استنباط آماری

Statistical Inference

رگرسیون خطی ساده و همبستگی :

Simple Linear Regression and Correlation

مقدمه :

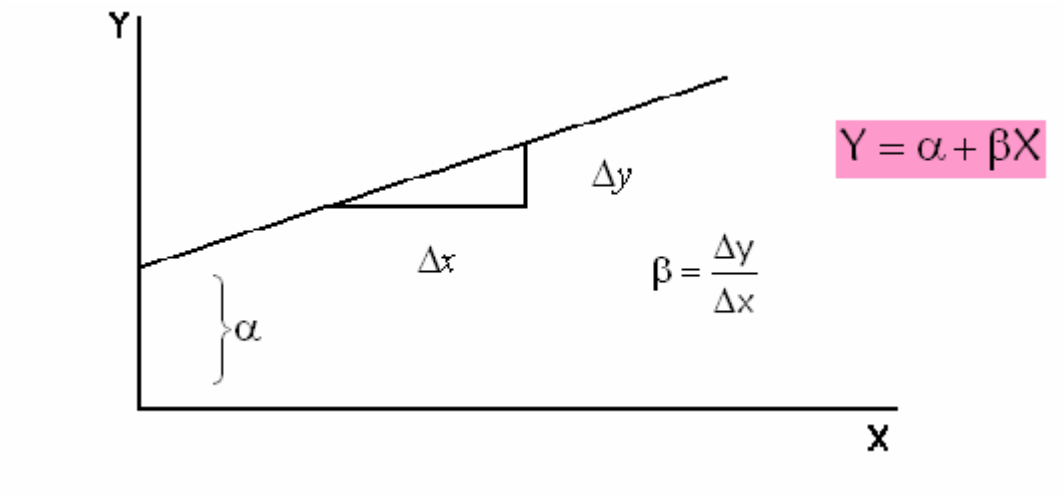
رگرسیون شاخه‌ای از علم آمار است که استفاده ار آن به نحو وسیعی در اکثر زمینه‌های علمی معمول شده است. با مطالعه یک جامعه آماری چنین به نظر می‌رسد که بین صفات متغیر آن جامعه کم و بیش ارتباط وجود دارد و گاهی نیز مشاهده می‌شود که تغییرات یک متغیر بطور مستقیم یا معکوس در تغییرات متغیر دیگر مؤثر است. بعنوان مثال، در اقتصاد رگرسیون برای اندازه‌گیری و یا تخمین روابط بین متغیرهای اقتصادی مورد استفاده قرار می‌گیرد و یا مثلاً بین قد و وزن افراد در یک جامعه رابطه مستقیم وجود دارد و یا بین دو صفت تحصیلات و تعداد اولاد رابطهٔ معکوس وجود دارد. بعنوان مثال تئوری اقتصاد عنوان می‌کند که میزان تقاضا بستگی به قیمت، درآمد و چند عامل دیگر دارد. عرضه نیروی کار در رابطه با میزان مزد پرداختی است و مصرف شخصی تابعی از درآمد قابل تصرف. از آنجا که این روابط فقط فرضیه‌هایی را در خصوص رفتارهای اقتصادی بیان می‌کنند، اقتصاددانان، داده‌های آماری،یعنی مشاهدات دنیای واقعی را به کار می‌گیرند تا صحت و سقم تئوریهای اقتصادی را آزمون کنند.

تابع خطی :

برآورد رابطه بین دو متغیر، امکان پذیر نخواهد بود مگر آنکه ابتدا فرض کنیم رابطه بین دو متغیر دارای فرم خاصی است. یکی از معمول‌ترین این فرمها،تابع خطی ساده است. یک چنین توابعی در اقتصاد از اهمیت بسیاری برخوردارند، زیرا کار کردن با آنها نسبتاً ساده است و اغلب می‌توانند بعنوان تقریبی از توابع غیرخطی بکار روند. فرم ریاضی یک تابع خطی ساده بصورت زیر است :

$$Y = \alpha + \beta x$$

که در آن مقادیر α و β ثابت هستند. ضریب α که عرض از مبدأ نامیده می‌شود، مقدار Y به ازاء X مساوی صفر را نشان می‌دهد. ضریب β که نمایانگر شیب خط است، میزان تغییرات Y را به ازای یک واحد تغییر در X مشخص می‌کند. در شکل زیر یک تابع خطی ساده که بصورت خطی مستقیم است، ترسیم شده است :



