Machine learning with R
الانحدار اللوجستي
Logistic Regression

- يستخدم الانحدار اللوجستي عندما يكون المتغير التابع y متغيرا ثنائيا يأخذ قيمتين فقط يرمز للأولى وهي وقوع الحدث بالرمز 1 وذلك باحتمال قدره (P) بينما يرمز للثانية وهي عدم وقوع الحدث بالرمز 0 وذلك باحتمال يساوي(1-p) ، فيما لا يضع قيود ا على أنواع المتغيارت المستقلة Xi والتي يمكن لها أن تكون متصلة أو فئوية أو خليط من الاثنين كما أنه لا يشترط اعتدالية توزيعها.

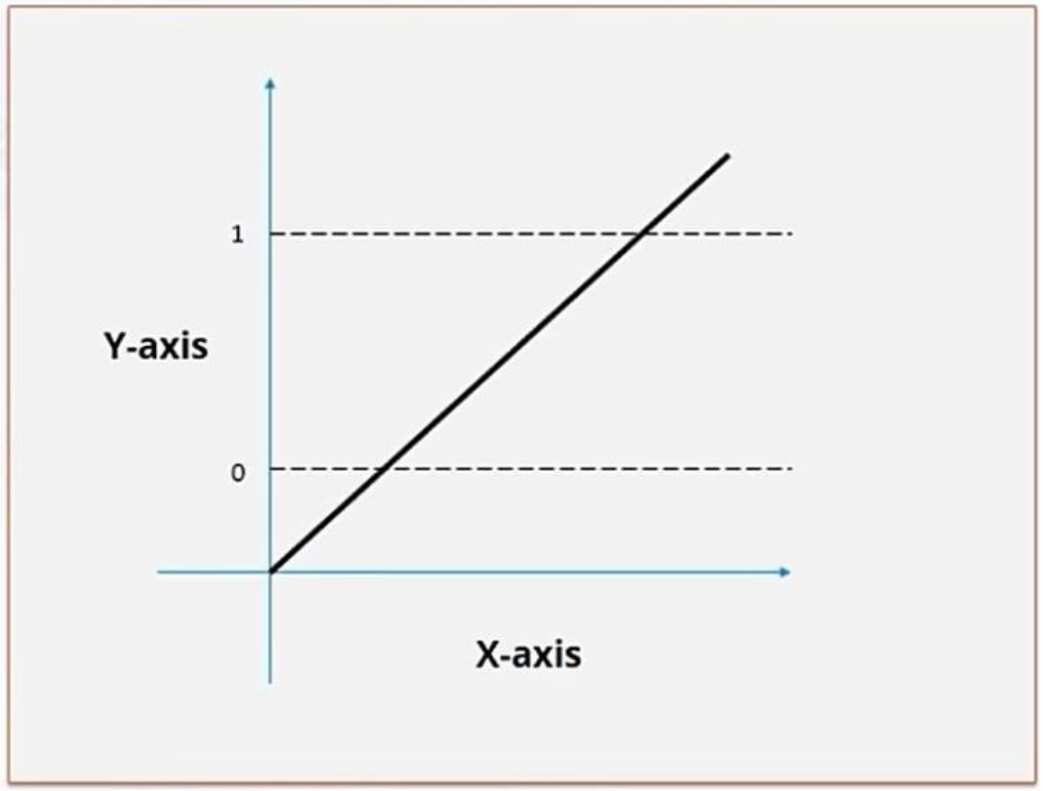
- وحيث أن معادلة الانحدار الخطي البسيط تكون على الشكل:
-
$$Y|X = \beta_0 + \beta_1 X + e$$
-
- حيث يعني الرمز Y/X : المتغير التابع بشرط حدوث المتغير المستقل .
- وبافتارض أن الخطأ العشوائي e يتبع التوزيع الطبيعي بمتوسط 0 وانحارف معياري مقداره $\sigma_{Y|X}$ أي أن $e \sim N(0, \sigma_{Y|X})$ ، فإن المتغير التابع Y يتبع التوزيع الطبيعي بمتوسط $\mu_{Y|X}$ وانحراف معياري $\sigma_{Y|X}$ أي أن $Y \sim N(\mu_{Y|X}, \sigma_{Y|X})$ وذلك لكل قيمة من قيم المتغير المستقل .
- ونظرا لأن $E(e) = 0$ ، لذا فإن القيمة المتوقعة للمتغير التابع Y عند قيمة معينة للمتغير المستقل X تكون على الشكل التالي:
-
$$E(Y/X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

إلا أنه ولعدم إمكانية تطبيق الانحدار الخطي البسيط في حالة المتغير التابع ثنائي القيمة (0,1) نتيجة لما يلي:

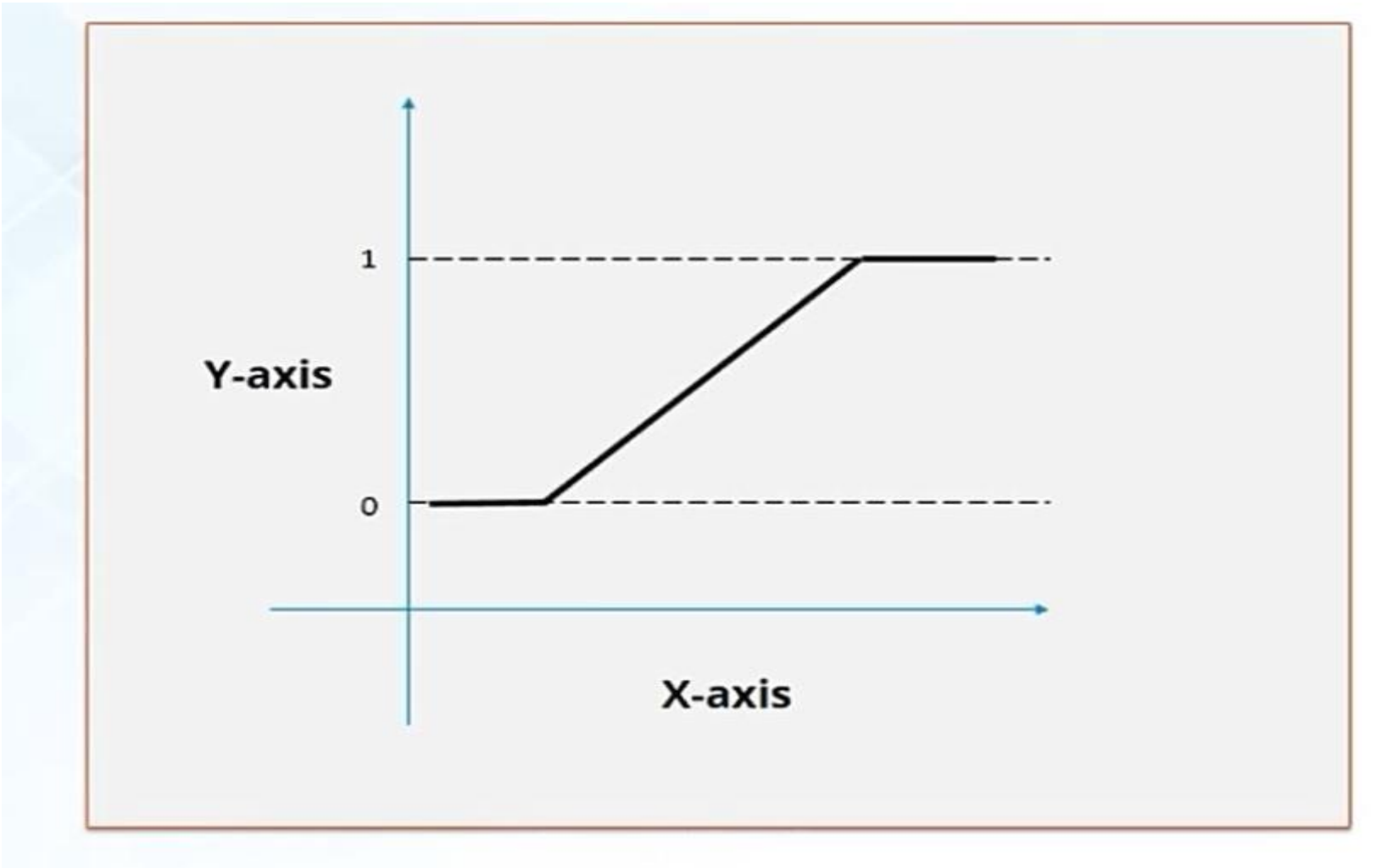
- أن تباين المتغير التابع (Y) يتغير بتغير قيم المتغير المستقل (X).
- أن تباين الخطأ لا يتوزع وفق التوزيع الطبيعي.
- أن القيم المقدرة لا يمكن تفسيرها بوصفها احتمالات ذلك لأن قيمها لا تتراوح بين (0،1).

لماذا لا يتم استخدام الانحدار الخطي

- بما ان قيم المتغير التابع تكون بين 0 و 1 فيجب ان يكون المنحنى مقطوع نحو تلك القيمتين



