغیر باشد، برای بررسی ارتباط بین دو متغیر می‌توان از نمودار پراکنش استفاده کرد. برای رسم این نمودار یکی از متغیرها روی محور طولی و متغیر دیگر روی محور عرضی در نظر گرفته شده و نقاط با مختصات حاصل از مقادیر دو متغیر رسم می‌شوند. فرض کنید ارتفاع و وزن ۱۰ بوته از یک گونه گیاهی به صورت زوج‌های

(۳,۱٤)، (۹,۳۱)، (۱۱,۳۲)، (۱۰,۳۳)، (۱۵,۳۵)، (۲۰,٤۰)، (۸,۲۸)، (۵,۲٤)، (۱۷,۳٦)، (۱٦,۳۲) اندازه‌گیری شده است. نمودار پراکنش این داده‌ها در نمودار ۱۱-۱ نشان داده شده است.



نمودار ۱۱-۱. نمودار پراکنش ارتفاع و وزن ۱۰ بوته از یک گونه گیاهی.

۱۱-۲ کواریانس

کواریانس دو متغیر تصادفی x و y به صورت

$$E(x-\mu_x)(y-\mu_y)=E(xy)-\mu_x\mu_y$$

تعریف و با نماد $cov(x,y)$ یا σ_{xy} نشان داده می‌شود. کواریانس معیاری برای سنجش رابطه خطی بین دو متغیر تصادفی است. حاصلضرب $(x-\mu_x)(y-\mu_y)$ ممکن است مثبت یا منفی باشد، پس امید ریاضی آن یعنی σ_{xy} بر حسب اینکه مثبت یا منفی باشد، بیان می‌کند که x و y به طور متوسط و بصورت خطی در یک جهت یا در دو جهت مخالف حرکت می‌کنند. کواریانس در حالت کلی پارامتری مربوط به جامعه است و تنها با در دست داشتن توزیع احتمالی توأم متغیرهای تصادفی (x,y) قابل محاسبه است. از این رو در کاربردها از برآورد نمونه‌ای آن استفاده می‌شود. اگر نمونه‌ای تصادفی از جامعه به صورت $(x_1,y_1),\cdots,(x_n,y_n)$ در اختیار باشد، در اینصورت

$$s_{xy}=\hat{\sigma}_{xy}=\frac{SP_{xy}}{n-1}=\frac{\sum(x-\overline{x})(y-\overline{y})}{n-1}$$

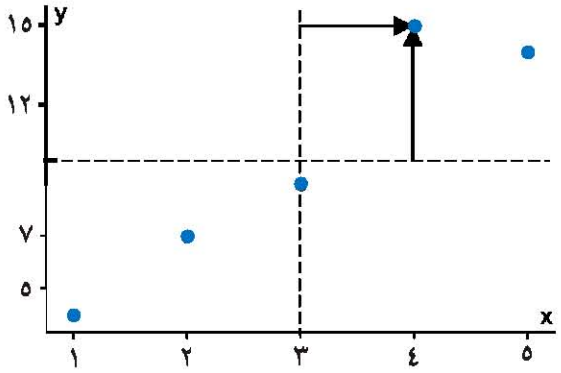
برآوردگری ناأریب برای σ_{xy} است. فرض کنید در آزمایشی مقادیر محصول به ازای مقادیر مختلف کود به صورت جدول ۱۱-۱ بوده است.

جدول ۱۱-۱. مقادیر مشاهده شده عملکرد تحت مقادیر مختلف کود.

x (میزان کود)	۱	۲	۳	٤	٥
y (میزان عملکرد)	٤	۷	۹	۱٥	۱٤

کواریانس بین دو متغیر x (میزان کود) و y (میزان محصول) برای نمونه مورد بررسی به صورت زیر برآورد می‌شود.

$$s_{xy} = \frac{\sum(x-\overline{x})(y-\overline{y})}{n-۱} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{n-۱} = \frac{۱۷٥ - \frac{۱٥\times٤۹}{٥}}{٤} = ۷$$



نمودار ۱۱-۲. نمودار پراکنش مقادیر مشاهده شده محصول (y) تحت مقادیر مختلف کود (x).

پراکنش نقاط مربوطه در این مثال نشان می‌دهد که دو متغیر گرایش به همراهی با هم دارند (نمودار ۱۱-۲). در این مثال حاصلضرب $(x-\mu_x)(y-\mu_y)$ برای هر کدام از نقاط مقداری مثبت به دست می‌آید. در نتیجه اگر مقدار $(x-\mu_x)(y-\mu_y)$ برای همه نقاط محاسبه و با هم جمع شوند، میانگین آنها که کواریانس نام دارد، می‌تواند به عنوان ملاکی برای میزان هم‌تغییری خطی x و y مد نظر باشد. مثبت بودن کواریانس در این مثال نشان می‌دهد که متغیرهای x و y بصورت خطی در یک راستا حرکت می‌کنند.

کواریانس شاخصی است که ارتباط خطی بین دو متغیر را منعکس می‌کند و بصورت

$$\begin{aligned}\sigma_{xy} &= E(x-\mu_x)(y-\mu_y)\\ &= E(xy)-\mu_x\mu_y\end{aligned}$$

تعریف می‌شود. برآورد نمونه‌ای کواریانس بصورت زیر است.

$$\hat{\sigma}_{xy} = s_{xy} = \frac{SP_{xy}}{n-1} = \frac{\sum(x-\overline{x})(y-\overline{y})}{n-1}$$

در عمل به جای کواریانس از ضریب همبستگی پیرسون برای نشان دادن میزان ارتباط بین دو متغیر استفاده می‌شود. زیرا کواریانس به واحد اندازه‌گیری وابسته بوده و با توجه به میانگین نمونه‌ها می‌تواند هر مقداری را اختیار کند.

۱۱-۳ ضریب همبستگی پیرسون

ضریب همبستگی پیرسون یا ضریب همبستگی خطی شدت و جهت همبستگی خطی بین دو متغیر تصادفی را نشان می‌دهد و با اعمال تغییراتی در تعریف کواریانس بصورت

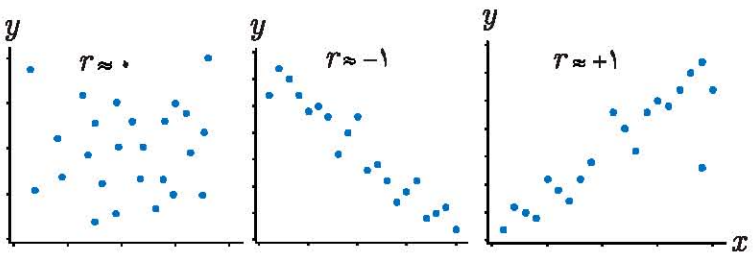
$$\rho_{(x,y)} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

تعریف می‌شود. ضریب همبستگی برخلاف کواریانس به واحد اندازه‌گیری وابسته نیست و مقادیر آن محدود به فاصله (۱و۱-) است. ضریب همبستگی نیز کمیتی مربوط به جامعه است، از اینرو در کاربردها از برآورد نمونه‌ای آن استفاده می‌شود. اگر $(x_1,y_1),\cdots,(x_n,y_n)$ نمونه‌ای تصادفی از جامعه باشد، در اینصورت

$$r_{(x,y)} = \hat{\rho}_{(x,y)} = \frac{s_{xy}}{s_x s_y}$$

برآوردگری مناسب برای $\rho_{(x,y)}$ است. مقدار ضریب همبستگی نزدیک به ۱+ نشان‌دهنده وجود ارتباط خطی مثبت بین x و y است، یعنی با افزایش x مقدار y نیز افزایش می‌یابد و برعکس. در صورتی که r نزدیک به ۱- باشد، بین دو متغیر ارتباط خطی منفی وجود داشته و با افزایش یکی

از آنها دیگری کاهش می‌یابد. مقادیر r نزدیک به صفر نیز بیانگر عدم وجود ارتباط خطی بین دو متغیر است (نمودار ۱۱-۳).



نمودار ۱۱-۳. نمودار پراکنش x و y به ازای مقادیر مختلف ضریب همبستگی.

مثال ۱۱-۱. در آزمایشی به منظور بررسی ارتباط جوانه‌زنی بذر با میزان دما، درصد جوانه‌زنی در اثر ۵ دمای مختلف پس از مدت مشخصی به دست آمده است. فرض کنید نتایج نمونه مورد بررسی به صورت زیر بوده است.

دما (x)	۱	۲	۳	٤	٥
درصد جوانه‌زنی (y)	۱	۱	۲	۲	٤

مقدار r برای داده‌های مثال ۱۱-۱ به صورت زیر به دست می‌آید.

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

که در آن:

$$SP_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = ۳۷ - \frac{۱۰ \times ۱۵}{۵} = ۳۷ - ۳۰ = ۷$$

$$SS_x = \sum x_i^{۲} - \frac{(\sum x_i)^{۲}}{n} = ۵۵ - \frac{۱۵^{۲}}{۵} = ۵۵ - ٤٥ = ۱۰$$

$$SS_y = \sum y_i^{۲} - \frac{(\sum y_i)^{۲}}{n} = ۲٦ - \frac{۱۰^{۲}}{۵} = ۲٦ - ۲۰ = ٦$$

لذا r برابر با

$$r = \frac{۷}{\sqrt{۱۰ \times ٦}} = ۰/۹۰٤$$

خواهد بود. این مقدار که از روی نمونه محاسبه شده است تخمینی از ضریب همبستگی جامعه (ρ) می‌باشد. ممکن است آزمون فرض $H_0 : \rho = 0$ در مقابل $H_1 : \rho \neq 0$ مد نظر باشد. بر اساس فرض صفر، متغیرهای x و y بصورت خطی ناهمبسته‌اند، یعنی بین آنها رابطه خطی وجود ندارد. البته به این معنی نیست که بین این دو متغیر هیچگونه رابطه دیگری نیز وجود ندارد.