در این تابع Y متغیر وابسته و X متغیر مستقل نامیده می‌شود. بعنوان مثال تصور کنید که $\alpha = 4$ و $\beta = 3$ است. در این صورت خواهیم داشت :

$$Y = 4 + 3X$$

اگر X به اندازه یک واحد تغییر کند، Y به اندازه ۳ واحد تغییر خواهد کرد. هنگامی که β مثبت است خط صعودی و چنانچه منفی باشد، خط نزولی است. درصورتی که β صفر باشد، خط موازی $Y = \alpha$ با محور Xها است.

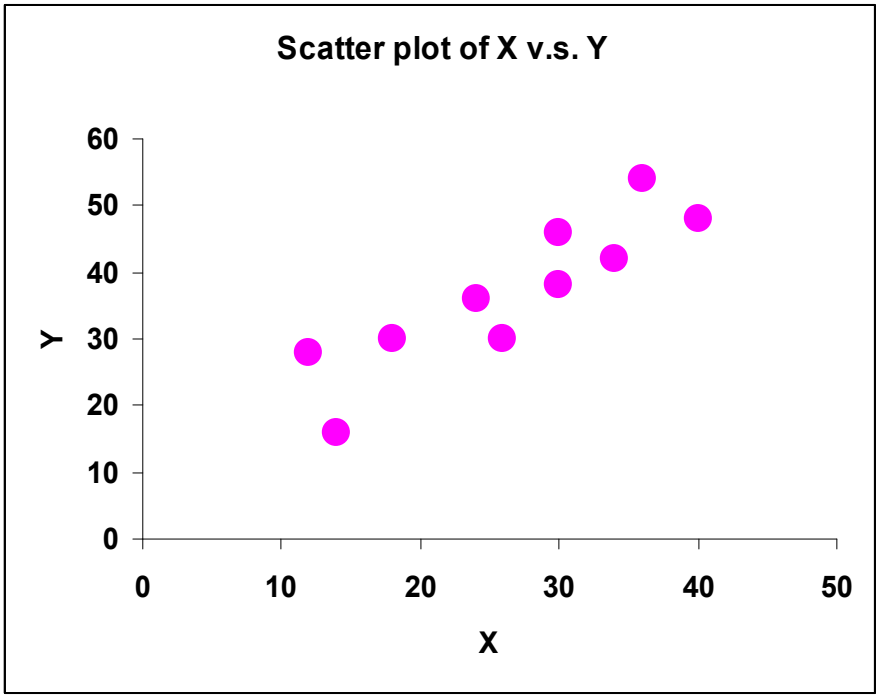
توجه داشته باشید که کلیه مقادیر X و Y که در تابع فوق صدق می‌کند، همگی بر روی یک خط راست واقع است.

برآورد تابع خطی ساده :

