الانحدار و الارتباط الخطي
Correlation and Linear Regression

مقدمة:

تقابلنا كثيرا في الحياة العملية مواقف تتضمن متغيرين (ظاهرتين) أو اكثر ويكون المطلوب معرفة ما اذا كان هناك علاقة بين هذه المتغيرات وما هو شكل هذه العلاقة؟ وايضا كيفية التنبؤ بإحد هذين المتغيرين في حالة معرفتنا بالمتغير الاخر. فكثيرا ما تجدين في بعض المجلات معادلة الطول مع الوزن فاذا اردت ان تعرفي الوزن المثالي ادخلى طولك في المعادلة ليظهر وزنك المثالي، وقد توصلوا إلى هذه المعادلة أو إلى هذه الصيغة بدراسة العلاقة ما بين المتغيرين الطول والوزن على مجموعة من الافراد.

وفي هذا الباب سنتعرف على العلاقة بين متغيرين X و Y بدلالة:

1- الانحدار والذي يفيدنا في:

- تحديد شكل العلاقة بين المتغيرين رياضيا وبيانيا (خط الانحدار).
- توضيح اتجاه العلاقة بين المتغيرين.
- التنبؤ بقيمة أحد المتغيرين بدلالة المتغير الآخر.

2- الارتباط ويفيدنا في:

- تحديد مقدار درجة العلاقة بين المتغيرين.
- تحديد اتجاه العلاقة بين المتغيرين.

وسنهتم في دراستنا بالعلاقة الخطية (فقط) بين المتغيرين X و Y، وفيما يلي توضيح ذلك:

أولاً: الانحدار الخطي البسيط Simple Linear Regression:

إذا كان لدينا متغيران فقط. المتغير X و هو متغير يتم تحديد قيمه من قبل الباحث أو الشخص الذي يقوم بالدراسة و لتكن هذه القيم هي $x_1, x_2, \dots, x_n$ و هو متغير رياضي يسمى **بالمتغير المستقل independent variable** . يرافق المتغير X متغير آخر Y و يفرض أن قيمه هي $y_1, y_2, \dots, y_n$ و يسمى **بالمتغير التابع dependent variable** و هو متغير عشوائي لأن نتيجته غير محددة و تعتمد على قيم المتغير المستقل. و إذا نظرنا إلى البيانات في صورة أزواج مرتبة (X,Y) يكون لدينا مجموعة من أزواج المشاهدات هي $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

في اغلب الحالات يكون الشكل الدقيق للعلاقة بين المتغيرين غير معروف وتكون المشكلة هي تحديد شكل هذه العلاقة وتقدير الثوابت التي تتضمنها ثم استخدام ما نحصل عليه من نتائج في التنبؤ بقيمة Y إذا تحددت قيمة X. ولتحديد نوع العلاقة بين المتغيرين (X,Y) نستخدم أزواج المشاهدات بتمثيلها على المستوى فنحصل على ما يسمى **بشكل الانتشار Scatter diagram** الذي يبين طبيعة العلاقة بين المتغيرين (X,Y).

و الانحدار هو أسلوب يمكن بواسطته تقدير قيمة أحد المتغيرين بمعلومية قيمة المتغير الآخر عن طريق معادلة الانحدار، و له أنواع.

أنواع الانحدار:

- **الانحدار الخطي البسيط:** فكلمة "بسيط" تعنى أن المتغير Y يعتمد على متغير مستقل واحد هو X وكلمة "خطي" تعنى أن العلاقة بين المتغيرين (X, Y) علاقة خطية.
- **الانحدار المتعدد:** إذا كان المتغير Y يعتمد على أكثر من متغير مستقل.
- **الانحدار الغير خطي:** إذا كانت العلاقة بين المتغير Y والمتغيرات المستقلة غير خطية كأن تكون من الدرجة الثانية أو أسية.

بعد تمثيل الأزواج المرتبة بالمستوى نحصل على شكل الانتشار فإذا اظهر الشكل الانتشاري للبيانات أن هناك علاقة خطية بين المتغيرين من النوع:

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad , i = 1,2,\dots,n$$

و تمثل معادلة خط مستقيم من الدرجة الأولى وستمثل معادلة خط انحدار Y/X . و تتكون قيمة المتغير y_i من المقدارين:

- $\beta_0 + \beta_1 x_i$ هو المقدار الناتج عن العلاقة بين المتغيرين (X,Y)، وتسمى β_0, β_1 **بمعالم العلاقة الخطية** حيث:

  β_0 هي الجزء الذي يقطعه الخط من محور Y .

