الفصل السادس عشر

الارتباط والانحدار الخطي البسيط

Correlation and Simple Linear Regression

16.1. مقدمة Introduction

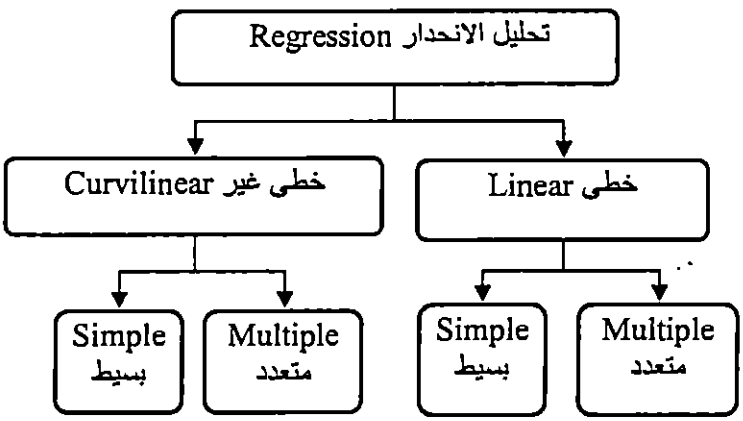
من خلال دراستنا في الأجزاء السابقة تعرفنا كيف يمكن التعامل مع البيانات الكثيرة، الغير معبرة والتي قد لا تعطي صورة واضحة عن حالة العملية إذا عرضت بطريقة جافة وبصورتها الخام، وتعرفنا كيف يمكننا معالجتها وتحويلها إلي مقاييس رقمية Measures وأشكال بيانية Graphs يسهل من خلالها الاستدلال عما يحدث في سلوك العملية من حيث اقترابها من الهدف المنشود أو ابتعادها عنه. وكل ذلك كان ينطبق علي بيانات لمتغير واحد One Variable، فمثلاً لمعرفة شكل توزيع مرتبات العاملين في شركه ما ، فإن المتغير في هذه الحالة هو المرتب) ، كنا نقوم بجمع عينات عن نفس المتغير المراد قياسه، ثم نقوم بحساب مقاييس مثل المتوسط Mean والانحراف المعيارى Standard Deviation، والمدى Range، ونرسم الأشكال Graphs مثل الدائرة Pie Chart، والساق والورقة Stem And Leaf والهستوجرام Histogram ومخطط الصندوق Box Plot، ثم من خلال المقاييس الرقمية والأشكال يمكننا دارسة هذا المتغير وتحليله ومعرفة الكثير عن خصائصه.

ولكن ماذا يكون الوضع لو كان لدينا متغيرين اثنين Two Variables أو أكثر ولا يمكن التعامل مع كل منها علي انفراد؟ وفى هذه الحالة سيكون مطلوبا التعرف على نوع وشكل العلاقة أو الارتباط بين هذه المتغيرات، هل هي علاقة عكسية؟ أم طردية ؟ وما مدى قوة هذه العلاقة؟ هل هي قوية أم ضعيفة؟ وهل هى حقيقية أم وهمية؟ ويحدث ذلك مثلاً عند دراسة تأثير درجة الحرارة والضغط على جودة المنتج النهائى، أو تغيير نمط الصيانة والتشغيل على أداء معدة، أو تأثير درجة حرارة الجو ومساحة العمل على سلوك الموظفين وإنتاجيتهم، أو معرفة توقيت وفاة الشخص من درجة حرارة جسمه بعد الوفاة، أو تأثر الأداء بتغير موقع العمل وتوقيته، فى هذه الحالة و باستخدام الانحدار Regression سنعرف كيف يمكننا التعامل مع هذا الموقف والتغلب عليه، وذلك بإيجاد مقاييس رقمية Measures إحصائية تبين قوة العلاقة، ورسم مخططات بيانية تبين شكل واتجاه تلك العلاقة،(مثلما حدث في حالة المتغير الواحد).

16.2. ما هو تحليل الانحدار؟ What Is Regression?

تحليل الانحدار Regression طريقة إحصائية لتحليل العلاقة بين متغيرين أو أكثر، وإيجاد المعادلة التي تحكم العلاقة بين هذه المتغيرات، بعض هذه المتغيرات يسمى تابعا Dependant وهو فى الغالب يمثل النتيجة Result أو الأثر Effect، ويرسم على المحور الصادي الرأسي Y، والمتغيرات الأخرى تسمى مستقلة Independent أو Predictor أو سبب Cause، وترسم على المحور السيني الأفقى X.

