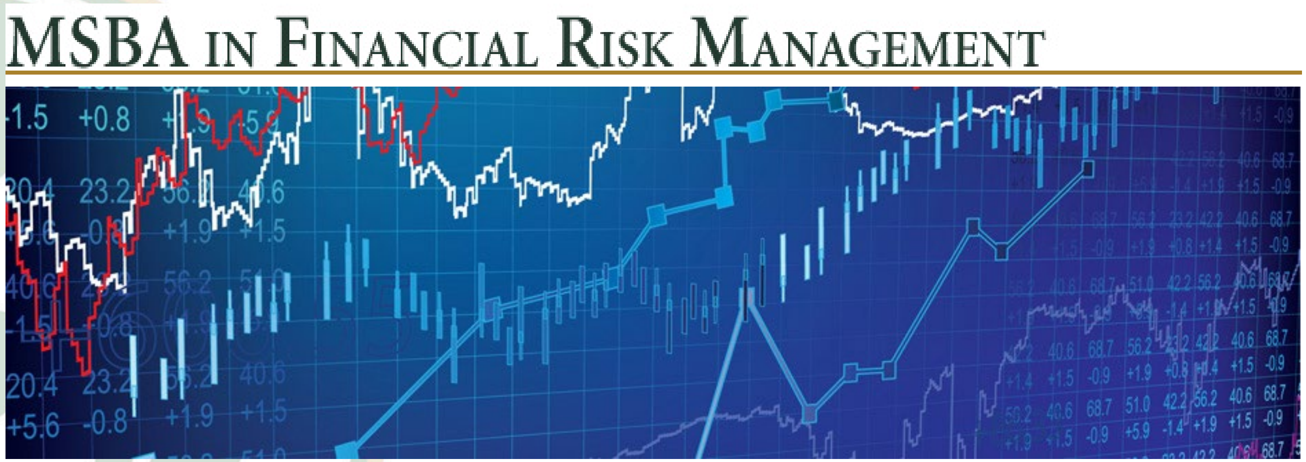
Financial Risk Management



Tianyang Wang CSU

Data Scientists in Finance

- Data Scientists Are A New Kind Of Statistician With Clout.
- If you want to work as a data scientist in finance, you will probably need most (if not all) of the following attributes:
- A first class degree in mathematics/statistics, computer science, physics, engineering or subject with significant mathematical content.
- An ability to program in multiple languages (both compiled and interpreted) such a C/C++, S (e.g. as implemented in R), Matlab, Python and/or Java.
- Good database skills (i.e. at least SQL programming) in any classical RDBMS (for example, MySQL, PostgreSQL, Oracle, SQL Server).
- An adeptness with handling time series data from Bloomberg, Reuters or any of the myriad financial data streams available.

Data Scientists in Finance

- There are also two very important characteristics of people doing data science jobs in finance which are less frequently discussed.
- - Firstly, you'll need to be able to communicate mathematical ideas well both verbally and visually to non-specialists.
- Secondly, you'll need to know how to harness their mathematical training to solve genuine commercial problems.

Data Scientists in Finance

- Alongside all this, you'll need a good understanding of optimization (underpinned by solid linear algebra and calculus learnt in school), of statistical inference, simulation, multivariate analysis and proper data visualization.
- If you possess such training, then understanding techniques such as support vector machines, neural networks, random forests and gradient boosting are merely a hop, skip and a jump away. I might just throw in NLP as well.
- With all this, your data science career will be underway. Good luck!

