Logistic Regression

Logistic Regression

- Logistic Regression **(LR)** is a statistical measure to link the Independent variables (X) with a Bernoulli output (Y) [0/1]

- LR is an extension of the Linear Regression

- Models the probability of an event occurring (Y) based on the Independent variables (x_1, x_2,… x_n) **that are numeric or categorical** in nature

- Estimate the probability that an event happens for any given combination of independent variables

- Classify observations in a particular category

Examples

- Will a potential customer get a bank loan - Get / Not get
- Allergic to a particular drug – Allergic / Not allergic
- Student will get admission in college – Will get / will not get

- Goal of LR is to estimate the probability **p**.
- This estimate of p is represented as $\hat{p}$
- The values of $\hat{p}$ lie between 0 and 1
- **Logit** is the name of function that links the X-variables with the probabilities (Y)

- Logit is defined as the natural log of the odds ratio

  logit(p) = log(p / 1-p)

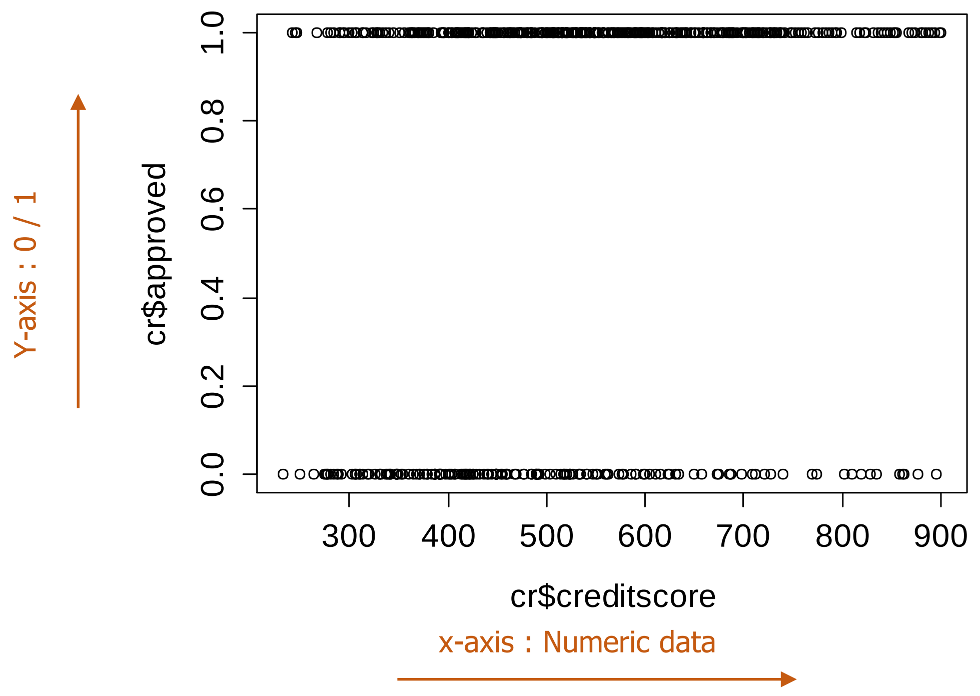
- In this equation, the probabilities lie along the X-axis

