الفصل التاسع

ملاءمة منحنٍ Curve Fitting:

تعدّ مسألة ملاءمة منحنٍ لبيانات تجريبية من المسائل الشائعة والمستخدمة على نطاق واسع في شتى المجالات. فلو كان لدينا البيانات $\{(x_i, y_i);\ i = 1,2,...,n\}$، وبفرض أن العلاقة بينهما مُعبّر عنها من خلال النموذج الرياضي:

$$y_i = g(x_i) + e_i;\ i = 1,2,3,...,n \qquad (1)$$

إذ إن e_i هو خطأ عشوائي (ضوضاء) Random Error (Noise) ، أما $g(.)$ فهي دالة ما، قد تكون مجهولة وقد تكون معلومة.

كما هو واضح فإن النموذج (1) يتضمن جزأين: الأول $g(x)$ ويمثل الجزء الحتمي Deterministic Part، وهو عبارة عن دالة رياضية مُعرّفة بدلالة المتغير x. أما الجزء الثاني فهو غير حتمي (تصادفي) والذي يمثل الخطأ العشوائي e. إن الخطأ العشوائي (الضوضاء) يمثل التغيرات الناجمة عن العوامل الفجائية والطارئة والتي هي خارج السيطرة. وعادة يعدّ e متغيرا عشوائيا معدله صفر وتباينه ثابت، يرمز له σ_e^2. لقد تناول الفصل السابق وبالتفصيل كيفية تقدير شكل الدالة $g(.)$، لذا فان هذا الفصل يبدأ من حيث انتهى الفصل السابق، ويتخصص بتقدير معلمات Parameters المعادلة الرياضية المناسبة للدالة $g(.)$ والتي سوف تعدّ معلومة.

طريقة المربعات الصغرى

:The Method of Least Squares

إنّ طريقة المربعات الصغرى هي طريقة قياسية وعامة تُستخدم على نطاق واسع لملاءمة نماذج رياضية لبيانات Data أو مشاهدات Observations معينة. يعود اكتشاف هذه الطريقة إلى الرياضي الفذ كاوس Carl Friedrich Gauss (1777–1855) بحدود العام 1795 وكان عندئذ في سن الثامنة عشرة من عمره.

المربعات الصغرى للبيانات الإحصائية

: Least Squares for Statistical Data

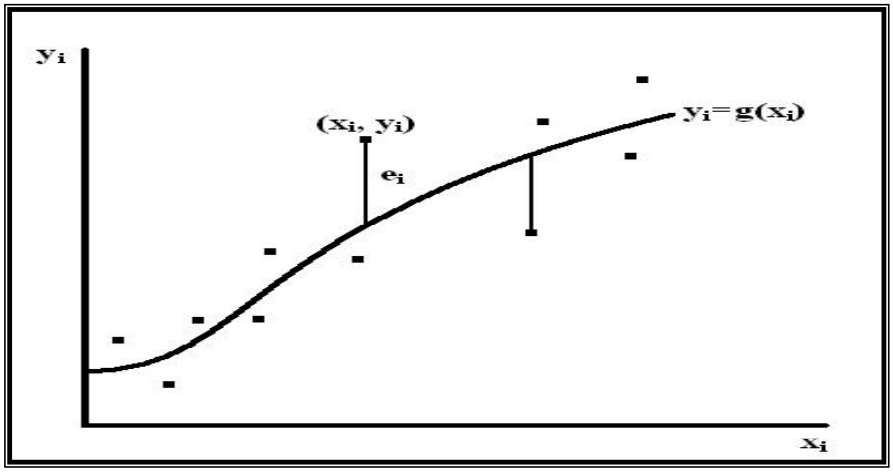
لنفرض أن لدينا زوجاً من المتغيرات (x,y)، إذ إن x يمثل المتغير المستقل Independent Variable، في حين أن y يمثل المتغير المعتمد Dependent Variable. ولنفرض أن البيانات $\{(x_i, y_i);\ i = 1,2,...,n\}$، مأخوذة من هذين المتغيرين وأن العلاقة بينهما مُعبّر عنها من خلال النموذج الرياضي (9-1). كما نفرض الآن أن $g(x)$ دالة معروفة لدينا.