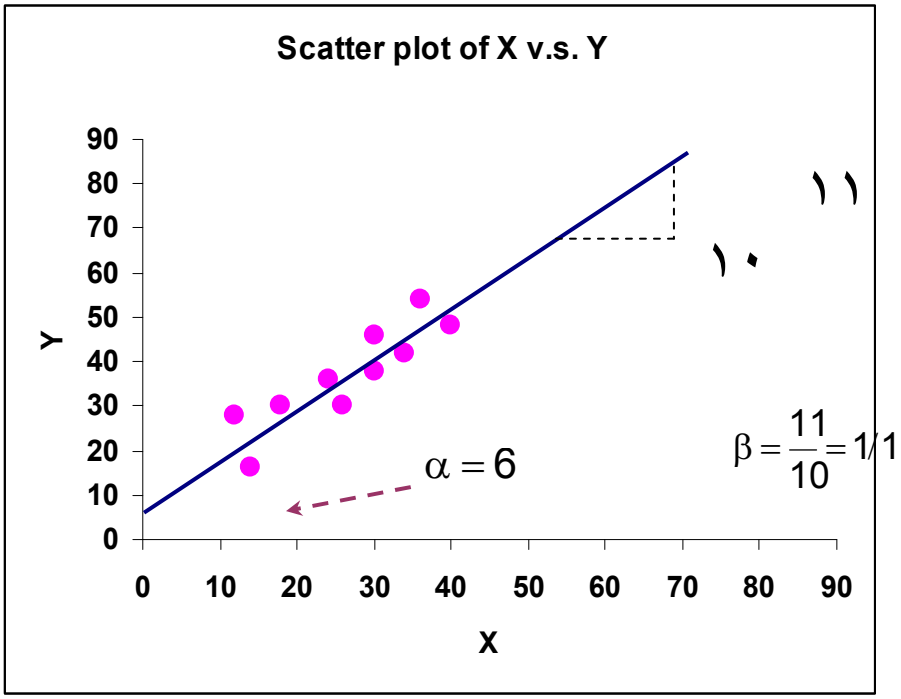
فرض کنید می‌خواهیم رابطه بین دو متغیر، مثلاً رابطه جمعیت (X) و فروش یک کالا (Y) را بررسی کنیم. برای این منظور تعداد ۱۰ مشاهده در اختیار داریم که در جدول زیر آمده است :

	Y	X
منطقه	**تعداد کالای فروش رفته**	**جمعیت به هزار نفر**
١	**٥٤**	**٣٦**
٢	**٣٠**	**٢٦**
٣	**٢٨**	**١٢**
٤	**٤٨**	**٤٠**
٥	**٣٦**	**٢٤**
٦	**٣٠**	**١٨**
٧	**٣٨**	**٣٠**
٨	**٤٦**	**٣٠**
٩	**١٦**	**١٤**
١٠	**٤٢**	**٣٤**

برای نشان دادن رابطه بین X و Y می‌توان ابتدا نقاط را بصورتی که در شکل زیر آمده است بر روی یک نمودار پراکنش مشخص کرد. آنچه مسلم است رابطه‌ای دقیق و ساده بین دو متغیر X و Y دیده نمی‌شود. بعبارت دیگر نقاط همگی بر یک خط راست قرار نمی‌گیرند. ولی این تمایل به وضوح دیده می‌شود که با افزایش X متغیر Y نیز افزایش پیدا می‌کند. بدین معنی که فروش در نقاط پرجمعیت بیشتر است. ما می‌خواهیم این تمایل را بصورت یک تابع خطی ساده نمایش دهیم.



یک راه انجام آن است که بصورت نظری و با کمک چشم خطی را از میان این نقاط بگونه‌ای عبور دهیم که احساس شود بین نقاط در دو طرف خط توازن برقرار شده است. این خط در شکل زیر ترسیم شده است.



ملاحظه می‌شود که در این نمودار $\alpha = 6$ و $\beta = 1/1$ است. تابع خطی برآورد شده می‌تواند اکنون بصورت $Y = 6 + 1/1X$ نوشته شود. معمول است که در رابطه برآورد شده بجای Y نماد $\hat{Y}$ نوشته شود. در این‌صورت Y میزان مشاهده شده را نشان می‌دهد، درصورتی که $\hat{Y}$ نمایانگر تخمین Y است که از تابع خطی برآورد شده حاصل شده است.

$$\hat{Y} = 6 + 1/1X$$

اکنون می‌توانیم با توجه به رابطه فوق میزان فروش را در منطقه‌ای مثلاًبا ۲۰ هزار نفر جمعیت تخمین زد :

$$\hat{Y} = 6 + 1/1(20) = 28$$

نکته قابل توجه این است که خط فوق را به کمک چشم بین نقاط رسم کردیم. بعید است که شخص دیگری به کمک چشم خطی را رسم کند که تا اندازه‌ای با خط فوق متفاوت باشد. البته تفاوت در رابطه با نقاط فوق نمی‌تواند خیلی زیاد باشد. ولی چنانچه پراکندگی نقاط بیشتر باشد، احتمال دارد که خط رسم شده از نظر عرض از مبدأ و شیب کاملاً متفاوت باشد. بنابراین لازم به نظر می‌رسد که برای انتخاب مناسبترین خط معیار دقیق‌تری را مورد توجه قرار داد.

