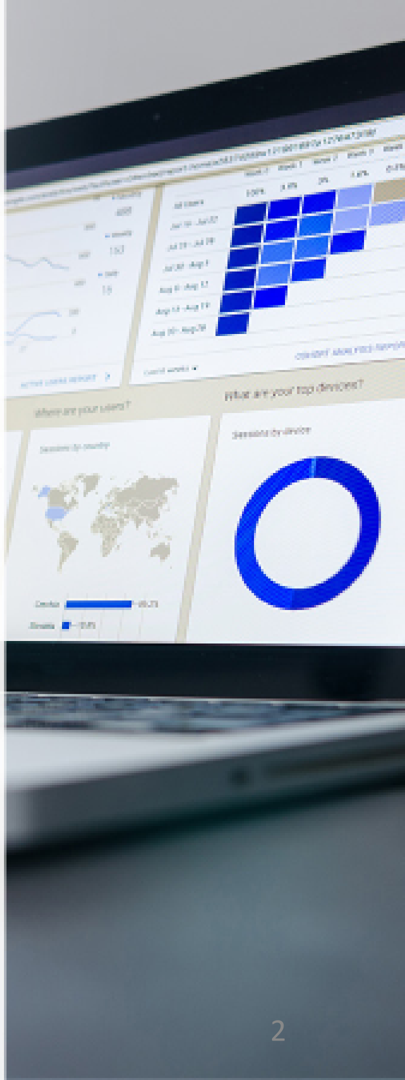
Data Analytics & Visualization

T.E. AI and DS - Jan 2023

Dr. M. Vijayalakshmi

Module 2B- **Regression Models**

- **Logistic Regression**: Logistic Response function and logit, Logistic Regression and GLM, Generalized Linear model,
- Predicted values from Logistic Regression, Interpreting the coefficients and odds ratios
- Linear and Logistic Regression: similarities and Differences
- Assessing the models.



logistic regression

- There are many applications in real life for which the dependent variable is "limited." Examples include:
  - An email is a spam
  - It’ll rain today
  - An individual will purchase a car
  - An online transaction is fraudulent
  - A contestant will win an election
  - A group of users will buy a product
  - A promotional email receiver is a responder or non-responder
- Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable: coded 0 or 1

Regression - Review

- Regression analysis is a type of predictive modeling technique which is used to find the relationship between a dependent variable (usually known as the “Y” variable) and either one independent variable (the “X” variable) or a series of independent variables.
- When two or more independent variables are used to predict or explain the outcome of the dependent variable, this is known as multiple regression.

Background

- Simple linear regression
  - Relationship between numerical response and a numerical or categorical predictor
- Multiple regression
  - Relationship between numerical response and multiple numerical and/or categorical predictors
- What we haven't seen is what to do when the predictors are
  - nonlinear, complicated dependence structure, etc.
- or when the response
  - is categorical, count data, etc.

Definition

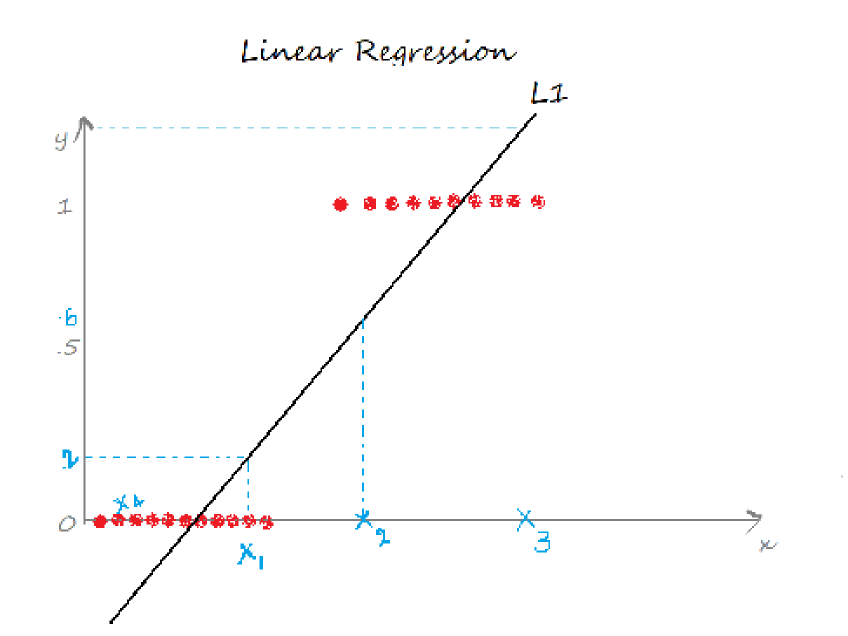
- Logistic regression is a statistical method that is used for building machine learning models where the dependent variable is dichotomous: i.e. binary.
- Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables.
- The independent variables can be continuous, nominal, discrete, ordinal, or of interval type.

Types of Logistic Regression

- **Binary logistic -** the dependent variable is binary in nature. For example, the output can be Success/Failure, 0/1 , True/False, or Yes/No.
- **Multinomial logistic regression** is used when you have one categorical dependent variable with two or more unordered levels (i.e two or more discrete outcomes. For example, let's imagine that you want to predict what will be the most-used transportation type in the year 2030. The transport type will be the dependent variable, with possible outputs of train, bus, tram, and bike (for example).
- **Ordinal logistic regression** is used when the dependent variable (Y) is ordered (i.e., ordinal). The dependent variable has a meaningful order and more than two categories or levels. Examples of such variables might be t-shirt size (XS/S/M/L/XL), answers on an opinion poll (Agree/Disagree/ Neutral), or scores on a test (Poor/Average/Good).

Problem with SLR

- Let us try linear regression to solve a binary class classification problem.
- Assume we have a dataset that is linearly separable and has the output that is discrete in two classes (0, 1).
- Draw a straight line(the best fit line) L1 such that the sum of distances of all the data points to the line is minimal.
- The equation of the line L1 is y=mx+c, where m is the slope and c is the y-intercept.



Problem with SLR

We define a threshold T = 0.5, above which the output belongs to class 1 and class 0 otherwise.

$$y = mx + c, \text{ Threshold } T = 0.5$$

$$y = \begin{cases} 1, & mx+c \geq 0.5 \\ 0, & mx+c < 0.5 \end{cases}$$

Case 1: the predicted value for x1 is ≈0.2 which is less than the threshold, so x1 belongs to class 0.

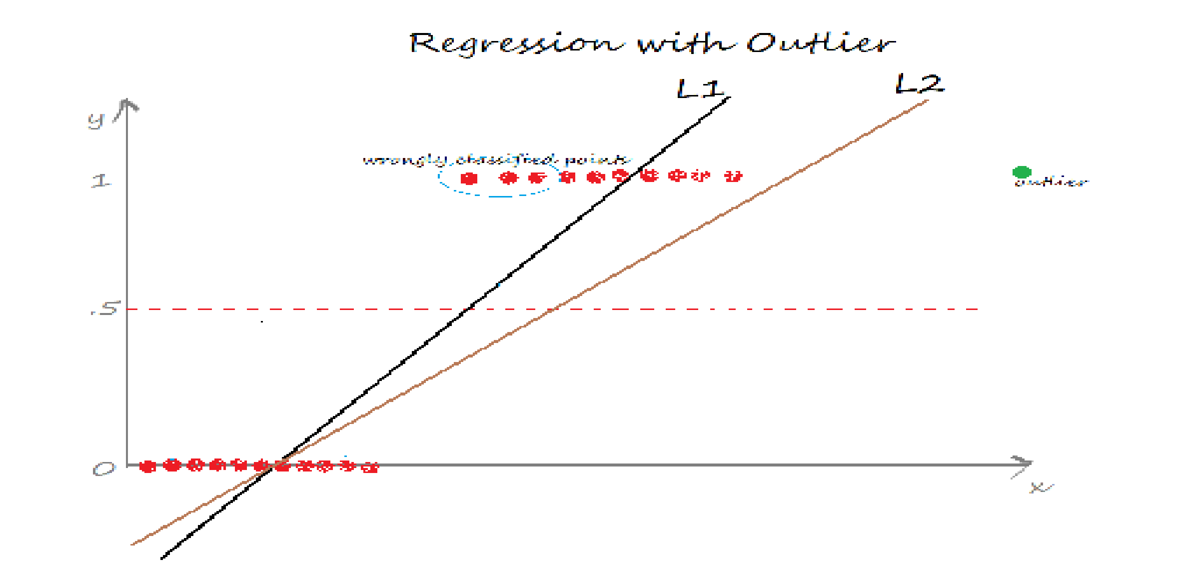
Case 2: the predicted value for the point x2 is ≈0.6 which is greater than the threshold, so x2 belongs to class 1.

