Logistic Regression الانحدار اللوجستي

التنبؤ: Prediction: هو تقدير القيمة المستقبلية لمتغير بناء على معرفة قيم متغير آخر، ويفيدنا في:

* تحديد العلاقة لشكل العلاقة بين المتغيرين رياضيا وبيانيا (رسم خط الانحدار)
* توضيح اتجاه العلاقة بين المتغيرين.
* التنبؤ بقيمة أحد المتغيرين بدلالة المتغير الآخر.

الانحدار: Regression: هو أسلوب يمكن بواسطته تقدير قيمة أحد المتغيرين لمعلومية قيمة المتغير الآخر عن طريق معادلة الانحدار، وله أنواع:

الانحدار الخطي البسيط: فكلمة بسيط تعني أن المتغير التابع Y يعتمد على متغير مستقل واحد وهو X وكلمة خطي تعني أن العلاقة بين المتغيرين X، Y علاقة خطية.

الانحدار الخطي المتعدد: إذا كان المتغير Y يعتمد على أكثر من متغير مستقل فنحن أمام الانحدار الخطي المتعدد.

الانحدار غير الخطي: إذا كانت العلاقة بين المتغير Y والمتغيرات المستقلة غير خطية مثل أن تكون من الدرجة الثانية، أو أسية أو لوغاريتمية.

الانحدار الخطي البسيط: بعد تمثيل الأزواج المرتبة بالمستوى نحصل على شكل الانتشار فإذا أظهر الشكل الانتشاري للبيانات أن هناك علاقة خطية بين المتغيرين نقوم بتقدير خط الانحدار Y على X بواسطة العلاقة:

$$\hat{Y} = a + bx$$

حيث a: ثابت الانحدار أو الجزء المقطوع من Y، b: ميل خط الانحدار

وتحسب القيمتان a و b من العلاقتان التاليتين:

$$b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} \qquad a = \frac{\sum Y - b\sum X}{n}$$

لإيجاد قيمة مقدرة جديدة Y_h نعوض بقيمة معلومة للمتغير المستقل ولتكن X_h في معادلة تقدير خط الانحدار $\hat{Y} = a + bx$

ملاحظات هامة:

* ميل الخط يمثل كمية التغير في Y المناظرة للتغير في X بمقدار وحدة واحدة

*إشارة معامل الانحدار تدل على نوع الارتباط (طردي أو عكسي)

مثال: لدراسة علاقة الاستهلاك المحلي (Y) بالانتاج (X) لمادة الفوسفاط (مليون قنطار) خلال عشرة سنوات يوضحها الجدول التالي:

Y	6	8	9	8	7	6	5	6	5	5
X	10	13	15	14	9	7	6	6	5	5

أوجد معادلة الانحدار الخطي البسيط وتوقع قيمة الاستهلاك عندما يصل الانتاج إلى 16 مليون قنطار

X	Y	XY	X^2
10	6	60	100
13	8	104	169
15	9	135	225
14	8	112	196
9	7	63	81
7	6	42	49
6	5	30	36
6	6	36	36
5	5	25	25
5	5	25	25
Σ	90	65	632	948

942

$$b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \frac{6320 - 90.65}{9420 - 90^2} = \frac{6320 - 5850}{9420 - 8100}$$

$$\boxed{b = 0,36}$$

$$a = \frac{\sum Y - b\sum X}{n} = \frac{65 - (0,36)(90)}{10}$$

$$\boxed{a = 3,26}$$

وتكون معادلة خط الانحدار $\hat{y} = 3.26 + 0,36x$

لتوقع قيمة الاستهلاك للعائلة عندما يصل الانتاج إلى 16 مليون قنطار، يكون لدينا $x_h = 16$ بالتعويض في معادلة خط الانحدار يكون لدينا

$$\hat{y}_h = a + bx_h$$
$$= 3,26 + 0,36(16) = 9,02$$

الانحدار اللوجستي: الانحدار اللوجستي هو نموذج يستعمل للتنبؤ باحتمالية وقوع حدث ما وذلك بملاءمة البيانات على منحنى لوجستي،