- لا يمكن استخدام صيغة واحدة لحل هذه المشكلة لهذا نلجأ للانحدار اللوجستي

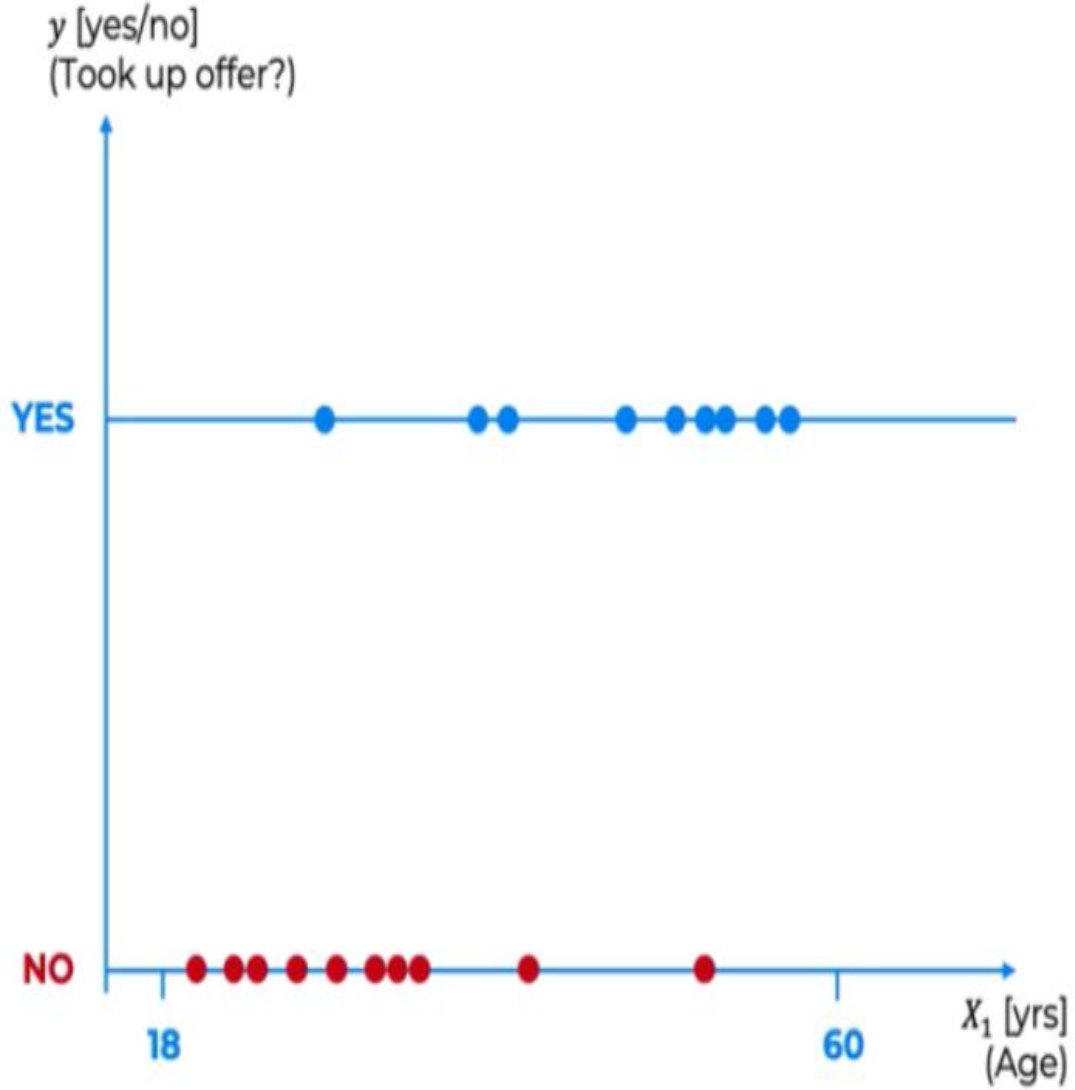


Logistic regression: predict a categorical dependent variable from a number of independent variables.

~

Will purchase health insurance: Yes / No

Age



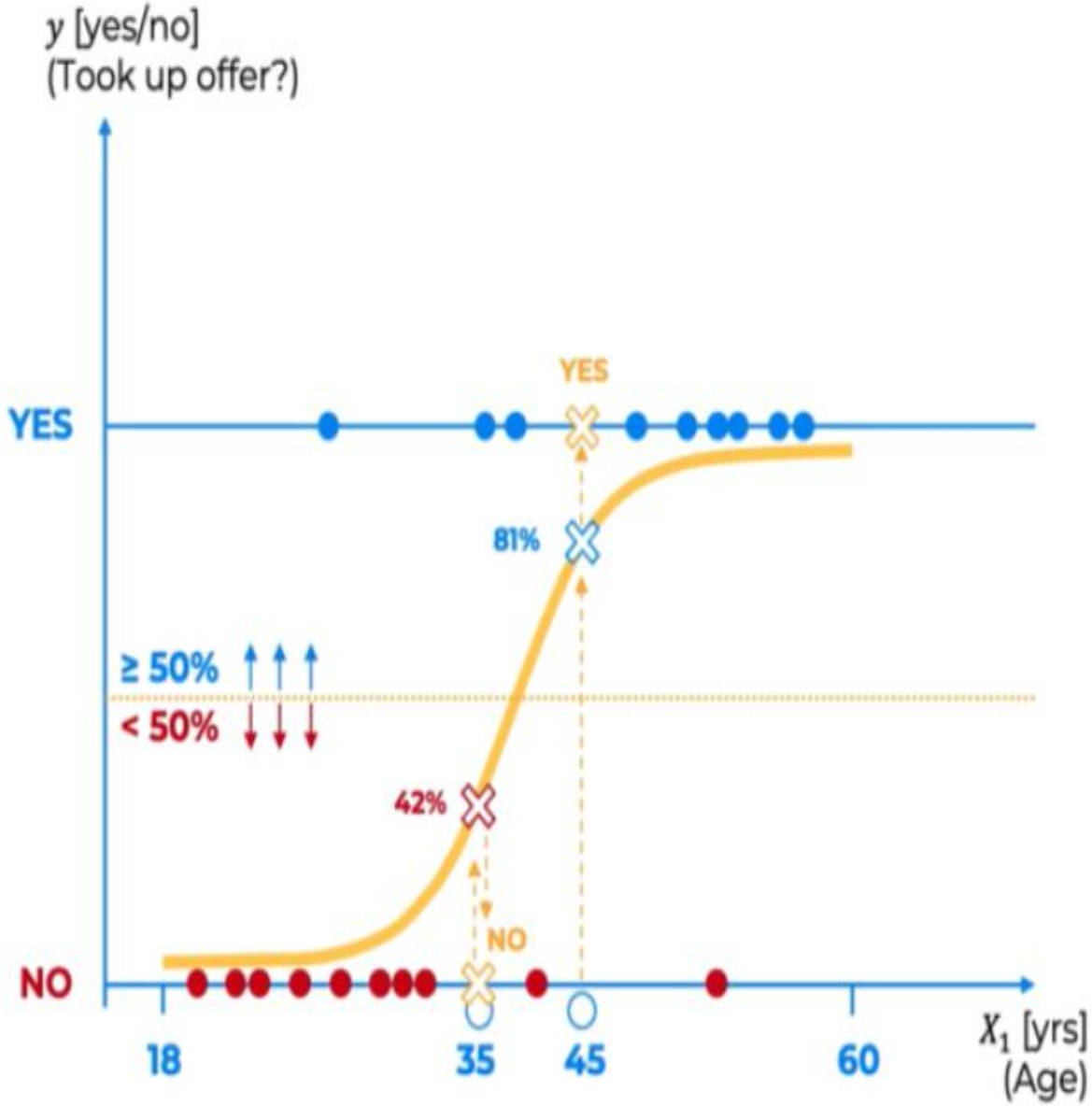
Logistic regression: predict a categorical dependent variable from a number of independent variables.

~

Will purchase health insurance: Yes / No

Age

$$\ln \frac{p}{1-p} = b_0 + b_1 X_1$$



$\sim$

Will purchase health insurance: Yes / No

Age

Income

Level of Education

Family or Single

$$\ln \frac{p}{1-p} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

- يتم استخدام نموذج اللوجيت الذي يعالج المشاكل السابقة حيث يمكن كتابته في حالة وجود متغير مستقل واحد كالتالي:

$$Log_e\left(\frac{P}{1}\right) = \hat{\beta}_0 + \hat{\beta}_1 X \qquad (3)$$

- وبصورة أخرى:

$$\therefore \left(\frac{P}{1-p}\right) = e^{\hat{\beta}_0+\hat{\beta}_1 X} \quad \rightarrow \quad P = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}}$$

حيث:

P هو احتمال وقوع الحدث محل الاهتمام أي احتمال النجاح.