So far so good, yeah!

Case 3: the predicted value for the point x3 is beyond 1.

Case 4: the predicted value for the point x4 is below 0.

The predicted values for the points x3, x4 exceed the range (0,1) which doesn't make sense because the probability values always lie between 0 and 1. And our output can have only two values either 0 or 1. Hence, this is a problem with the linear regression model.



- L2 is the new best-fit line after the addition of an outlier.
- If we closely observe, some of the data points are wrongly classified.
- Thus this increases the error term
- This again is a problem with the linear regression model.

SLR Assumptions - Review

- Linearity: The relationship between X and the mean of Y is linear.
- Homoscedasticity: The variance of residual is the same for any value of X (Constant variance of errors).
- Independence: Observations are independent of each other.
- Normality: The error(residuals) follow a normal distribution.

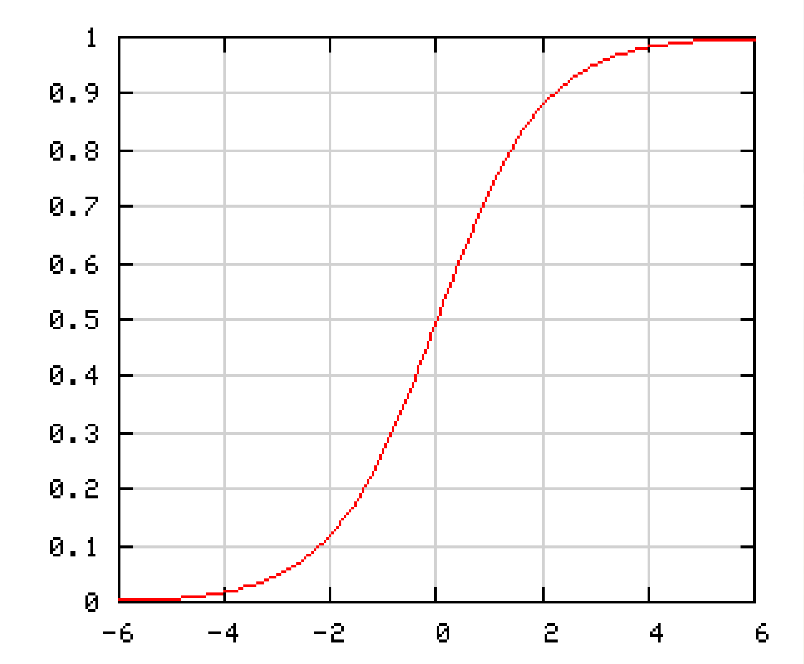
Disadvantages SLR 1

- First, what is our dependent variable, conceptually? It is the probability of y=1. But we only observe y=0 and y=1.
- If we use SLR, we'll get predicted values that fall between 0 and 1—which is what we want—but we'll also get predicted values that are greater than 1, or less than 0. That makes no sense.

Further Disadvantages

- Noted that classification is not normally distributed which is violated assumption 4: **Normality**.
- Moreover, both mean and variance depend on the underlying probability. Any factor that affects the probability will change not just the mean but also the variance of the observations, which means the variance is no longer constantly violating the assumption 2: **Homoscedasticity**.
- As a result, we cannot directly apply linear regression because it won't be a good fit.

Logistic Regression—s shaped curve

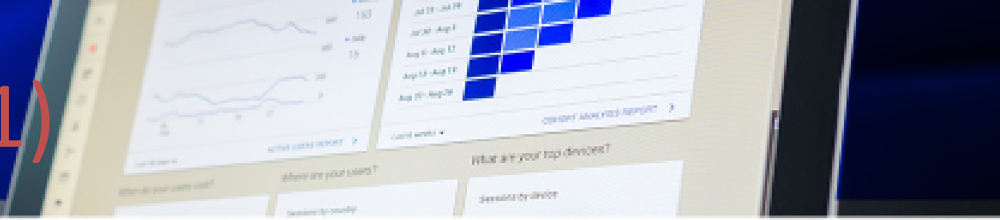


- The reality is that many choice functions can be modeled by an S-shaped curve.
- Therefore (much as when we discussed linear transformations of the X variable), it makes sense to model a non-linear relationship.

Binary Logistic Regression

- We have seen the challenges posed by SLR to model a situation with categorical outputs.
- Logistic regression is a regression model
- When the output is one of two categories it is called Binary Logistic regression
- Classification also can give a binary response Y or N, However, simply guessing “yes” or “no” is pretty crude — especially if there is no perfect rule.
- We need Something which takes noise into account, and doesn’t just give a binary answer, will often be useful.
- In short, we want probabilities — which means we need to fit a stochastic model.
- Logistic Regression is an easily interpretable classification technique that gives the probability of an event occurring, not just the predicted classification. It also provides a measure of the significance of the effect of each individual input variable, together with a measure of certainty of the variable's effect

Logistic regression (1)



Age and signs of coronary heart disease (CD)

Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

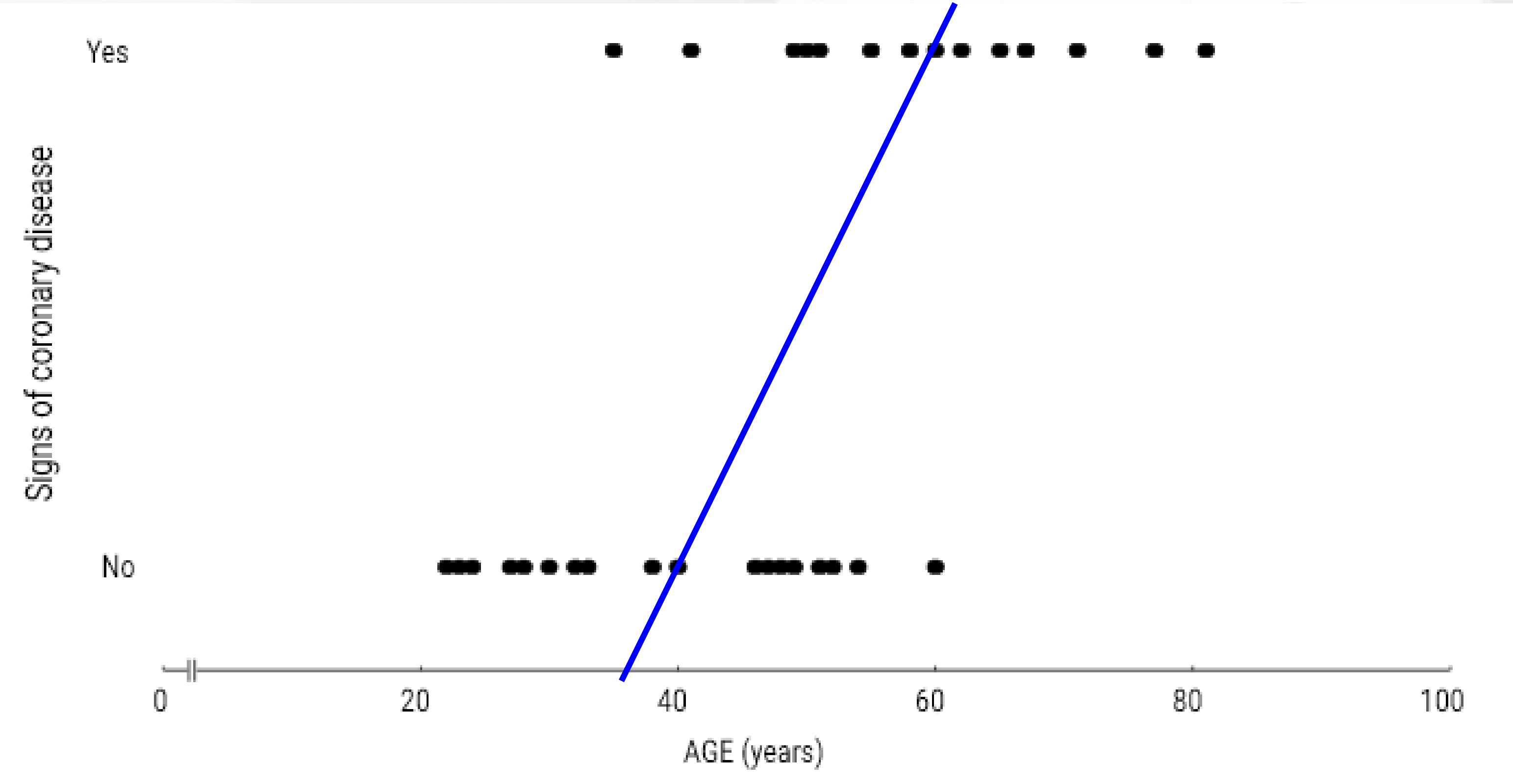
How can we analyse these data?