TOP 10 Machine Learning Algorithms

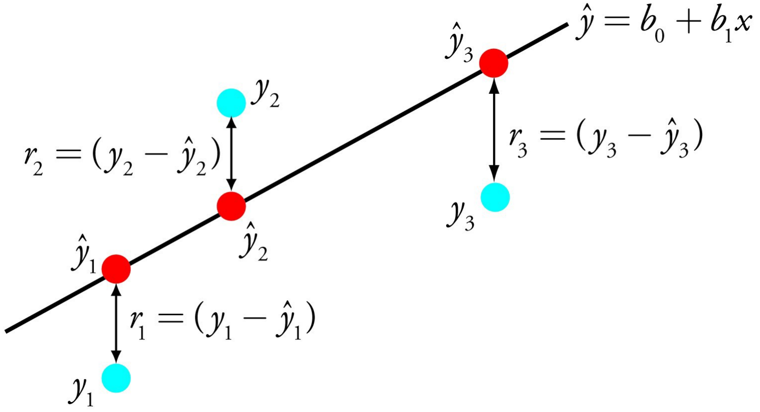
1—Linear Regression

1—Linear Regression

- Linear regression is perhaps one of the most well-known and well-understood algorithms in statistics and machine learning.
- Predictive modeling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explainability. We will borrow, reuse and steal algorithms from many different fields, including statistics and use them towards these ends.
- The representation of linear regression is an equation that describes a line that best fits the relationship between the input variables (x) and the output variables (y), by finding specific weightings for the input variables called coefficients (B).

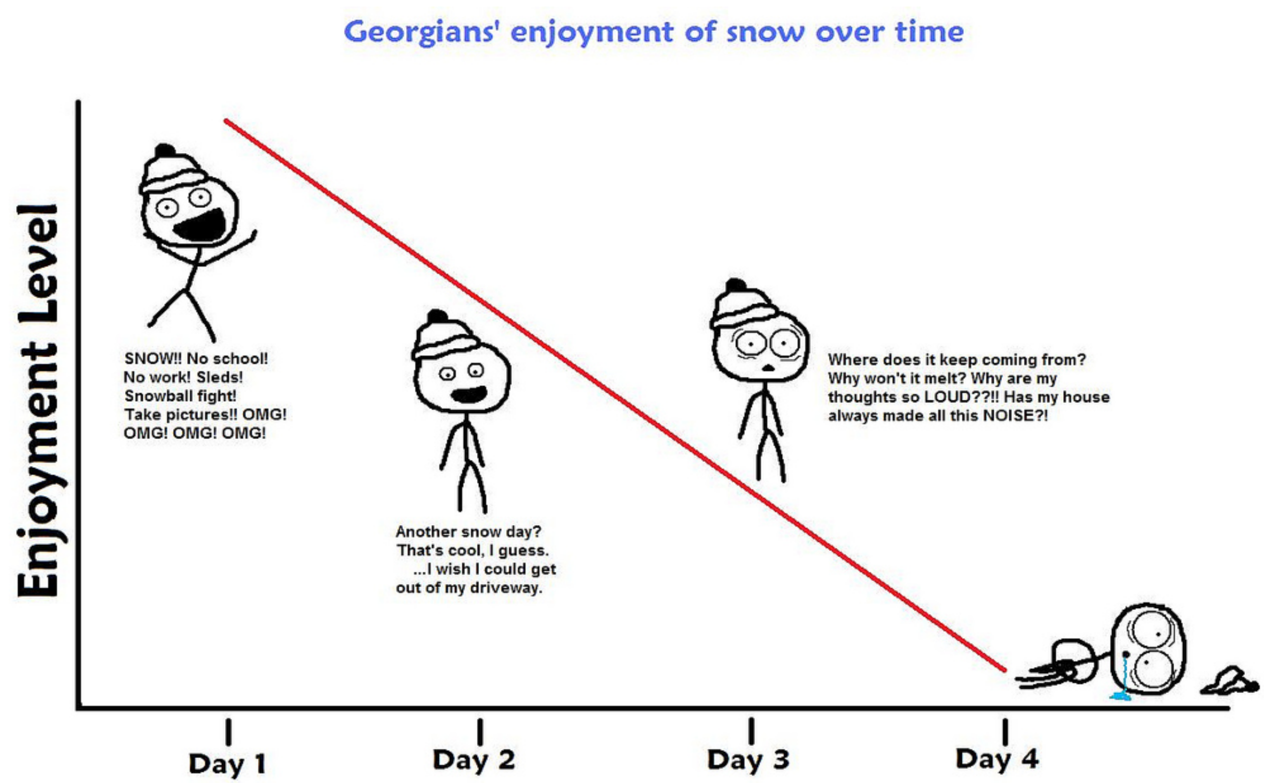
1—Linear Regression

- Different techniques can be used to learn the linear regression model from data, such as a linear algebra solution for ordinary least squares and gradient descent optimization.
- Linear regression has been around for more than 200 years and has been extensively studied. Some good rules of thumb when using this technique are to remove variables that are very similar (correlated) and to remove noise from your data, if possible. It is a fast and simple technique and good first algorithm to try.



