الفصل السادس

الانحدار الخطي وتحليل الارتباط
Linear Regression and Correlation Analysis

نهتم في بحوث الهندسة الطبية الحيوية أو التصميم، في كثير من الأحيان، فيما إذا كانت هناك ارتباط بين متغيرين، أو مجتمعين إحصائيين، أو عمليتين. ويمكن لهذه الارتباطات أن تعطينا معلومات حول العمليات البيولوجية الأساسية في الحالات العادية والمرضية وبالتالي تساعدنا على نمذجة العمليات، مما يسمح لنا بالتنبؤ بسلوك إحدى العمليات مع الأخذ في الاعتبار حالة عملية أخرى مرتبطة.

بالأخذ في الاعتبار مجموعتين من العينات، X و Y، فإننا نطرح السؤال التالي: "هل المتغيران أو العمليتان العشوائيتان، X و Y، مرتبطتان؟" وبعبارة أخرى، هل يمكن نمذجة y كدالة خطية لـ x بحيث يكون:

$$\hat{y} = mx + b$$

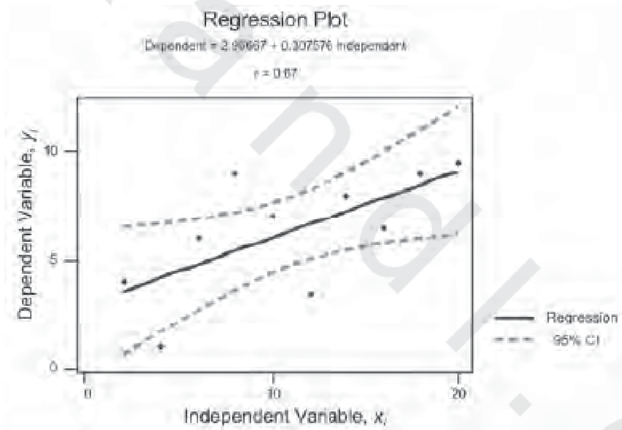
وبالنظر إلى المخطط التالي لبيانات تجريبية (الشكل ٦.١)، حيث تم رسم مجموعة البيانات، y_i، مقابل مجموعة البيانات، x_i:

نلاحظ أن البيانات تقترب من الوقوع على خط مستقيم. وهناك ميل إلى أن تصبح اتجاه بحيث تزداد y بما يتناسب مع الزيادات في x وهدفنا هو تحديد الخط (النموذج الخطي) الذي يلائم هذه البيانات بأفضل ما يمكن ومقدار تقارب نقاط

البيانات المُقاسة بالنسبة إلى الخط الذي تم ملاءمته (الذي تم إنشاؤه بواسطة النموذج). وبعبارة أخرى، إذا كان الخط الذي تم نمذجته ملائمًا جدًّا للبيانات، فإننا نثبت أن y يمكن نمذجتها بدقة كدالة خطية لـ x، وبالتالي، يمكننا التنبؤ بـ y مع الأخذ في الاعتبار x باستخدام النموذج الخطي.

إن المدخل لملاءمة خط يتنبأ بالعملية y بأفضل ما يمكن من العملية x، هو إيجاد البارامترات m و b، التي تقلّل إلى الحد الأدنى من الخطأ بين بيانات النموذج والبيانات الفعلية بطريقة المربعات الصغرى:

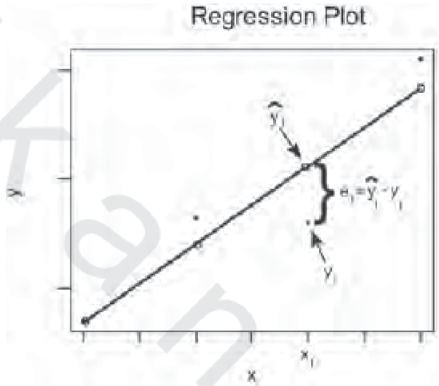
$$\min\left[(y-\hat{y})^2\right]$$



الشكل (٦,١). نتائج خط الانحدار مُطبَّقة على العينات الموضحة بالمخطط المبعثر (النقاط السوداء). يوضح الخط الأسود المتصل خط الملاءمة الأفضل (بارامترات النموذج المذكورة فوق المخطط) على النحو الذي حدده خط الانحدار توضح المنحنيات الحمراء المنقطة فترة الثقة للميل (slope). وأخيراً، فإن القيمة r هي معامل الارتباط.

