الانحدار الخطي البسيط

Linear Regression

يختص الانحدار الخطي البسيط بإيجاد المعادلة الرياضية التي تربط بين المتغيرين (X) و(Y) ، وسمي بالانحدار الخطي لأن العلاقة خطية (معادلة خط مستقيم) ويتم التأكد من ذلك من رسم شكل الانتشار، وايضا سمي الانحدار البسيط ؛ لأنه اختص بدراسة العلاقة بين متغيرين اثنين فقط ، و لإيجاد المعادلة التي تربط بين المتغيرين ، نحدد اولاً ايهما المتغير التابع ونرمز له بالرمز (Y) وايهما المستقل ونرمز له بالرمز (X) .

المعادلة التي تربط بين المتغيرين تسمى معادلة انحدار المتغير التابع (Y) على المتغير المستقل(X) ويكون شكل معادلة خط الانحدار (المستقيم) للمجتمع على الشكل : $Y=\beta_0+\beta_1X+E_i$.

وللعينة يكون شكل المعادلة : $Y=b_0+b_1X+e_i$.

وهنا تم اعتبار ان المتغير التابع هو المتغير (Y)وأن المتغير المستقل هو المتغير(X)في بعض التطبيقات يطلب من الاحصائي تحديد المتغير التابع من المتغير المستقل .

وحيث أن β_0 المعامل الثابت اما

$\beta_1 =$ فهي تسمى ميل الانحدار أو معامل الانحدار، وهي تمثل مقدار التغير في المتغير التابع(Y)عند زيادة قيمة المتغير (X) بمقدار وحدة واحدة .

عندما تكون البيانات مأخوذة من عينة مسحوبة من المجتمع الاحصائي ، يتم تقدير نموذج الانحدار ويتم التعبير عن شكل معادلة الانحدار المقدرة كالاتي :

$$\hat{Y}_i=\hat{b}_0+\hat{b}_1X+e_i$$

$\hat{Y}$: تمثل القيمة المقدرة للمتغير التابع للمجتمع (الظاهرة التي تتاثر بالمتغير المستقل) .

X :تمثل قيمة أو قيم المتغير المستقل المسحوبة من المجتمع الاحصائي(الظاهرة المؤثرة على المتغير تابع)

$\hat{b}_0$: المعامل الثابت .

$\hat{b}_1$: معامل الانحدار أو ميل الانحدار.

e_i : يمثل انحراف القيم التقديرية $\hat{Y}$ عن القيم الحقيقية Y ويمكن التعبير عن هذه العلاقة بالصيغة التالية $e_i=Y-\hat{Y}$.

تقدير معامل (ميل) الانحدار باستخدام المربعات الصغرى

يتم استخدام طريقة المربعات الصغرى في تقدير ميل الانحدار المجهول ، حيث تقوم هذه الطريقة بتقليل مجموع مربعات انحرافات القيم التقديرية عن القيم الحقيقية .

$$e_i=Y_i-\hat{Y}_i \quad , \hat{Y}_i=b_0+b_1X_i.$$

و بالتعويض عن قيمة $\hat{Y}_i$ ، بتربيع الطرفي وادخال المجموع ، ثم المفاضلة بالنسبة b_0 ومساواتها بالصفر، وادخال المجموع بالقسمة على2- نحصل على :

$$\sum_{i=1}^{n} Y_i = n\hat{b}_0 + \hat{b}_1 \sum_{i=1}^{n} X_i \ldots\ldots\ldots\ldots\ldots(1)$$

وتربيع طرفي e_i ثم المفاضلة بالنسبة لb_1 , ومساواتها بالصفر والقسمة على2- نحصل على :

$$\sum_{i=1}^{n} X_iY_i = \hat{b}_0 \sum_{i=1}^{n} X_i + \hat{b}_1 \sum_{i=1}^{n} X_i^2 \ldots\ldots(2)$$

وبإجراء الاختصارات والاضافات والعمليات الرياضية نصل الى الصيغة النهائية :

$$\hat{b}_0 . \frac{n\sum_{i=1}^{n} xiyi - \sum_{i=1}^{n} xi \sum_{i=1}^{n} yi}{n\sum_{i=1}^{n} xi^2 - (\sum_{i=1}^{n} xi)^2} =$$