16.3. أنواع تحليل الانحدار Types of Regression



شكل رقم 16- 1 الانواع المختلفة للانحدار

وينقسم الانحدار Regression طبقا و الشكل 16-1 إلى عدة أنواع:

1. الانحدار الخطى البسيط Simple Linear Regression:عندما يكون لدينا متغير مستقل واحد على المحور الأفقى مع وجود المتغير التابع على المحور الرأسي، وتكون المعادلة فى هذه الحالة $Y = B_0 + B_1 X_1 + e$ ، مثل دراسة العلاقة بين العمر (X) وضغط الدم (Y)، أو دراسة العلاقة بين الدعاية (X) وحجم المبيعات (Y).
2. الانحدار الخطى المتعدد Multiple Linear Regression:عندما يكون لدينا أكثر من متغير مستقل على المحور الأفقى Causes، وتكون المعادلة فى هذه الحالة $Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + \ldots\ldots\ldots + B_n X_n + e$ ، مثل دراسة العلاقة بين العمر X_1 ووزن الجسم X_2 والموقف من التدخين X_3 وتأثير ذلك على نسبة الكولسترول فى الدم Y.

3. الانحدارالبسيط الغير خطى Simple Curvilinear Regressionعندما يكون لدينا متغير مستقل واحد على المحور الأفقى، وظهور علاقة تربيعية من الدرجة الثانية Quadratic مثل دراسة العلاقة بين المسافة اللازمة لإيقاف السيارة Y وسرعتها X_1، وتكون المعادلة فى هذه الحالة:

$$Y = B_0 + B_1 X_1 + B_2 X_2^2 + \ldots\ldots + e$$

4. الانحدار المتعدد الغير خطى Multiple Curvilinear Regression عندما يكون لدينا أكثر من متغير مستقل على المحور الأفقى Causes وظهور علاقات متداخلة مثل دراسة العلاقة بين مقدار التأمين على الحياة Y وكل من درجة الخطورة X_1 والعمر X_2 ، وتكون المعادلة فى هذه الحالة:

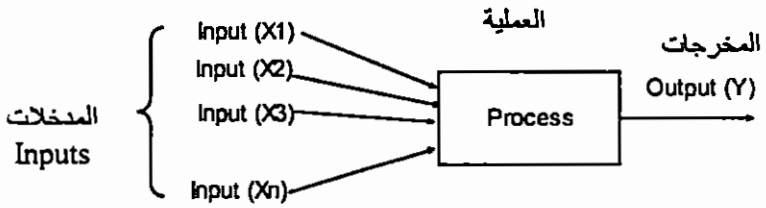
$$Y = B_0 + B_1 X_1 + B_2 X_2^2 + B_3 X_1 X_2 \ldots\ldots + e$$

حيث إن Y هي المتغير التابع Dependant أو النتيجة Result أو الأثرEffect ، و X_n هي المتغيرات المستقلة Independent أو الأسباب Causes، و B_0 هي نقطة تقاطع خط الانحدار مع المحورالرأسى، و B_n هي معاملات الانحدار للمتغيرات المختلفة، و e هي قيمة الخطأ التعويضي أو الفرق بين القيم الحقيقية والقيم التى يمكن استنتاجها من المعادلة، ويلاحظ أننا فى حالات كثيرة نحاول تحويل النوعين الثالث والرابع إلى النوعين الأول والثاني أى تحويل الغير خطىCurvilinear إلى خطى Linear باستخدام التحويل اللوغاريتمى مثلا.

16.4. استخدامات تحليل الانحدار Uses of Regression

يستخدم تحليل الانحدار Regression فى عدة أغراض منها:

1. التنبؤ Prediction بشكل العلاقة بين عدة متغيرات، ومحاولة إيجاد المعادلة التي تحكم هذه العلاقة، أى معرفة هل هناك علاقة بين هذه المتغيرات أم لا، وما نوع هذه العلاقة؟ هل هي طردية أم عكسية؟ وهل هي خطية أم غير خطية؟ والتنبؤ من العمليات الهامة فى مجال التخطيط واتخاذ القرارات Planning and Decision Making، ويجب أن يكون هذا التنبؤ والتوقع بمحاذير أهمها أن يكون التوقع داخل المدى المتوافر فيه البيانات فقط وهذا ما يطلق عليه Entrapolation، فإذا تجاوزنا المدى المتوافر فيه البيانات فإن ذلك ما يطلق عليه Extrapolations وهو غير مستحب وغير محبذ لأنه غالبا ما ينطوي على نسبة من الخطأ.



شكل رقم 16- 2 تأثر مخرجات اى عملية بمدخلاتها

2. السيطرة والتحكم Control فمعرفة وتحديد المعادلة التي تحكم العلاقة بين عدة متغيرات Input X's، ومدى تأثير كل متغير على النتيجة Output Y، كما فى شكل 16-2 ، يمنحنا القدرة على التحكم فى المتغير التابع أو الناتج النهائي.