Examining Relationship

- Two purposes of the linear regression line:
  - to **estimate the average** value of y at any specified value of x
  - to **predict the value** of y for an **individual**, given that individual's x value

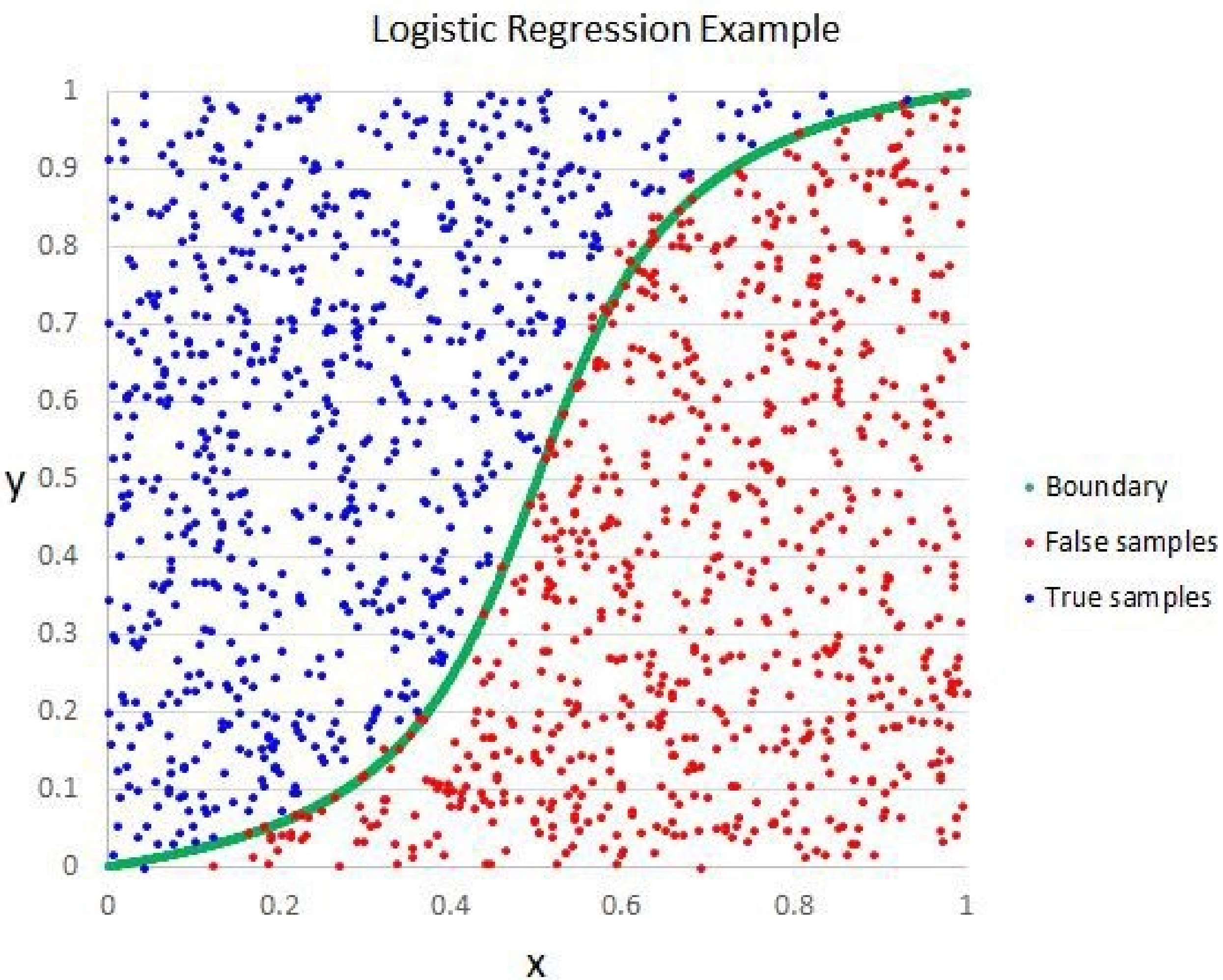


2—Logistic Regression

2—Logistic Regression

- Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values).
- Logistic regression is like linear regression in that the goal is to find the values for the coefficients that weight each input variable. Unlike linear regression, the prediction for the output is transformed using a non-linear function called the logistic function.
- The logistic function looks like a big S and will transform any value into the range 0 to 1. This is useful because we can apply a rule to the output of the logistic function to snap values to 0 and 1 (e.g. IF less than 0.5 then output 1) and predict a class value.

2—Logistic Regression



2—Logistic Regression

- Because of the way that the model is learned, the predictions made by logistic regression can also be used as the probability of a given data instance belonging to class 0 or class 1. This can be useful for problems where you need to give more rationale for a prediction.
- Like linear regression, logistic regression does work better when you remove attributes that are unrelated to the output variable as well as attributes that are very similar (correlated) to each other. It's a fast model to learn and effective on binary classification problems.