برازش خط مستقیم به روش حداقل مربعات (کمترین مجذورات) :

Least-Square Method

آمارگران بهترین برازش را آنچنان خطی تعریف می‌کنند که مجموع مربهات خطا (Sum of Squares of Errors)، کمترین مقدار ممکن را داشته باشد. خطا عبارت است از فاصلهٔ عمودی بین مقدار واقعی مشاهده شده و مقداری که برای آن از خط برازش داده شده بدست می‌آید. مقدار خطا را معمولاً با حرف e نمایش می‌دهند.

برای هر مجموعه‌ای از مشاهدات آماری، خطوط مختلف دارای مجموع مربعات خطای متفاوتی، یعنی $\sum e_i^2$های متفاوتی خواهند بود. بهتریان خط برازش داده شده آنچنان خطی است که در آن $\sum e_i^2$ دارای کمترین مقدار باشد. این خط به نام خط حداقل مربعات نامیده می‌شود. αو β مربوط به خط حداقل مربعات بصورت زیر بدست می‌آیند.

$$\hat{\beta} = \frac{\sum XY - n\overline{X}\overline{Y}}{\sum X^2 - n\overline{X}^2} \quad \text{و} \quad \hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$$

اکنون به عنوان یک مثال، مقدار αو β را برای جمعیت و میزان فروش بدست می‌آوریم. داریم :

$$\hat{\beta} = \frac{\sum XY - n\overline{X}\overline{Y}}{\sum X^2 - n\overline{X}^2} = \frac{105560 - 10(26/4)(36/8)}{77680 - 10(26/4)^2} = 1/053106$$

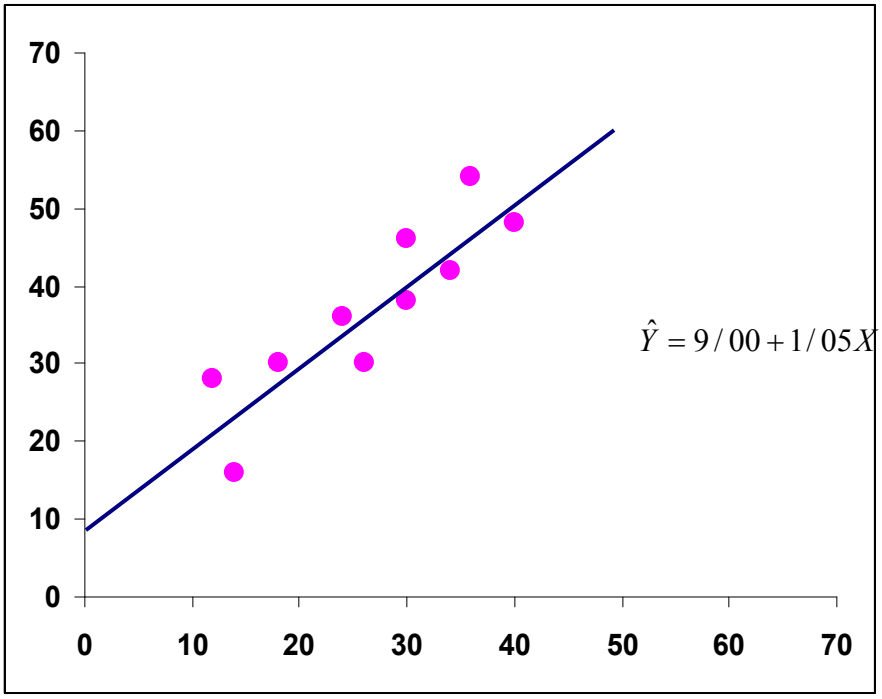
و

$$\hat{\alpha} = 36/8 - 1/053(26/4) = 8/988002$$

با گرد کردن مقادیر αو β داریم :

$$\hat{Y} = 9/00 + 1/05X$$

و این خط در شکل زیر رسم شده است :


70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
$\hat{Y} = 9/00 + 1/05X$