در صورتی که مقدار ρ نزدیک به ۱+ یا ۱- باشد، توزیع آماره r متقارن نبوده و از این رو نمی‌توان با استفاده از الگوهایی نظیر توزیع t و z اقدام به آزمون فرض نمود. فقط در حالتی که ρ صفر باشد، توزیع آماره r، یک توزیع نرمال با میانگین صفر و واریانس $(1-r^2)\Big/(n-2)$ خواهد بود. پس اگر آزمون $H_0 : \rho = 0$ مد نظر باشد، آماره t به صورت

$$t = \frac{r-\rho}{\sigma_r} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

محاسبه شده و با t جدول با درجه آزادی $n-2$ مقایسه می‌شود. اگر ρ برابر با صفر نباشد، توزیع r نامتقارن و دارای کجی خواهد بود. پس در مواردی که فرضی غیر از $H_0 : \rho = 0$ مورد آزمون باشد، نمی‌توان از آماره t یا z در آزمون فرض استفاده کرد. در چنین مواردی r ابتدا با استفاده از تبدیل $z' = 0.5 \ln\frac{1+r}{1-r}$ به متغیر جدیدی با توزیع نرمال تبدیل می‌شود تا امکان انجام آزمون با استفاده از توزیع z به وجود آید. سپس آزمون‌های آماری بر روی متغیر نرمال حاصل z'، انجام می‌شود. ثابت می‌شود که z' دارای توزیع تقریباً نرمال با انحراف معیار $\sqrt{1/(n-3)}$ می‌باشد. همچنین می‌توان ابتدا فاصله اطمینان $(1-\alpha)$ را برای z' محاسبه نموده و سپس با استفاده از رابطه زیر آن را به فاصله اطمینان r برگرداند.

$$r = \frac{e^{2z'}-1}{e^{2z'}+1}$$

مثال ۱۱-۲. ضریب همبستگی بین درصد پروتئین و درصد چربی، در ۵۰ مورد اندازه‌گیری روی بذر سویا، برابر با ۰/۵۲۷ برآورد شده است. فرض $H_0 : \rho = 0/5$ را در مقابل $H_1 : \rho \neq 0/5$ در سطح ۰/۰۵ بیازمایید.

با استفاده از رابطه $z' = 0/5 \ \ln\frac{1+r}{1-r}$ مقادیر تبدیلی بصورت $z'_{0/527}$ و $z'_{0/5}$ بدست می‌آیند، لذا

$$z = \frac{z'_r - z'_\rho}{\sigma_{z'}} = \frac{z'_{0/527} - z'_{0/5}}{\sqrt{\frac{1}{n-3}}} = \frac{0/5864 - 0/549}{\sqrt{\frac{1}{47}}} = 0/2$$

چون مقدار آماره z از z جدول (۱/۹۶) کوچکتر است، دلیل کافی برای رد فرض $\rho = 0/5$ در سطح ۰/۰۵ وجود ندارد.

همچنین می‌توان فاصله اطمینان $1-\alpha$ مثلاً ۰/۹۵ را برای z' بصورت

$$z'_r \pm z_{\alpha/2} s_{z'} = 0/5864 \pm 1/96(0/146) = 0/5864 \pm 0/286$$

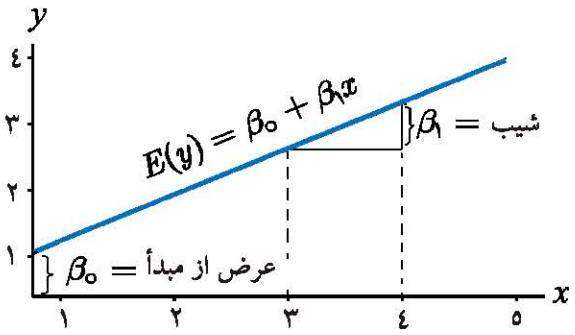
محاسبه کرد. بدین معنی که با احتمال ۰/۹۵، z' در فاصله ۰/۳ تا ۰/۸۸ قرار دارد. با تبدیل z' به r می‌توان حدود اعتماد ۹۵٪ را برای r به دست آورد. مقدار r برای $z' = 0/3$، برابر با ۰/۲۹ و برای $z' = 0/88$ برابر با ۰/۷ می‌باشد. بنابراین با احتمال ۹۵٪، فاصله ۰/۲۹ تا ۰/۷۰ ضریب همبستگی جامعه را در بر دارد.

۱۱-۴ رگرسیون خطی ساده

رگرسیون خطی ساده ارتباط خطی دو متغیر را بررسی می‌کند. در رگرسیون خطی ساده رابطه بین متغیر مستقل x و متغیر پاسخ y توسط معادله یک خط مستقیم در محور مختصات نشان داده می‌شود. معادله چنین خطی به صورت $E(y) = \beta_0 + \beta_1 x$ نوشته می‌شود که در آن،

β_0 عرض از مبدأ خط بوده و نقطه‌ای است که در آن، خط رگرسیون محور y را قطع می‌کند.

β_1 نیز شیب خط رگرسیون است و بیانگر میزان افزایش در متغیر y به ازای یک واحد افزایش در متغیر x می‌باشد.



نمودار ۱۱-٤. مدل رگرسیون خطی ساده.

به عنوان مثال درصد جوانه‌زنی بذر با افزایش دما رابطه مستقیم دارد. چون میزان جوانه‌زنی (y) به دما بستگی دارد، آن را متغیر پاسخ یا وابسته می‌نامند. میزان جوانه‌زنی به ازای یک سطح دما در تکرارهای مختلف ثابت نبوده و به دلیل تأثیر عوامل ناشناخته، مقداری متغیر است. اگر میزان دما به طور دلخواه توسط محقق تعیین شود، از نوع داده‌های آزمایشگاهی و در صورتی که محقق نقشی در تعیین آن نداشته باشد، از نوع داده‌های مشاهده‌ای خواهد بود که در هر دو صورت متغیر دما را متغیر مستقل می‌نامند (برای مثال داده‌های مربوط به دما و میزان جوانه زنی برای چند گیاه به طور اتفاقی ثبت شده یا میزان جوانه زنی یک گونه گیاهی در مناطق با دماهای مختلف اندازه‌گیری شده و محقق می‌خواهد اثر دما را بر آن بررسی کند).

مثال ۱۱-۳. فرض کنید در یک آزمایش، درصد جوانه‌زنی در اثر ۵ دمای مختلف پس از مدت مشخصی به صورت زیر بوده است.

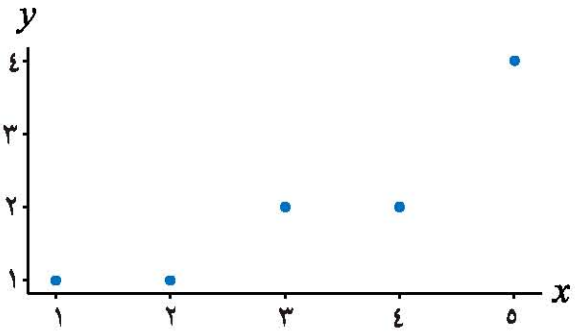
دما (x)	١	٢	٣	٤	٥
درصد جوانه‌زنی (y)	١	١	٢	٢	٤

نمودار پراکنش درصد جوانه‌زنی به ازای دماهای مختلف (نمودار ۱۱-۵) نشان می‌دهد که دما بر روی میزان جوانه‌زنی اثر دارد. چگونگی این اثر به وسیله معادله‌ای که x و y را به هم ربط می‌دهد توصیف می‌شود. برآورد یک معادله، هم‌ارز برازاندن یک خط به این نقاط است، که خط رگرسیون y بر حسب x نامیده می‌شود و می‌تواند در پیش‌بینی درصد جوانه‌زنی (y) به ازای دماهای مختلف (x) بکار رود. معادله کلی خط رگرسیونی برای میانگین درصد جوانه‌زنی در جامعه به صورت $E(y) = \beta_0 + \beta_1 x$ نشان داده می‌شود که در آن β_0 و β_1 مقادیری ثابت هستند. بدیهی است که مقدار مشاهده شده متغیر وابسته (درصد جوانه‌زنی) به ازای یک مقدار

مشخص دما (متغیر مستقل) دقیقاً برابر با مقدار پیش‌بینی شده نبوده و مقداری خطا وجود خواهد داشت. لذا مدل رگرسیون را می‌توان به صورت

$$y = \beta_0 + \beta_1 x + \varepsilon$$

نیز نوشت، که در آن ε نشاندهنده خطا در تعیین y بر اساس x است. پارامترهای β_0 و β_1 ضرایب رگرسیونی نامیده شده و بر اساس مشاهدات نمونه برآورد می‌شوند.



نمودار ۱۱-۵. نمودار پراکنش مربوط به مثال ۱۱-۳ بر روی محور مختصات.

۱۱-۵ برازش خط رگرسیونی

برازش یک مدل رگرسیونی به مجموعه‌ای از مشاهدات، به معنی برآورد پارامترهای مدل بر اساس آن مجموعه از مشاهدات است. خطوط مختلفی را می‌توان از بین نقاط مربوط به نمونه مورد بررسی در مثال ۱۱-۳ عبور داد، ولی مناسب‌ترین آنها خطی است که به ازای آن، مجموع توان‌های دوم خطا حداقل باشد (نمودار ۱۱-۶). معادله خط مذکور که بر اساس اطلاعات نمونه به دست می‌آید، به صورت $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ نشان داده می‌شود. $\hat{y}$ برآورد مقدار y به ازای هر x و $\hat{\beta}_1$ برآورد β_1 است. اگر $(x_1, y_1), \cdots, (x_n, y_n)$ نمونه‌ای تصادفی باشد، برآورد مجموع توان‌های دوم خطا بصورت

$$SS_E = \sum \left(y_i - \hat{y}\right)^2 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x)]^2$$

و برآورد ضریب رگرسیون بصورت

$$\hat{\beta}_1 = \frac{SP_{xy}}{SS_x}, \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

می‌باشد. در مثال ۱۱-۳، این مقادیر بصورت

$$SP_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = ۳۷ - \frac{۱۰ \times ۱۵}{۵} = ۳۷ - ۳۰ = ۷$$

$$SS_x = \sum x_i^{۲} - \frac{(\sum x_i)^{۲}}{n} = ۵۵ - \frac{۱۵^{۲}}{۵} = ۵۵ - ۴۵ = ۱۰$$

$$\hat{\beta}_۱ = \frac{SP_{xy}}{SS_x} = \frac{۷}{۱۰} = ۰/۷$$

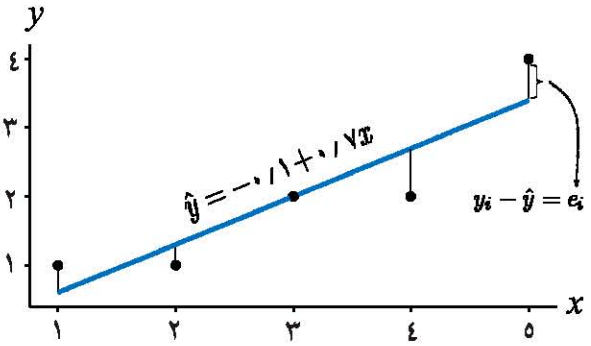
$$\hat{\beta}_۰ = \overline{y} - \hat{\beta}_۱ \overline{x} = \frac{\sum y_i}{۵} - \hat{\beta}_۱ \frac{\sum x_i}{۵} = \frac{۱۰}{۵} - ۰/۷(\frac{۱۵}{۵}) = ۲ - ۰/۷(۳) = -۰/۱$$

بدست می‌آیند. لذا معادله خط رگرسیون برآورد شده به صورت

$$\hat{y} = \hat{\beta}_۰ + \hat{\beta}_۱ x = -۰/۱ + ۰/۷x$$

است. با استفاده از این معادله می‌توان مقدار y را به ازای هر مقدار مشخص x پیش‌بینی کرد (نمودار ۱۱-۶). مثلاً به ازای $x = ۲$ مقدار y به صورت زیر پیش‌بینی خواهد شد.

$$\hat{y} = -۰/۱ + ۰/۷x = -۰/۱ + ۰/۷(۲) = ۱/۳$$



نمودار ۱۱-۶. خط برازنده شده به داده‌های مثال ۱۱-۳.