(1-P) هو احتمال وقوع الحدث ليس محل الاهتمام أي احتمال الفشل.

نسبة الترجيح للحدث محل الاهتمام $\left(\frac{P}{1-p}\right)$

اللوغاريتم الطبيعي Log e= 2.718281

وبذلك يمكن كتابة معادلة الانحدار في حالة وجود عدد kمن المتغي ارت المستقلة على الشكل:

$$Log_e\left(\frac{P}{1-p}\right) = \hat{\beta}_0 + \sum_{i=1}^{k}\hat{\beta}_i X_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$$

وتسمى المعادلة (2(نموذج الانحدار اللوجستي.

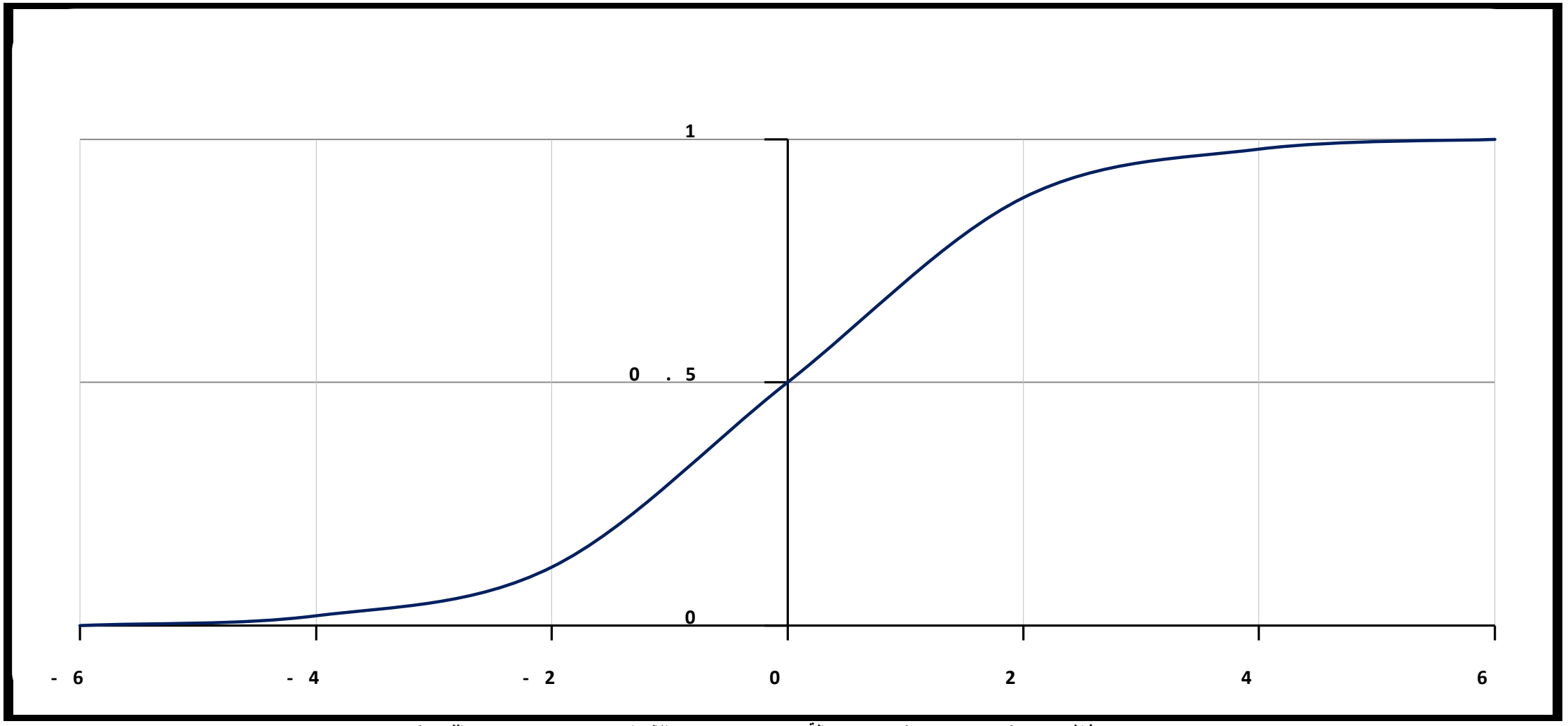
$$\therefore \boldsymbol{Logit} = \boldsymbol{Log_e(odds)} = \boldsymbol{Log_e}\left(\frac{P}{1-p}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_1 +$$

$$\hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$$

هذا وسيتم اختصار استخدام الرمز Log وذلك للتعبير عن اللوغاريتم الطبيعي

.

والدالة اللوجستية كما يتضح في الشكل (0) هي دالة متصلة يتراوح مداها بين (0,1) حيث تقترب من الصفر كلما اقترب الطرف الأيمن للدالة من - ∞ كما تقترب من الواحد كلما اقترب هذا الطرف من+ ∞



شكل (6) : دالة الانحدار اللوجستي

تعريف

- الانحدار اللوجستي هو نموذج إحصائي ينتمي لنماذج الانحدار الخطي ناتج من نمذجة متغير ثنائي الحد بدلالة مجموعة من المتغيرات العشوائية المتوقعة، رقمية كانت أو فئوية. يستخدم الانحدار اللوجستي للتنبؤ باحتمالية وقوع حدث ما بمعرفة إضافية لقيم متغيرات يمكن أن تكون مفسرة أو مرتبطة بهذا الحدث.