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} yi}{n} + \hat{b}_0 \frac{\sum_{i=1}^{n} xi}{n} \qquad , \qquad \hat{b}_1 = \bar{Y} + b_0 \bar{X}$$

فرضيات نموذج الانحدار الخطي البسيط :

الافتراض هو ان معطيات المتغير المستقل قادرة على اظهار تأثيرها في تغير المتغير التابع

كما أن قيم المتغير المستقل تكون مستقلة عن بعضها ، ويمكن صياغة الفرض كالاتي

$\sum_{i=1}^{n}(xi - \bar{x})^2 \neq 0$ كما ان قيم المتغير المستقل تكون مستقل عن بعضها ،ويمكن صياغة الفرض كالات

*الخطأ العشوائي يتبع التوزيع الطبيعي ، وتوزيع المعاينةy يتبع التوزيع الطبيعي وايضا المتغير التابع e_i * الخطأ العشوائي

* الخطأ العشوائي لمعالم نموذج الانحدار تتبع ايضا التوزيع الطبيعي $e_i \sim N(0, S^2)$.

* القيمة المتوقعة للخطأ العشوائي مساوي للصفر$E(ei) = 0$.

* تباين حد الخطأ العشوائي ثابت لكل قيم المتغير العشوائي $V(e_i) = E(e_i^2) - E(e_i)^2$

* القيمة التي يأخذها الخطأ العشوائي في فترة ما تكون غير متعلقة بقيمة اخرى في فترة اخرى

$$COV(ei,ej) = 0 \qquad j=1\ldots\ldots n \qquad i=1\ldots\ldots n$$

*ملاحظة :

يمكن اختبار فرضيات نموذج الانحدار الخطي البسيط باستخدام اختبار F.

ان تحليل الانحدار هو وسيلة إحصائية تفيد في دراسة العلاقة بين متغيرين او اكثر بشكل معادلة بحيث يكون احد هذه المتغيرات متغير معتمد والأخر متغير مستقل يؤثر في المتغير المعتمد ويمكن تقسيم الانحدار الى قسمين:

1- انحدار خطي بسيط Simple linear Regression
2- انحدار خطي متعدد Multiple linear Regression

ونظرا لضيق الوقت سوف نتناول الانحدار الخطي البسيط فقط.

في الانحدار الخطي البسيط: يكون هناك متغيرين فقط احدهما مستقل xi والأخر معتمد yi

اما في الانحدار الخطي المتعدد فيكون هناك عدة متغيرات مستقلة ومتغير معتمد واحد فقط

نموذج الانحدار(معادلة الانحدار)

$$\hat{y}_i = \hat{B}_0 + \hat{B}_1 x_i$$

حيث ان : $\hat{Y} =$ المعتمد التغير $= dependent\ variable$

$x_i =$ المتغير المستقل $= independent\ variable$

$\hat{B}_0\ ,\ \hat{B}_1 =$ معاملات الانحدار

نقطة التقاطع مع المحور $\hat{B}_0 = Y$

ميل خط الانحدار $\hat{B}_1 =$

اما المعادلات الخاصة بمعاملات الانحدار فتحسب وفق القوانين التالية:

$$\hat{B}_1 = \frac{\sum (xi-\bar{x})(yi-\bar{y})}{\sum (xi-\bar{x})^2}$$

او

$$\hat{B}_1 = \frac{\sum xiyi - \frac{(\sum xi)(\sum yi)}{n}}{\sum xi^2 - \frac{(\sum xi)^2}{n}}$$

يمكن استخدام احد القانونين فقط

$$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x}$$

مثال // البيانات التالية تمثل x المطر والنبات الطبيعي y والمطلوب حساب معادلة انحدار y/x ؟

	xi	yi
	30	73
	20	50
	60	128
	80	170
	40	87
	50	108
	60	135
	30	69
	70	148
	60	132
المجموع	500	1100

الحل نجد

	Xi^2	Yi^2	xiyi
	900	5329	2190
	400	2500	1000
	3600	16384	7680
	6400	28900	13600
	1600	7569	3480
	2500	11664	5400
	3600	18225	8100
	900	4761	2070
	4900	21904	10360
	3600	17424	7920
المجموع	28400	134660	61800