تتلخص فكرة طريقة المربعات الصغرى باختيار المعلمات التي تؤدي إلى تقليل Minimizing أخطاء الملاءمة إلى اقل ما ممكن. لذا فإن الحل الناجم عن طريقة المربعات الصغرى يجعل مجموع مربعات الأخطاء عن النموذج المُلائم اقل ما ممكن Minimum. إن النموذج المُلائم على وفق هذه الطريقة يكون الأفضل على وفق معيار معروف جيدا يعرف بمعيار معدل مربع الخطأ Mean Square Error والذي يرمز له عادة MSE.

طالما فرضنا أن البيانات $\{(x_i, y_i);\ i = 1,2,...,n\}$ تُحقق النموذج الرياضي (1)، لذا فإن كل زوج من هذه البيانات يحقق المعادلة:

$$y_i = g(x_i) + e_i;\ \ i = 1,2,...,n$$

إذ إن $e_i = y_i - g(x_i)$ يمثل مقدار انحراف النقطة (x_i, y_i) عن المنحني $y_i = g(x_i)$، ويسمى الباقي (الراسب) Residual (جمعه بواقي أو رواسب)، انظر الشكل اللاحق. إن هذه البواقي قد تكون موجبة (إذا وقعت النقطة فوق المنحني)، أو سالبة (إذا وقعت النقطة تحت المنحني). وطالما أن الذي يُهمّنا هو مقدار القرب من المنحني (بغض النظر عن الإشارة)، لذا فإننا نحتاج إلى التخلص من إشارة البواقي السالبة. ويمكن انجاز ذلك إما بأخذ القيم المطلقة لها، أو بتربيعها. وتتبنى طريقة المربعات الصغرى الخيار الثاني لأنه الأيسر رياضيا. لذا فإن المبدأ المُعتمد في طريقة المربعات الصغرى أن يكون مجموع مربعات بواقي النموذج اقل ما يمكن.



الشكل (9-2): طريقة المربعات الصغرى.

نموذج الانحدار الخطي البسيط

:Simple Linear Regression Model

كما هو معروف فان الصيغ الخطية (التي تكون على شكل خط مستقيم) هي أبسط الصيغ الرياضية. لذا لو فرضنا أن الجزء الحتمي من النموذج (1) على شكل معادلة خط مستقيم، فإن البيانات $\{(x_i, y_i);\ i = 1,2,...,n\}$ سوف تُحقق النموذج الرياضي الآتي:

$$y_i = a + bx_i + e_i;\ \ i = 1,2,...,n \qquad (2)$$

وكما هو واضح فإن هذا النموذج يتضمن معلمتين مجهولتين هما a و b. إن هدفنا الذي نروم إليه هو تقدير هاتين المعلمتين باستخدام طريقة المربعات الصغرى.

نفرض أن Q يمثل مجموع مربعات البواقي، أي إن:

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 .$$

ولكي يكون Q أقل ما يمكن، نأخذ مشتقته الجزئية بالنسبة إلى كل من a و b، ثم نساوي المقدار الناتج بالصفر:

$$\frac{\partial Q}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0,$$

$$\frac{\partial Q}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0.$$

وبقسمة طرفي المعادلتين الأخيرتين على 2- ثم تبسيطهما تنتج المعادلتان الآتيتان اللتان تُعرفان بالمعادلتين الطبيعيتين Normal Equations:

$$\sum_{i=1}^{n} y_i = na - b\sum_{i=1}^{n} x_i,$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} {x_i}^2.$$

وبحل المعادلتين الأخيرتين آنيا نحصل على مُقدّري المربعات الصغرى:

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sum x_i^2 - n\,\bar{x}^2}, \qquad (3a)$$

$$a = \bar{y} - b\bar{x}. \qquad (3b)$$

إذ إن $\bar{x} = \sum_{i=1}^{n} x_i / n$ يمثل معدل قيم x، في حين أن $\bar{y} = \sum_{i=1}^{n} y_i / n$ يمثل معدل قيم y.

ملاحظة:

1- إن نموذج الانحدار الخطي البسيط يمثل معادلة أفضل خط مستقيم على وفق معيار معدل مربع الخطأ MSE. ويطلق عادة على هذا النموذج معادلة انحدار y على x Regression Equation of y on x.