  β_1 هي ميل الخط المستقيم أو **معامل الانحدار regression coefficient**.
- e_i هو مقدار الخطأ العشوائي في تحديد قيمة y_i ولا يمكننا التحكم فيه. ويفترض دائماً أن المقادير العشوائية $e_i \, , i = 1, 2, \dots, n$ هي متغيرات عشوائية .

و توجد عدة طرق جبرية للحصول على أفضل خط مستقيم يمر بالنقاط ويمثل العلاقة بين المتغيرين بأفضل ما يمكن و أفضل هذه الطرق هي

طريقة المربعات الصغرى: Method of Least Squares

عادة ما تكون المعلمتان β_0, β_1 مجهولتان وبالتالي نستخدم مشاهدات العينة $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ في تقديرهما، و بالتالي نحصل على تقدير لخط انحدار Y على X. وباستخدام الطرق الرياضية نتوصل لتقدير قيمتي β_0, β_1 من المعادلتين التاليتين:

$$b_1 = \frac{\left(\sum xy / n\right) - \bar{x}\,\bar{y}}{S_x^2} \quad ,$$

$$b_0 = \bar{y} - b_1\,\bar{x}$$

Where

$$\bar{x} = \frac{\sum x}{n} \qquad , \quad \bar{y} = \frac{\sum y}{n}$$

$$S_x^2 = \frac{\sum x^2}{n} - \bar{x}^2$$

وبالتالي تقدير خط انحدار Y على X هو

$$\hat{y} = b_0 + b_1 x$$

مثال 1:

قام أحد مدراء التسويق بدراسة لتحديد ما اذا كانت هناك علاقة بين عمر الشخص (X) وعدد المجلات التي يشترك فيها هذا الشخص (Y) فكانت لدينا البيانات التالية:

عمر الشخص (X)	21	26	33	35	48	50	55	64
عدد الاشتراكات (Y)	4	0	3	1	3	0	2	6

والمطلوب:

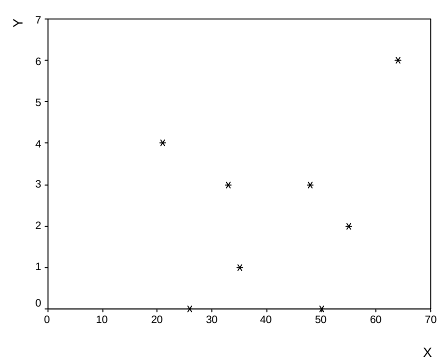
أ) ارسمي الشكل الانتشاري لهذه البيانات.

ب) قدري معادلة انحدار عدد الاشتراكات (Y) على عمر الشخص (X)، ثم مثلي هذه العلاقة بيانيا.

ج) قدري عدد الاشتراكات التي من الممكن ان يشترك بها شخص ما يبلغ من العمر 45 سنة.

الحل:

أ)-



ب)-

	X	Y	X^2	Y^2	XY
	21	4	441	16	84
	26	0	676	0	0
	33	3	1089	9	99
	35	1	1225	1	35
	48	3	2304	9	144
	50	0	2500	0	0
	55	2	3025	4	110
	64	6	4096	36	384
Total	332	19	15356	75	856

$$\bar{x} = \frac{\sum x}{n} = \frac{332}{8} = 41.5$$

$$\bar{y} = \frac{\sum y}{n} = \frac{19}{8} = 2.375$$

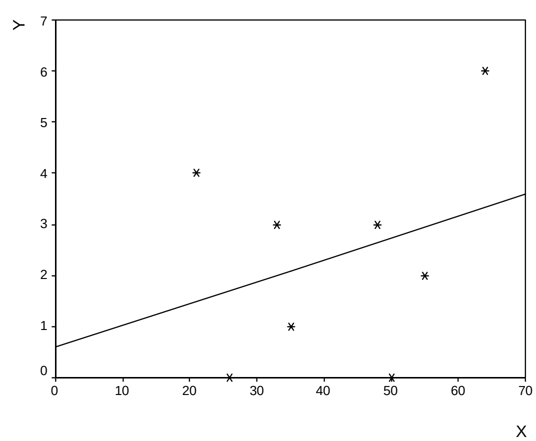
$$S_x^2 = \frac{\sum x^2}{n} - \bar{x}^2 = \frac{15356}{8} - (41.5)^2 = 197.25 \Rightarrow S_x = \sqrt{197.25} = 14.04$$