بما ان معادلة الانحدار $\hat{y}_i = \hat{B}_0 + \hat{B}_1 x_i$

نحتاج الى

إيجاد

$$\bar{x} = \frac{\sum x_i}{n} = \frac{500}{10} = 50$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{1100}{10} = 110$$

$$\hat{\boldsymbol{B}}_1 = \frac{\sum \boldsymbol{xiyi} - \frac{(\sum \boldsymbol{xi})(\sum \boldsymbol{yi})}{\boldsymbol{n}}}{\sum \boldsymbol{xi}^2 - \frac{(\sum \boldsymbol{xi})^2}{\boldsymbol{n}}} = \frac{\mathbf{6180} - \frac{(\mathbf{500})(\mathbf{100})}{\mathbf{10}}}{\mathbf{28400} - \frac{(\mathbf{500})^2}{\mathbf{10}}} = 2$$

$$\hat{\boldsymbol{B}}_0 = \bar{\boldsymbol{y}} - \hat{\boldsymbol{B}}_1 \bar{\boldsymbol{x}} = \mathbf{10} - (\mathbf{2})(\mathbf{50}) = \mathbf{10}$$

$$\hat{y}_i = \hat{B}_0 + \hat{B}_1 x_i$$

$$\hat{y}_i = 10 + 2x_i$$

نموذج الانحدار الخطي المتعدد

وتكون صيغته كالتالي : $i=1.......k$ $Y_i= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} +.... \beta_K X_{Ki} +e_i$ والنموذج الممثل للعينة المسحوبة من نفس المجتمع الاحصائي للدراسة :

$Y_i =b_0 +b_1X_{1i} +b_2X_{2i} +.....+b_kX_{Ki}$

أي أن المتغير التابع قد يتأثر بأكثر من متغير مستقل واحد كما هو الحال في المثال الذي يكون فيه وزن الطفل هو المتغير Y .

بينما عمر الطفل هو المتغير المستقل (المفسر) ، ولكن يمكن اضافة متغير آخر و هو طول الطفل كمتغير مفسر ثاني حيث يصبح :

$Y_i= b_0+b_1X_{1i}+ b_2X_{2i}+e_i$

و هكذا النموذج يصبح X_2 هو طول الطفل و X_1 عمر الطفل .

عندما تكون المتغيرات المستقلة $(X_1,X_2,....X_p)$ تساوي صفر، فان β_0 يمثل متوسط المتغير التابع Y ،

حيث أن المعامل β_1 يمثل الاثر الصافي أو المباشر لتغير (X_1) بوحدة واحدة على القيمة المتوقعة للمتغير Y

مع ثبات باقي المتغيرات ، في حين أن المعامل β_2 تمثل التغير في القيمة المتوقعة للمتغير التابع Y ،

ناتج عن تغير المتغير المستقل(X_2) بوحدة واحدة بفرض ثبات المتغيرات $(X_1,X_3....X_P)$ ،وهكذا يستمر التفسير لباقي معاملات الانحدار الجزئية .

تقدير معالم النموذج

وتستخدم طريقة المربعات الصغرى في تقدير معالم نموذج الانحدار الخطي المتعدد ،ويأخذ نموذج الانحدار المقدّر من العينة المقابلة لمعادلة انحدار المجتمع الشكل :

$$Y_i =b_0+ b_1X_{1i} + b_2X_{2i} +........b_pX_{pi} + e_i \qquad i=1,2.....n$$

حيث أن :

$(b_0,b_1,....b_p)$ هي القيم المقدرة للمعالم $(\beta_0, \beta_1,.... \beta_P)$

e_i : الباقي و هو الفرق بين القيمة الفعلية للمتغير التابع و القيمة المقدرة لها للمشاهدة رقم (i).

n :عدد المشاهدات (حجم العينة) ، ويمكن كتابة المعادلة في شكل مصفوفة :

$$
\begin{pmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ b_p \end{pmatrix}
\quad
\begin{pmatrix} 1 & X_{11} & X_{21} & X_{31} & . & X_{P1} \\ 1 & X_{12} & X_{22} & X_{32} & . & X_{P2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & X_{1n} & X_{2n} & X_{3n} & & X_{Pn} \end{pmatrix}
\begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{pmatrix}
\quad
\begin{pmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{pmatrix}
$$

ويمكن باختصار كتابة نموذج الانحدار كالتالي :

$$e + b\ Y= X$$

حيث أن :

b= متجه عمودي من الدرجة $((p+1)\times 1)$ يحتوي على قيم المعاملات المقدرة .