وبعبارة أخرى، كما هو موضح في الشكل (٦,٢)، فإن لكل قيمة مُقاسة للمتغير المستقل، x، سوف يكون هناك قيمة مُقاسة للمتغير التابع، y، بالإضافة إلى

القيمة المتوقعة أو التي تم نمذجتها لـ y، يُرمز لها بـ $\hat{y}$ ، التي سوف يحصل عليها المرء إذا تم استخدام المعادلة $\hat{y} = mx+b$ للتنبؤ بـ y. تدل المعادلة $e_i = y_i - \hat{y}_i$ على الأخطاء التي تحدث في كل زوج من (x_i, y_i) عندما لا تكون القيمة التي تم نمذجتها مطابقة تماماً للقيمة المُتوقعة بسبب عوامل لم يتم أخذها في الاعتبار في النموذج (الضجيج، والآثار العشوائية، وعدم الخطية).



الشكل (٦,٢). يمكن استخدام خط الانحدار لتقدير الخط المستقيم الذي "يلائم" على أفضل وجه نقاط البيانات المُقاسة (الدوائر المملوءة). تمثل x_i و y_i في هذا التوضيح المتغيرات المُقاسة المستقلة وغير المستقلة (التابعة)، على التوالي. يتم استخدام خط الانحدار لنمذجة المتغير التابع، y، كدالة خطية للمتغير المستقل، x. إن الخط المستقيم الذي يمر عبر نقاط البيانات المُقاسة هو نتيجة الارتداد الخطي حيث يتم تقليل الخطأ، e_i، إلى الحد الأدنى على كامل نقاط البيانات بين القيمة المتوقعة (الدوائر المفتوحة) للمتغير التابع، $\hat{y}_i$، والقيمة المُقاسة للمتغير التابع، y_i .

في محاولة لملاءمة الخط إلى البيانات التجريبية، فإن هدفنا هو تقليل هذه الأخطاء، e_i ، إلى الحد الأدنى بين القيم المُقاسة والمُتوقعة للمتغير التابع، y. إن الطريقة المستخدمة في خط الانحدار والعديد من تقنيات النمذجة الطبية الحيوية الأخرى هي إيجاد بارامترات النموذج، مثل m و b، التي تقلّل من مجموع الأخطاء المُربعة، e_i^2 ، إلى الحد الأدنى.

بالنسبة لخط الانحدار، فإننا نسعى إلى تقدير المربعات الصغرى لـ m واستخدام التقريب التالي:

لنفترض أنه لدينا N عينة لكل من العمليات x و y. نحاول التنبؤ بـ y من x المُقاسة باستخدام النموذج التالي:

$$\dot{y} = mx + b$$

إن الخطأ في التنبؤ في كل نقطة بيانات، x_i، هو

$$error_i = y_i - \hat{y}_i$$

وفي طريقة المربعات الصغرى، فإننا نختار m و b للتقليل من مجموع الأخطاء المُربعة إلى الحد الأدنى:

$$(y_i - \hat{y}_i)^2 \qquad \text{عندما } i=1 \text{ إلى } N$$

لإيجاد حل شكل متقارب لـ m و b، نستطيع كتابة صيغة لمجموع الأخطاء المُربعة:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

ثم نستبدل $\hat{y}_i$ بـ $(mx_i + b)$، نموذجنا، وإجراء عمليات التربيع [3, 5].

يمكننا بعد ذلك أخذ مشتقات الصيغة أعلاه بالنسبة إلى m ومن ثم مرة أخرى بالنسبة إلى b. إذا وضعنا الصيَغ المُشتقة إلى الصفر لإيجاد القيم الدنيا، سيكون لدينا معادلتان بمجهولين، m و b، ويمكننا ببساطة استخدام الجبر للحل بالنسبة للبارامترات المجهولة، m و b. وسوف نحصل على الصيَغ التالية لـ m و b بدلالة x_i و y_i المُقاسة:

$$m = \frac{\sum_{i=0}^{N-1} x_i y_i - \left(\sum_{i=0}^{N-1} x_i\right)\left(\sum_{i=0}^{N-1} y_i\right) \Big/ N}{\sum_{i=0}^{N-1} x_i^2 - \left(\sum_{i=0}^{N-1} x_i\right)^2 \Big/ N}$$