- Compare mean age of diseased and non-diseased

  - Non-diseased: 38.6 years
  - Diseased: 58.7 years ($p<0.0001$)

- Linear regression?

Dot-plot: Data from Table



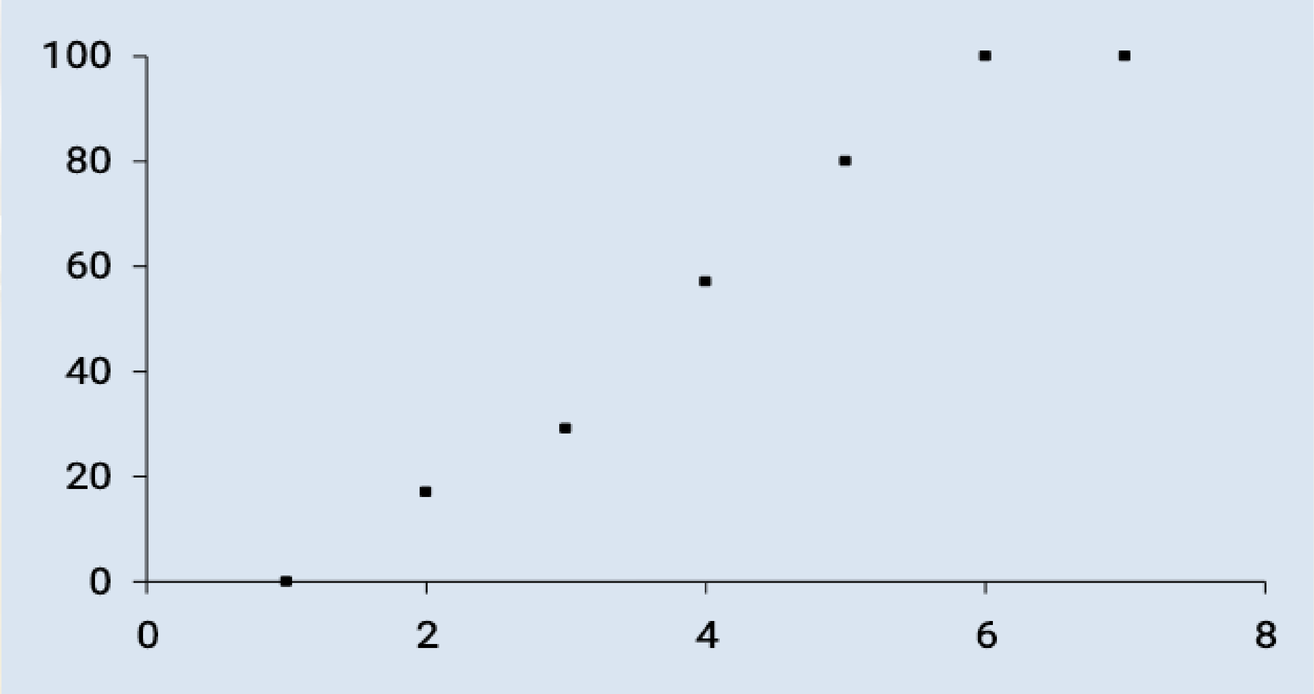
Logistic regression (2)

Prevalence (%) of signs of CD according to age group

		Diseased	
Age group	# in group	#	%
20 - 29	5	0	0
30 - 39	6	1	17
40 - 49	7	2	29
50 - 59	7	4	57
60 - 69	5	4	80
70 - 79	2	2	100
80 - 89	1	1	100