- But, probabilities need to be along the Y-axis – **Inverse logit(p)**

- So, Inverse of the above function gives the **Sigmoid function**

  $\mathbf{logit^{-1}(x) = (e^{x} / 1 + e^{-x})}$

  x = linear combination of independent variables in the coefficients



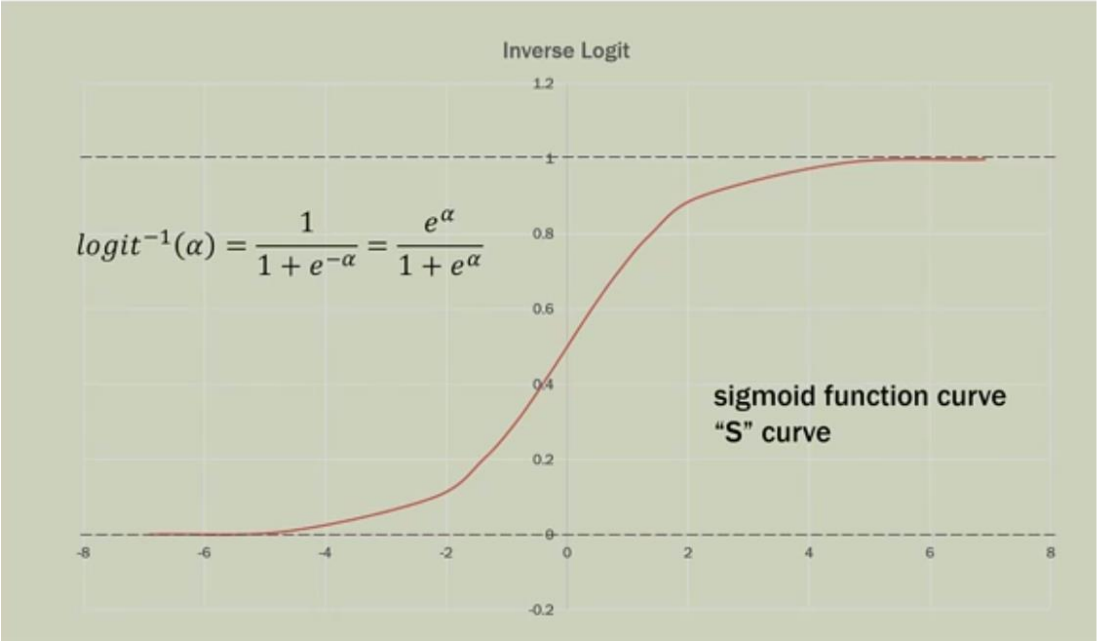


Image courtesy : Statistics 101

Develop an estimated regression equation

- that fits the Inverse Logit model
- Use the coefficients returned by the equation, plot them to get the S-graph

A note on Logistic Regression Coefficients

- Regression coefficients of LR are calculated using the Maximum Likelihood Estimation (MLE)
- MLE ≡ Least Squares of Linear Regression

MLE is trying to find out the optimal value for the Mean or Standard Deviation for a given distribution of a dataset

$$\text{logit}(p) = \log(p / 1-p) = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots\ldots\ldots\ldots \quad (1)$$

Taking antilog on both sides in (1)

$$p/1-p = e^{a + bx}$$

Solving for p using algebra, we get

$$\hat{p} = \frac{e^{a+bx}}{1 + e^{a+bx}} \qquad \hat{p} = \frac{e^{a+b_1x_1 + b_2x_2 + \ldots + b_nx_n}}{1 + e^{a+b_1x_1 + b_2x_2 + \ldots + b_nx_n}}$$