16.5. الفرق بين الانحدار والارتباط Difference Between Regression and Correlation

يعد الانحدار Regression و الارتباط Correlation من الطرق الإحصائية لتحليل العلاقة بين متغيرين أو أكثر، وإيجاد المعادلة التي تحكم العلاقة بين هذه المتغيرات، والفرق بينهما أنه فى حالة الانحدار Regression يمكننا تحديد قيم المتغيرات المستقلة قبل إجراء التجربة، فمثلاً لدراسة العلاقة بين جرعات دواء للأنفلونزا والقضاء على المرض، فإننا نقوم بتحديد الجرعات مسبقاً قبل التجربة، ثم نبحث تأثيرها على المرض. أما فى حالة الارتباط Correlation فلا يكون لنا الحرية فى تحديد قيم أى من المتغيرات المستقلة إذ تكون جميعها خاضعة للمؤثرات العشوائية الخارجة عن السيطرة، ويمكن لأى منها اتخاذ أى قيمة من القيم الممكنة لها، فمثلاً لدراسة العلاقة بين مستوى الكوليسترول فى الدم وبين وزن الجسم، فإننا نأخذ عينة عشوائية من المرضى، ثم نقيس الوزن ومستوى الكوليسترول، أى أن المتغيرين عشوائيان.

وعموما فإنه فى حدود دراستنا فإن هذا الفرق لن يعنى الكثير لنا إلا فى طريقة تصميم عملية جمع البيانات فقط.

16.6. البيانات المزدوجة والارتباط Biverate Data & Correlation

البيانات المزدوجة Biverate Data مجموعة من البيانات عند جمعها يتم جمع متغيرين عشوائيين في آن واحد، مثل أطوال الرجال ومقاسات أحذيتهم، أو أوزان النساء ومعدل التنفس في الدقيقة، أي يوجد متغيرين أحدهما X والآخر Y وقد يوجد علاقة بين هذين المتغيرين أو لا توجد.

وهذا النوع من البيانات Biverate يمكن تمثيله بيانياً عن طريق منحنى الانتشار Scatter Diagram (لمزيد من المعلومات عن منحنى الانتشار ، يرجى الرجوع إلى الفصل التاسع)

وفيه يتم رسم البيانات المستقلة Dependent Data أو السبب Cause علي المحور الأفقى X، ورسم البيانات التابعة Independent Data أو النتيجةEffect علي المحور الرأسي Y، والمثال 15-1 يوضح تلك الفكرة.

مثال رقم 16- 1:

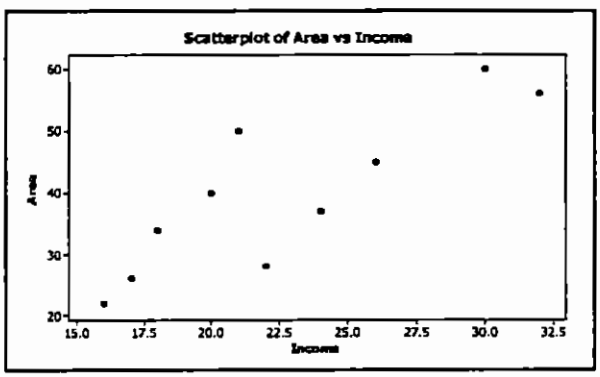
تود إحدى الشركات العقارية تحليل السوق العقارى فى إحدى المناطق، ولذا فقد قامت بعمل مسح Survey لدراسة العلاقة بين دخل الأسرة (X) ومساحة منزلها بالمتر المربع (Y) ، وتم جمع عينات عشوائية كانت نتيجتها كما بالجدول 16-1 كالتالي :

جدول رقم 16- 1 دخل الأسرة (X) ومساحة منازلهم بالمتر المربع (Y)

Income (X)	الدخل بالألف جنيه	22	26	45	37	28	50	56	34	60	40
Square meters (Y)	مساحة المنزل بالمتر المربع	16	17	26	24	22	21	32	18	30	20

و لتسهيل فهم الموضوع في البداية، سنفترض أن العلاقة بين المتغيرين المراد دراستهما هي علاقة خطية بسيطة Simple Linear، وسنقوم برسم هذه العلاقة، ثم نحاول إثبات مدى صحة هذه العلاقة وهل هي علاقة حقيقية أم لا ؟ ثم نحاول تجسيد قيمة الارتباط بين هذين المتغيرين وتحويلها إلى مقياس رقمى يحدد قوة هذه العلاقة.

بمساعدة Minitab يمكننا رسم مخطط الانتشار Scatter Diagram (يرجى مراجعة موضوع مخطط الانتشار فى الفصل التاسع من هذا الكتاب) فيتضح أنه كلما زاد الدخل الشهري للأسرة زادت مساحة منزلها، ولذا فإنه يعطى إنطباعا باحتمال وجود علاقة طردية موجبة Positive Relation أو علاقة مباشرة Direct Relation كما يظهر فى الشكل 16-3.