يستخدم الانحدار اللوجستي عدة متغيرات منبئة (مستقلة Predictors) والتي يمكن أن تكون تسبيبية أو فئوية أو ترتيبية أو اسمية مقابل متغير متنبأ به (تابع) واحد. مثال احتمال إصابة فرد ما من عدم إصابته بالربو يمكن التنبؤ بها من خلال معرفة معلومات حول الجانب الوراثي والمناعي للفرد وتقديره درجة التحسس لديه .... إن الانحدار اللوجستي يقوم بتحليل العلاقة بينما مجموعة من المتغيرات المستقلة ومتغير تابع يكون تصنيفيا Categorical أو ترتيبيا ordinal من خلال تقدير احتمال وقوع حدث ما عن طريق تركيب منحنى لوجستي

مجال استخدام الانحدار اللوجستي:

يوظف الانحدار اللوجستي في ما في الدراسات التي:

* تهدف إلى التنبؤ وشرح قيم متغيرات كيفية، حيث يعتبر أنسب أسلوب احصائي في الدراسات التي تهدف إلى بنائها نماذج تنبؤية ويكون المتغير التابع منها متغيرا كيفيا.
* كذلك يستخدم الانحدار اللوجستي بهدف التنبؤ باحتمالات

حدث ما يكون في وضعية خطر، حيث يشاع استخدام الانحدار اللوجستي في المجال الصحي والطبي قصد التنبؤ بإمكانية الإصابة بمرض ما من عدمه..

* كما يستخدم في التسويق لحساب ميل المستهلك إلى شراء منتج ما أو امتناعه عن الشراء.

شروط تطبيق الانحدار اللوجستي:

* أن يكون المتغير التابع متغيرا اسميا ثنائيا لصنف الانحدار الثنائي الحدين، أو تصنيفيا متعدد الحدود أو رتبيا بالنسبة للانحدار اللوجستي متعدد الحدود، أو رتبيا بالنسبة للانحدار اللوجستي الترتيبي.
* أن يحتوي النموذج على أكثر من متغير مستقل يكون في المستوى التصنيفي أو الترتيبي أو الفئوي أو النسبي.
* أن يحتمل المتغير التابع احتمالا واحدا لكل فرد بمعنى لا يكون الفرد في فئتين في نفس الوقت.
* عدم وجود التعددية الخطية multicollinearity بمعنى عدم وجود ارتباطات قوية بين المتغيرات المستقلة

أنواع الانحدار اللوجستي:

1- الانحدار اللوجستي الترتيبي: يتميز بتفسير أثر المتغيرات المبنية (المستقلة) باختلاف مستويات قياسها على الاستجابات الترتيبية بمعنى أن يكون المتغير التابع متغيرا ترتيبيا.

2- الانحدار اللوجستي ثنائي الحدين أيضا بالانحدار اللوجستي الثنائي ويستخدم في تفسير أثر المتغيرات المفسرة على الاستجابات الثنائية بمعنى تفسير قدرة مجموعة من المتغيرات المستقلة (المبنية) ذات المستويات المختلفة على التنبؤ بمتغير تابع ثنائي التفرع (ذكر، أنثى / مريض، صحيح)

لذلك فالقيم لا ماذا تكون موجودة (إيجابية) وتأخذ القيمة 1 أو غير موجودة (سلبية) وتأخذ القيمة (0) وتأتي معادلة الانحدار اللوجستي الثنائي الحدين كما يلي:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

Y = المتغير التابع (الثنائي)

b_0 = يمثل الحد الثابت في معادلة الانحدار اللوجستي

$b_n X_n$ عينة اللوجيت للمتغيرات المستقلة

3- الانحدار اللوجستي المتعدد الحدود: يتم استخدامه في حالة كون المتغير التابع يتكون من أكثر من فئتين تصنيفيتين أو اسميتين.

$B_0 + B_1 X \longrightarrow Age(X)$

$Logit(X) = g(X) = B_0 + B_1 X$

يستخدم الانحدار اللوجيستي للتنبؤ بمتغير ثنائي (يأخذ القيم 0، 1) 0 يعني الفشل، 1 يعني النجاح.
المتغير التابع يأخذ قيمتين فقط
مشكلة استخدام الانحدار الخطي البسيط للبيانات الثنائية

---

* ليس هناك شيء يضمن أن قيم المتغير التابع لن تزيد عن 1 ولن تقل عن 0
* إن غالبية النقاط لن تقع على خط الانحدار نظرا لطبيعة درجات المتغير التابع.