$$S_y^2 = \frac{\sum y^2}{n} - \bar{y}^2 = \frac{75}{8} - (2.375)^2 = 3.73 \Rightarrow S_y = \sqrt{3.73} = 1.93$$

$$b_1 = \frac{\left(\sum xy/n\right) - \bar{x}\ \bar{y}}{S_x^2} = \frac{(856/8) - (41.5)(2.375)}{197.25} = .043$$

$$b_0 = \bar{y} - b_1\ \bar{x} = 2.375 - (.043)(41.5) = .591$$

$$\therefore \hat{y} = .591 + .043\ x$$



ج)- $\hat{y} = .591 + (.043)(45) = 2.526 \approx 3$

أي أن الشخص الذي عمره 45 قد يشترك بثلاث أعداد تقريبا من المجلة.

ثانياً:الارتباط:Correlation :

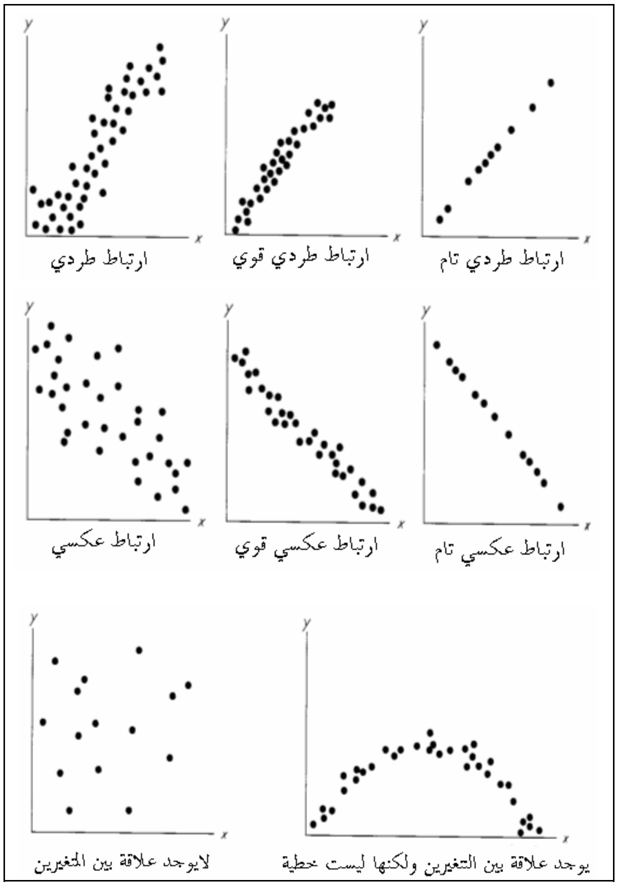
بعد تمثيل أزواج المشاهدات على المستوى يتكون لدينا شكل الانتشار Scatter diagram الذي يبين طبيعة العلاقة بين المتغيرين (X,Y). و الجدول التالي يوضح أنواع الارتباط واتجاه العلاقة وشكل الانتشار لكل نوع:

درجة وقوة الارتباط:

الارتباط التام: بمعنى اذا عرفت قيمة احد المتغيرين عرفنا قيمة المتغير الاخر بالضبط. مثال: محيط المربع = 4 × طول الضلع أو y=4x، هي علاقة رياضية تامة اذا عرفنا طول الضلع لاي مربع نستطيع معرفة محيطه بالضبط.

الارتباط المنعدم: بمعنى اذا عرفنا قيمة احد المتغيرين استحال علينا معرفة قيمة المتغير الاخر. مثال على ذلك العلاقة بين الطول والتحصيل الدراسي.

الارتباط الغير تام: يتبع فيه احد المتغيرين المتغير الاخر في تغيره إلى حد ما وهو النوع الاكثر شيوعا في اغلب العلاقات التي تصادفنا.