شكل رقم 16- 3 رسم بيانات المثال 16-1

ولتأكيد هذا الاحتمال يجب تحويله إلي كمية Quantity أو قيمة Value، ولذا نحاول رسم خط مستقيم Linear يتوسط هذه القراءات ستكون معادلته:

$$Y = B_0 + B_1 X_1 + e$$

ومن المتوقع وجود خطأ في استنتاج الارتباط والعلاقة بين المتغيرين، لعدة أسباب:

1. أننا أفترضنا أن العلاقة خطية ولها معادلة خط مستقيم وهذا فرض غير مؤكد.
2. عندما أخذنا العينات في البداية أفترضنا أنها تمثيل وتعبير عن المجتمع وهذا افتراض نعرف أيضا أنه ليس صحيحا تماما.

وبالتالي فلابد أن نسلم بأنه سيكون هناك خطأ في استنتاج شكل هذه العلاقة بدقة، حتى وأن كان هذا الخطأ صغيراً.

ولتوضيح ذلك دعنا ننظر إلي القيمة الأولي من العينات التي تم تجميعها وهي 22 & 16، سنجد أنها تخبرنا بوجود رجل دخله الشهري هو (22000 جنيه) ومساحة شقته هي (160 متر مربع)، ولكن و بعد أن رسمنا مخطط الانتشار Scatter، ورسمنا الخط المستقيم الذي يتوسط النقاط المرسومة، ولو حاولنا من الرسم معرفة مساحة شقة الشخص الذي دخله 22000 جنيه، فسوف نقوم برسم خط رأسي عند قيمة 22000 حتي يقطع الخط المستقيم المعبر عن العلاقة، ثم نمد خط أفقي حتي يقطع المحور الرأسي عند قيمة تكون هي قيمة مساحة شقة ذلك الشخص، وحينئذ سنجد أنها ليست 160 مترا مربعا كما هي الحقيقة ، أي ستكون إما أكبر من 160 أو أقل من 160، وهذا ما نعني به الخطأ في استنتاج العلاقة بدقة، كذلك لو حاولنا معرفة مساحة شقة شخص دخله 51000 ، وهي قيمة ليست من قيم العينات المأخوذة ، فإننا بالتالي نتوقع أن القيمة التي سنقوم باستنتاجها من المعادلة، أو من الرسم ليست صحيحة تماما، أي سيوجد خطأ في الاستنتاج.

و فيما يلي سنقوم باستخدام النموذج الإحصائى Statistical Model لفهم العلاقة بين المتغيرين، وما يصحب ذلك من افتراضات Assumptions، وإختبار هذه الفرضيات Test of Hypothesis لتحديد ما إذا كان هذا النموذج Model جيد أم لا، وما مستوى الثقة Confidence Levelفيه، وما مدى اعتمادنا عليه.

وحيث إن استنتاج شكل العلاقة في هذه المرحلة قد يترتب عليه اتخاذ قرارات مهمة في مرحلة لاحقة، فإنه ينبغي التأكد من أن هذه العلاقة هي علاقة حقيقية وليست ظاهرية، وسيكون هذا من خلال عدة معايير ومقاييس رقمية لتحديد قوة هذا الارتباط ومنها معامل

الارتباط Coefficient of Correlation، و مجموع المربعات Sum of Squares ، وخط أقل المربعات Least Square Line كما سيتضح فيما بعد.

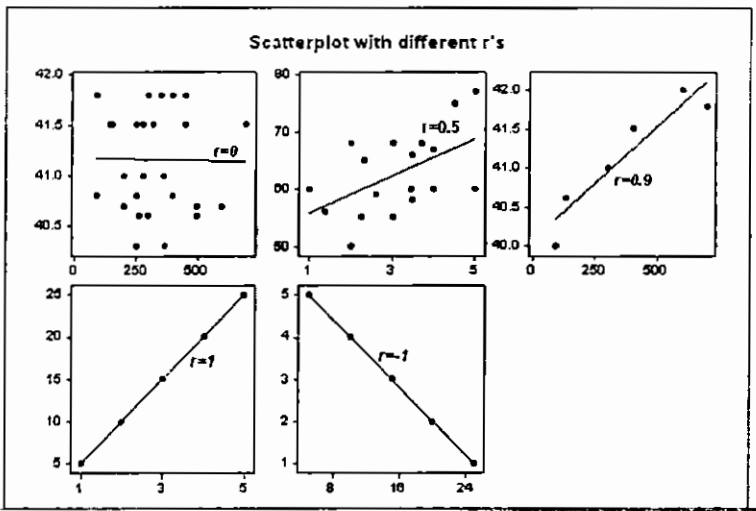
16.7. معامل الارتباط الخطى Coefficient of Linear Correlation (r)