۱۱-۶ فرض‌های اولیه تحلیل رگرسیونی

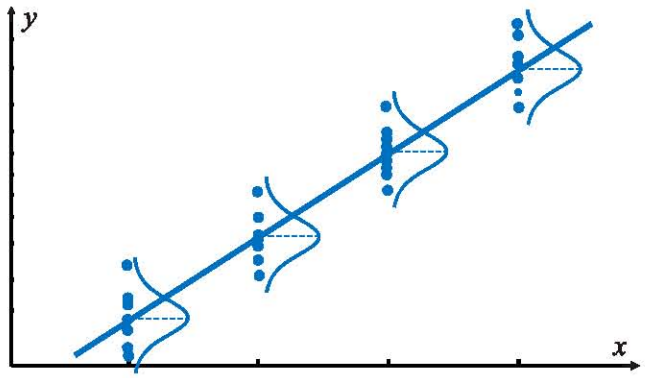
در تحلیل رگرسیونی، برآورد پارامترهای مدل رگرسیونی بر اساس روش حداقل مجموع توان‌های دوم با در نظر گرفتن فرضیات اولیه‌ای در مورد جمله خطا صورت می‌پذیرد. این فرضیات را می‌توان به صورت زیر بیان کرد.

۱- میانگین توزیع خطاها برابر با صفر است. به این معنی که خطاهای مثبت و منفی همدیگر را خنثی کرده و در صورت تکرار آزمایش به ازای یک مقدار مشخص x، میانگین مقادیر y برابر با $E(y) = \beta_0 + \beta_1 x$ خواهد بود.

۲- واریانس خطا به ازای هر مقدار x ثابت و برابر با مقداری مانند σ^2 است.

۳- خطاها از یکدیگر و از x ناهمبسته می‌باشند. به عبارت دیگر مقدار خطا (ε) به ازای یک مقدار y هیچ اثری بر روی خطای مرتبط با مقدار دیگری از y ندارد (نمودار ۱۱-۷).

بطور خلاصه، خطاها متغیرهایی تصادفی و ناهمبسته با میانگین صفر و واریانس ثابت تلقی می‌شوند. نکته قابل توجه آن است که برای برآورد پارامترها، فرض نرمال بودن خطاها ضروری نیست. زیرا روش کمترین توان‌های دوم روشی جبری است و به توزیع خاصی وابسته نیست. البته برای استنباط در مورد ضرایب رگرسیونی یعنی برای انجام آزمون فرض و ساختن فاصله اطمینان، فرض نرمال بودن خطاها ضرورت پیدا می‌کند.



نمودار ۱۱-۷. توزیع احتمال خطا به ازای مقادیر مختلف متغیر مستقل.

۱۱-۷ برآورد خطای مدل

هر اندازه تغییرات داده‌ها نسبت به خط رگرسیون برآورد شده بیشتر باشد، خطاهای تصادفی و در نتیجه واریانس خطای مدل رگرسیونی بیشتر شده و برآوردهای $\hat{\beta}_0$ و $\hat{\beta}_1$ و همچنین مقدار پیش‌بینی شده $\hat{y}$ به ازای یک مقدار مشخص x دقت کمتری خواهند داشت. به همین دلیل است که واریانس خطا (σ^2) در رابطه برآورد فاصله اطمینان و نیز محاسبه آماره آزمون به چشم می‌خورد. غالباً در عمل σ^2 نامعلوم است و لذا از طریق نمونه برآورد می‌شود. این برآورد که با s^2

نشان داده می‌شود، برابر است با مجموع توان‌دوم فواصل نقاط از خط رگرسیون تقسیم بر درجه آزادی. از آنجا که در ایجاد مدل رگرسیونی، دو پارامتر (β_0 و β_1) باید برآورد شود، $n-2$ درجه آزادی برای خطا باقی می‌ماند، لذا:

$$s^2 = \frac{SS_E}{df_E} = \frac{SS_E}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$$

$$SS_E = \sum(y_i - \hat{y}_i) = SS_y - \hat{\beta}_1 SP_{xy} = SS_y - \hat{\beta}_1^2 SS_x$$

$$SS_y = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

در مثال ۱۱-۳ در مورد میزان دما (x) و میزان جوانه‌زنی (y)، مجموع توان‌های دوم خطا به صورت زیر به دست می‌آید.

x	y	$\hat{y} = -۰/۱ + ۰/۷x$	$y_i - \hat{y}$	$(y_i - \hat{y})^2$
۱	۱	۰/٦	٤/۰	۰/۱٦
۲	۱	۱/۳	-۰/۳	۰/۰۹
۳	۲	۲/۰	۰	۰
٤	۲	۲/۷	-۰/۷	۰/٤۹
۵	٤	۳/٤	۰/٦	۰/۳٦
			مجموع خطاها = ۰	$SS_E = ۱/۱$

$$s^2 = \frac{SS_E}{n-2} = \frac{۱/۱}{۳} = ۰/۳٦٦۷$$

انحراف معیار خطا که خطای معیار مدل رگرسیونی نیز نامیده می‌شود، برابر است با $s = \sqrt{۰/۳٦٦۷} = ۰/٦۰۵$ و نشان می‌دهد که در صورت نرمال بودن توزیع خطاها، حدود ۹۵٪ مقادیر مشاهده شده y در فاصله $2s$ از مقادیر پیش‌بینی شده قرار می‌گیرند. در مثال ۱۱-۳ همه مقادیر y در فاصله (۲×۰/٦۰۵) از $\hat{y}$ قرار گرفته‌اند.

جدول ۱۱-۲. حل مثال ۱۱-۳ و خروجی مربوطه در برنامه Minitab.

مراحل:

۱. مشاهدات مربوط به متغیر مستقل (x) و وابسته (y) را در ستون‌های اول و دوم صفحه داده‌ها وارد کنید.

۲. وارد مسیر مقابل شوید: **Stat > Regression > Regression...**

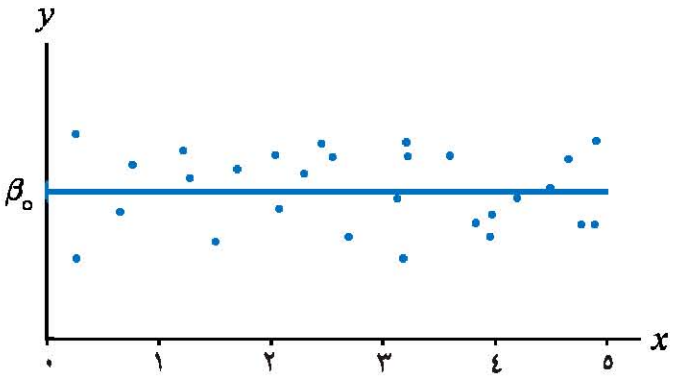
۳. در مقابل گزینه Response ستون دوم و در مقابل گزینه Predictors ستون اول (متغیر مستقل) را انتخاب کنید.

۴. بر روی OK کلیک کنید.

```
The regression equation is
y = - 0.100 + 0.700 x

Predictor      Coef  SE Coef      T      P
Constant   -0.1000   0.6351  -0.16  0.885
x           0.7000   0.1915   3.66  0.035

S = 0.605530   R-Sq = 81.7%   R-Sq(adj) = 75.6%

Analysis of Variance

Source           DF      SS      MS      F      P
Regression        1  4.9000  4.9000  13.36  0.035
Residual Error    3  1.1000  0.3667
Total             4  6.0000
```

۱۱-۸ بررسی کارایی مدل رگرسیونی

در عمل ضرایب رگرسیونی تنها بر اساس یک نمونه تصادفی برآورد می‌شوند. لذا اعتماد به این ضرایب بدون در نظر گرفتن توزیع برآوردگرهای مربوط به آنها صحیح نیست. به عبارتی تنها پس از انجام آزمون فرض و کسب اطمینان از معنی‌داری این ضرایب، می‌توان مدل بدست آمده را به جامعه تحت بررسی تعمیم داد. به کمک آزمون فرض‌ها می‌توان مؤثر بودن یا نبودن مدل رگرسیونی در پیش‌بینی مقادیر متغیر وابسته بر اساس مقادیر متغیر مستقل را آزمود. اگر متغیر x در مدل $y = \beta_0 + \beta_1 x + \varepsilon$ نقشی در پیش‌بینی مقادیر y نداشته باشد، شیب خط رگرسیون برابر با صفر خواهد بود، زیرا در این صورت میانگین y در اثر افزایش یا کاهش x تغییری نخواهد کرد (نمودار ۱۱-۸).



نمودار ۱۱–۸ مدل رگرسیونی $y = \beta_0 + \varepsilon$ که در آن شیب خط (β_1) برابر با صفر است.

برای بررسی این موضوع می‌توان فرض $H_0 : \beta_1 = 0$ را در مقابل $H_1 : \beta_1 \neq 0$ آزمود. چون انجام آزمون بدون اطلاع از توزیع برآوردگرهای ضرایب رگرسیونی (β_1 و β_1) ممکن نسیت، در این مرحله، فرض نرمال بودن مانده‌ها نیز به فرضیات اولیه مدل رگرسیونی افزوده می‌شود. در اینصورت $\hat{\beta}_1$ دارای توزیع نمونه‌ای نرمال با میانگین β_1 و انحراف معیار $\sigma_{\hat{\beta}_1} = \sigma / \sqrt{SS_x}$ خواهد بود. چون $\sigma_{\hat{\beta}_1}$ از روی نمونه برآورد می‌شود، لذا فرض‌های مذکور را می‌توان به کمک آماره

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\frac{s}{\sqrt{SS_x}}}$$

مورد آزمون قرار داد، که دارای توزیع t با n–۲ درجه آزادی است. در صورتی که قدر مطلق آماره t از t جدول برای سطح معنی‌داری مورد نظر و درجه آزادی $n - ۲$ بزرگتر باشد فرض صفر رد خواهد شد.