Dot-plot: Data from Table

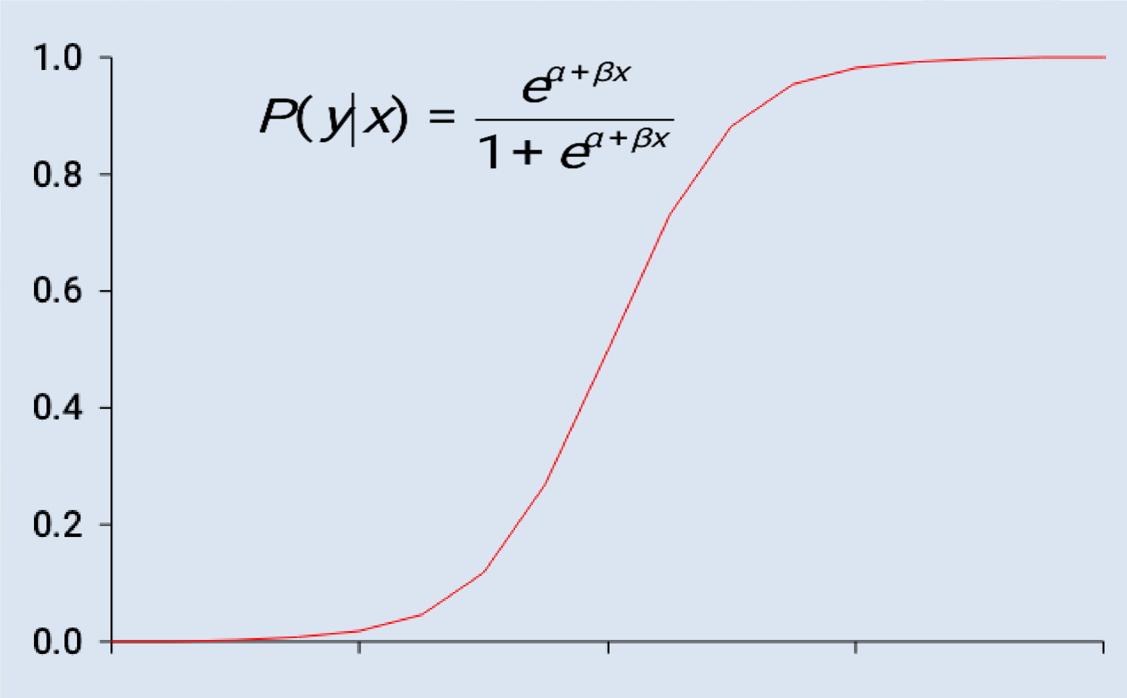
Diseased %



Age group

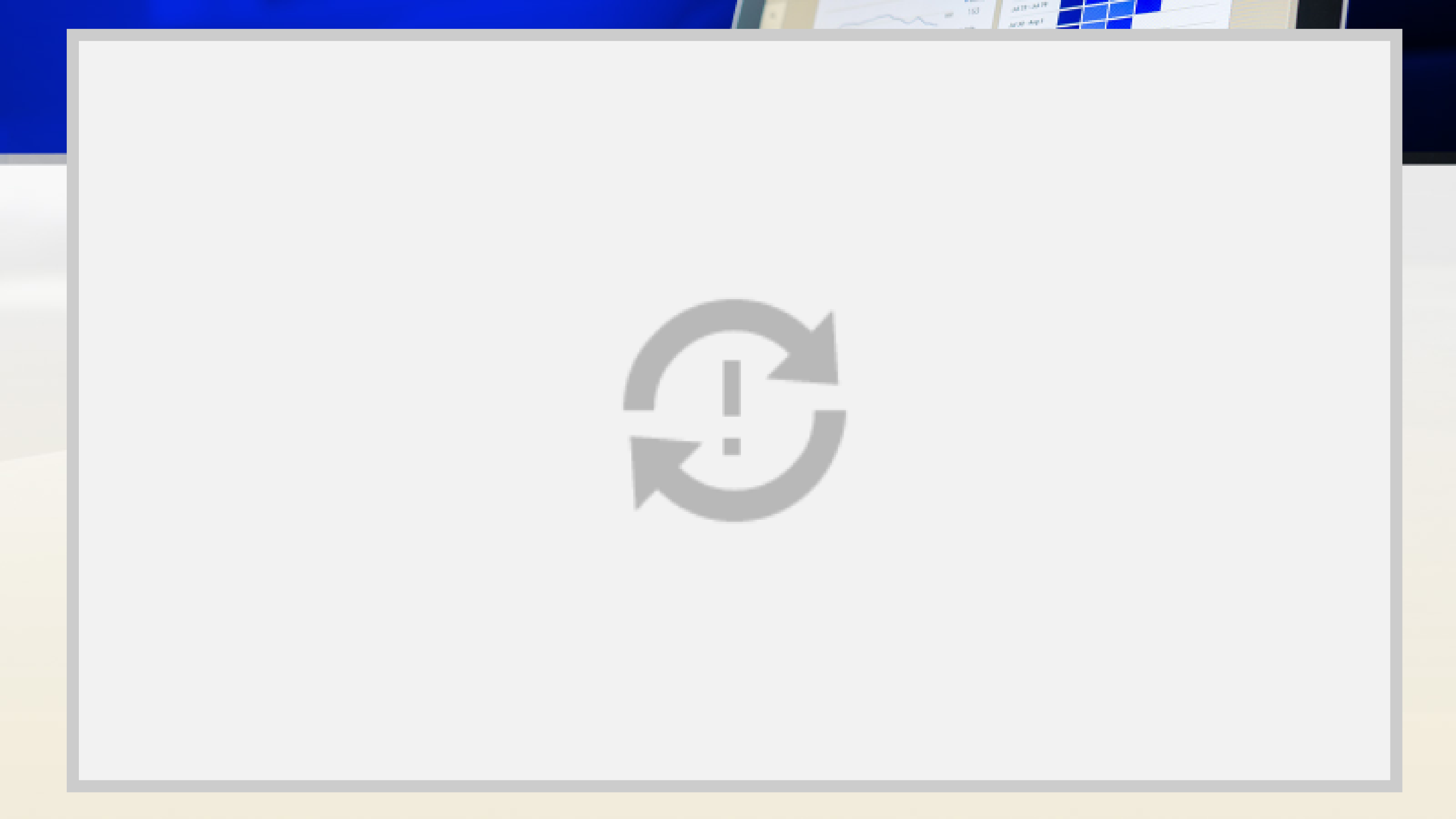
Logistic function (1)

Probability of disease



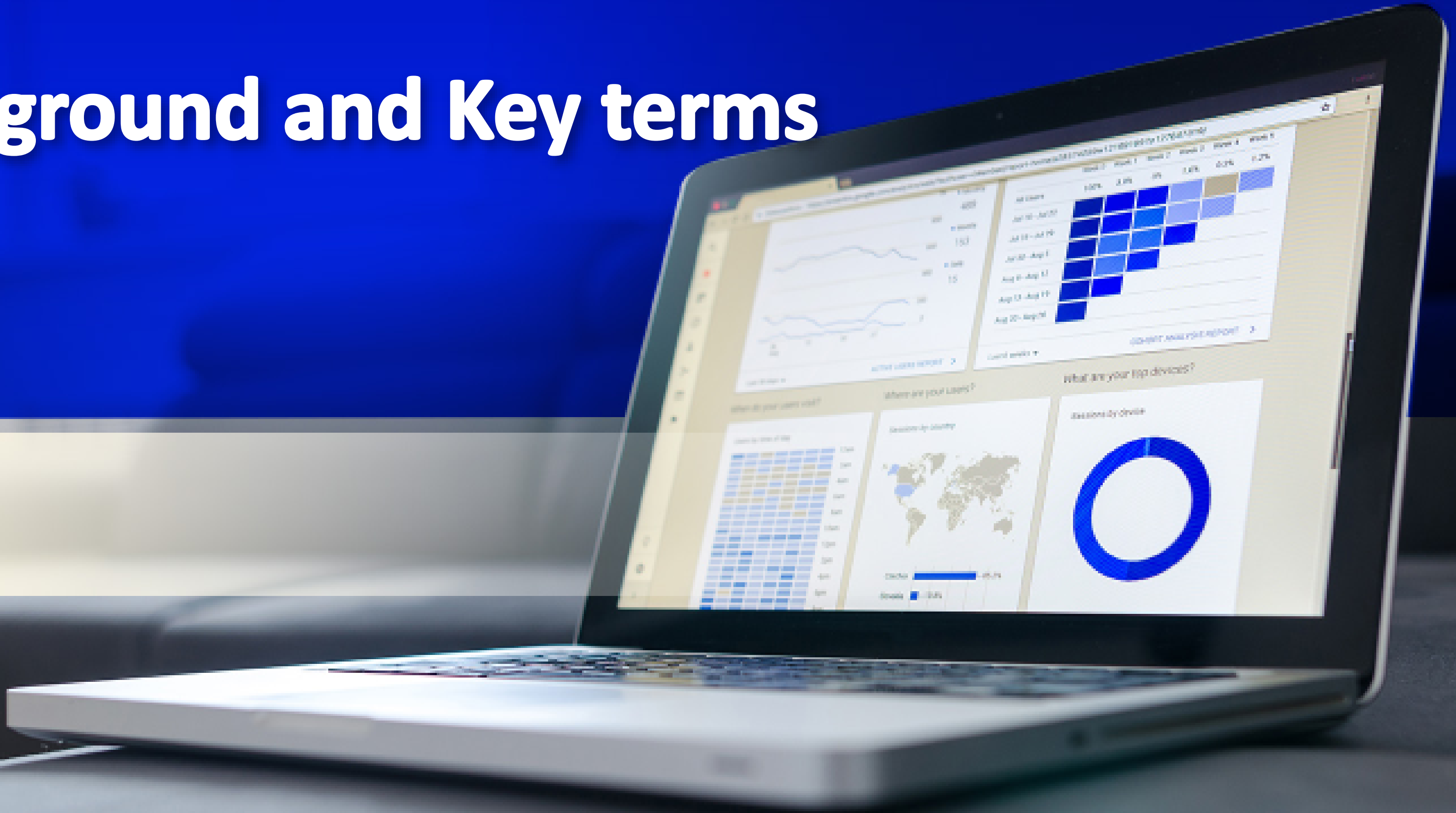
x

Binary Response & Logistic Regression



Logistic regression

Background and Key terms



Sample Data

Favored Stock		Less Favored Stock	
Success	Size	Success	Size
1	1	0	1
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	0	0	0
1	0	0	0

Contingency Table

Type of Stock	Large	Small	Total
Preferred	10	2	12
Not Preferred	1	11	12
Total	11	13	24

Basic Concepts

- Probability
  - Probability of being a preferred stock = 12/24 = 0.5
  - Probability that a company's stock is preferred given that the company is large = 10/11 = 0.909
  - Probability that a company's stock is preferred given that the company is small = 2/13 = 0.154

Odds and Probability

- Odds(Event) = Prob(Event)/(1-Prob(Event))
- Prob(Event) = Odds(Event)/(1+Odds(Event))

Concepts … contd.