مما سبق يتضح لنا أنه حتي بعد رسم مخطط الانتشار Scatter، ورسم خط مستقيم فإننا لا زلنا نحتاج دليل آخر يعبر عن مدى قوة هذه العلاقة وهذا الارتباط، وأحد هذه الأدلة هو معامل الارتباط Coefficient of Correlation أو معامل ارتباط بيرسون ونعبر عنه بالرمز r ، وهو معيار يقيس مقدار التباين والتأثير الذي يطرأ على Y عندما يتغير X بمقدار معين، ويتم حساب قيمته من صيغة تبدو معقدة ولكن لا تنزعج فهي فعلاً بسيطة، وسيقوم الكمبيوتر بتحمل مشقة حسابها بينما سيكون عليك تفسير قيمتها.

$$r = \sum (x-\bar{x})(y-\bar{y}) \Big/ \sqrt{(x-\bar{x})^2} * \sqrt{(y-\bar{y})^2}$$

$$r = \left[\sum xy - \left(\sum x\right)\left(\sum y\right)/n\right] \Big/ \left[\sqrt{\sum x^2 - \left(\sum x\right)^2/n} * \sqrt{\sum y^2 - \left(\sum y\right)^2/n}\right]$$

ملاحظات علي معامل الارتباط كما يظهرها الشكل 16-4:

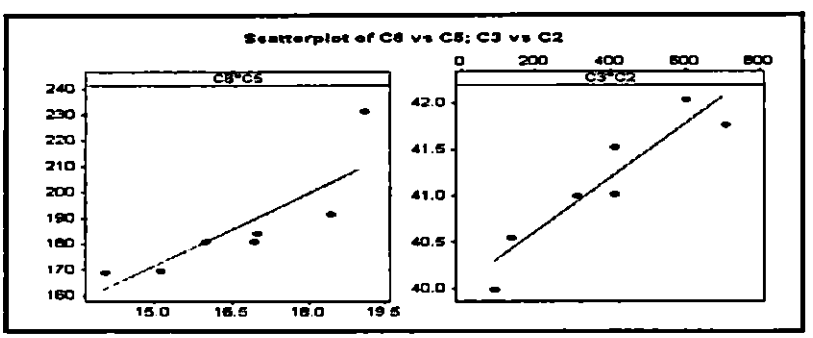


شكل رقم 16- 4 العلاقة بين قيمة المعامل r وشكل العلاقة بين المتغيرين

- هذا المعامل يقيس قوة العلاقة بين X & Y، و تتراوح قيمته من 1- حتي 1، وتدل القيمة 1- على وجود ارتباط خطى تام عكسي، والقيمة صفر تعنى عدم وجود علاقة، بينما

القيمة 1 تعنى وجود ارتباط خطى تام طردي، ويجب الانتباه الشديد إلى أن وجود ارتباط خطى بين متغيرين لا يعنى بالضرورة وجود السببية بين هذين المتغيرين، أى قد يكون المتغيران قد تأثرا بمتغير ثالث كان هو السبب في حدوثهما معا، فمثلاً إذا حسبنا معامل الارتباط بين معدل وقوع حوادث السيارات وبين معدل شراء الشماسي في فصل الشتاء، لوجدنا أن هناك ارتباطا بينهما وقد يكون معامل الارتباط كبيرا، فهل نستطيع الجزم بأن أحد المتغيرين هو سبب للآخر؟ كلا ولكن يمكن أن يعزى الأمر إلى سبب ثالث وهو معدل سقوط الأمطار في فصل الشتاء.

- كلما زادت القيمة المطلقة له $|r|$ دل ذلك علي قرب النقط من الخط المرسوم، وبالتالي دل على دقة المعادلة الممثلة لهذه العلاقة.
- إذا اقتربت قيمة $|r|$ من الصفر دل ذلك علي عدم وجود علاقة بين المتغيرين محل الدراسة Variables Are Uncorrelated وفي هذه الحالة سيظهر مخطط الانتشار Scatter كأنه نقاط عشوائية لا تدل على وجود علاقة.
- إذا كانت القيمة المطلقة $|r| = 1$ دل ذلك علي دقة الفرضية بوجود علاقة خطية، وفي هذه الحالة يمر الخط بجميع النقاط وتكون المتغيرات مرتبطة ارتباطا تاما Variables Are Perfectly Correlated.
- إشارة القيمة R تدل علي اتجاه العلاقة فالإشارة الموجبة تدل علي علاقة طردية موجبةPositive Direct Correlation والإشارة السالبة تدل علي علاقة عكسية سالبة Negative Inverse Correlation.
- القيمة المطلقة $|r|$ لا تعبر عن ميل الخط الممثل فكلا الخطين الظاهرين فى شكل رقم 16-5 لهما نفس معامل الارتباط R= 0.9 وميلهما مختلف.