مثال ۱۱–٤. در مثال ۱۱–۳، فرض $H_0 : \beta_1 = 0$ را در مقابل $H_1 : \beta_1 \neq 0$ در سطح $\alpha = ۰/۰۵$ بیازمایید.

$$t = \frac{۰/۷ - ۰}{\frac{۰/٦۰۵۵}{\sqrt{۱۰}}} = \frac{۰/۷}{۰/۱۹} = ۳/٦۵$$

چون قدر مطلق مقدار آماره t از $t_{۰/۰۲۵,۳} = ۳/۱۸۲$ بزرگتر است، پس شیب خط رگرسیون برابر با صفر نیست. فاصله اطمینان ۰/۹۵ برای شیب خط به صورت

$$\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1} = ۰/۷ \pm ۳/۱۸۲(۰/۱۹) = ۰/۷ \pm ۰/۶$$

است. بدین معنی که فاصله ۰/۱ تا ۱/۳ با احتمال ۰/۹۵ شیب واقعی خط رگرسیون جامعه یعنی (β_1) را در بر دارد.

۱۱-۹ ضریب تعیین

همواره قسمتی از تغییرات کل، $SS_y = \sum(y_i - \overline{y})^2$، مربوط به خطا یعنی $SS_E = \sum(y_i - \hat{y})^2$ بوده و در صورتی که همه نقاط بر روی خط رگرسیون برآورد شده قرار گیرند، مجموع توان‌های دوم خطا صفر خواهد شد. لذا SS_y را می‌توان به دو بخش SS_E و $SS_y - SS_E$ تقسیم نمود. مقدار $SS_y - SS_E$ در واقع بخشی از تغییرات کل است که توسط مدل رگرسیون تبیین می‌شود و به همین دلیل مجموع توان‌های دوم رگرسیون (SS_R) نام دارد. ضریب تعیین (R^2) بیان می‌کند که چه مقدار از کل تغییرات توسط مدل رگرسیونی تبیین می‌شود. هر چه این مقدار بیشتر باشد رابطه بین x و y قوی‌تر است. ضریب تعیین در رگرسیون خطی ساده برابر است با مربع ضریب همبستگی و بصورت

$$R^2 = \frac{SS_y - SS_E}{SS_y} = \frac{SS_R}{SS_y}$$

برآورد می‌شود. مقدار ضریب تعیین در مثال ۱۱-۳ به صورت

$$R^2 = \frac{SS_y - SS_E}{SS_y} = \frac{۶ - ۱/۱}{۶} = ۰/۸۱۶$$

بدست می‌آید. پس می‌توان گفت که ۰/۸۱ تغییرات جوانه‌زنی (متغیر y) توسط دما (متغیر x) توجیه می‌شود.

۱۱-۱۰ بررسی کارایی مدل رگرسیونی با آزمون F

همانگونه که در نمودار ۱۱-۵ نشان داده شده است، اگر فرض $H_0 : \beta_1 = 0$ درست باشد خط رگرسیون افقی بوده و در نتیجه مجموع توان‌های دوم رگرسیون برابر با صفر خواهد بود. به عبارت دیگر میانگین توان‌های دوم رگرسیون نیز برآوردی از واریانس خطا (σ^2) می‌باشد. اگر H_0 درست نباشد مجموع توان‌های دوم رگرسیون برآوردی از واریانس خطا به اضافه واریانس رگرسیون خواهد بود. برای آزمون فرض $H_0 : \beta_1 = 0$ با استفاده از توزیع F کافی است که نسبت واریانس رگرسیونی به واریانس خطا محاسبه شده و با F جدول مربوط به درجات آزادی ۱ (مربوط به رگرسیون) و $n-2$ (مربوط به خطا) مقایسه شود. تأیید بزرگتر بودن واریانس رگرسیون نسبت به واریانس خطا بیانگر معنی‌دار بودن رگرسیون یا ضریب رگرسیون خواهد بود.

در مثال ۱۱-۳ میانگین توان‌های دوم خطا برابر با ۰/۳٦٦۷ بدست آمد که از تقسیم توان‌های دوم خطا (۱/۱) بر درجه آزادی آن (۳-۲=n-۲) بدست آمده است. چون رابطه $S_{yy} = SS_R + SS_E$ برقرار است، بنابراین مجموع توان‌های دوم رگرسیون در مثال ۱۱-۳ برابر با ٤/۹ = ۱/۱ − ٦ است. البته SS_R را می‌توان مستقیماً بصورت

$$\sum(\hat{y} - \overline{y})^2 = \hat{\beta}_1^2 \sum(x_i - \overline{x})^2 = \hat{\beta}_1^2 SS_x = \hat{\beta}_1 SP_{xy}$$

$$= (0/7)^2 \times 10 = 0/7 \times 7 = 4/9$$

محاسبه نمود. با توجه به اینکه SS_R رگرسیون دارای یک درجه آزادی است، SS_R برابر با MS_R خواهد بود. لذا نسبت دو واریانس تحت عنوان آماره F برابر با

$$F = \frac{MSR}{MSE} = \frac{4/9}{0/3667} = 13/36$$

است. با توجه به اینکه آماره F محاسبه شده از $F_{0/05,(1,3)} = 10/12$ بزرگتر است، پس رگرسیون معنی‌دار بوده و شیب خط برابر با صفر نیست. نتایج آزمون F را می‌توانید در خروجی برنامه Minitab (جدول ۱۱-۲) نیز مشاهده کنید.

۱۱-۱۱ پیش‌بینی بر اساس مدل رگرسیونی

پس از برآورد مدل مناسب برای ارتباط خطی موجود بین میزان کود و عملکرد، می‌توان از آن در پیش‌بینی استفاده کرد. پیش‌بینی تنها برای مقادیری از متغیر مستقل مجاز است که در دامنه xهای بکار رفته در برازش مدل، واقع باشند. زیرا در خارج از این فاصله اعتبار مدل از بین می‌رود و پیش‌بینی بر اساس خط رگرسیونی ممکن است دور از واقعیت باشد. برای مثال، میزان کود بر میزان محصول اثر مستقیم دارد به این معنی که افزایش کود باعث فراوانی محصول می‌شود ولی واضح است که مصرف بیش از حد کود هم کاهش میزان محصول را در پی خواهد داشت. لذا مدل رگرسیونی که بر اساس سطوح مناسب کود ساخته شود پیش‌بینی خوبی برای سطوح بالای کود ارائه نمی‌دهد.

پیش‌بینی می‌تواند به دو صورت باشد. می‌توان به ازای یک مقدار x، میانگین همه مقادیر y یعنی $E(y)$ را پیش‌بینی نمود یا اینکه به ازای همان مقدار x، مقدار y را فقط در یک آزمایش منفرد پیش‌بینی کرد. حالت اول در مثال ۱۱-۳ می‌تواند برآورد میانگین درصد جوانه‌زنی در دمای ۴°C در چندین تکرار باشد. مثال حالت دوم برآورد درصد جوانه‌زنی در دمای ۴°C در یک تکرار است. در هر دو مورد درصد جوانه‌زنی پیش‌بینی شده به صورت

$$\hat{y} = \beta_0 + \beta_1 x = -0.1 + 0.7(4) = 2.7$$

بدست می‌آید. تنها اختلاف بین دو نوع برآورد در دقت آنها می‌باشد. انحراف معیار میانگین جامعه همه مقادیر ممکن y (خطای معیار $\hat{y}$) به ازای مقدار ثابت $x = x_0$ برابر

$$\sigma_{\hat{y}} = \sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{SS_x}}$$

است. که در آن σ انحراف معیار خطا می‌باشد. این در حالی است که انحراف معیار برآورد $\hat{y}$ (خطای معیار پیش‌بینی) برابر

$$\sigma_{(y-\hat{y})} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{SS_x}}$$

است. فاصله اطمینان $\alpha - ۱$ برای میانگین y به ازای $x = x_۰$ به صورت زیر به دست می‌آید که در آن $t_{\alpha/۲}$ بر اساس $n - ۲$ درجه آزادی از جدول t استخراج می‌شود و s نیز برآوردی از σ است که از روی نمونه محاسبه می‌شود. در مثال ۱۱-۳ فاصله اطمینان ۰/۹۵ برای میانگین yها به ازای $x = ٤$ به صورت زیر است:

$$\hat{y} \pm t_{\alpha/۲} s \sqrt{\frac{۱}{n} + \frac{(x_۰ - \overline{x})^۲}{SS_x}} = ۲/۷ \pm ۳/۱۸۲(۰/٦۰۵)\sqrt{\frac{۱}{۵} + \frac{(٤ - ۳)^۲}{۱۰}} = ۲/۷ \pm ۱/۰۵$$

که در آن $\alpha/۲$ برابر با ۰/۰۲۵ در نظر گرفته شده است. لذا فاصله اطمینان ۰/۹۵ برای میانگین درصد جوانه‌زنی در دمای $^{\circ}C$ ٤ بصورت ۱/٦۵ تا ۳/۷۵ می‌باشد. این فاصله با استفاده از نمونه موجود به دست آمده است و در صورتی که با استفاده از یک نمونه بزرگتر اطلاعات بیشتری حاصل می‌شد، فاصله اطمینان کوتاه‌تری محاسبه می‌شد.

فاصله پیش‌بینی $\alpha - ۱$ برای یک y منفرد به ازای $x = x_۰$ به صورت

$$\hat{y} \pm t_{\alpha/۲} s \sqrt{۱ + \frac{۱}{n} + \frac{(x_۰ - \overline{x})^۲}{SS_x}}$$

می‌باشد. در مثال ۱۱-۳ فاصله پیش‌بینی ۰/۹۵ برای یک y منفرد به ازای $x = ٤$ به صورت

$$\hat{y} \pm t_{\alpha/۲} s \sqrt{۱ + \frac{۱}{n} + \frac{(x_۰ - \overline{x})^۲}{SS_x}} = ۲/۷ \pm ۳/۱۸۲(۰/٦۰۵)\sqrt{۱ + \frac{۱}{۵} + \frac{(٤ - ۳)^۲}{۱۰}} = ۲/۷ \pm ۲/۱۹$$

به دست می‌آید. لذا اگر یک بار دیگر دمای $^{\circ}C$ ٤ بررسی شود، درصد جوانه‌زنی با احتمال ۰/۹۵ بین ۰/۵۱ تا ٤/۸۹ قرار خواهد گرفت. در صورت استفاده از یک نمونه بزرگتر فاصله پیش‌بینی دقیقتری (با طول کوتاهتر) به دست خواهد آمد. در جدول ۱۱-۳ فاصله اطمینان برای $E(y)$ و فاصله پیش‌بینی ۹۵٪ برای y با استفاده از برنامه Minitab برای چند مقدار مختلف x نشان داده شده است.