اتجاه الارتباط أو العلاقة:

اذا كان المتغيرين مرتبطين فانه يمكن تزايد احدهما يؤدي إلى زيادة الاخر ويسمى هذا بالارتباط الطردي، اما اذا ادى زيادة احدهما إلى تناقص الاخر والعكس وهذا ما يعرف بالارتباط العكسي.

للإرتباط معاملان وهما:

1. معامل الارتباط الخطي Coefficient of Correlation:

معامل الارتباط الخطي او ما يسمى بمعامل بيرسون Pearson وهو يقيس خطية العلاقة بين المتغيرين (X,Y) ويرمز له بالرمز r . و يعطى بالصورة التالية:

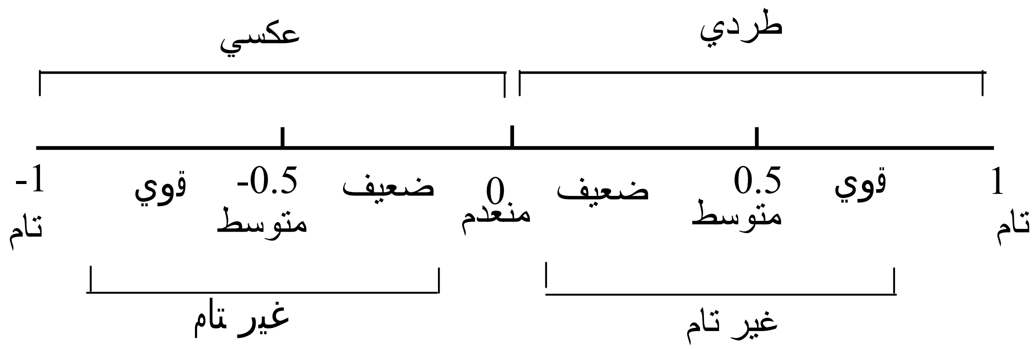
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

وسنستخدم في حساباتنا الصيغة التالية للتسهيل وهي تبسيط للصورة السابقة:

$$r = \frac{(\sum x_i y_i / n) - \bar{x}\ \bar{y}}{S_x\ S_y} \quad ; \quad -1 \le r \le 1$$

حيث
$$\bar{x}=\frac{\sum x}{n}\quad,\qquad \bar{y}=\frac{\sum y}{n}$$
$$S_x=\sqrt{\frac{\sum x^2}{n}-\bar{x}^2}\quad;\qquad S_y=\sqrt{\frac{\sum y^2}{n}-\bar{y}^2}$$

وهو دائما ياخذ القيم من -1 إلى 1، اي ان $-1\le r\le 1$. فاذا كانت العلاقة بين المتغيرين طردية قوية تقترب قيمة r من 1، بينما اذا كانت العلاقة بين المتغيرين عكسية قوية تقترب قيمة r من -1، وفي حالة عدم وجود علاقة أو ارتباط خطي بين المتغيرين أو كانت هذه العلاقة ضعيفة فان قيمة r تقترب من الصفر أو تساوي الصفر.



<u>مثال مدراء التسويق:</u>

د) احسبي معامل الارتباط وحددي نوعه.

<u>الحل:</u>

د)-
$$r=\frac{\left(\sum xy\setminus n\right)-\bar{x}\;\bar{y}}{S_x\;S_y}=\frac{(856/8)-(41.5)(2.375)}{(14.04)(1.93)}=\frac{8.4375}{27.0972}=0.311$$

نلاحظ أن العلاقة بين عمر الشخص وعدد المجلات التي يشترك بها علاقة طردية ضعيفة

ملحوظة:

- تتفق إشارة معامل الارتباط r مع إشارة معامل الانحدار b_1 . ويمكن أيجاد معامل الارتباط r بدلالة معامل الانحدار b_1 بالقانون:

$$r=\frac{S_x}{S_y}.b_1$$

مثال (2):

البيانات التالية عن أسعار السلع X و الكميات المطلوبة منها Y:

$n = 8$, $\sum xy = 27416$, $\sum x = 528$, $\sum y = 408$,
$\sum x^2 = 35584$, $\sum y^2 = 21176$

المطلوب: 1)- ادرسي العلاقة بين السعر والكمية وحددي نوعها واتجاهها.