- Odds
  - Odds of a preferred stock = 12/12 = 1
  - Odds of a preferred stock given that the company is large = 10/1 = 10
  - Odds of a preferred stock given that the company is small = 2/11 = 0.182

Logistic Regression

- Take Natural Log of the odds:
  - ln(odds(Preferred|Large)) = ln(10) = 2.303
  - ln(odds(Preferred|Small)) = ln(0.182) = -1.704

- Combining these relationships
  - ln(odds(Preferred|Size)) = -1.704 + 4.007*Size
  - Log of the odds is a linear function of size
  - The coefficient of size can be interpreted like the coefficient in regression analysis

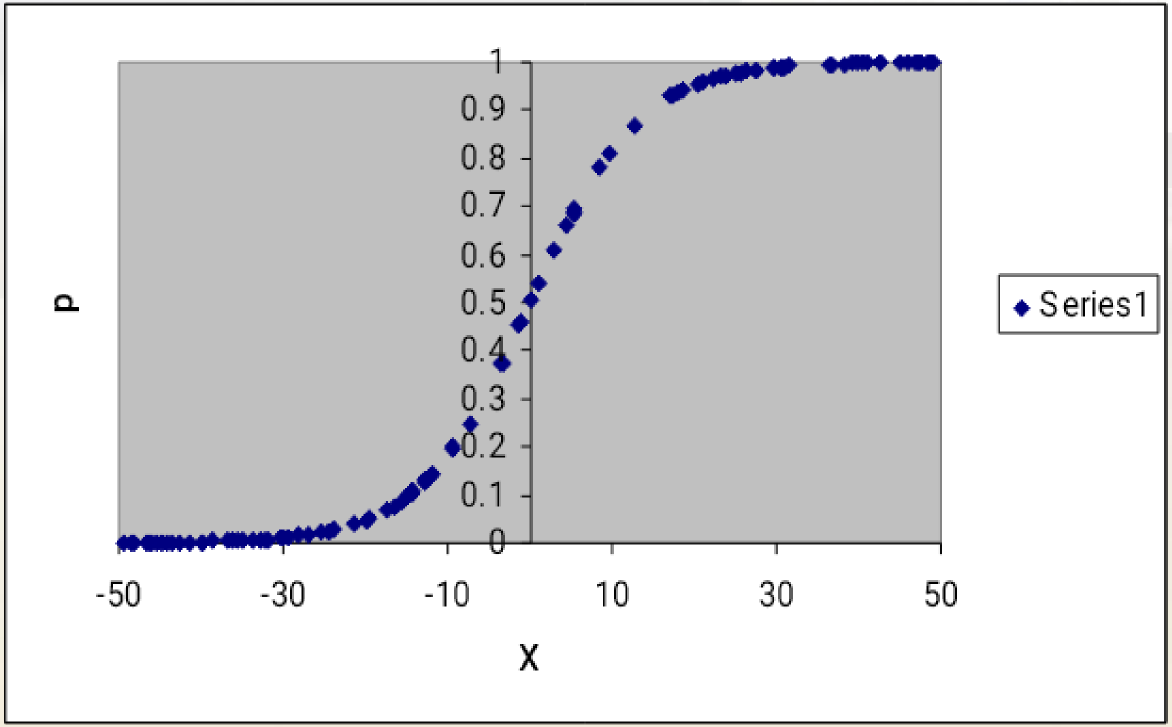
Interpretation

- Positive sign $\Rightarrow$ ln(odds) is increasing in size of the company i.e. a large company is more likely to have a preferred stock vis-à-vis a small company
- Magnitude of the coefficient gives a measure of how much more likely

General Model

- $\ln(\text{odds}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_K$ (1)
- Recall:
  - Odds = p/(1-p)
- $\ln(p/1-p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X$ (2)
- p = $$\frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$$
- p = $$\frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1)}}$$

Logistic Function



Estimation

- Coefficients in the regression model are estimated by minimizing the sum of squared errors
- Since, p is non-linear in the parameter estimates we need a non-linear estimation technique
  - **Maximum-Likelihood Approach**
  - Non-Linear Least Squares

Maximum Likelihood Approach

- Conditional on parameter β, write out the probability of observing the data
- Write this probability out for each observation
- Multiply the probability of each observation out to get the joint probability of observing the data condition on β
- Find the β that maximizes the conditional probability of realizing this data

Logistic Response Function and Logit

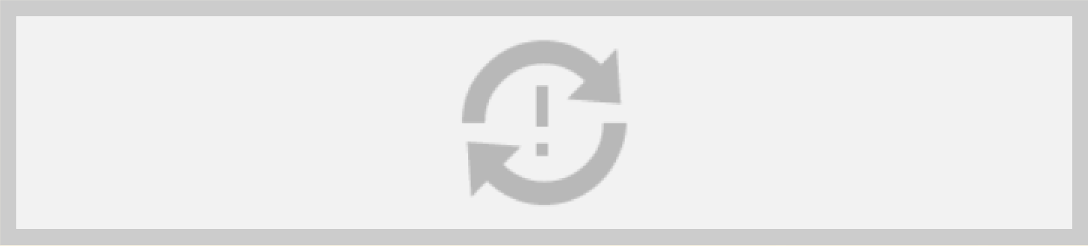
- The key ingredients are the *logistic response function* and the *logit*, in which we map a probability (which is on a 0–1 scale) to a more expansive scale suitable for linear modeling.
- The first step is to think of the outcome variable not as a binary label, but as the probability *p* that the label is a “1.” Naively, we might be tempted to model *p* as a linear function of the predictor variables:

Logistic Response Function and Logit

- However, fitting this model does not ensure that *p* will end up between 0 and 1, as a probability must.
- Instead, we model *p* by applying a *logistic response* or *inverse logit* function to the predictors:
- This transform ensures that the *p* stays between 0 and 1.

Logistic Response Function and Logit

- To get the exponential expression out of the denominator, we consider *odds* instead of probabilities.
- Odds, are the ratio of “successes” (1) to “nonsuccesses” (0).
- In terms of probabilities, odds are the probability of an event divided by the probability that the event will not occur.
- For example, if the probability that a horse will win is 0.5, the probability of “won’t win” is (1 – 0.5) = 0.5, and the odds are 1.0.



Logistic Response Function and Logit

- We can obtain the probability from the odds using the inverse odds function:

- We combine this with the logistic response function, shown earlier, to get:

Logistic Response Function and Logit

- Finally, taking the logarithm of both sides, we get an expression that involves a linear function of the predictors:
- The log-odds function, also called logit function, maps the probability p from (0, 1) to any value (+∞ to - ∞)
- The transformation circle is complete; we have used a linear model to predict a probability, which, in turn, we can map to a class label by applying a cutoff rule
  - any record with a probability greater than the cutoff is classified as a 1.

The function that maps a probability to a scale suitable for a linear model (logit)

The logistic transformation (1)

- This type of relationship is described by a special formula.
  - Remember, if the relationship was linear then the equation is just:

$$\boxed{\pi = \alpha + \beta X}$$

  - But the relationship on the graph is actually described by:

$$\boxed{\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X}$$

The logistic transformation (2)

$$\log\left(\frac{\pi}{1-\pi}\right) = a + \beta X$$

This is just the *odds*.