جدول ۱۱-۳. حل مثال ۱۱-۳ و خروجی مربوطه در برنامه Minitab.

مراحل:

۱. مشاهدات مربوط به متغیر مستقل (x) و وابسته (y) را در ستون‌های اول و دوم صفحه داده‌ها وارد کنید.

۲. وارد مسیر **Stat > Regression > Regression...** شوید:

۳. در مقابل گزینه Response ستون دوم و در مقابل گزینه Predictors ستون اول (متغیر مستقل) را انتخاب کنید.

۴. بر روی گزینه Options کلیک کرده و در پنجره مربوطه در مقابل گزینه Predictions intervals for new... ستون مربوط به متغیر مستقل را وارد کنید (البته به جای آن می‌توان یک مقدار جدید x را وارد کرد).

۴. بر روی گزینه OK کلیک کنید.

```
New
Obs     Fit  SE Fit        95% CI            95% PI
  1   0.600   0.469  (-0.893; 2.093)  (-1.838; 3.038)
  2   1.300   0.332  ( 0.245; 2.355)  (-0.897; 3.497)
  3   2.000   0.271  ( 1.138; 2.862)  (-0.111; 4.111)
  4   2.700   0.332  ( 1.645; 3.755)  ( 0.503; 4.897)
  5   3.400   0.469  ( 1.907; 4.893)  ( 0.962; 5.838)
```

البته باید توجه داشت که نباید از مدل برای پیش‌بینی درصد جوانه‌زنی به ازای مقادیر دمای کمتر از ۱ و بیشتر از ۵ درجه استفاده کرد. زیرا چنین مقادیری خارج از دامنه داده‌هایی هستند که مدل بر اساس آنها برآورد شده است. ممکن است درصد جوانه‌زنی به ازای دمای کمتر از ۱ و بیشتر از ۵ درجه، پراکندگی متفاوتی پیدا کند.

مثال ۱۱-۵. جدول زیر عملکرد گندم را بر حسب مقادیر مختلف کود نشان داده است. به فرض آنکه ۵۵۰ گرم کود مصرف شود، مطلوب است **الف)** فاصله اطمینان ۹۵٪ میانگین میزان محصول در همه کرت‌ها **ب)** فاصله اطمینان ۹۵٪ میزان محصول فقط در یک کرت گندم.

x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$
۱۰۰	۴۰	-۳۰۰	-۲۰	۹۰۰۰۰	۴۰۰	۶۰۰۰	۴۲/۳	-۲/۳	۵/۲۹
۲۰۰	۵۰	-۲۰۰	-۱۰	۴۰۰۰۰	۱۰۰	۲۰۰۰	۴۸/۲	۱/۸	۳/۲۴
۳۰۰	۵۰	-۱۰۰	-۱۰	۱۰۰۰۰	۱۰۰	۱۰۰۰	۵۴/۱	-۴/۱	۱۶/۸۱
۴۰۰	۷۰	۰	۱۰	۰	۱۰۰	۰	۶۰/۰	۱۰/۰	۱۰۰/۰
۵۰۰	۶۵	۱۰۰	۵	۱۰۰۰۰	۲۵	۵۰۰	۶۵/۹	-۰/۹	۰/۸۱
۶۰۰	۶۵	۲۰۰	۵	۴۰۰۰۰	۲۵	۱۰۰۰	۷۱/۸	-۶/۸	۴۶/۲۴
۷۰۰	۸۰	۳۰۰	۲۰	۹۰۰۰۰	۴۰۰	۶۰۰۰	۷۷/۷	۲/۳	۵/۲۹

$$\sum x = ۲۸۰۰ \qquad \sum y = ٤۲۰ \qquad \sum xy = ۱۸٤۵۰۰ \qquad \sum x^۲ = ۱٤۰۰۰۰۰ \qquad \overline{x} = ٤۰۰ \qquad \overline{y} = ٦۰$$

$$\hat{\beta}_۱ = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^{\,۲} - \frac{(\sum x_i)^۲}{n}} = \frac{۱۸٤۵۰۰ - \frac{۲۸۰۰ \times ٤۲۰}{۷}}{۱٤۰۰۰۰۰ - \frac{۲۸۰۰^۲}{۷}} = \frac{۱٦۵۰۰}{۲۸۰۰۰۰} = ۰/۰۵۹$$

$$\hat{\beta}_۰ = \overline{y} - \hat{\beta}_۱ \overline{x} = ٦۰ - (۰/۰۵۹)٤۰۰ = ۳٦/٤۳, \quad s^۲ = \frac{\sum (y - \hat{y})^۲}{n - ۲} = ۳۵/۵, s = ۵/۹٦$$

الف)

$$\hat{y} \pm t_{\alpha/۲} s \sqrt{\frac{۱}{n} + \frac{(x_i - \overline{x})^۲}{SS_x}} = \hat{y} \pm t_{۰/۰۲۵} ۵/۹٦ \sqrt{\frac{۱}{۷} + \frac{(۵۵۰ - \overline{x})^۲}{SS_x}}$$

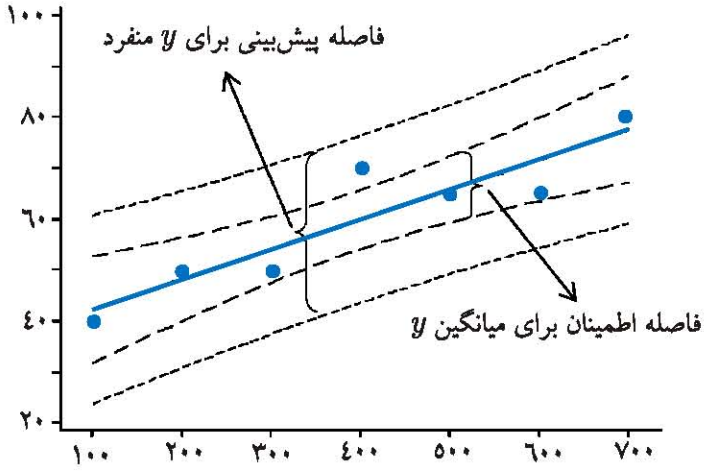
$$٦۸/۸ \pm ۲ / ۵۷۱(۵/۹٦)\sqrt{\frac{۱}{۷} + \frac{(۵۵۰ - ٤۰۰)^۲}{۲۸۰۰۰۰}} = ٦۸/۸ \pm ۲/۵۷۱(۲/۸۱٦) = ٦۸/۸ \pm ۷/۲٤$$

ب)

$$\hat{y} \pm t_{\alpha/۲} s \sqrt{۱ + \frac{۱}{n} + \frac{(x_i - \overline{x})^۲}{SS_x}} = \hat{y} \pm t_{۰/۰۲۵} ۵/۹٦ \sqrt{۱ + \frac{۱}{۷} + \frac{(۵۵۰ - \overline{x})^۲}{SS_x}}$$

$$٦۸/۸ \pm ۲/۵۷۱(۵/۹٦)\sqrt{۱ + \frac{۱}{۷} + \frac{(۵۵۰ - ٤۰۰)^۲}{۲۸۰۰۰۰}} = ٦۸/۸ \pm ۲/۵۷۱(٦/۵۹) = ٦۸/۸ \pm ۱٦/۹۷٤$$

طول این فاصله (۱٦/۹۷٤) بیش از دو برابر طول فاصله بدست آمده در قسمت الف (۷/۲٤) است. ملاحظه می‌شود پیش‌بینی یک مشاهده منفرد، مشکل‌تر از پیش‌بینی یک میانگین است. نمودار ۱۱-۹ فاصله اطمینان را برای میانگین y و فاصله پیش‌بینی برای y منفرد را نشان می‌دهد.



نمودار ۱۱–۹. فاصله پیش‌بینی ۹۵٪ برای یک y منفرد و فاصله اطمینان ۹۵٪ برای میانگین y به ازای مقادیر مختلف x در مثال ۱۱–۵.

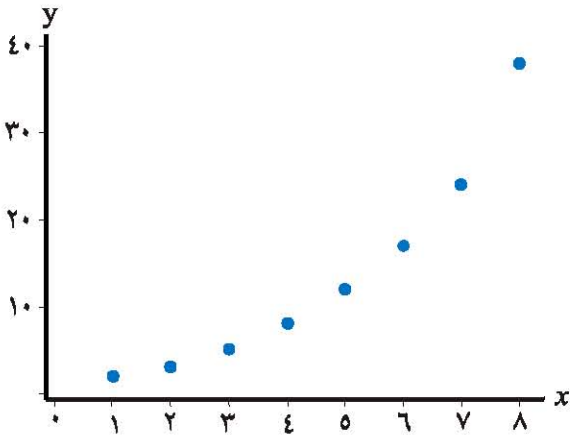
۱۱–۱۲ تبدیل داده‌ها در رگرسیون

گاهی رابطه بین دو متغیر x و y خطی نبوده و نقاط مربوطه در نمودار پراکنش در راستای یک خط مستقیم قرار نمی‌گیرند و لذا مقدار ضریب تعیین (R^2) در رگرسیون خطی کوچک خواهد بود. در چنین مواردی، ممکن است رابطه غیر خطی را بتوان به کمک یک تبدیل به یک رابطه خطی تبدیل کرد. تبدیل داده‌های یک متغیر با توجه به روند تغییرات آن انجام می‌شود. برای مثال اگر مقادیر y در مقایسه به مقادیر x با سرعت بیشتری افزایش یابند، ممکن است x و $\ln y$ یک رابطه خطی با هم داشته باشند یا ممکن است روند افزایش yها طوری باشد که x و $\sqrt{y}$ یک رابطه خطی ایجاد کند.

مثال ۱۱–۶. فرض کنید میزان فروش یک کالای خاص طی ماه‌های مختلف به صورت زیر بوده است.

x (ماه)	۱	۲	۳	٤	٥	٦	۷	۸
y (میزان فروش)	۲	۳	٥	۸	۱۲	۱۷	۲٤	۳۸

نمودار پراکنش داده‌ها نشان می‌دهد که نقاط در راستای یک خط مستقیم قرار ندارند. لذا رابطه خطی بین دو متغیر وجود ندارد و استفاده از مدل رگرسیون خطی برای بررسی رابطه بین این دو متغیر مناسب نخواهد بود. کم بودن مقدار R^2 نیز گواهی بر این ادعا است (جدول ۱۱–٤).



نمودار ۱۱-۱۰. نمودار پراکنش داده‌های مثال ۱۱-٦.