شكل رقم 16- 5 خطان لهما نفس قيمة "R " والميل مختلف

مثال رقم 16- 2:

ما معني أن R = 0.82 ؟

الحل : والإجابة على هذا السؤال أنه يوجد ارتباط قوي بين المتغيرين The Two Factors Are Highly Correlated، ويعنى ذلك أيضا أن 82% من التباين الحادث فى المتغير التابع Y بسبب المتغير المستقل X.

16.8. تباين خط الانحدار Variance of Regression Line

تعتبر قيمة مجموع المربعات Sum of Squares من المقاييس الرقمية التي تعبر عن مدي قرب مجموعة النقاط المسجلة من الخط الممثل للعلاقة، (لحساب قيمة مجموع المربعات فاننا نقيس الفرق بين كل نقطة والخط الممثل للعلاقة، ثم نربع هذه القيم ثم نجمعها جبريا)، وكلما زادت قيمتها دل ذلك علي وجود تبعثر و تباين بين النقاط وبين الخط، ومجموع المربعات ما هو إلا مجموع مربعات الفروق بين القيم الحقيقية والقيم المستنتجة للنقط، وفى بعض الأحيان نعبر عنها بمجموع المربعات الناتجة عن الخطأ Sum of Squares Due to Error" SSE"، ولها أهمية كبرى في تقدير وحساب قيمة كل من معامل ارتباط بيرسون r و تباين خط الانحدار Variance of Regression Line كما يلي:

$$\text{Sum of Squares Due To Errors} = SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

$$\text{Sum of Squares for X} = SSx = \sum\left(x-\bar{x}\right)^2 = \sum x^2 - n\bar{x}^2$$

وهى معادلة تبين مدي قرب مجموعة X'sمن الخط

$$\text{Sum of Squares for Y} = SSy = \sum\left(y-\bar{y}\right)^2 = \sum y^2 - n\bar{y}^2$$

وهى معادلة تبين مدي قرب مجموعة Y'sمن الخط

Sum of Cross Product for XY Xy=SCPxy=

$$x_y = SPC_{xy} = \left(x-\bar{x}\right)\left(y-\bar{y}\right) = \sum xy - n\bar{x}\bar{y}$$

ويمكن حساب قيمة "تباين خط الانحدار " من المعادلة التالية : (-

$$\text{Variance of Regression Line} = S_e^2 = \frac{SSE}{DF} = \frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{n-2}$$

والجذر التربيعي لهذه القيمة يعطى " تقدير الخطأ المعياري"، وكلما كانت قيمته صغيرة دل ذلك على اقتراب النقاط من الخط الممثل (Alan H. Kvanli,1996).

ويمكن حساب قيمة معامل الارتباط الخطى r أيضا من المعادلة التالية:

$$r = \frac{SCPxy}{\left[\sqrt{Sxx} * \sqrt{Syy}\right]}$$

16.9. معامل التباين Covariance Coefficient

وهو مقياس آخر لبيان قوة الارتباط بين متغيرين هما X علي المحور الأفقي، وY علي المحور الرأسي ويكتب COV(X, Y)، وهو يشبه معامل الارتباط إلي حدٍ بعيد.

$$COV(XY) = \frac{1}{n-1}\Sigma(x-\bar{x})(y-\bar{y}) = \frac{1}{n-1}SCPxy$$

ومن المعادلة السابقة ، يتبين وجود علاقة جبرية بين r & COV

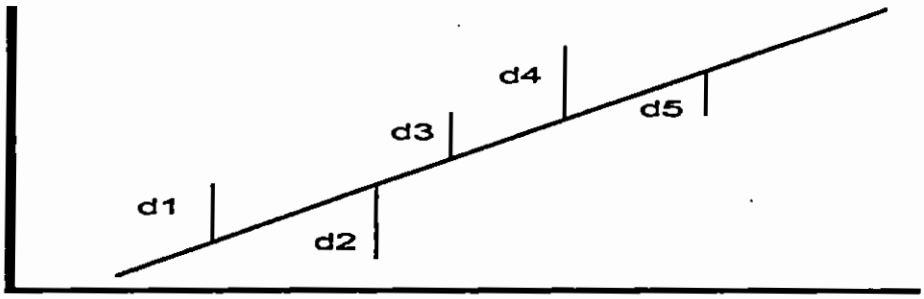
$$r = \text{Sample Correlation Between X and Y} = \frac{COV(XY)}{S_x S_y}$$

$$\text{Where Sx = Standard Deviation of the X Values} = \sqrt{\frac{Sxx}{n-1}}$$

$$\text{Sy = Standard Deviation of the Y Values} = \sqrt{\frac{Syy}{n-1}}$$

16.10. خطوط المربعات الصغرى Least Squares Lines

إذا تأكدنا بما لا يدع مجالا للشك بأن العلاقة بين المتغيرين هي علاقة خطية، فمن أين لنا أن نضمن أن مكان الخط المعبر عن النموذج Model هو أفضل الأماكن؟ هذا ما سنحاول أن نعرفه في السطور القليلة القادمة.