As the probability increases (from zero to 1), the odds increase from 0 to infinity.

The log of the odds then increases from –infinity to +infinity.

So if β is ‘large’ then as X increases the log of the odds will increase steeply.

The steepness of the curve will therefore increase as β gets bigger.

Logistic Regression

Brief Review

Logistic Regression

In logistic regression:

- Outcome variable is binary
- Purpose of the analysis is to assess the effects of multiple explanatory variables, which can be numeric and/or categorical, on the outcome variable.

Requirements for Logistic Regression

The Following need to be specified:

- An outcome variable with two possible categorical outcomes (1=success; 0=failure).
- Estimating the probability *P* of the outcome variable.
- Linking the outcome variable to the explanatory variables.
- Estimating the coefficients of the regression equation, as well as their confidence intervals.
- Testing the goodness of fit of the regression model.

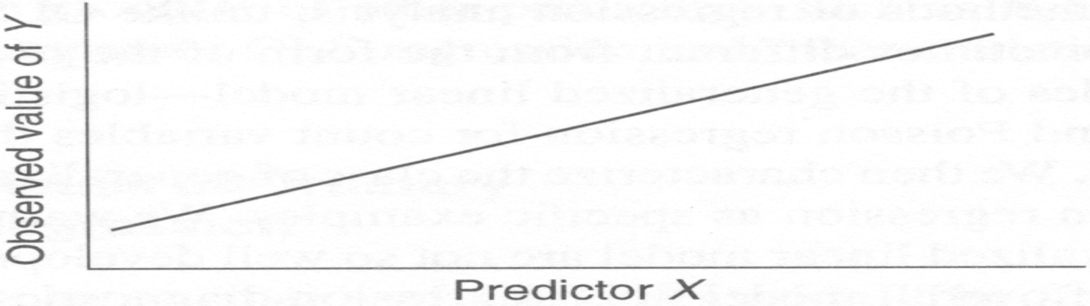
Measuring the Probability of Outcome

- The probability of the outcome is measured by the odds of occurrence of an event.
- If P is the probability of an event, then (1-P) is the probability of it not occurring.
- Odds of success = P / 1-P

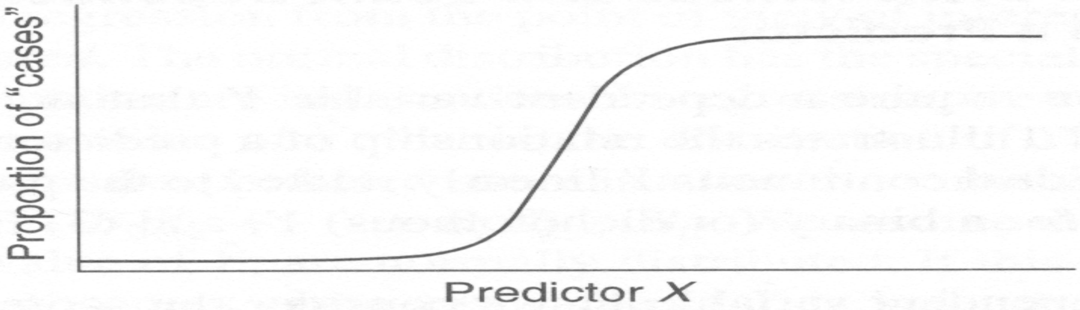
$$\frac{P}{1-P}$$

The logistic function

(A) For a continuous outcome variable Y, the numerical value of Y at each value of X.



(B) For a binary outcome variable, the proportion of individuals who are "cases" (exhibit a particular outcome property) at each value of X.



The logistic function

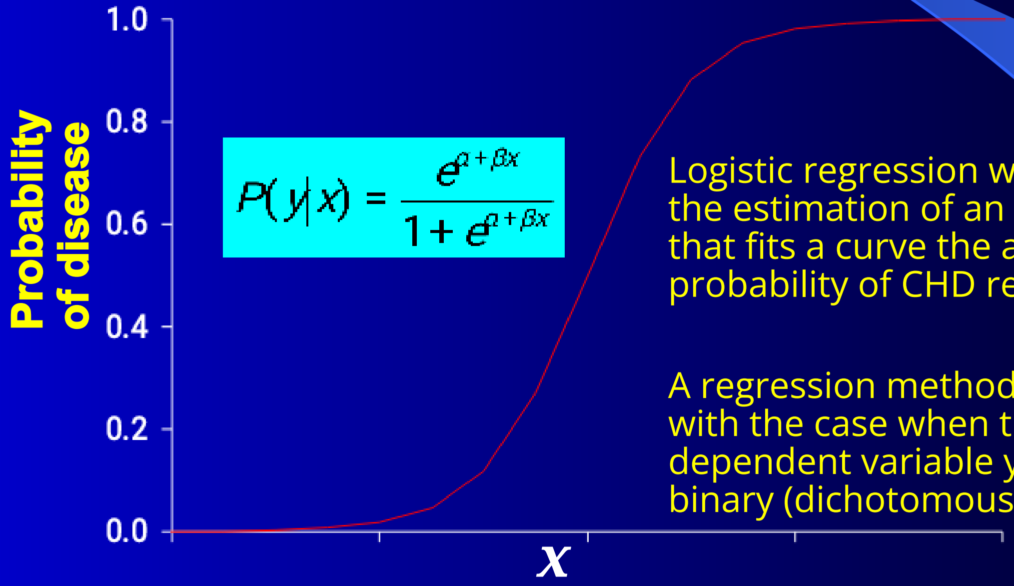
$$\widehat{Y_i} = \frac{e^u}{1+e^u}$$

- Where Y-hat is the estimated probability that the ith case is in a category and u is the regular linear regression equation:

$$u = A + B_1 X_1 + B_2 X_2 + \cdots + B_K X_K$$

Logistic function

For a response variable y with p(y=1)= P and p(y=0) = 1- P



$$P(y|x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

Logistic regression will allow for the estimation of an equation that fits a curve the age/ probability of CHD relationship

A regression method to deal with the case when the dependent variable y is binary (dichotomous)

The logistic function

- Change in probability is not constant (linear) with constant changes in X
- This means that the probability of a success (Y = 1) given the predictor variable (X) is a non-linear function, specifically a logistic function

The logistic function

- It is not obvious how the regression coefficients for X are related to changes in the dependent variable (Y) when the model is written this way
- Change in Y(in probability units)|X depends on value of X. Look at S-shaped function

The Logistic Regression

- The joint effects of all explanatory variables put together on the odds is
  - Odds = P/1-P = $e^{\alpha + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp}$

- Taking the logarithms of both sides
  - Log{P/1-P} = $\log^{\alpha+\beta 1X1+\beta 2X2+\ldots+\beta pXp}$
  - Logit P = α+β1X1+β2X2+..+βpXp

- The coefficients β1, β2, βp are such that the sums of the squared distance between the observed and predicted values (i.e. regression line) are smallest.

The Logistic Regression

- Logit p = $\alpha + \beta 1X1 + \beta 2X2 + .. + \beta pXp$
- α represents the overall disease risk
- $\beta 1$ represents the fraction by which the disease risk is altered by a unit change in X1
- $\beta 2$ is the fraction by which the disease risk is altered by a unit change in X2
- ……. and so on.
- What changes is the log odds. The odds themselves are changed by e^{β}
- If $\beta = 1.6$ the odds are $e^{1.6} = 4.95$

Fitting this model (1)

- So that's what we want to do, but how do we do it?
  - With SLR we tried to minimize the squares of the residuals, to get the best fitting line.
  - This doesn't really make sense here (remember the errors won't be normally distributed as there's only two values).
- We use something called *maximum likelihood* to estimate what the β and α are.