- بتعريف الدالة اللوجستية، و هي قيمه احتمالية تأخذ قيم بين صفر وواحد.

$$f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}}$$

تقدير دالة الانحدار اللوجستي

$$\text{Logit}(Y) = \log\left(\frac{Y}{1-Y}\right) = \frac{e^{\beta_\circ + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_\circ + \beta_1 x_1 + \beta_2 x_2}}$$

Here,

$\beta_\circ$ = Constant Coefficient

β_1 = Coefficient of x1

β_2 = Coefficient of x2

x_1 = Independent variable

x_2 = Independent variable

e = Euler's Number

P(Y) = Probability that Y equals 1

تطبيقات نموذج الانحدار اللوجستي

1. في مجالات الطب والإحصاء الحيوي : مثلا احتمال حدوث نوبة قلبية عند شخص ما خلال فترة زمنية معينة حسب المعرفة القبلية ببعض المعلومات الديمغرافية)عمره أو جنسه مثلا(أو الطبية)مؤشراته البدنية أو الصحية أو الغذائية(أو الوبائية)سلوكياته كالتدخين مثلا.

2.الصيدلة: في تقدير رد الفعل والمقارنة بين نجاعة الأدوية.

•3.التأمينات: لفرز وتقسيم مجموعات العملاء حسب المخاطر ومدى قابلية جذبهم لمنتجات تأمين معينة.

•4.المجال البنكي: خصوصا في تنقيط العملاء أثناء دراسة ملفات القروض.

•5.التسويق: حساب توقعات ميل المستهلك إلى شراء منتج ما أو امتناعه عن الشراء.

•6.في سبر الآراء والعلوم السياسية: مثلا للتنبؤ بقرار التصويت في الانتخابات اعتمادا على تنميط قبلي للمصوتين (مستواهم الاجتماعي، توجهاتهم السياسية، مستواهم التعليمي)

مثال: mtcars

- Henderson and Velleman (1981) comment in a footnote to Table 1: ‘Hocking [original transcriber]'s noncrucial coding of the Mazda's rotary engine as a straight six-cylinder engine and the Porsche's flat engine as a V engine, as well as the inclusion of the diesel Mercedes 240D, have been retained to enable direct comparisons to be made with previous analyses.’

- Source
- Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391–411.

- Motor Trend Car Road Tests
- The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

- A data frame with 32 observations on 11 (numeric) variables.

- [, 1] **mpg** **Miles/(US) gallon**
- **[, 2] cyl Number of cylinders**
- **[, 3] disp Displacement (cu.in.)**
- **[, 4] hp Gross horsepower**
- **[, 5] drat Rear axle ratio**
- **[, 6] wt Weight (1000 lbs)**
- **[, 7] qsec 1/4 mile time**
- **[, 8] vs Engine (0 = V-shaped, 1 = straight)**
- **[, 9] am Transmission (0 = automatic, 1 = manual)**
- **[,10] gear Number of forward gears**
- **[,11] carb Number of carburetors**

Let's take a sample dataset in R, which is called **mtcars.** Our aim is to predict whether a car will have a V-engine or a Straight engine based on our inputs.

Key

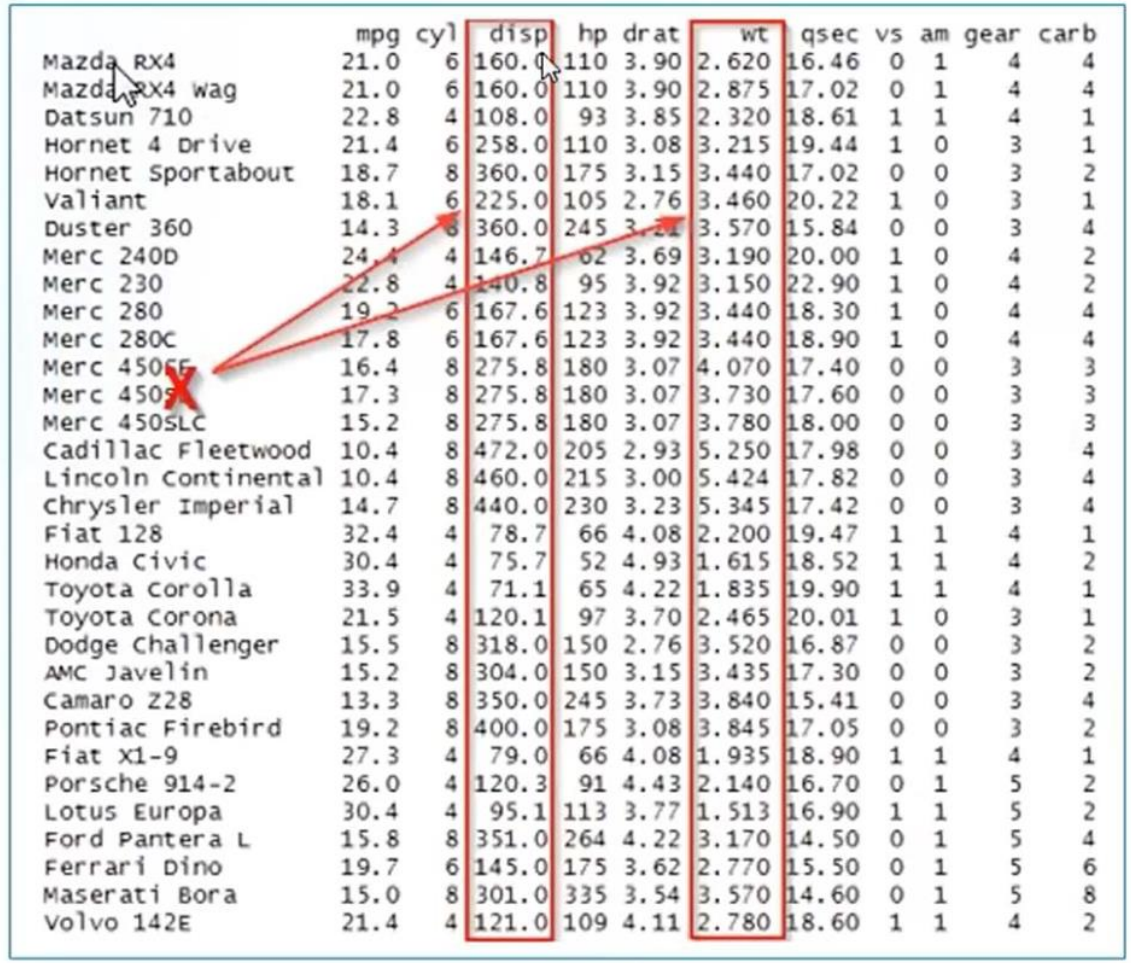
- **Mpg** - Miles/US Gallon
- **Cyl** – Number of cylinders
- **Disp** – Displacement of car
- **Hp** – Gross horsepower
- **Drat** – Rear axle ratio
- **Wt** – Weight (lb/1000)
- **Qsec** – 1/4 mile time
- **Vs** – V Engine
- **Am** – Transmission Type
- **Gear** – Number of forward gears
- **Carb** - Number of carburetors