شكل رقم 16- 6 منحنى الانتشار ومكان الخط الممثل للبيانات (Alan H. Kvanli,1996)

فبالنظر إلى شكل 16-6 نجد خط وحيد إفترضنا أنه يعبر العلاقة ، ولكي يكون مكان هذا الخط أفضل الأماكن المحتملة، فلابد أن يكون D1 & D2 & D3 & أقل ما يمكن (لكي يكون الخطأ في الاستنتاج أقل ما يمكن)، وبالتدقيق سنجد أن D1 هي الفرق بين متوسط القراءات وقراءة النقطة الأولي هي:

$$d_1 = \bar{y}_1 - y_1$$

وكذلك

$$d_2 = \bar{y}_2 - y_2$$

وكذلك

$$d_3 = \bar{y}_3 - y_3$$

ولو رجعنا قليلا لطريقة حساب التباين (يرجى الرجوع إلى الفصل الحادى عشر)، سنجد أننا كنا نربع هذه القيمة لكل نقطة ثم نقوم بحساب التباين الكلى Total Variance من خلال التعويض في المعادلة :

$$(TotalVaria\ nce)^2 = (Var_1)^2 + (Var_2)^2 + (Var_3)^2 + ...$$

وهذا ما سنفعله الآن، فلكي يكون الخط معبرا عن النموذج Model بصورة قوية فيجب أن تكون القيم "d " الموضحة فى شكل 15-7 أقل ما يمكن.

$$\sum d^2 = d_1^2 + d_2^2 + d_3^2 + ...$$

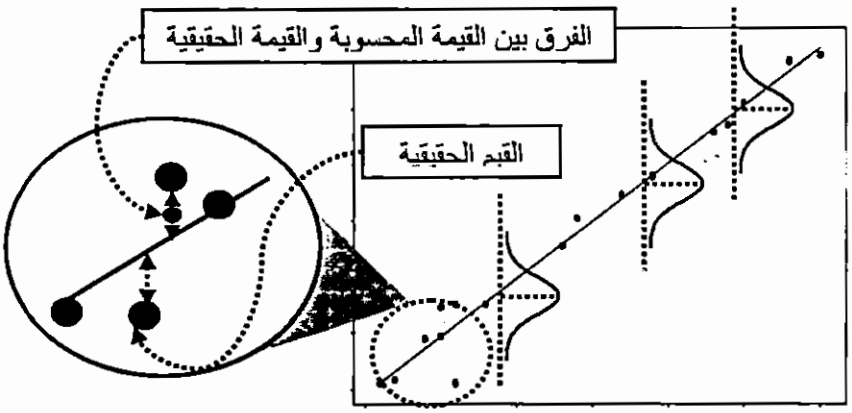
وأي خط سيحقق هذا الشرط سيكون هو الخط الأكثر تعبيرا عن العلاقة ، و أعتقد أنه بات الآن واضحا سبب تسمية الطريقة بخط المربعات الأصغر Least Squares Line، أي الخط الذي يعطي أقل تباين Variance.

وبفرض أن هذا الخط سيخضع للمعادلة $\hat{y} = B_1 X + B_0 + e$

فإذا تم تحديد قيمة B0 & B1 فإنه يمكننا استنتاج المعادلة، ويتم حساب هاتين القيمتين من خلال معادلات رياضية كثيرة، وسيقوم الكمبيوتر أيضا بدلا منا بحساب كل من B0 & B1 من العلاقة:

$$B_1 = \frac{SCP_{xy}}{SS_x} \quad \& \quad B_0 = \hat{y} - B_1 \bar{x}$$

وبالتالي يمكننا تحديد مكان الخط الأكثر تعبيرا عن العلاقة ونطلق عليه اسم النموذج Model ، وكذلك يمكننا تحديد ميل هذا الخط.



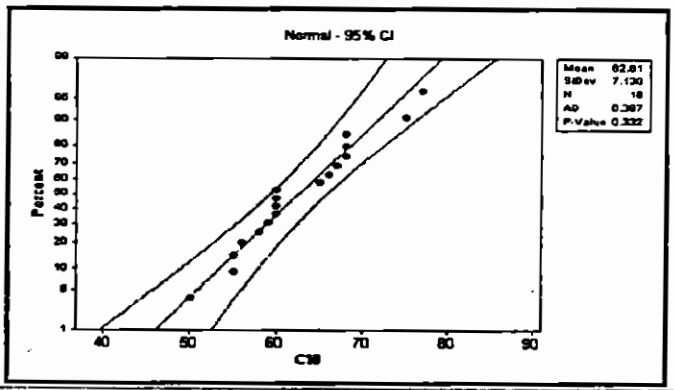
شكل رقم 16- 7 الفرق بين القيم الحقيقة للعينات وبين القيم المستنتجة من المعادلة