Fitting this model (2)

- *Maximum likelihood* is an *iterative process* that estimates the best fitted equation.
    - The iterative bit just means that we try lots of models until we get to a situation where tweaking the equation any further doesn't improve the fit.
    - The maximum likelihood bit is kind of complicated, although the underlying assumptions are simple to understand, and very intuitive. The basic idea is that we find the coefficient value that makes the observed data most likely.

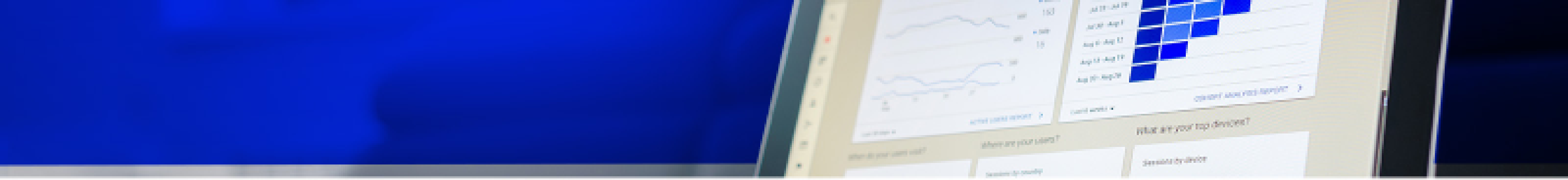
Maximum Likelihood Estimation (MLE)

- 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 (p_1, p_2, ..., p_n) that occur in the sample:

  $$L = \text{Prob}\ (p_1 * p_2 * * * p_n)$$

- The higher the L, the higher the probability of observing the ps in the sample.

Logistic regression



Recall the **simple linear regression** model:

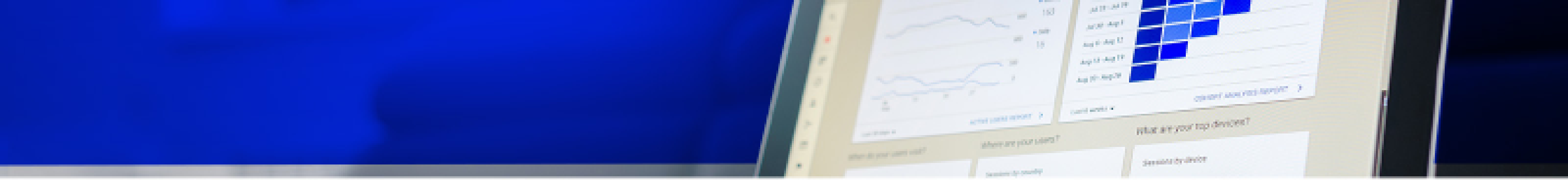
$$y = b_0 + b_1 x + e$$

where we are trying to predict a continuous dependent variable *y* from a continuous independent

This model can be extended to **Multiple linear regression** model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + + \beta_p x_p + \varepsilon$$

Here we are trying to predict a continuous dependent variable *y* from a several continuous dependent variables x_1 , x_2 , … , x_p .



Now suppose the dependent variable *y* is **binary.**

It takes on two values "Success" (1) or "Failure" (0)

We are interested in predicting a *y* from a continuous dependent variable *x*.

This is the situation in which **Logistic Regression** is used

Example

We are interested how the success (y) of a new antibiotic cream is curing "skin problems" and how it depends on the amount (x) that is applied daily.

The values of y are 1 (Success) or 0 (Failure).

The values of x range over a continuum

The logisitic Regression Model

Let p denote $P[y = 1] = P[\text{Success}]$.
This quantity will increase with the value of x.

The ratio: $\dfrac{p}{1-p}$ is called the **odds ratio**

This quantity will also increase with the value of x, ranging from zero to infinity.

The quantity: $\ln\left(\dfrac{p}{1-p}\right)$

is called the **log odds ratio**

Example: odds ratio, log odds ratio

Suppose a die is rolled:

Success = "roll a six", $p = 1/6$

The **odds ratio** $\dfrac{p}{1-p} = \dfrac{\frac{1}{6}}{1-\frac{1}{6}} = \dfrac{\frac{1}{6}}{\frac{5}{6}} = \dfrac{1}{5}$

The **log odds ratio**

$$\ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{1}{5}\right) = \ln(0.2) = -1.69044$$

The logisitic Regression Model

Assumes the **log odds ratio** is linearly related to *x*.

i. e. : $$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

In terms of the **odds ratio**

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x}$$

The logisitic Regression Model

Solving for p in terms x.