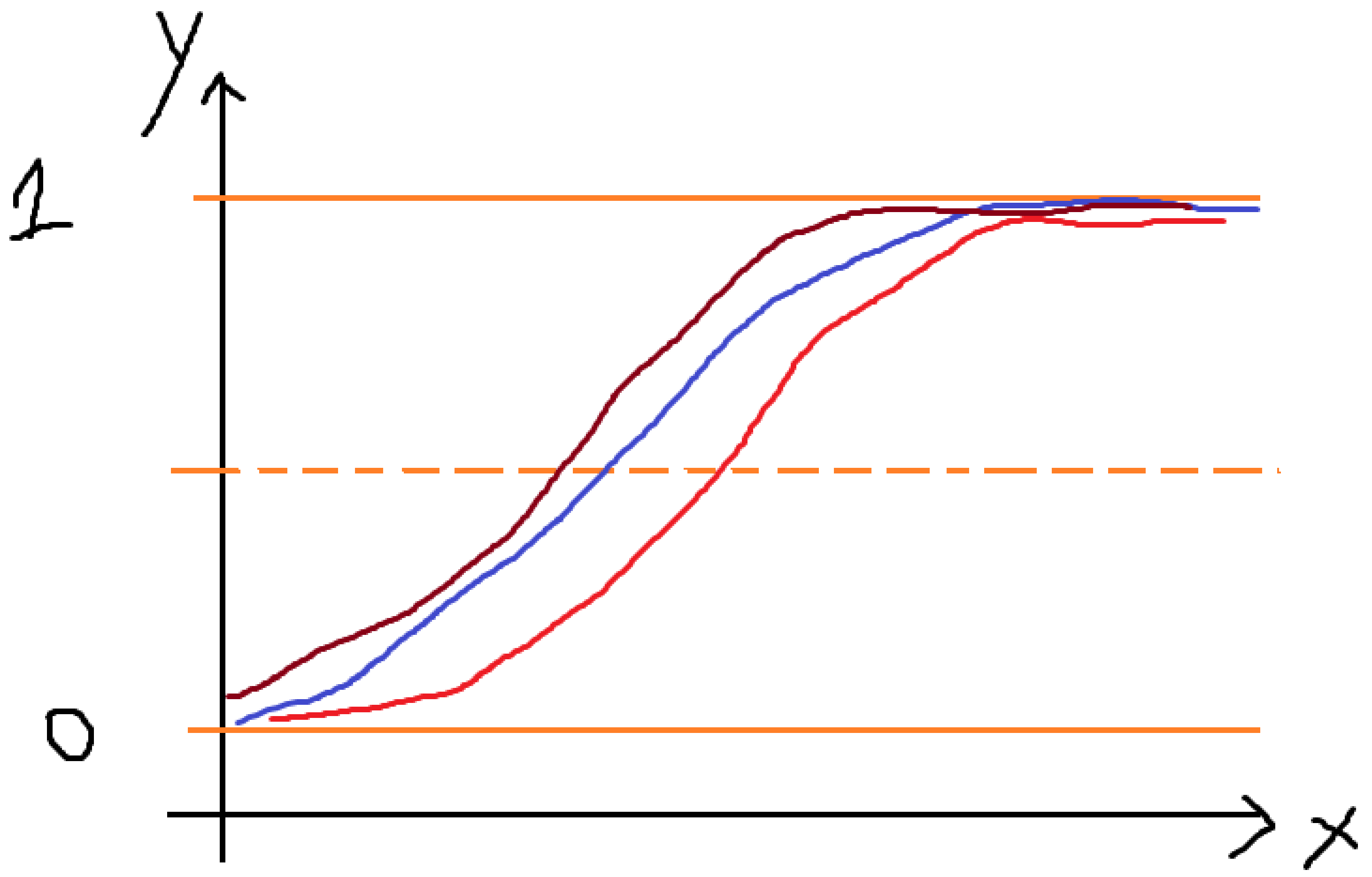
Estimated Logistic Regression Equation

p	logit(p)
0.5	0
0.6	+ve
0.2	-ve

where
p/1-p → odds ratio
$x_1, x_{2..}$ → independent variables
(RHS) → link function to determine a non-linear relation in a linear way
b → coefficients.

Eg:
if b = 1.12624, then exp(b) = 3.084
(the odds ratio)
1 unit increase in x multiplies the odds of event happening (Y) by 3.084

In LogisticRegression, we select the best fit curve



Odds, Odds Ratio

- Logistic Regression results are interpreted using the concept of odds

Odds

The ratio of the probability of success and failure

Assume probability of an event occurring = 0.8 (success)

∴ probability of failure = 0.2 (1-0.8)

Odds of success

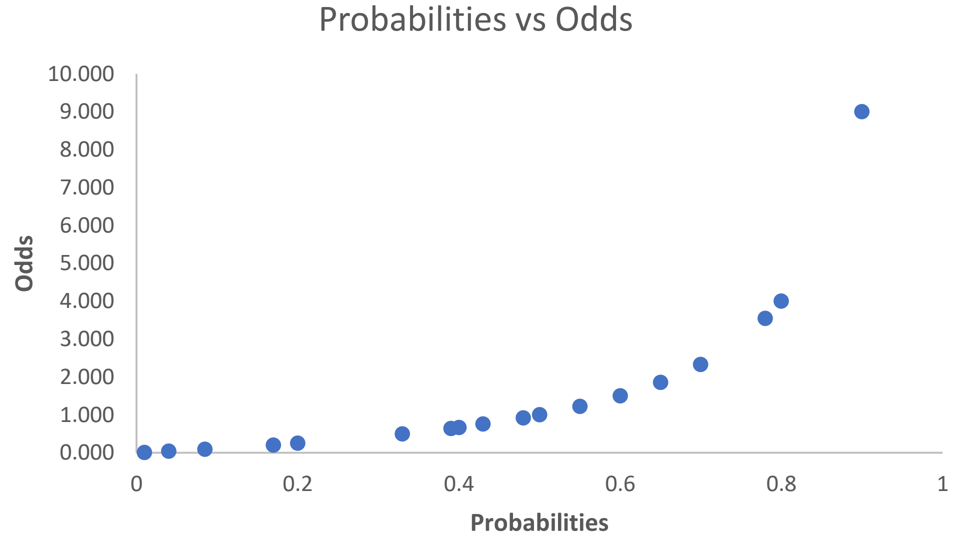
= ratio of (probability of success / probability of failure)