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

For now, let's take disp and wt as our primary independent variables. Why? We'll be discussing it in our next section.

Key

- **Mpg** - Miles/US Gallon
- **Cyl** – Number of cylinders
- **Disp** – Displacement of car
- **Hp** – Gross horsepower
- **Drat** – Rear axle ratio
- **Wt** – Weight (lb/1000)
- **Qsec** – 1/4 mile time
- **Vs** – V Engine
- **Am** – Transmission Type
- **Gear** – Number of forward gears
- **Carb** - Number of carburetors

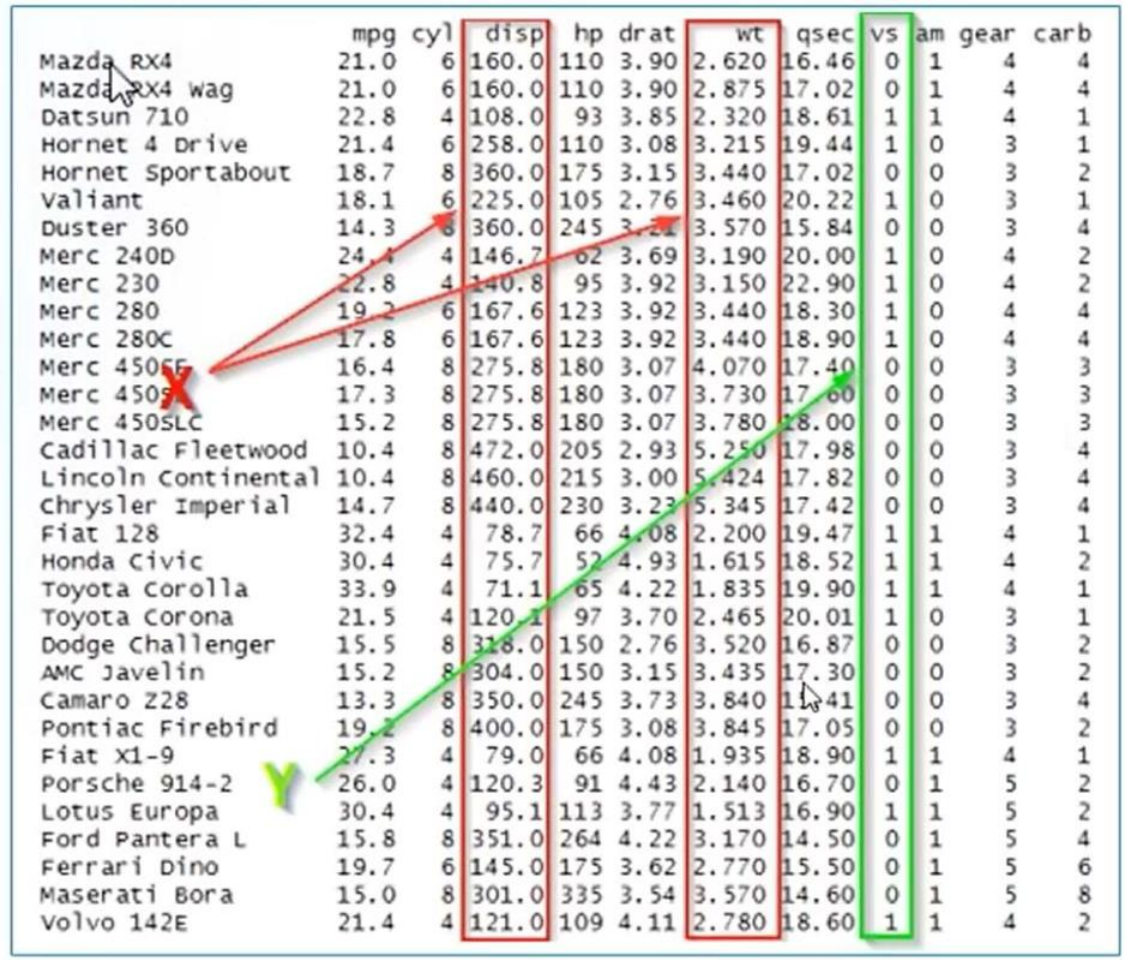


	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

Since our aim is to know which engine will fit, the engine will either be V – type or not, i.e either 1 or 0. Therefore, our dependent variable is Y.

Key

- **Mpg** - Miles/US Gallon
- **Cyl** – Number of cylinders
- **Disp** – Displacement of car
- **Hp** – Gross horsepower
- **Drat** – Rear axle ratio
- **Wt** – Weight (lb/1000)
- **Qsec** – 1/4 mile time
- **Vs** – V Engine
- **Am** – Transmission Type
- **Gear** – Number of forward gears
- **Carb** - Number of carburetors



	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

Before creating the model, we divide our dataset into training and testing.