هذه المعادلة لا تصلح للتنبؤ $\hat{y} = aX + b$ واحيانا

بالمتغير التابع، فنحن نحتاج إلى معادلة تعطي نتيجتين القيم التي تخرجها المعادلة قيم موجبة ومحصورة بين الصفر والواحد

لذلك سنستخدم الدالة الأسية التي تعطي مخرجات موجبة

$e = 2,718282$

هذه العلاقة تعطي الشرط 1 ← قيما موجبة

$\hat{y} = e^{aX+b}$ ← الشرط 2

$$\hat{y} = \frac{e^{aX+b}}{1 + e^{aX+b}}$$

$$\hat{y} = \frac{1}{1+e^{-(ax+b)}}$$

$$0 \leq \hat{y} \leq 1$$

$$\frac{\hat{y}}{1-\hat{y}} \Rightarrow +\infty \text{ و } 0$$

الدالة محصورة بين

نحن نريدها بين $-\infty, +\infty$

$$\log\left(\frac{y}{1-y}\right) = \ln\left(\frac{y}{1-y}\right)$$

$$\ln\left(\frac{y}{1-y}\right) = ax + b$$

y احتمال وقوع الحدث، $1-y$ احتمال عدم وقوع الحدث

(الحدث المتمم)

$\left(\frac{y}{1-y}\right)$ Odds Ration تسمى نسبة الترجيح

a تمثل المقدار الذي يزيد في اللوغاريتم نسبة الترجيح عندما يزيد (ينقص) المتغير المستقل درجة واحدة.

* إذا كانت النسبة $\left(\frac{y}{1-y}\right)$ أصغر من الواحد أي أن احتمالية وقوع الحدث أقل من احتمالية عدم وقوعه ولوغاريتم نسبة الترجيح يكون سالبا

* والعكس بالعكس الملاحظة السابقة

* إذا كانت نسبة الترجيح = 1 أي أن احتمالية وقوع الحدث تساوي احتمالية عدم وقوعه ولوغاريتم نسبة الترجيح يكون مساويا للصفر

وفقا للمعادلة إذا ازاد المتغير X درجة واحدة فإن نسبة الترجيح تزداد بمقدار $\exp(\beta)$ وليس المتغير التابع الذي يزيد بهذه النسبة.

- إذا كانت β سالبة فهذا يعني علاقة عكسية بين المتغيرين فإن قيمة $\exp(\beta)$ ستكون أقل من (1)
- إذا كانت قيمة β موجبة فهي علاقة طردية بين المتغيرين فإن $\exp(\beta)$ تكون أكبر من (1)