وهنا لا بد من الإجابة علي سؤال هام جدا وهو: هل الخط الذي حددناه بالرغم من كونه الأكثر تعبيراً عن النموذج Model، هو الذي يعبر تعبيراً مطلقا عن كل النقاط؟

الإجابة لا بالرغم من كل هذا المجهود الذي بذلناه، لماذا لا ؟ لأن بعض النقاط حتي ولو قليلة لا تقع علي الخط نفسه، وهذا جزء من الخطأ Error، وهو القيمة E في المعادلة المفترضة، وهناك جزء آخر من الخطأ يرجع إلي التقريب في حساب مكان الخط وميله لأننا حسبنا Bo B1 & من عينة، وبمعادلة تقريبية Empirical وعليه فيوجد خطآن:

1. الخطأ الأول مجموع مربعات الانحدار Sum of Squares of Regression "SSR"، وهو خطأ في تحديد مكان الخط وميله، وقيمته تتحدد من العلاقة $SSR = \frac{S_{xy}^2}{SS_x}$ وهو جزءان:

- جزء من الخطأ Error في B_0، وبسببه فإن الخط يمكن أن يتحرك لأعلي أو لأسفل مكوناً محلاً هندسياً لمكان الخط.
- وجزء من الخطأ Error في B1 (ميل الخط) وبسببه فإن الخط يتأرجح مع عقارب الساعة أو عكسها مرتكزا على نقطة منتصف ذلك الخط المرسوم، ومحصلة هاتين الحركتين (كنتيجة للخطأ فى الانحدار Error of Regression) تنتج المساحة المحصورة بين الخطين المنحنيين، والتي تمثل المحل الهندسي للخط المعبر عن العلاقة بين المتغيرين، ويوضحها الشكل 8-16.



شكل رقم 16- 8 المحل الهندسى للمعادلة النموذج الممثلة للبيانات

ومعنى ذلك أن أى خط داخل هذه المساحة، من المحتمل أن يعبر عن النقاط ولكن بنسب متفاوتة من الخطأ Error، والكمبيوتر عندما يقوم بحساب القيمة ورسم الشكل فإنه يختار الخط الأكثر تعبيرا بدرجة ثقة 95 % Confidence Level أي أن هناك خطأ في حدود 5% .

2. الخطأ الثاني مجموع مربعات الخطأ " SSE Sum of Squares of Errors "، والناتج عن القيم المتبقية Residuals وهو الفرق بين القيمة الحقيقية للنقط وبين القيمة

المحسوبة من معادلة الخط المستقيم المستنتجة، وقيمته تتحدد من العلاقة

$$SSE = S_{yy} - \frac{S_{xy}^2}{SS_x}$$

وبعد أن أثبتنا مبدئيا وجود ارتباط وعلاقة محتملة من خلال مخطط الانتشار بين المتغيرين، فهل يعني ذلك أن هذه العلاقة حقيقية ؟ إن مما يدهشنا لأول وهلة أن الإجابة على هذا السؤال لا، لأن ثبوت وجود علاقة بين المتغيرين من خلال مخطط الانتشار لا يعني أن هذه العلاقة حقيقية، وسيعتمد الحكم على حقيقية هذه العلاقة بشكل كبير على خبرة فريق العمل.

والمثال الواضح على ذلك هو مثال زيادة مبيعات الشماسي، وزيادة معدل الحوادث في فصل الشتاء ففي هذا المثال نجد أن مخطط الانتشار يشير إلى وجود علاقة، بالرغم من أننا متأكدون أنه لا توجد علاقة مباشرة، وإنما العلاقة هي وجود عامل ثالث وهو فصل الشتاء، ففيه يكثر المطر فيزداد الطلب على الشماسي، كما يزداد معدل الحوادث بسبب سقوط الامطار على الطرق وفى أماكن السير.

وإذا كان مثال الشماسي والحوادث سهل الفهم، وفيه أدركنا أنه فعلاً لا توجد علاقة بين المتغيرين بالرغم من ظهور مؤشر علاقة في مخطط الانتشار، فما العمل إذا كان الموضوع قيد البحث غير واضح وليس بهذه السهولة ؟ هل هناك مقياس أو طريقة أستطيع بها تأكيد أو نفي هذه العلاقة ؟

نعم: الوسيلة الفعالة لإثبات ذلك هي طريقة تحليل التباين "Analysis of Variance ANOVA " فهي من خلال مقاييس محددة، ستوضح وبدرجة ثقة عالية، ما إذا كانت هذه العلاقة حقيقية أم لا، وهذا ما سنتناوله فى الفصل السابع عشر.