= 0.8/0.2

= 4 : 1

Odds can be

- In Favour
- Against



Odds increase as probability increases and vice versa

p	1-p	odds (p/1-p)	logodds
0.01	0.99	0.010	-4.59512
0.04	0.96	0.042	-3.17805
0.085	0.915	0.093	-2.37627
0.17	0.83	0.205	-1.58563
0.2	0.8	0.250	-1.38629
0.33	0.67	0.493	-0.70819
0.39	0.61	0.639	-0.44731
0.4	0.6	0.667	-0.40547
0.43	0.57	0.754	-0.28185
0.48	0.52	0.923	-0.08004
0.5	0.5	1.000	0
0.55	0.45	1.222	0.200671
0.6	0.4	1.500	0.405465
0.65	0.35	1.857	0.619039
0.7	0.3	2.333	0.847298
0.78	0.22	3.545	1.265666
0.8	0.2	4.000	1.386294
0.9	0.1	9.000	2.197225
0.99	0.01	99.000	4.59512

Why log odds ?

- It is usually difficult to model a variable which has restricted range, such as probability. This transformation is an attempt to get around the restricted range problem. It maps probability ranging between 0 and 1 to log odds ranging from negative infinity to positive infinity.

- Log of odds is one of the easiest to understand and interpret. This transformation is called logit transformation.

Feature Selection

Salient features / Independent variables in Logistic regression can be determined by the following methods:

- Summary of the model
- Recursive Feature Elimination
- Information value (IV)
- step(<model>) function using AIC score (in R)

How is Logistic regression different from Linear Regression ?

#	Linear Regression	Logistic Regression
1	Linear Regression makes a few **assumptions** on the data	Logistic Regression does not make these assumptions
2	Uses the general linear equation $y = a + \sum(b_i x_i) + \epsilon$ y -> continuous dependent variable – any value x_i -> continuous / binary variables	Uses the same basic Linear equation $y = e^{a+bx} / 1 + e^{a+bx}$ y -> continuous dependent variable – Dichotomous (0/1) x_i -> continuous / binary variables
3	Change in x = change in y	Change in x = change in odds of y
4	Uses LSE (Least Square Error)	Uses MLE (Maximum Likelihood Estimation)
5	Eg: BMI can predict Blood Pressure	Eg: BMI can predict the odds of being a diabetic

For a binary distribution (Logistic Regression), why can’t we use Linear Regression ?

- The linear regression model is based on an assumption that the outcome is continuous, with errors (e), which are normally distributed.
  If the outcome variable is binary this assumption is clearly violated.

- For a binary outcome the mean is the probability of a 1, or success. If we use linear regression to model a binary outcome it is quite possible to have a fitted regression that can give predicted values for some observations more than (0,1) range

Interpreting the Logistic Regression output

```
Call:
glm(formula = admit ~ ., family = binomial, data = training_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6915  -0.9117  -0.6167   1.1011   2.1731

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.916494   1.357630  -2.148   0.0317 *
gre          0.002092   0.001309   1.598   0.1101
gpa          0.773460   0.397327   1.947   0.0516 .
prestige    -0.670156   0.155945  -4.297 1.73e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 360.03  on 279  degrees of freedom
Residual deviance: 324.68  on 276  degrees of freedom
AIC: 332.68

Number of Fisher Scoring iterations: 3
```

*The coefficient for a variable (gre) says that, holding the other variables constant (gpa and prestige), what is the rate of change of odds of getting a **yes***

e.g: keeping gpa and prestige fixed, the odds of getting an admission in a college with the gre score alone is exp(0.002092) = 1.002094 = 0.209% (which is very less)
[abs(1-exp(0.002092))*100]