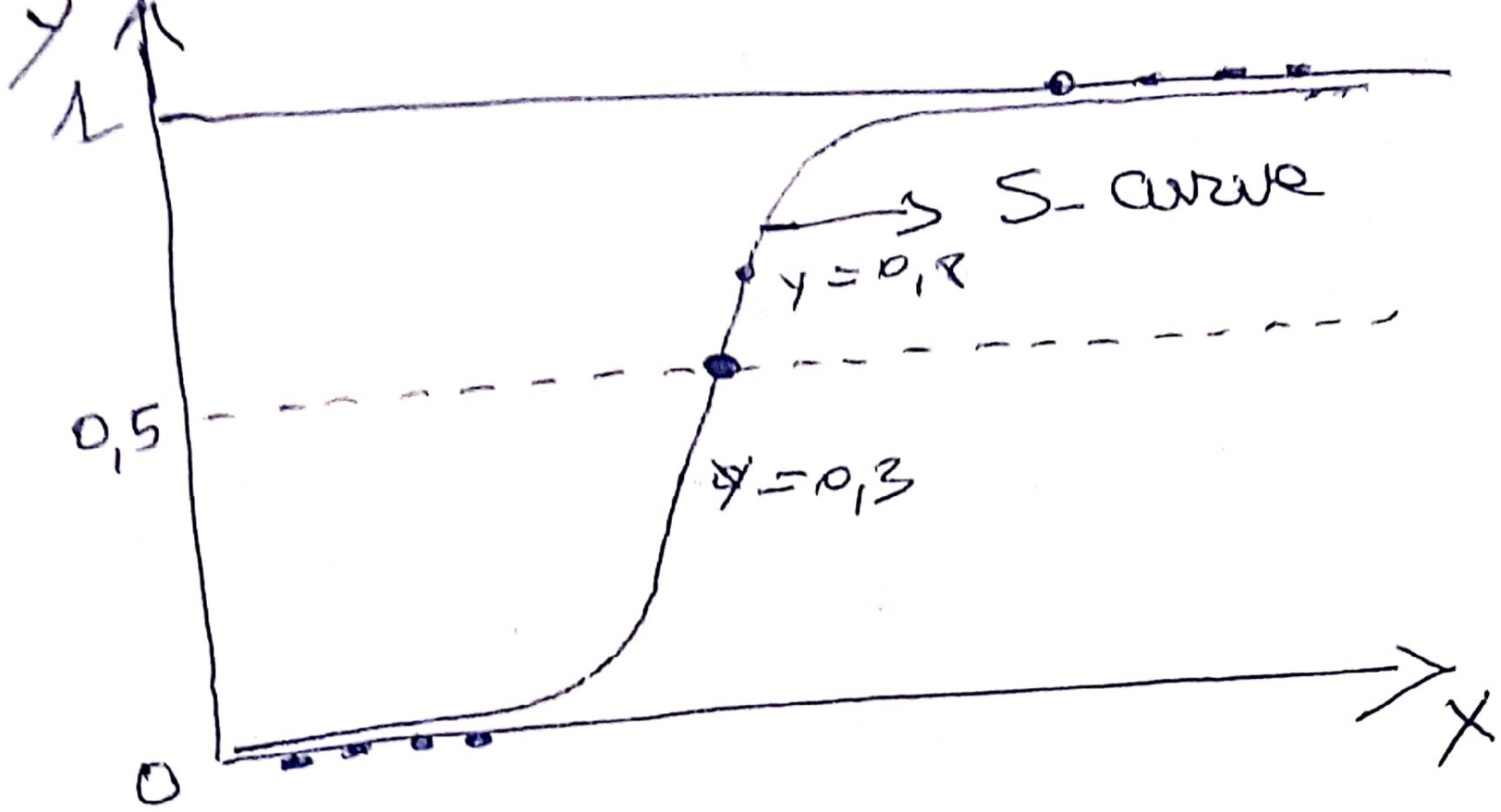
تعطى مجموعة البيانات للنجاح/الرسوب في امتحان لـ 5 طلبة في الجدول أدناه - إذا كانت المعادلة التالية:

$$\text{Log}_e(\text{Odds}) = -64 + 2 \times \text{hours}$$

Hours studies	Results (1 = Pass, 0 = Fail)
29	0
15	0
33	1
28	1
39	1

1- ماهو احتمال النجاح لطالب درس لمدة 33 ساعة

2- كم على الأقل يجب أن يدرس الطالب لضمان احتمال/فرصة بمعدل احتمالية يزيد عن 95٪؟



: دالة

$$S(X) = \frac{1}{1+e^{-x}}$$

$$\log(Odds) = z = -64 + 2 \times hours$$

$$z = -64 + 2 \times 33$$

$$= -64 + 66 = 2$$

$$P = \frac{1}{1+e^{-2}}$$

$$P = \frac{1}{1+e^{-2}} = 0,88$$

هذا يعني أن الطالب الذي يدرس لمدة 33 ساعة فإن هناك احتمال معدله 88% أن هذا الطالب سينجح في الامتحان

2-

$$P = \frac{1}{1+e^{-z}} = 0,95$$

$$\Rightarrow 0,95 \times (1+e^{-z}) = 1$$

$$0,95e^{-z} = 1 - 0,95$$

$$e^{-z} = \frac{0,05}{0,95} = 0,0526$$

$$\ln(e^x) = X$$

$$\ln(e^{-z}) = \ln(0,0526)$$

$$-z = -2,94 \Rightarrow \boxed{z = 2,94}$$