Training Dataset

80 %

20%

Testing Dataset

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

X

Y

بعد القيام بعملية التقدير باستخدام طريقة الامكانية العظمى Maximum Likelihood Estimation(MLE)

نحصل على المعلمات التالية:

β_0

β_1

β_2

```
Call:
glm(formula = vs ~ wt + disp, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7083  -0.3406   0.4944   0.6033   1.6998

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.83010    2.67220   0.685   0.4934
wt           1.09428    1.59890   0.684   0.4937
disp        -0.02529    0.01499  -1.687   0.0915 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 29.065  on 20  degrees of freedom
Residual deviance: 17.095  on 18  degrees of freedom
AIC: 23.095

Number of Fisher Scoring iterations: 5
```

glm(VS ~ wt + disp, data = train, family = "binomial")

summary(model)

1. Building a Generalized Linear Model (GLM):

model <- glm(**vs ~ wt + disp, data = train, family = "binomial")**

This line fits a generalized linear model using the glm() function.

Vs ~. : This formula specifies that the dependent variable is VS, and we want to model it using all other variables in the train data frame as independent variables (predictors). The . acts as a shorthand for including all terms.

data = train: This argument specifies the data frame (train) from which the model will be trained.

family = "binomial": This argument indicates that we're dealing with a binomial regression model, suitable for predicting binary outcomes like admission (admit/not admit).

2. Summarizing the Model:

summary(model)

This line calls the summary() function on the fitted model object (model).

The summary() function will print detailed information about the model, including:

Coefficients: Estimates (values) and standard errors for each predictor variable in the model.

Significance levels (p-values): To assess if the association between a predictor and the outcome is statistically significant.

Deviance and null deviance: Measures of goodness-of-fit for the model.

Interpretation:

The summary will provide insights into the strength and significance of the relationships between each predictor and Y.

Important Notes:

Make sure there are no highly correlated predictor variables in your data, as this can lead to multicollinearity issues in the model.

Consider checking for influential outliers that might affect the model's fit.

Let's take a value from the test dataset

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2

$\beta_0 = 1.83010$

$\beta_1 = 1.09428$

$\beta_2 = -0.02529$

$e = 2.7183$

$X1 = 120.3$

$X2 = 2.140$

Substituting Values

Logit (Y)

$$= \frac{0.9849}{1.9849} = 0.4962$$

Probability of 'vs' being '1' = 0.4962

We will assume the **threshold** to be 0.5

Example 2

glm(diabet ~ ., data = train, family = "binomial")

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.785848   1.458604  -7.395 1.42e-13 ***
npreg         0.097192   0.073604   1.320   0.1867
glu           0.042215   0.007157   5.898 3.68e-09 ***
bp           -0.006346   0.015657  -0.405   0.6853
skin          0.006775   0.024235   0.280   0.7798
bmi           0.090286   0.035178   2.567   0.0103 *
ped           0.764312   0.517721   1.476   0.1399
age           0.043896   0.023445   1.872   0.0612 .
---
```

*** - 99.9% confident

** - 99% confident

* - 95% confident

. - 90% confident

Null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean)

Residual deviance shows how well the response variable is predicted with inclusion of independent variables.

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.97282    1.39227  -7.881 3.24e-15 ***
npreg         0.10033    0.07325   1.370 0.170817
glu           0.04199    0.00710   5.914 3.34e-09 ***
bmi           0.09177    0.02613   3.513 0.000444 ***
ped           0.76849    0.51710   1.486 0.137237
age           0.04079    0.02204   1.850 0.064249 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 318.85  on 247  degrees of freedom
Residual deviance: 203.65  on 242  degrees of freedom
AIC: 215.65

Number of Fisher Scoring iterations: 5
```

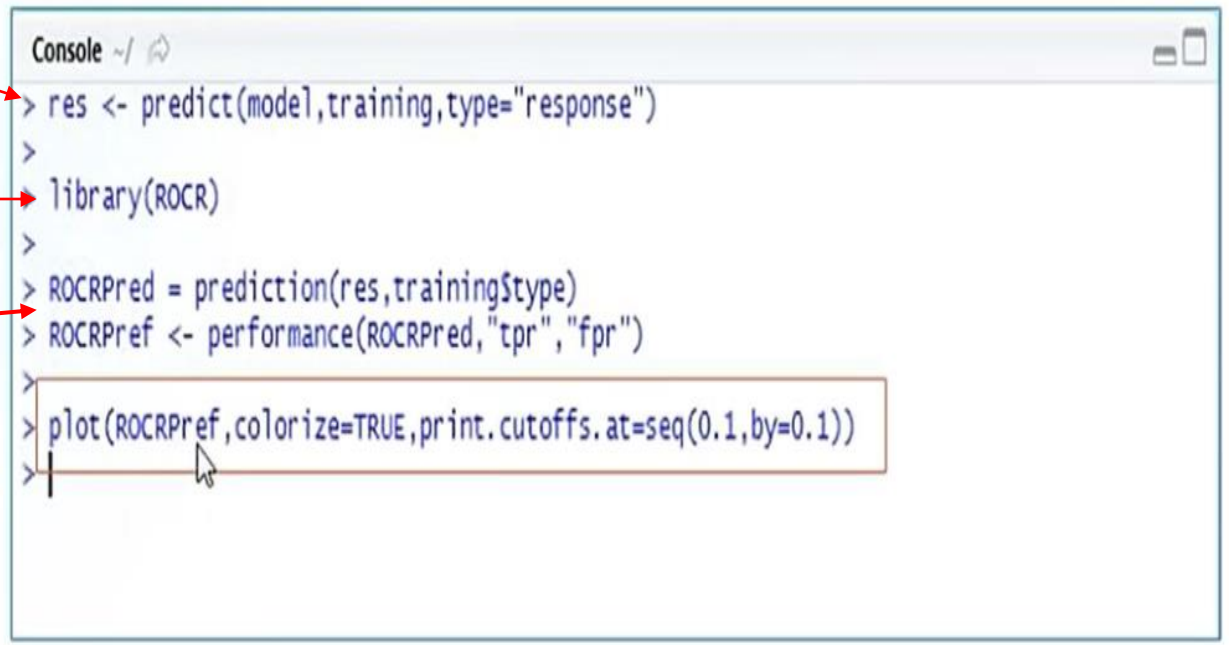
Let us take a sample for our significant fields and see whether our patient is diabetic or not based on the model that we created.