- تمثل المعلمة a المقطع الصادي Intercept (مقدار المسافة التي تقطعها المعادلة على محور الصادات عندما يكون x=0)، في حين تمثل المعلمة b ميل Slope المستقيم $\left({}^{dy}/_{dx} = b\right)$.

3- النقطة $(\bar{x}, \bar{y})$ دائما تحقق معادلة الانحدار (تقع عليها).

4- يمكن الاستفادة من هذا النموذج للتنبؤ بقيم المتغير y إذا عُلمت قيم المتغير x ولا يجوز العكس. فإذا أريد التنبؤ بقيم x من قيم y فيجب أولا عمل معادلة انحدار x على y بشكل x=c+dy+e، ثم إيجاد قيم c و d كما يأتي:

$$d = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum y_i^2 - n\bar{y}^2}, \qquad (4a)$$

$$c = \bar{x} - d\bar{y}.. \qquad (4b)$$

المثال (1):

افرض أن لديك البيانات :

X	2	5	9	14
Y	15	23	28	36

أ- لائم نموذج انحدار خطي بسيط للمتغير y على x لهذه البيانات واحسب معدل مربع الخطأ. استخدم النموذج المُلائم للتنبؤ بقيمة المتغير y عندما تكون قيمة المتغير x هي 40.

الحل:

أ-

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{901 - 4(7.5)(25.5)}{306 - 4(7.5)^2} = \frac{136}{81} = 1.6790,$$

$$a = \bar{y} - b\bar{x}. = 25.5 - 1.6790(7.5) = 12.9074.$$

والجدول الآتي يوضح حساب بواقي النموذج:

i	x_i	y_i	$\hat{y}_i = 12.9074 + 1.6790x_i$	$e_i = y_i - \hat{y}_i$
1	2	15	16.2654	-1.2654
2	5	23	21.3024	1.6976
3	9	28	28.0184	-0.0184
4	14	36	36.4134	-0.4134

(1) مجموع مربعات الأخطاء (البواقي):

$$\sum_{i=1}^{n} e_i^2 = (-1.2654)^2 + \ldots + (-0.4134)^2 = 4.6543.$$

(2) معدل مربعات الأخطاء Mean-Square-Error (MSE):

$$\text{معدل مربعات الأخطاء} = \frac{\text{مجموع مربعات الأخطاء}}{\text{حجم العينة} - \text{عدد معلمات النموذج المقدرة}}.$$

أي إن:

$$MSE = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{4.6543}{4-2} = 2.32715.$$

علما بأن مقام معدل مربعات الأخطاء يسمى درجات الحرية Degrees of Freedom (df).

(3) جذر معدل مربعات الأخطاء Root-Mean-Square-Error (RME):

$$RME = \sqrt{MSE} = \sqrt{2.32715} = 1.5255.$$

إن RME يمثل معدل بعثرة (تشتت) نقاط لبيانات عن معادلة الانحدار، وبلغة الإحصاء فإن RME يمثل القيمة التقديرية للانحراف المعياري للخطأ العشوائي e في النموذج (9−1)، أي إن $RME = \hat{\sigma}_e$.

ب− لاستخدام النموذج المُلائم للتنبؤ نقوم بالتعويض عن قيمة المتغير x بالقيمة المطلوبة x=40 مع إهمال الخطأ العشوائي (لان معدله صفر):

$$\hat{y} = E(y/x=40) = 12.9074 + 1.6790(40) = 80.0674.$$

إذ إن $E(y/x=40)$ ترمز إلى التوقع الشَرطي للمتغير y مُعطى أن x=40.

الواجب

إذا كان المتغير X يمثل وزن الدجاجة (باون)، و المتغير Y يمثل كمية الغذاء المستهلك من قبل الدجاجة (باون):

x	4.6	5.1	4.8	4.4	5.9	4.7	5.1	5.2	4.9	5.1
y	87.1	93.1	89.8	91.4	99.5	92.1	95.5	99.3	63.4	94.4

لائم معادلة انحدار خطي لهذه البيانات.

Ans. y=55.26+7.69x, $\hat{\sigma e} = 2.3656$