

۶-۵ همبستگی

آنچه که تا به حال در رگرسیون خطی مطرح شد بر این فرض استوار بود که x یک مقدار ثابت است و به ازای مقدار x متغیر پاسخ یک متغیر تصادفی مانند y است. اگر x و y هر دو متغیرهای تصادفی باشند و قصد بررسی ارتباط خطی بین آن دو را داشته باشیم باید از مفهوم همبستگی^۱ استفاده کنیم.

در فصل ۱ مفهوم $cov(X, Y)$ تعریف شد که معیاری برای وجود رابطه بین دو متغیر تصادفی است. چنانچه بررسی رابطه خطی مدنظر باشد از ضریب همبستگی ρ استفاده می کنند. ρ به صورت زیر تعریف میشود.

$$\rho = \frac{cov(X, Y)}{\sigma_x \sigma_y}$$

ثابت میشود که همواره: $-1 \leq \rho \leq 1$. ρ معیاری برای شدت رابطه خطی بین دو متغیر تصادفی است.

اگر ρ مثبت باشد نشان دهنده، وجود رابطه خطی مستقیم بین دو متغیر تصادفی (مانند میزان فروش و هزینه تبلیغات) است. اگر ρ منفی باشد دلالت بر وجود رابطه خطی معکوس بین دو متغیر تصادفی است مانند رابطه بین مقدار فروش و قیمت کالا. اگر دو متغیر تصادفی مستقل باشند، آنگاه $\rho = 0$ ولی عکس آن برقرار نیست. یعنی اگر $\rho = 0$ باشد نمی توان استقلال دو متغیر تصادفی را نتیجه گرفت و تنها می توان نتیجه گرفت که رابطه خطی برقرار نیست و ممکن است روابط تابعی دیگری غیر از خط راست برقرار باشد.

ضریب ρ را می توان براساس داده های نمونه ای به صورت زیر برآورد کرد:

$$r = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \hat{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \hat{y})^2}}$$

r ضریب همبستگی خطی نامیده می شود. نزدیک بودن r به $+1$ و -1 دلیل بر قوی بودن رابطه خطی بین دو متغیر تصادفی است.

مثال ۶-۳ فرض کنید که بخواهیم بر مبنای داده های زیر تعیین کنیم که آیا وابستگی خطی بین زمان لازم بر حسب دقیقه، برای تکمیل یک فرم معین توسط یک منشی در صبح یا عصر وجود دارد یا نه؟

۱۰,۵	۸,۴	۶,۶	۹,۰	۷,۱	۱۰,۹	۹,۴	۷,۰	۹,۶	۸,۲	X (صبح)
۱۲,۳	۸,۴	۶,۳	۹,۲	۷,۶	۱۱,۳	۸,۵	۶,۹	۹,۶	۸,۷	Y (عصر)

ضریب همبستگی نمونه ای را محاسبه و آن را تعبیر کنید.

^۱ Correlation

حل. از داده ها به دست می آوریم $n = 10$ و $\sum x^2 = 771,35$ ، $\sum y = 88,8$ ، $\sum y^2 = 819,34$ ، و $\sum xy = 792,92$ ، به طوری که

$$\begin{aligned} S_{xx} &= 771,35 - \frac{1}{10} (88,8)^2 = 19,661 \\ S_{yy} &= 819,34 - \frac{1}{10} (88,8)^2 = 30,796 \\ S_{xy} &= 792,92 - \frac{1}{10} (88,8)(88,8) = 23,024 \end{aligned}$$

و

$$r = \frac{23,024}{\sqrt{(19,661)(30,796)}} = 0,936$$

نزدیک بودن مقدار r به ۱ دلالت بر آن می کند که وابستگی مثبتی بین زمان لازم برای انجام وظیفه معینی توسط یک منشی در صبح و در عصر وجود دارد، و این از پراکنش نگار شکل ۵.۱۴ نیز آشکار است.

۵-۶-۱ ضریب تعیین R^2

یک معیار مهم برای پی بردن بر شدت رابطه خطی ضریب تعیین یا R^2 است که در واقع مربع ضریب همبستگی نمونه است.

R^2 را می توان طبق رابطه r صفحه قبل به صورت زیر نوشت:

$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{S_{xy}}{S_{xx}} \frac{S_{xy}}{S_{yy}} = \beta_1 \frac{S_{xy}}{S_{yy}} = \frac{SSR}{S_{yy}}$$

S_{yy} بیانگر کل تغییرات مشاهدات y است و SSR تعبیر جالبی دارد و به عنوان تغییراتی از y که توسط مدل رگرسیون شناسایی شده است تعبیر می شود. بنابراین $\frac{SSR}{S_{yy}}$ را نسبتی از تغییرات کل y ها که در اثر بستگی x به y قابل توضیح است. هر چه مقدار R^2 به ۱ نزدیکتر باشد شدت رابطه خطی بیشتر است.

مثال ۴-۵. در مثال ۳-۵ مقدار $R^2 = (0,936)^2 = 0,876$ به دست می آید یعنی تقریباً ۸۸ درصد از تغییر در y ها در اثر بستگی داشتن به x قابل توضیح است و توانسته ایم ۸۸ درصد از تغییرات y را با مدل رگرسیون خطی ساده تعیین کنیم. با افزودن متغیرهای مستقل دیگر ممکن است بتوانیم ۱۲ درصد تغییرات باقی مانده را هم توضیح دهیم. این مبحثی است که رگرسیون چند متغیره نامیده می شود.