Store the predicted values for training dataset in 'res' variable

Import the library for the ROCR package

Define the 'ROCRPred' and and 'ROCRPerf' variables

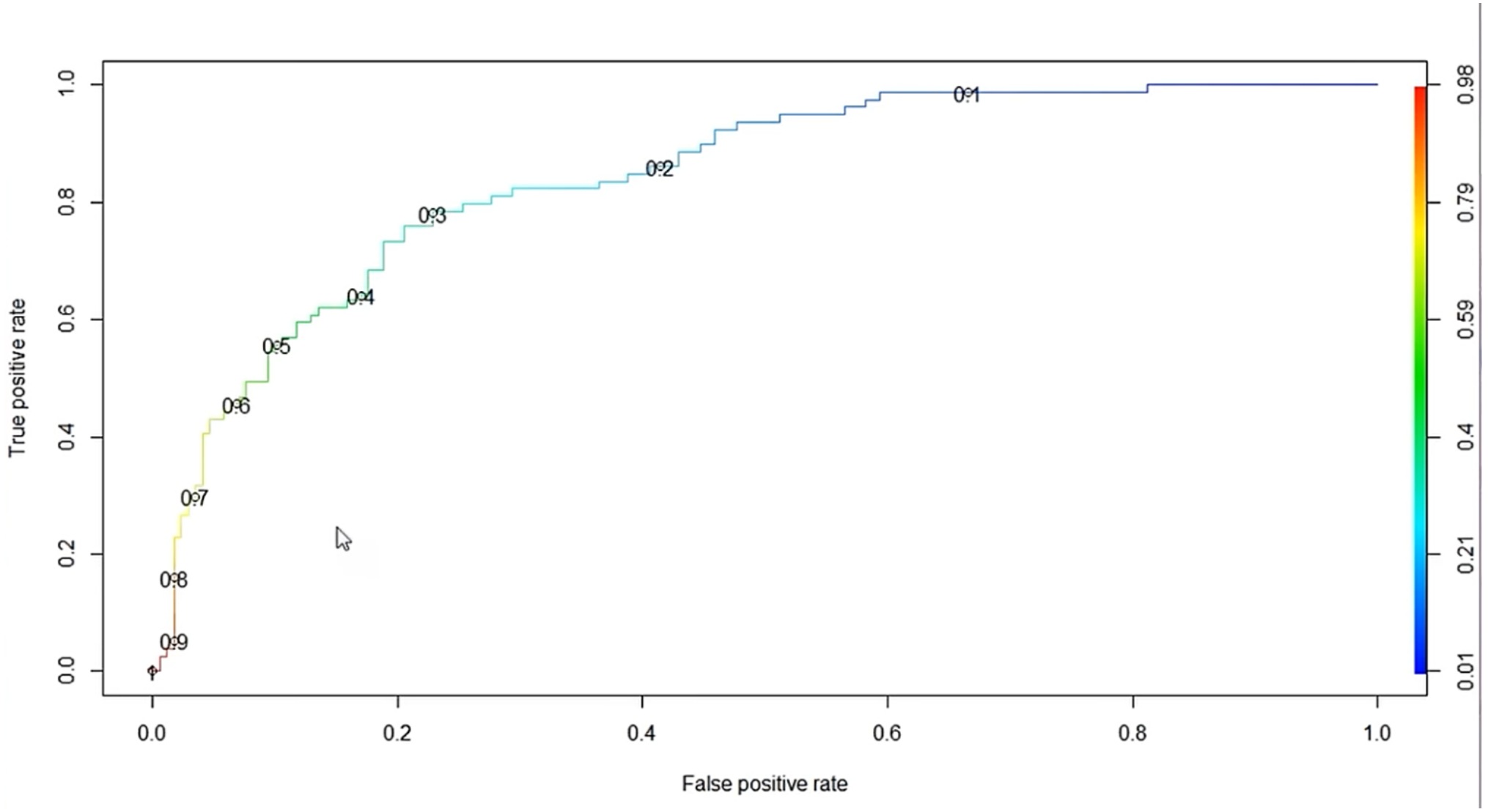
Plot the graph!



Confusion Matrix:

A confusion matrix is a table that visually summarizes the performance of a classification model. It shows how many data points were correctly and incorrectly classified for each category.

ROC CURVE


True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.98
0.79
0.59
0.4
0.21
0.01
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

ROC Curve:

An ROC curve is a graphical tool used to evaluate the performance of binary classification models like your logistic regression model for predicting admission status.

It plots the True Positive Rate (TPR) on the y-axis and the False Positive Rate (FPR) on the x-axis.

Interpretation:

TPR (Recall): The proportion of actual “yes" cases that the model correctly predicted as "Admit." A higher TPR indicates the model is good at capturing true admits.

FPR: The proportion of actual "Not " cases that the model incorrectly predicted as “yes." A lower FPR indicates the model is good at avoiding false positives.

Ideal ROC Curve:

A perfect classifier would have an ROC curve that goes straight from the bottom left corner (0 FPR, 0 TPR) to the top left corner (0 FPR, 1 TPR) and then horizontally to the top right corner (1 FPR, 1 TPR). This represents a model that perfectly distinguishes between the two classes.

Thank you