الحل :

$$\bar{x} = \frac{\sum x}{n} = \frac{528}{8} = 66$$

$$\bar{y} = \frac{\sum y}{n} = \frac{408}{8} = 51$$

$$S_x = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2} = \sqrt{\frac{35584}{8} - (66)^2} = \sqrt{92} = 9.59$$

$$S_y = \sqrt{\frac{\sum y^2}{n} - \bar{y}^2} = \sqrt{\frac{21176}{8} - (51)^2} = \sqrt{46} = 6.78$$

$$r = \frac{\sum \frac{xy}{n} - \bar{x}\bar{y}}{S_x S_y}$$

$$r = \frac{\frac{27416}{8} - (66 \times 51)}{(9.59 \times 6.78)}$$

$$r = 0.938$$

يوجد علاقة طردية قوية بين السعر و الكمية المطلوبة .

2. معامل ارتباط الرتب " معامل سبيرمان" Spearman's rank correlation:

في حالة البيانات الوصفية (القابلة للترتيب) يتعذر علينا حساب معامل الارتباط بيرسون Pearson في هذه الحالة نلجأ إلى معامل ارتباط آخر للعالم Spearman و تتلخص طريقة حسابه في اعطاء رتب لمشاهدات كل من المتغيرين على النحو التالي:

- نبحث في قيم المتغير X عن أكبر قيمة فتعطى لها الرتبة (1) ثم القيمة الأصغر منها وتعطى لها الرتبة (2) و هكذا لباقي المشاهدات و نضع هذه الرتب في خانة جديدة تسمى (رتب X).
- نتبع نفس الطريقة مع قيم المتغير Y وتوضع الرتب في خانة جديدة تسمى (رتب Y).
- نضيف خانة تسمى (فرق الرتب d): d = رتب X - رتب Y.
- نضيف خانة اخرى d^2 و هي حاصل ضرب d x d ثم نوجد مجموعها $\sum d^2$.
- ثم بعد ذلك يحسب معامل الارتباط بين رتب المتغير X ورتب المتغير الاخر Y باستخدام العلاقة التالية :

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

n: تمثل حجم العينة(عدد المشاهدات).

ملحوظة:

يمكننا استخدام معامل ارتباط سبيرمان Spearman في حالة البيانات الكمية ايضا. اي اننا نستطيع ان نستخدم معامل ارتباط بيرسون وسبيرمان في حالة البيانات الكمية، اما في حالة البيانات الوصفية (الوصفية القابلة للترتيب فقط) نستطيع استخدام معامل سبيرمان Spearman فقط.

من مثال (1): مدراء التسويق:

هـ) – اوجدي معامل ارتباط الرتب سبيرمان.

X	Y	رتب X	رتب Y	d	d^2
21	4	8	2	6	36
26	0	7	7.5	- 0.5	0.25
33	3	6	3.5	2.5	6.25
35	1	5	6	-1	1
48	3	4	3.5	0.5	0.25
50	0	3	7.5	4.5	20.25
55	2	2	5	-3	9
64	6	1	1	0	0
					$\sum d^2 = 73$

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{(6)(73)}{(8)(64-1)}$$

$$= 1 - \frac{438}{504} = 1 - 0.87 = 0.13$$

نلاحظ أن : العلاقة طردية ضعيفة و هي نفس النتيجة التي تحصلنا عليها عند ايجاد الارتباط بمعامل بيرسون

مثال 2:

لدينا البيانات التالية لخمسة طالبات في مادتي الاحصاء والرياضيات، فهل يوجد علاقة بين تقديرات الطالبة في كلتا المادتين.

B	C	D	C	A	تقديرات الاحصاء
A	D	D	D	B	تقديرات الرياضيات

الحل:

d^2	d	رتب الرياضيات	رتب الاحصاء	تقديرات الرياضيات(Y)	تقديرات الاحصاء(X)
1	-1	2	1	B	A
0.25	-0.5	4	3.5	D	C
1	1	4	5	D	D
0.25	-0.5	4	3.5	D	C
1	1	1	2	A	B
3.5	0				

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{(6)(3.5)}{(5)(25-1)} = 1 - 0.175 = 0.825$$

اذن العلاقة طردية قوية بين تقديرات الطالبة للمادتين.