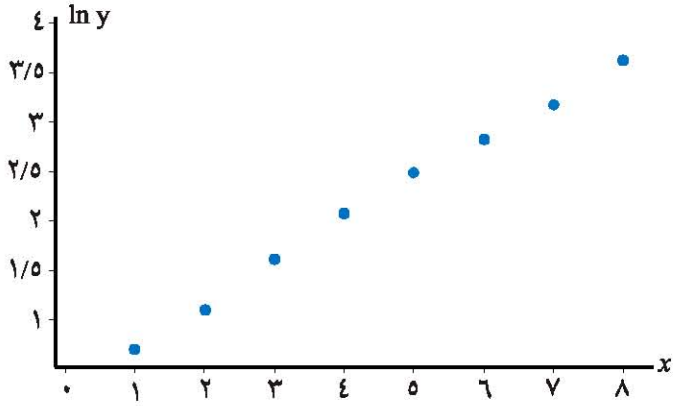
جدول ۱۱-٤. خروجی برنامه Minitab برای مثال ۱۱-٦.

```
The regression equation is
y = - 7.64 + 4.73 x

Predictor      Coef   SE Coef       T       P
Constant     -7.643     3.651   -2.09   0.081
Y            4.7262    0.7230    6.54   0.001

S = 4.68555    R-Sq = 87.7%    R-Sq(adj) = 85.6%

Analysis of Variance

Source           DF       SS       MS       F       P
Regression        1   938.15   938.15   42.73   0.001
Residual Error    6   131.73    21.95
```

در اینجا مقادیر y با سرعت بیشتری نسبت به مقادیر x افزایش یافته است. برای خطی کردن چنین رابطه‌ای ممکن است تبدیل y به $\ln y$ بصورت زیر مفید باشد.

x (سال)	۱	۲	۳	٤	٥	٦	۷	۸
Ln y	۰/٦۹	۱/۰۲	۱/٦۱	۲/۰۸	۲/٤۸	۲/۸۳	۳/۱۸	۳/٦٤

نمودار پراکنش داده‌ها (نمودار ۱۱-۱۱) نشان می‌دهد که مشاهدات تبدیل یافته در راستای یک خط مستقیم قرار دارند. ضریب تعیین رگرسیون خطی برای این مشاهدات تبدیل شده برابر با ۹۹/۷٪ است (جدول ۱۱-٥)، لذا یک رابطه خطی شدید بین مقادیر x و مقادیر $\ln y$ وجود دارد.



نمودار ۱۱-۱۱. نمودار پراکنش داده‌های مثال ۱۱-٦ پس از تبدیل لگاریتمی.

در این مثال رابطه رگرسیونی غیر خطی با استفاده از تبدیل لگاریتمی یکی از متغیرها به یک رابطه رگرسیونی خطی با ضریب تعیین بالا تبدیل شد. در صورتی که تبدیل لگاریتمی چاره‌ساز نباشد باید تبدیل‌های دیگری را برای داده‌های یکی یا هر دو متغیر آزمود. از تبدیل‌های معمول می‌توان به معکوس کردن داده‌ها $(1/y$ و یا $1/x)$، گرفتن ریشه دوم $(y^{0.5}$ یا $x^{0.5})$ یا ریشه چهارم از داده‌ها $(y^{0.25}$ یا $x^{0.25})$ اشاره کرد. در هر صورت تبدیلی مناسبتر است که به ازای آن، مقدار R^2 رگرسیون خطی برآورد شده به یک نزدیکتر باشد. جدول ۱۱-٦ برخی مدل‌های غیر خطی و تبدیل‌های خطی کننده متناظر آنها را نشان می‌دهد.

جدول ۱۱-۵. خروجی برنامه Minitab برای داده‌های بالا.

```
The regression equation is:  Ln y = 0.322 + 0.418 x

Predictor       Coef    SE Coef       T      P
Constant     0.32224    0.05045    6.39  0.001
C1          0.417679   0.009990   41.81  0.000

S = 0.0647432   R-Sq = 99.7%   R-Sq(adj) = 99.6%

Analysis of Variance

Source           DF       SS       MS        F      P
Regression        1   7.3272   7.3272  1748.02  0.000
Residual Error    6   0.0252   0.0042
Total             7   7.3523
```

جدول ۱۱-۶. برخی مدل‌های خطی شونده و تبدیل‌های خطی کننده آنها.

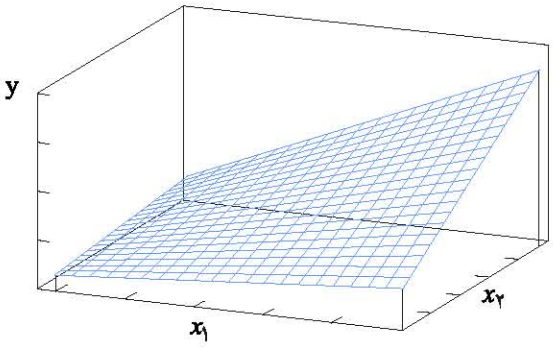
مدل غیر خطی	تبدیل	فرم خطی
$y = \alpha x^{\beta}$	$y' = \log y,\ \ x' = \log x$	$y' = \log \alpha + \beta x'$
$y = \alpha e^{\beta x}$	$y' = \ln y$	$y' = \ln \alpha + \beta x$
$y = \alpha + \beta \log x$	$x' = \log x$	$y = \alpha + \beta x'$
$y = \dfrac{x}{\alpha x - \beta}$	$y' = \dfrac{1}{y},\ \ x' = \dfrac{1}{x}$	$y' = \alpha + \beta x'$

۱۱-۱۳ رگرسیون چندگانه

با توجه به اینکه در عمل غالباً دو یا چند متغیر بر روی متغیر وابسته تأثیر می‌گذارند، بررسی تأثیر همزمان و خطی دو یا چند متغیر مستقل بر متغیر وابسته جالب خواهد بود. برای مثال میزان محصول گندم نه فقط به مقدار کود، بلکه به میزان بارندگی، میزان درجه حرارت، میزان نور و غیره بستگی دارد. در این صورت کمبود نور یا حرارت، عملکرد را محدود می‌کند حتی اگر کود کافی نیز در دسترس گیاه قرار گرفته باشد. اگر متغیر وابسته بصورت خطی تحت تأثیر k متغیر مستقل باشد، معادله رگرسیون خطی چند متغیره را می‌توان بصورت

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

نوشت که در آن $\beta_0, \beta_1, \cdots, \beta_k$ پارامترهای مدل و ε جمله خطا را نشان می‌دهد. این معادله تعمیمی از حالت خطی ساده است. معادله $E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ که معادله میانگین پاسخ یا رویه پاسخ نامیده می‌شود در حالت کلی نشان‌دهنده یک رویه در فضای $k+1$ است. در حالت خاص که $k = 2$ باشد، معادله $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ نشاندهنده صفحه‌ای در فضای سه بعدی است (نمودار ۱۱-۱۲).



نمودار ۱۱-۱۲. صفحه میانگین پاسخ به ازای $k = ۲$ یعنی $E(y) = \beta_۰ + \beta_۱x_۱ + \beta_۲x_۲$.

برآورد مدل رگرسیونی در حالت چند متغیره همانند حالت تک متغیره با استفاده از روش کمترین توان‌های دوم و با مینیمم کردن مجموع توان‌های دوم خطا صورت می‌پذیرد. در حالت دو متغیره می‌توان نوشت

$$SS_E = \sum\big(y_i - \hat{y}\big)^۲ = \sum[y_i - (\hat{\beta}_۰ + \hat{\beta}_۱x_۱ + \hat{\beta}_۲x_۲)]^۲$$

اگر مشتق‌های جزئی نسبت به $\hat{\beta}_۰$، $\hat{\beta}_۱$ و $\hat{\beta}_۲$ محاسبه و برابر صفر قرار داده شوند، برآوردگرهای ضرایب رگرسیونی از حل همزمان معادلات

$$\begin{aligned} n\hat{\beta}_۰ + \hat{\beta}_۱\sum x_۱ + \hat{\beta}_۲\sum x_۲ &= \sum y \\ \hat{\beta}_۰\sum x_۱ + \hat{\beta}_۱\sum x_۱^۲ + \hat{\beta}_۲\sum x_۱x_۲ &= \sum x_۱y \\ \hat{\beta}_۰\sum x_۲ + \hat{\beta}_۱\sum x_۱x_۲ + \hat{\beta}_۲\sum x_۲^۲ &= \sum x_۲y \end{aligned}$$

بدست می‌آیند. در واقع روش محاسبه ضرایب مشابه روش رگرسیون ساده است با این تفاوت که در اینجا محاسبات پیچیده‌تر خواهند بود، به ویژه اگر بیش از دو متغیر در مدل وجود داشته باشد. به همین دلیل در اینجا به بررسی رگرسیون چندگانه با استفاده از برنامه آماری Minitab پرداخته می‌شود.

مثال ۱۱-۷. میزان عملکرد ۱۵ کرت از یک نوع گیاه بر اساس میزان کود (x_1) و تعداد دفعات آبیاری (x_2) در جدول زیر آورده شده است. معادله رگرسیون را برآورد کنید به طوری که بتوان با داشتن میزان کود و تعداد دفعات آبیاری، عملکرد را پیش‌بینی کرد.

دفعات آبیاری (x_2)	کود (x_1)	میزان عملکرد (y)
۱۱	۵۰	۲۸۴۱
۸	۲۱	۱۸۷۶
۱۰	۳۸	۲۹۳۴
۱۰	۱۸	۱۵۵۲
۱۲	۴۳	۳۰۶۵
۱۲	۶۵	۳۶۷۰
۵	۵۰	۲۰۰۵
۸	۴۸	۳۲۱۵
۸	۱۷	۱۹۳۰
۶	۷۰	۲۰۱۰
۹	۲۰	۳۱۱۱
۹	۲۹	۲۸۸۲
۵	۱۵	۱۶۸۳
۷	۱۴	۱۸۱۷
۱۳	۶۰	۴۰۶۶

معادله رگرسیون برازش شده همانگونه که در خروجی Minitab نشان داده شده است $\hat{y} = ۲۷۷ + ۱۵/۳x_1 + ۱۹۵x_2$ می‌باشد. در این مثال $\hat{\beta}_1$ برابر با ۱۵/۳ و $\hat{\beta}_2$ برابر با ۱۹۵ به دست آمده است. این بدان معنی است که با ثابت بودن تعداد دفعات آبیاری، هر واحد کود عملکرد را به اندازه ۱۵/۳ واحد و با ثابت بودن میزان کود، هر بار آبیاری عملکرد را به اندازه ۱۹۵ واحد افزایش می‌دهد. همچنین $\hat{\beta}_0$ برابر با ۲۷۷ به دست آمده است.