و

$$b = \overline{y} - m\overline{x}$$

حيث

$$\bar{y}=\frac{1}{N}\sum_{i=0}^{N-1}y_i \quad \text{و} \quad \bar{x}=\frac{1}{N}\sum_{i=0}^{N-1}x_i$$

وبالتالي، بعد حصولنا على البيانات المُقاسة، يمكننا ببساطة استخدام المعادلات لـ m و b لإيجاد الخط، أو النموذج الخطي، ذي الملاءمة الأفضل.

معامل الارتباط The Correlation Coefficient

من المهم إدراك أن خط الانحدار يلائم خطًّا ما إلى أي مجموعتين من البيانات بغض النظر عن مدى جودة نمذجة البيانات بواسطة النموذج الخطي. وحتى لو كانت البيانات، عندما يتم رسمها كمخطط مبعثر، لا تبدو أنها تشبه الخط بشيء، فإن خط الانحدار سوف يلائم الخط للبيانات. وكمهندسين طبيين حيويين، علينا أن نسأل: "ما مدى جودة "ملاءمة" البيانات المُقاسة للخط المُقدَّر من خلال خط الانحدار ؟"

إن أحد مقاييس مدى جودة ملاءمة البيانات التجريبية للنموذج الخطي هو معامل الارتباط. يأخذ معامل الارتباط، r، قيمة بين -1 و 1 ويشير إلى مدى جودة ملاءمة النموذج الخطي للبيانات.

يمكن تقدير معامل الارتباط، r، من البيانات التجريبية x_i و y_i، وذلك باستخدام المعادلة التالية:

$$r=\frac{\sum_{i=0}^{N-1}(x_i-\bar{x})(y_i-\bar{y})}{\left[\sum_{i=0}^{N-1}(x_i-\bar{x})^2\sum_{i=0}^{N-1}(y_i-\bar{y})^2\right]^{1/2}}$$

حيث

$$\bar{y}=\frac{1}{N}\sum_{i=0}^{N-1}y_i \quad \text{و} \quad \bar{x}=\frac{1}{N}\sum_{i=0}^{N-1}x_i$$

من المهم أن نلاحظ أن $r = 0$ لا يعني أن العمليتين، x و y، مستقلتان. إنه يشير فقط إلى أن أي تبعية بين x و y ليست موصوفة أو مُنمذجة جيداً بواسطة علاقة خطية. ويمكن أن تكون هناك علاقة غير خطية بين x و y. إن $r = 0$ يعني ببساطة أن x و y غير مترابطتين بالمعنى الخطي. وهذا يعني أن المرء قد لا يتوقع y من x باستخدام النموذج الخطي، $y = mx + b$.

إن أحد المقاييس المتصلة بمعامل الارتباط، r، هو معامل التحديد، R^2، الذي هو ملخص الإحصائية التي تخبرنا عن مدى جودة ملاءمة نموذج انحدار للبيانات. ويمكن استخدام R^2 كمقياس لجودة ملاءمة أي نموذج ارتداد، وليس فقط خط الانحدار. وبالنسبة لخط الانحدار، فإن R^2 هو مربع معامل الارتباط، وله قيمة بين 0 و 1. يخبرنا معامل التحديد عن مقدار التغير في البيانات الذي يمكن تفسيره بواسطة بارامترات النموذج كجزء من مجموع التغير في البيانات.

ومن المهم معرفة أن الميل المُقدَّر للملاءمة الأفضل ومعامل الارتباط هو الإحصائيات التي قد تكون أو لا تكون مهمة. وبالتالي، قد يتم إجراء الاختبارات t لاختبار ما إذا كان الميل المُقدَّر من خلال ملاءمة خطية يختلف كثيراً عن الصفر [3]. وبالمثل، قد يتم إجراء الاختبارات t لاختبار ما إذا كان معامل الارتباط يختلف كثيراً عن الصفر. وأخيراً، فقد نحسب أيضاً فترات الثقة للميل المُقدَّر [3].