X مصفوفة بيانات العينة من الدرجة $((p+1)\times n)$ حيث n عدد المشاهدات (حجم العينة) .

e= متجه عمودي من الدرجة $n\times 1$ ويحتوي على البواقي .

وتستخدم طريقة المربعات الصغرى لتقدير معالم نموذج الانحدار المتعدد، وتنص على تصغير مجموع مربعات البواقي الى ادنى قيمة له $\min \sum_{i=1}^{n} e_i^2$.

. ويتم تقدير معالم نموذج الانحدار بحيث تكون الدالة دالة صغرى ،وبإجراء العمليات الرياضية والاختصارات نحصل على مقدر المربعات الصغرى .

$$\sum_{i=1}^{n} ei^2 = \sum_{i=1}^{n}(y\text{i} - b_0 - b_1X_{11} - b_2X2_i - \ldots\ldots - b_pX_{pi})$$

$$e^Te = \begin{bmatrix} e_1 & e_2 & e_3 & . & . & . & e_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

$$e^21 + e^22 + e^23 + \ldots\ldots\ldots + e^2n = \sum_{i=1}^{n} ei^2$$

اذا :

$$e^Te =(Y-Xb)^T(Y-Xb\ )$$

$$e^Te = Y^TY - 2b^TX^TY + b^TX^TXb \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

ومساواة الناتج بالصفر b وبمفاضلة المعادلة رقم 1 بالنسبة ل

$$\frac{\partial e^Te}{\partial b} = -2X^TY + 2X^TXb = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

$$X^TXb = X^TY \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

نحصل على: $(X^TX)^{-1}$ وبضرب المعادلة (3) ف

$$(X^TX)^{-1} X^TXb = (X^TX)^{-1}X^TY$$

$$Ib = (X^TX)^{-1}X^TY \quad , \qquad \hat{b} = (X^TX)^{-1}X^TY$$

نموذج الانحدار متعدد الحدود(Polynomial Regression Models):

ويعرف نموذج الانحدار الخطي متعدد الحدود بانه دالة بأنه دالة يظهر فيها متغير مستقل اساسي عدد من المرات مرفوعا في كل مرة الى درجة أعلى . ويأخذ النموذج من الدرجة (p) الصيغة التالية :

$$Y_i = \beta_0 + \beta_1 X_i^{\,1} + \beta_2 X_i^{\,2} + \ldots + \beta_p X_i^{\,p} + U_i \qquad\qquad i = 1\ldots n$$

ويعتبر نموذج الانحدار متعدد الحدود حالة خاصة من نموذج الانحدار الخطي المتعدد ، وبما انه يوجد متغير اساسي واحد فانه يمكن تمثيل أي نموذج متعدد الحدود ، برسم شكل الانتشار بين المتغير التابع والمتغير المستقل بهدف تحديد منحنى يصف العلاقة بينهما .

وستقتصر دراستنا على نموذجيّ الانحدار من الدرجة الثانية والثالثة ؛ لان الدرجات الأعلى نادرا ما تستخدم في الواقع العملي فضلاً عن أن مشكلة الارتباط الخطي المتعدد تتفاقم بزيادة عدد الحدود . وبما أن هذه النماذج كما اسلفنا هي حالات خاصة من نموذج الانحدار الخطي المتعدد ، فانه يتم استخدام طريقة المربعات الصغرى الاعتيادية لتقدير معالم النموذج ، وكذلك يتم حساب الاحصاءات الاخرى كمعامل التحديد ،الخطأ المعياري وغيرها بنفس الصيغ المستخدمة لنموذج الانحدار الخطي المتعدد .

*اختبار معنوية نموذج الانحدار ككل

لاختبار معنوية كل المتغيرات المستقلة يتم اختبار الفرض التالي :

فرض العدم $\beta_0 = \beta_1 = \beta_2$: H_0 في مقابل الفرض البديل : ليست كل قيم المعالم مساوية للصفر .

ويكافئ هدا الاختبار اختبار F :

$$F_0 = \frac{ESS/P}{RSS/(n-p-1)} \sim F_{p,\,(n-p-1)}$$

ESS: مجموع مربعات الانحدار لنموذج الانحدار من الدرجة الثانية ز

RSS:مجموع مربعات البواقي للنموذج .

P: عدد المتغيرات المستقلة (X^2,X) .

n : عدد المشاهدات .