Logistic Regression

- Logistic regression is a variation of ordinary regression which is used when the dependent (response) variable is a dichotomous variable (i. e. it takes only two values, which usually represent the occurrence or non-occurrence of some outcome event, usually coded as 0 or 1) and the independent (input) variables are continuous, categorical, or both.
- For instance, in credit card company, the client default or not.

	A	B
1	Age	Default
2	20	0
3	23	1
4	24	1
5	25	0
6	25	0
7	26	1
8	26	0
9	28	1
10	28	0
11	29	0
12	30	1
13	30	1
14	30	1
15	30	1
16	30	0
17	30	1
18	32	1
19	32	1
20	33	1
21	33	1
22	34	0
23	34	0
24	34	1
25	34	1
26	34	1
27	35	1
28	35	0
29	36	1
30	36	1
31	36	1

The Linear Probability Model

- Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable: coded 0 (did not vote) or 1(did vote)
- In the OLS regression:

$$Y = \gamma + \varphi X + e \; ; \text{ where } Y = (0, 1)$$

  - The error terms are heteroskedastic
  - e is not normally distributed because Y takes on only two values
  - The predicted probabilities can be greater than 1 or less than 0

The Logistic Regression Model

Unlike ordinary linear regression, logistic regression does not assume that the relationship between the independent variables and the dependent variable is a linear one. Nor does it assume that the dependent variable or the error terms are distributed normally.

The "logit" model solves these problems:

$$\ln[p/(1-p)] = \alpha + \beta X + e$$

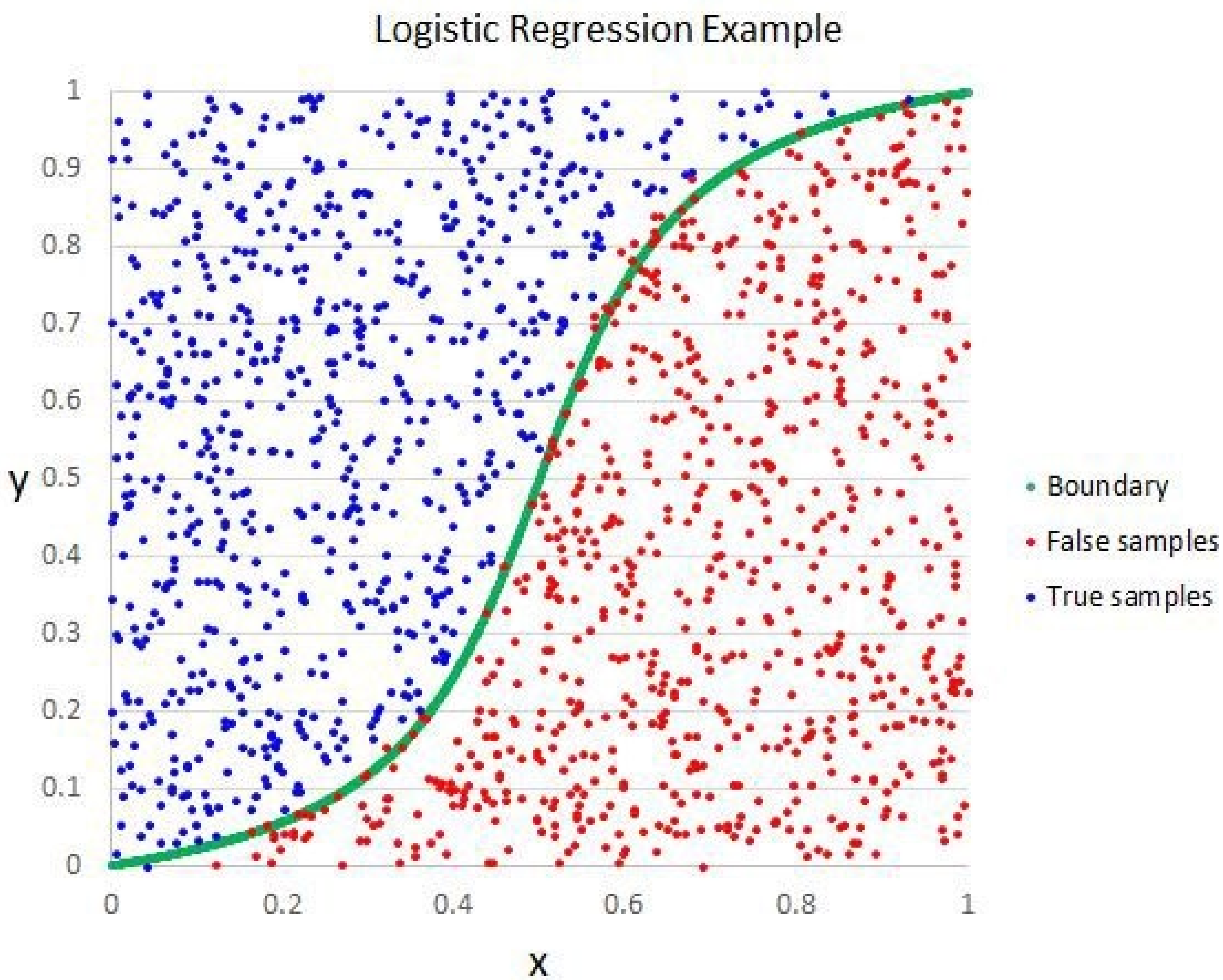
- p is the probability that the event Y occurs, p(Y=1)
- p/(1-p) is the "odds ratio"
- ln[p/(1-p)] is the log odds ratio, or "logit"