در قسمت بعدی خطای معیار ($SE\ \ Coef$) برای هر کدام از این ضرایب محاسبه شده است و به کمک آنها می‌توان فاصله اطمینان $1-\alpha$ را برای هر کدام از ضرایب به دست آورد. برای مثال فاصله اطمینان ۹۵٪ برای β_1 بر اساس رابطه $\hat{\beta}_1 \pm t_{\alpha/2} S_e$ برابر با $۱۵/۳ \pm ۲/۱۷۹(۶/۹۷)$ است. به عبارت دیگر با احتمال ۹۵٪ مقدار β_1 در فاصله ۰/۱۱ تا ۳۰/۴۸ قرار دارد. با توجه به اینکه این فاصله صفر را در بر نمی‌گیرد، لذا β_1 در سطح ۵٪ اختلاف معنی‌داری با صفر دارد.

جدول ۱۱-۷. حل مثال ۱۱-۷ و خروجی مربوطه در برنامه Minitab.

مراحل:

۱. داده‌های متغیر x_1، x_2 و y را در ستون‌های جداگانه وارد کنید.

۲. وارد مسیر مقابل شوید: **Stat > Regression > Regression...**

۳. در مقابل گزینه Response ستون y و در مقابل گزینه Predictors متغیرهای مستقل را انتخاب کنید.

۴. بر روی OK کلیک کنید.

```
The regression equation is
Y = 277 + 15.3 X1 + 195 X2

Predictor     Coef  SE Coef      T      P
Constant     276.8    485.5   0.57  0.579
X1          15.262    6.970   2.19  0.049
X2          195.40    54.07   3.61  0.004

S = 481.650   R-Sq = 67.7%   R-Sq(adj) = 62.3%

Analysis of Variance

Source           DF       SS       MS      F      P
Regression        2  5834582  2917291  12.58  0.001
Residual Error   12  2783846   231987
Total            14  8618428
```

در قسمت بعدی خروجی، انحراف معیار خطا ($S = ۴۸۱/۶۵$) و همچنین ضریب تعیین ($R^۲ = ۶۷/۷٪$) نشان داده شده است. انحراف معیار خطا میزان پراکندگی داده‌ها را حول رگرسیون نشان می‌دهد و در حالت کلی به صورت

$$S = \sqrt{\frac{SS_E}{df_E}} = \sqrt{\frac{\sum \left(y_i - \hat{y}\right)^۲}{n - k - ۱}}$$

محاسبه می‌شود که در آن، n برابر با تعداد مشاهدات و k تعداد متغیرهای مستقل است. درجه آزادی در این مثال ۱۲ می‌باشد. ضریب تعیین ($R^۲$) نیز که از تقسیم SS رگرسیون بر SS کل ($SS_y = SS\ Total$) به دست می‌آید بیانگر سهمی از تغییرات متغیر وابسته است که توسط رگرسیون توجیه می‌شود. مقدار $R^۲$ در اینجا برابر با ۶۷/۷٪ است. با توجه به اینکه $R^۲$ خیلی به ۱ نزدیک نیست پیش‌بینی عملکرد بر اساس تعداد آبیاری و میزان کود با استفاده از این مدل خیلی دقیق نخواهد بود. برنامه‌های آماری همچنین ضریب تعیین تصحیح شده یا $R^۲_{adj}$ را نیز به دست می‌دهند که از رابطه

$$R^۲_{adj} = ۱ - \frac{SS_E \ / \ (n - k - ۱)}{SS_y \ / \ (n - ۱)}$$

بدست می‌آید. بعضی از محققین استفاده از آماره R^2_{adj} را در مورد رگرسیون‌های چندگانه ترجیح می‌دهند، زیرا ضریب تعیین معمولی (R^2) همواره با اضافه شدن یک متغیر جدید به مدل رگرسیون افزایش می‌یابد یا حداقل کاهش نمی‌یابد. ضریب تعیین تصحیح شده با استفاده از میانگین توان‌های دوم به جای مجموع توان‌های دوم این مشکل را حل می‌کند. اگر تفاوت R^2 با R^2_{adj} زیاد باشد، متغیر یا متغیرهایی که مشارکت معنی‌داری در برازش ندارند در مدل رگرسیونی گنجانده شده‌اند. از این خاصیت R^2_{adj} ممکن است حتی برای گزینش متغیرهای مستقل مفید در مدل‌سازی استفاده شود.

تحلیل واریانس در خروجی Minitab (جدول ۱۱-۷) نیز نشان می‌دهد که تغییرات کل به دو قسمت مجموع توان‌های دوم خطا و مجموع توان‌های دوم رگرسیون تفکیک شده است.

۱۱-۱۴ رگرسیون چندجمله‌ای

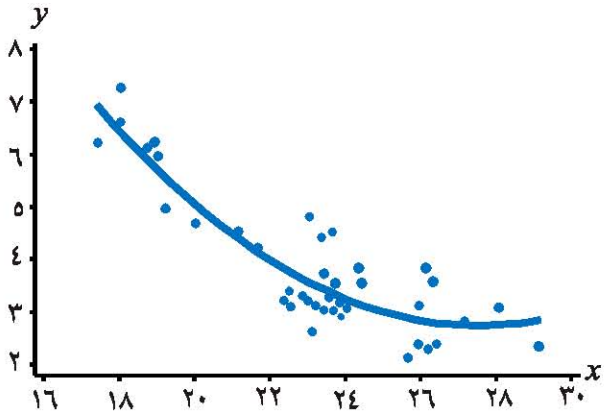
گاهی در رگرسیون ساده مشاهده می‌شود که نمودار پراکنش داده‌ها منحنی است و نمی‌توان تبدیل مناسبی برای خطی کردن رابطه به کار برد. در چنین مواردی متغیر وابسته y تابعی از درجات بالاتر x نیز می‌باشد. به اینگونه مدل‌ها رگرسیون چندجمله‌ای گفته می‌شود. برای مثال مدل چندجمله‌ای $y = \beta_0 + \beta_1 x + \beta_2 x^2$ رگرسیون درجه دو نامیده می‌شود و برای حالاتی مناسب است که با وجود یک متغیر مستقل، نمودار پراکنش داده‌ها ابتدا یک کاهش و سپس افزایش تدریجی یا ابتدا یک افزایش سپس کاهش تدریجی نشان می‌دهند. به مدل چندجمله‌ای $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$ نیز رگرسیون درجه سه گفته می‌شود و برای مواردی مناسب است که با وجود یک متغیر مستقل، نمودار پراکنش داده‌ها دو نقطه ماکسیمم و یک نقطه مینیمم یا دو نقطه مینیمم و یک نقطه ماکسیمم را نشان دهد.

تحلیل رگرسیون چندگانه همانند رگرسیون چندجمله‌ای است با این تفاوت که در اینجا یک متغیر مستقل وجود دارد. اگر داده‌های متغیر مستقل در ستون جدیدی به توان دو برسند و اثر x و x^2 بر متغیر وابسته در قالب رگرسیون چندگانه بررسی شود، مدل رگرسیون دوجمله‌ای ایجاد شده است.

مثال ۱۱–۸ در تحقیقی به منظور بررسی اثر میانگین حداقل دمای روزانه در یک ماه اول بعد از کاشت بر روی عملکرد برنج بر حسب تن در هکتار، داده‌های زیر به دست آمد.

میانگین دما (x)	عملکرد (y)	میانگین دما (x)	عملکرد (y)
۲۳/٤	۳	۲۹/۲	۲/۳
۲۳/٤	٤/٤	۲۸/۱	۳/۱
۲۳/۲	۳/۱	۲۷/۲	۲/۸
۲۳/۱	۲/٦	۲٦/٤	۲/٤
۲۳/۱	٤/۸	۲٦/۳	۳/٦
۲۳/۰	۳/۲	۲٦/۲	۲/۳
۲۲/۹	۳/۳	۲٦/۲	۳/۸
۲۲/٥	۳/۱	۲٦/۰	۳/۱
۲۲/٥	۳/٤	۲٥/۹	۲/٤
۲۲/٤	۳/۲	۲٥/۷	۲/۱
۲۱/۷	٤/۲	۲٤/٥	۳/٥
۲۱/۲	٤/٥	۲٤/٤	۳/۸
۲۰/۰	٤/۷	۲٤/۰	۳/۱
۱۹/۲	٥/۰	۲۳/۹	۲/۹
۱۹/۰	٦/۲	۲۳/۹	۳/۲
۱۹/۰	٦/۰	۲۳/۷	۳/۰
۱۸/۸	٦/۱	۲۳/۷	٤/٥
۱۸/۰	۷/۳	۲۳/۷	۳/٥
۱۸/۰	٦/٦	۲۳/٦	۳/۳
۱۷/٤	٦/۲	۲۳/٥	۳/۷

نمودار پراکنش نشان می‌دهد که داده‌ها در راستای یک منحنی طوری قرار گرفته‌اند که ممکن است رگرسیون درجه دو برازش خوبی بر نقاط داشته باشد.



نمودار ۱۱–۱۳. نمودار پراکنش داده‌های مربوط به عملکرد برنج در مقابل میانگین حداقل دمای روزانه و منحنی برازش داده شده بر آنها.

جدول ۱۱-۸ برازش منحنی رگرسیون درجه دو به داده‌های مثال ۱۱-۸.

مراحل:

۱. مشاهدات متغیر y و x را بترتیب در ستون‌های اول و دوم صفحه داده‌ها وارد کنید.

۲. مربع داده‌های ستون مربوط به متغیر x را تحت عنوان x_2 در ستون سوم وارد کنید. برای این منظور می‌توان از Calculator... موجود در منوی Calc استفاده کرد.

۳. وارد مسیر مقابل شوید: **Stat > Regression > Regression...**

۴. در مقابل گزینه Response ستون y و در مقابل گزینه Predictors متغیرهای ستون‌های x و x_2 را انتخاب کنید.

۵. بر روی OK کلیک کنید.

```
The regression equation is
y = 33.7 - 2.24 x + 0.0408 x2


Predictor       Coef   SE Coef      T      P
Constant      33.695     4.876   6.91  0.000
x            -2.2469    0.4299  -5.23  0.000
x2          0.040764  0.009391   4.34  0.000


S = 0.587621   R-Sq = 80.8%   R-Sq(adj) = 79.8%


Analysis of Variance

Source           DF      SS      MS      F      P
Regression        2  53.932  26.966  78.09  0.000
Residual Error   37  12.776   0.345
Total            39  66.708
```

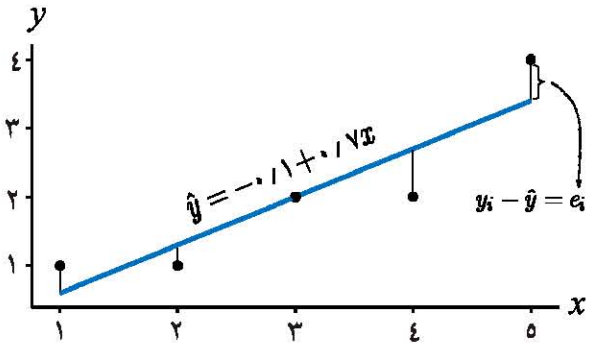
همانگونه که در خروجی Minitab نشان داده شده است معادله منحنی رگرسیون برازش شده بصورت $\hat{y} = ۳۳/۷ - ۲/۲٤x + ۰/۰٤x^۲$ می‌باشد. اگر به جای رگرسیون درجه دو، رگرسیون خطی ساده بر داده‌های متغیر وابسته و مستقل برازش داده می‌شد، ضریب تعیین برابر با ۷۱/۱٪ به دست می‌آمد، که خیلی کمتر از مقدار فعلی (۸۰/۸٪) است.