Null and Residual deviance

- **Null deviance**

How well the response variable is predicted by the model when only the intercept term is present

- **Residual deviance**

How well the response variable is predicted by the model when all the variables are included

- ND and RD are chi-square statistics with the *dof*
- In the example, the addition of 3 independent variables decreased the deviance from 360.03 to 324.68 (a reduction of 35.35) with a loss of 3 *dof*
- If Null deviance is small, the Null model explains the data well

- Likewise with Residual deviance

- **Fisher Scoring Iterations**

Number of iterations performed to get the best fit curve

Loss Function

- The loss function for Logistic Regression is called the **Log Loss / Cross Entropy**

- **Formula**
  - ***For a single training example***

    $$E(Loss) = -\{y\log(\hat{p}) + (1-y)\log(1-\hat{p})\}$$

  - ***For multiple training examples***

    $$E(Loss) = -\sum_{n=1}^{N}[y\log(\hat{p}) + (1-y)\log(1-\hat{p})]$$

  where

  y = actual class label (0 or 1)

  $\hat{p}$ = predicted y (probability values between 0 and 1)

  log = natural log

- In Logistic Regression, the output is 0 / 1
- Output (probabilities) are numbers between 0 and 1
- Hence, Logistic Regression Error cannot have a Gaussian Distribution

- **Incorrect prediction = Bigger cost**

$$E(Loss) = -\{y\log(\hat{p}) + (1-y)\log(1-\hat{p})\}$$

y (actual)	**$\hat{p}$ (pred)**	**log($\hat{p}$)**	**log(1-$\hat{p}$)**	**y*log($\hat{p}$) (a)**	**(1-y) log(1-$\hat{p}$) (b)**	**-(a+b)**
1	1.000	0.000	#NUM!	0.000	#NUM!	#NUM!
0	0.000	#NUM!	0.000	#NUM!	0.000	#NUM!
1	0.900	-0.105	-2.303	-0.105	0.000	0.105
1	0.500	-0.693	-0.693	-0.693	0.000	0.693
1	0.100	-2.303	-0.105	-2.303	0.000	2.303

Dummy variables

- Every independent factor variable is coded (also known as One-Hot encoding)
- Requires a Reference class value

- **Example:**
- Consider the following factor variables having the following values
- Text in red is the **"reference class"**
- For 'n' factor values, there will be n-1 dummy variables

Department	BusinessTravel	Gender
HR	**Frequently**	**Male**
R&D	Rarely	Female
Sales	None	
Admin		

- Codification of the factor variables will be as follows

Department

RD	Sales	Admin
1	0	0
0	1	0
0	0	1

BusinessTravel

Rarely	None
1	0
0	1

Gender

Female
1

Interpretation of Dummy variables

- Consider the "Titanic" dataset, where the factor variable **"SeatType"** has values:
  - **First**
  - **Second**
  - **Third**
- Reference class = "First"
- The Regression model (glm) outputs the following coefficients for the "SeatType":
  - Second = -1.270
  - Third = -2.241

This means that

The chances of survival of Second/Third class relative to the First class

exp(-1.270) = 0.2808 (odds)	exp(-2.241) = 0.1063 (odds)
• The odds of surviving in Second class is 0.2808 times the odds of surviving in the first class (other variables fixed) • 0.2808 - 1 = -0.7192 : The odds of surviving is 71.92% less for Third class passenger than for a First class passenger	• The odds of surviving in Third class is 0.1063 times the odds of surviving in the first class (other variables fixed) • 0.1063 - 1 = 0.8937 : The odds of surviving is 89.37% less for Third class passenger than for a First class passenger

Interpretation of Dummy variables

The chances of survival of Second and Third class

Coeff(Third) – Coeff(Second)
= -2.241 – (-1.270)
= -0.971

exp(-0.971) = 0.3787 (odds)

- The odds of surviving in Third class is 0.3787 times the odds of surviving in the Second class
- 0.3787 - 1 = -0.6212 : The odds of surviving is 62.12% less for Third class passenger than for a Second class passenger

Model evaluation

- Using the Confusion matrix, we can determine the goodness of a classification model using various measures