Logistic Regression

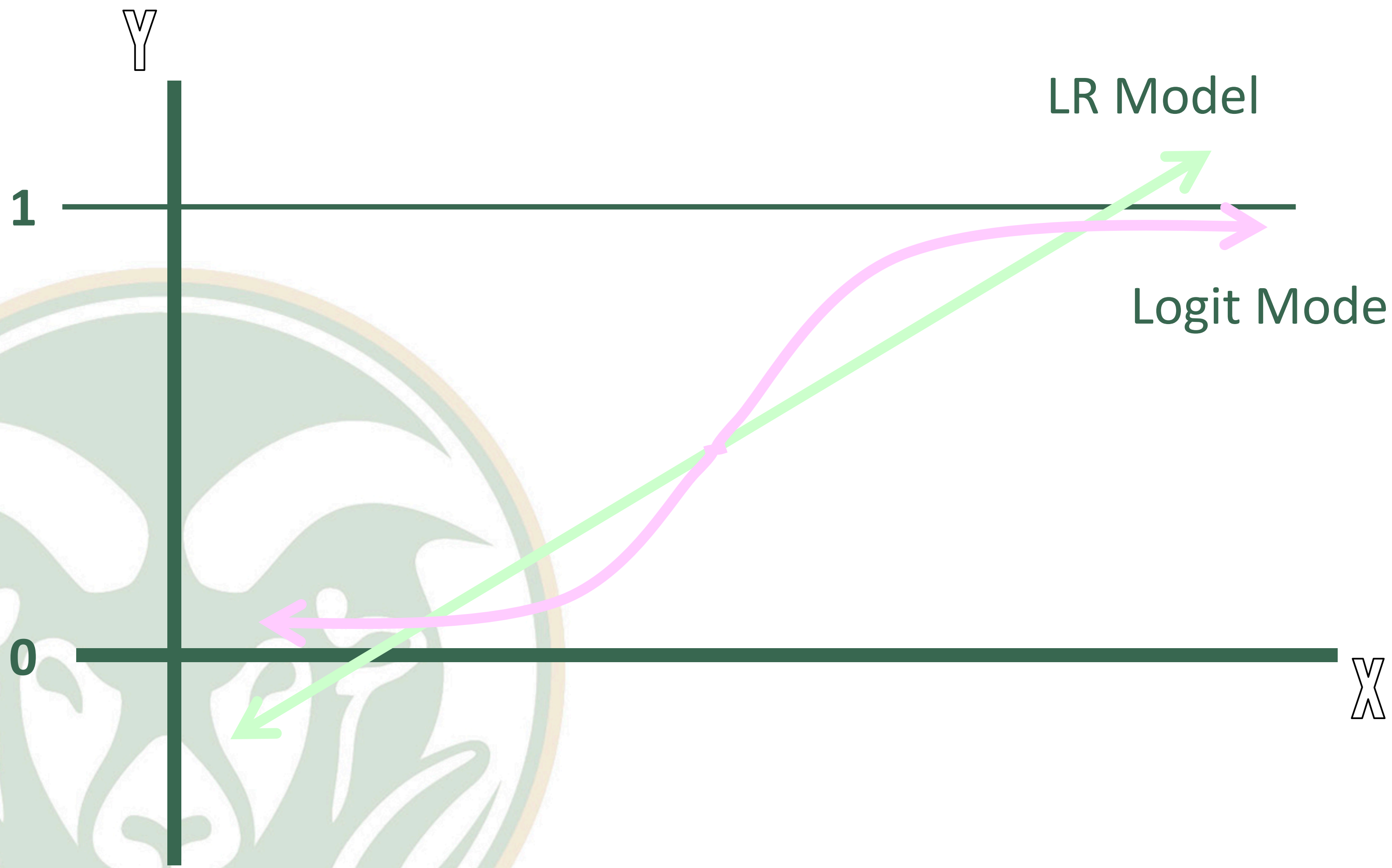
- Response - Presence/Absence of characteristic
- Predictor - Numeric variable observed for each case
- Model - $p(x) \equiv$ Probability of presence at predictor level x

$$p(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

- $\beta = 0 \Rightarrow$ P(Presence) is the same at each level of x
- $\beta > 0 \Rightarrow$ P(Presence) increases as x increases
- $\beta < 0 \Rightarrow$ P(Presence) decreases as x increases


Logistic Regression Example
y
x
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
Boundary
False samples
True samples

Comparing LR and Logit Models



Maximum Likelihood Estimation

- MLE is a statistical method for estimating the coefficients of a model.
- The likelihood function (L) measures the probability of observing the particular set of dependent variable values (p1, p2, ..., pn) that occur in the sample:

  L = Prob (p1* p2* * * pn)
- The higher the L, the higher the probability of observing the ps in the sample.
- MLE involves finding the coefficients (α, β) that makes the log of the likelihood function (LL < 0) as large as possible