۱۱-۱۵ بازبینی مدل آماری

به طور کلی در تحلیل رگرسیونی، مدلی بر اساس کمترین مجموع توان‌های دوم خطا برازش داده شده و بعضاً فرض‌های مختلفی مورد آزمون قرار می‌گیرد. غالباً لازم است که مدل رگرسیونی برآورد شده به طور دقیقتری بازبینی شود. این بازبینی با اهداف مختلفی انجام

می‌شود. مثلاً اینکه آیا می‌توان مدل مناسبتری که دارای خطای کمتر و دقت بیشتری باشد برآورد نمود؟ یکی از معمولترین کارها برای کشف عیوب مدل و ارائه راه حلی مناسب برای اصلاح آن، مشاهده نمودار پراکنش مانده‌ها می‌باشد. مانده‌ها در واقع انحرافات نقاط از رگرسیون برازش داده‌شده هستند و نمی‌توانند پس از برازش مدل بر داده‌ها تبیین شوند. برای مثال اگر عملکرد (متغیر وابسته) در مقادیر کودی ۱، ۲، ۳، ٤ و ٥ به ترتیب ۱، ۱، ۲، ۲ و ٤ باشد، مانده‌ها یا خطاها را می‌توان پس از برازش خط رگرسیون به صورت زیر به دست آورد.

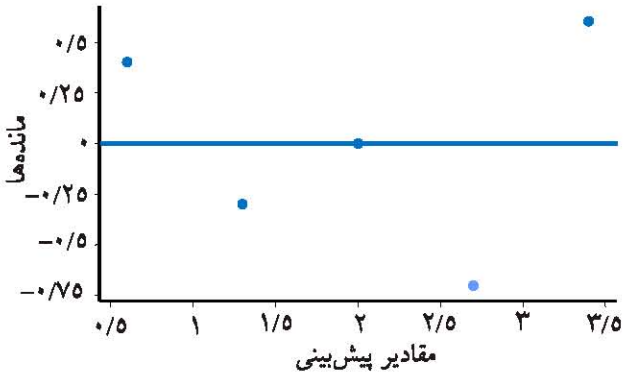
$$e_i = y_i - \hat{y}$$



نمودار ۱۱-۱٤. مانده‌ها (e_i ها) پس از برازش خط رگرسیون بر مشاهدات.

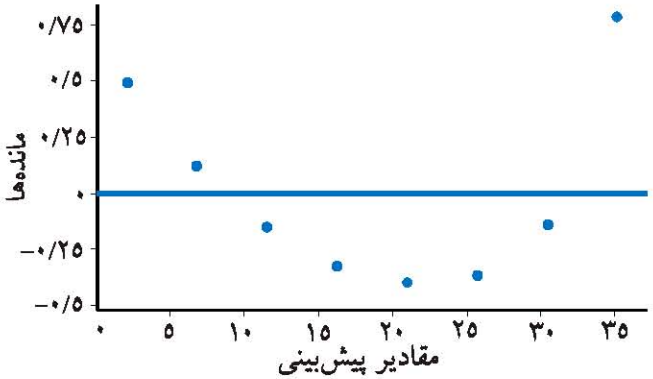
مهمترین فرض در هر مدلی این است که مانده‌ها مستقل از هم و دارای توزیع نرمال باشند. سایر فرض‌های رگرسیون در بند ۱۱-٦ آورده شده‌اند. در واقع همه این فرض‌ها دلالت بر این دارند که مانده‌ها نباید از هیچ الگوی مشخصی پیروی کرده و باید به صورت تصادفی مشاهده شوند. مشاهده هر گونه الگوی منظمی در مورد مانده‌ها برقراری فرض‌های رگرسیون را زیر سؤال می‌برد. روش معمول در بررسی مانده‌ها مشاهده نمودار پراکنش مانده‌ها در مقابل مقادیر پیش‌بینی شده می‌باشد.

در برنامه Minitab پس از رفتن به مسیر **Stat > Regression > Regression...** برای تجزیه رگرسیون، با کلیک روی گزینه **Graphs...** و انتخاب **Residuals versus fits** می‌توان نمودار پراکنش مانده‌ها را در مقابل مقادیر برآورد شده مشاهده نمود. نمودار پراکنش مانده‌ها برای درصد جوانه‌زنی به ازای دماهای مختلف پس از برازاندن خط رگرسیون ساده، در نمودار ۱۱-۱٥ نشان داده شده و هیچگونه روندی در آن مشاهده نمی‌شود.



نمودار ۱۱-۱۵. نمودار پراکنش ماندهها پس از برازش خط رگرسیون ساده بر دادهها.

حال اگر به نمودار پراکنش ماندههای مربوط به مثال ۱۱-٦ پس از برازاندن خط رگرسیون ساده نگاه کنید میبینید که به جای اینکه ماندهها حول محور $\hat{y}$ به صورت تصادفی پراکنده شده باشند، ابتدا یک کاهش سپس افزایش تدریجی دارند. در چنین مواردی مدل خطی برازش دادهشده مناسب نبوده و باید حالتهای دیگری از جمله مدلهای غیر خطی بررسی شوند. همچنین میتوان تبدیل مناسبی را در مورد دادههای یکی یا هر دو متغیر اعمال کرده و سپس رگرسیون خطی ساده را در مورد دادههای تبدیل شده به کار برد. به یاد داریم که این مشکل در مثال مذکور از طریق تبدیل لگاریتمی متغیر وابسته حل شد.



نمودار ۱۱-۱٦. نمودار پراکنش ماندهها پس از برازش خط رگرسیونی ساده بر دادههای مثال ۱۱-٦.

تمرین‌ها

۱۱–۱. داده‌های زیر را در نظر بگیرید. **الف)** دادها را بر روی محور مختصات نشان دهید. **ب)** مقادیر r و r^2 را محاسبه کنید. **ج)** آیا شواهد کافی مبنی بر وجود رابطه خطی بین x و y وجود دارد؟ در سطح معنی‌داری $\alpha = ۰/۰۱$ بررسی کنید.

x	۳	۵	۶	۴	۳	۷	۶	۵	۴	۷
y	۴	۳	۲	۱	۲	۳	۳	۵	۴	۲

۱۱–۲. در تحقیقی به منظور بررسی وجود ارتباط بین قطر درختان در فاصله ۱/۵ متری از زمین و ارتفاع درختان، ۱۲ درخت به طور تصادفی انتخاب و داده‌های زیر به دست آمد.

قطر درخت (cm)	۹/۸	۷/۲	۲۴/۰	۷/۸	۲/۵	۴۰/۲	۱۵/۵	۶۴/۴	۳۱/۵	۱۰/۸	۳/۰	۸/۳
ارتفاع درخت (m)	۱۰/۹	۹/۸	۲۰/۵	۱۱/۰	۵/۵	۲۰/۴	۱۷/۵	۲۶/۸	۲۵/۶	۱۲/۳	۷/۹	۱۳/۶

الف) مقدار r^2 را محاسبه و تفسیر کنید. **ب)** معادله خط رگرسیونی ارتفاع بر روی قطر درختان را با روش حداقل مجموع توان‌های دوم به دست آورده و پارامترهای برآورد شده را تفسیر نمائید. **ج)** مقادیر برازش داده شده و خطاها را برای هر یک از نقاط به دست آورده و آنها را بر روی محور مختصات نشان دهید. آیا نقاط در راستای یک خط راست قرار می‌گیرند؟ **د)** واریانس کل، خطا و رگرسیون را به دست آورده معنی‌دار بودن ارتباط خطی را از طریق آزمون F بررسی کنید. **ه)** انحراف معیار $\hat{\beta}_0$ و $\hat{\beta}_1$ را به دست آورده و با استفاده از آنها فاصله اطمینان ۹۵٪ را برای عرض از مبدأ و شیب خط جامعه پیدا کرده و نتایج را توضیح دهید. **و)** قطر یک درخت جدید اندازه‌گیری و مقدار آن ۱۰ سانتی‌متر بوده است. مقدار ارتفاع آن را پیش‌بینی کرده و فاصله پیش‌بینی ۹۵٪ را برای این ارتفاع به دست آورید. **ز)** فرض‌های رگرسیون را با استفاده از خط رگرسیون و نقاط مشاهده شده بر روی محور مختصات مورد بحث قرار دهید.

۱۱–۳. مقدار pH در ۱۷ نوع خاک با درصد رس (x_1)، درصد مواد آلی (x_2) و درصد کربنات کلسیم (x_3) متفاوت به صورت زیر بوده است. معادله رگرسیون را با استفاده از یک برنامه آماری برآورد کنید به طوری که بتوان با داشتن مقدار رس، مقدار مواد آلی و مقدار کربنات کلسیم، pH خاک را پیش‌بینی کرد. موارد مشاهده شده در خروجی برنامه را تفسیر نمائید.

درصد کربنات کلسیم (x_3)	درصد مواد آلی (x_2)	درصد رس (x_1)	pH خاک (y)
٦/١	٤/٣	٥١/١	٧/١
٠/٠	٢/٦	٢٢/٠	٥/٤
٢/٠	٣/٠	١٧/٠	٧/٠
٠/٠	٣/٠	١٦/٨	٦/١
٠/٠	٤/٠	٥/٥	٣/٧
٠/١	٣/٣	٢١/٢	٧/٠
١٦/٨	٣/٧	١٤/١	٧/٤
١٧/٣	٠/٧	١٦/٦	٧/٤
١٥/٦	٣/٧	٣٥/٩	٧/٣
١١/٩	٣/٣	٢٩/٩	٧/٥
٢/٨	٣/١	٢/٤	٧/٤
٦/٢	٢/٨	١/٦	٧/٤
٠/٣	١/٨	١٧/٠	٧/٣
٩/١	٢/٣	٣٢/٦	٧/٣
٠/٠	٤/٠	١٠/٥	٤/٠
٢٦/٠	٥/١	٣٣/٠	٧/١
٠/٠	١/٩	٢٦/٠	٥/٦