Example: Credit Card Default

	A	B	C	D	E	F	G	H	I
1	**Age**	**Default**	p(x)	1-p(x)	V(x)	ln V(x)		b0	b1
2	20	0	0.685347058	0.314652942	0.314652942	-1.15629		**1.663148**	**-0.04423**
3	23	1	0.656052371	0.343947629	0.656052371	-0.42151			
4	24	1	0.646003201	0.353996799	0.646003201	-0.43695		obj	**-65.2721**
5	25	0	0.635823405	0.364176595	0.364176595	-1.01012			
6	25	0	0.635823405	0.364176595	0.364176595	-1.01012			
7	26	1	0.625520567	0.374479433	0.625520567	-0.46917			
8	26	0	0.625520567	0.374479433	0.374479433	-0.98222			
9	28	1	0.604578221	0.395421779	0.604578221	-0.50322			
10	28	0	0.604578221	0.395421779	0.395421779	-0.9278			
11	29	0	0.593955934	0.406044066	0.406044066	-0.90129			
12	30	1	0.583244999	0.416755001	0.583244999	-0.53915			
13	30	1	0.583244999	0.416755001	0.583244999	-0.53915			
14	30	1	0.583244999	0.416755001	0.583244999	-0.53915			
15	30	1	0.583244999	0.416755001	0.583244999	-0.53915			
16	30	0	0.583244999	0.416755001	0.416755001	-0.87526			
17	30	1	0.583244999	0.416755001	0.583244999	-0.53915			
18	32	1	0.56159543	0.43840457	0.56159543	-0.57697			
19	32	1	0.56159543	0.43840457	0.56159543	-0.57697			
20	33	1	0.550676629	0.449323371	0.550676629	-0.59661			
21	33	1	0.550676629	0.449323371	0.550676629	-0.59661			
22	34	0	0.539708774	0.460291226	0.460291226	-0.7759			
23	34	0	0.539708774	0.460291226	0.460291226	-0.7759			
24	34	1	0.539708774	0.460291226	0.539708774	-0.61673			
25	34	1	0.539708774	0.460291226	0.539708774	-0.61673			

Multiple Logistic Regression

- Extension to more than one predictor variable (either numeric or dummy variables).

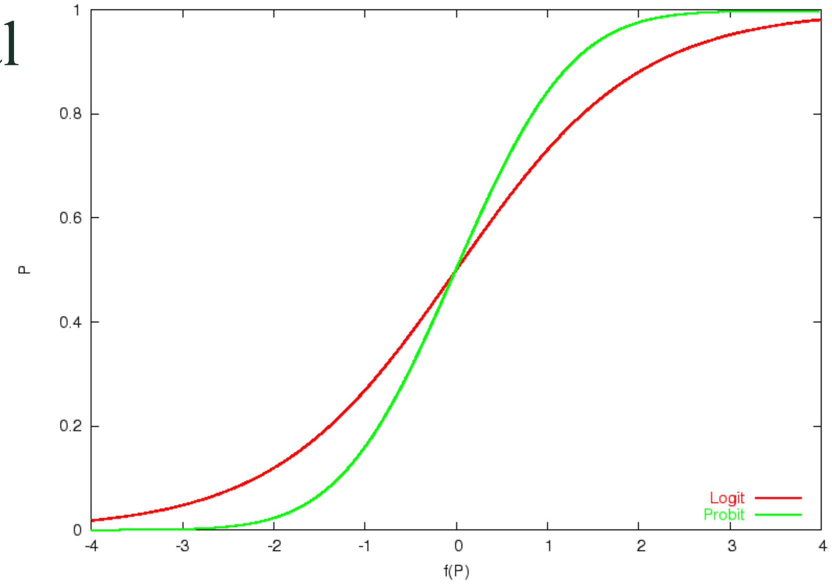
- With *p* predictors, the model is written:

$$p = \frac{e^{\alpha + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\alpha + \beta_1 x_1 + \cdots + \beta_p x_p}}$$

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$$

Normal (Probit) Regression

- ε is distributed as a standard normal
  - Mean zero
  - Variance 1



- Evaluate probability (y=1)
  - $Pr(y_i=1) = Pr(\varepsilon_i > - x_i \beta) = 1 - \Phi(-x_i \beta)$
  - Given symmetry: $1 - \Phi(-x_i \beta) = \Phi(x_i \beta)$
- Evaluate probability (y=0)
  - $Pr(y_i=0) = Pr(\varepsilon_i \leq - x_i \beta) = \Phi(-x_i \beta)$
  - Given symmetry: $\Phi(-x_i \beta) = 1 - \Phi(x_i \beta)$