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x}$$

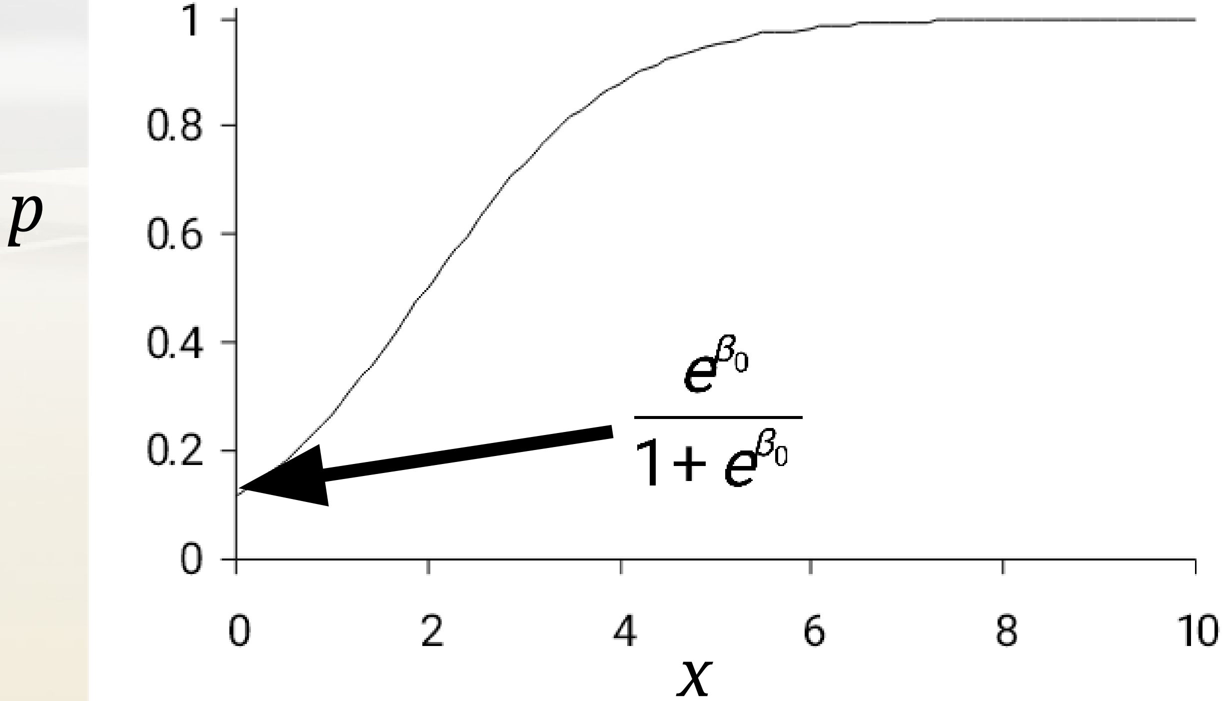
$$p = e^{\beta_0 + \beta_1 x}(1-p)$$

$$p + pe^{\beta_0 + \beta_1 x} = e^{\beta_0 + \beta_1 x}$$

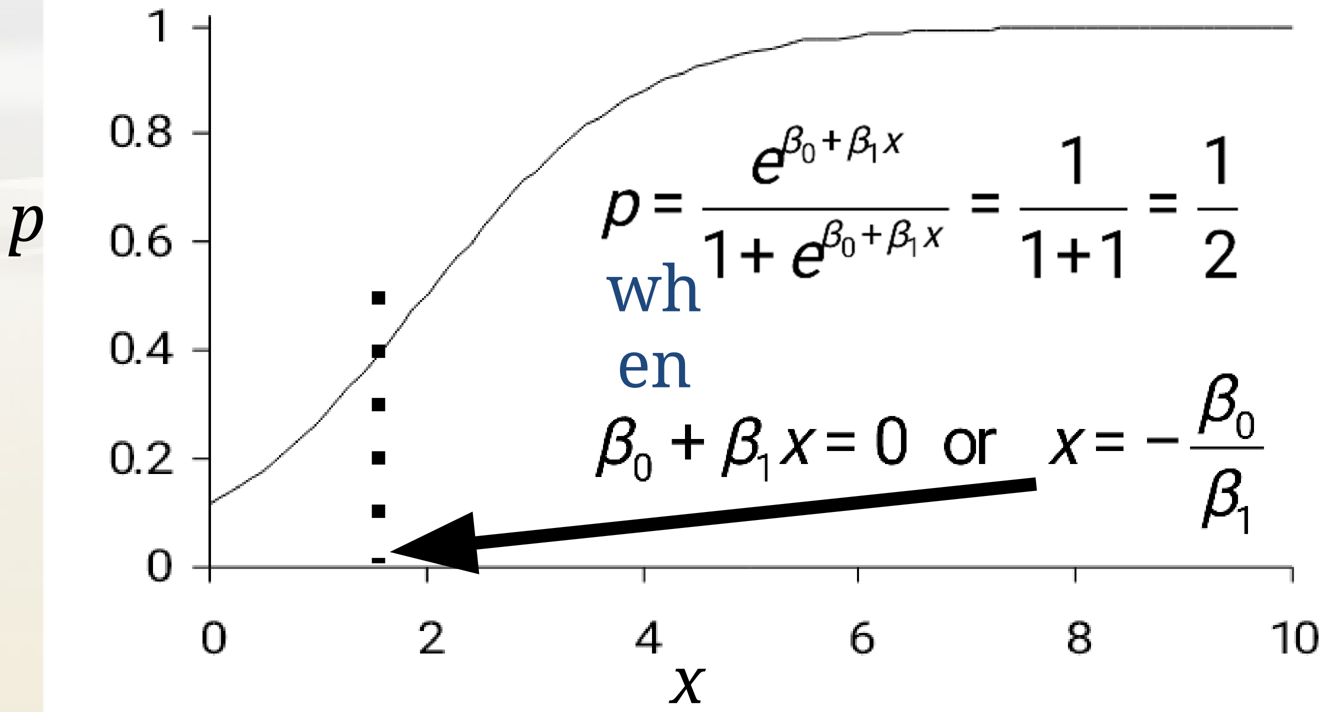
or

$$\boxed{p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}}$$

Interpretation of the parameter β_0 (determines the intercept)



Interpretation of the parameter β_1 (determines when p is 0.50 (along with β_0))



$$\frac{dp}{dx} = \frac{d}{dx}\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

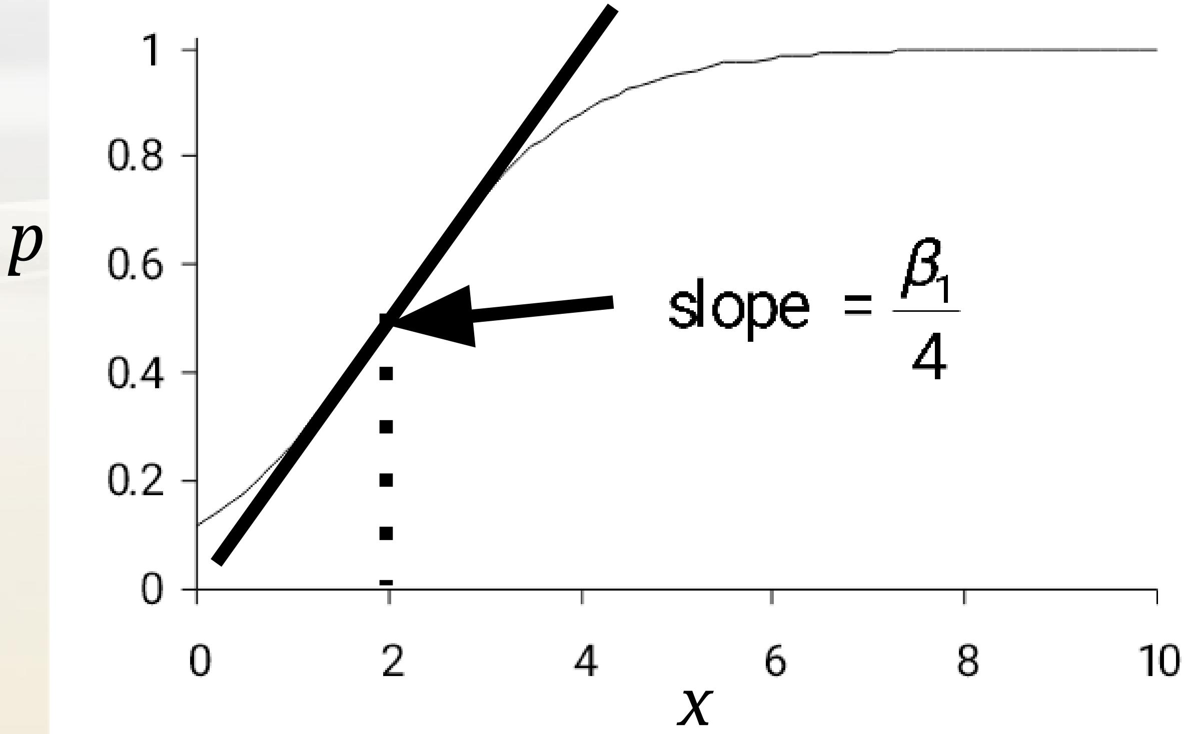
Also

$$= \frac{e^{\beta_0+\beta_1 x}\beta_1\left(1+e^{\beta_0+\beta_1 x}\right) - e^{\beta_0+\beta_1 x}\beta_1 e^{\beta_0+\beta_1 x}}{\left(1+e^{\beta_0+\beta_1 x}\right)^2}$$

$$= \frac{e^{\beta_0+\beta_1 x}\beta_1}{\left(1+e^{\beta_0+\beta_1 x}\right)^2} = \frac{\beta_1}{4}$$ when $$x = -\frac{\beta_0}{\beta_1}$$

$\frac{\beta_1}{4}$ is the rate of increase in *p* with respect to *x* when *p* = 0.50

Interpretation of the parameter β_1 (determines slope when p is 0.50)



The data

The data will for each case consist of

1. a value for x, the continuous independent variable
2. a value for y (1 or 0) (Success or Failure)

Total of n = 250 cases

case	x	y
1	0.8	0
2	2.3	1
3	2.5	0
4	2.8	1
5	3.5	1
6	4.4	1
7	0.5	0
8	4.5	1
9	4.4	1
10	0.9	0
11	3.3	1
12	1.1	0
13	2.5	1
14	0.3	1
15	4.5	1
16	1.8	0
17	2.4	1
18	1.6	0
19	1.9	1
20	4.6	1

case	x	y
230	4.7	1
231	0.3	0
232	1.4	0
233	4.5	1
234	1.4	1
235	4.5	1
236	3.9	0
237	0.0	0
238	4.3	1
239	1.0	0
240	3.9	1
241	1.1	0
242	3.4	1
243	0.6	0
244	1.6	0
245	3.9	0
246	0.2	0
247	2.5	0
248	4.1	1
249	4.2	1
250	4.9	1

Estimation of the parameters

The parameters are estimated by Maximum Likelihood estimation and require a statistical package or programmimg support like R Python etc

The parameter Estimates

	β	SE
X	1.0309	0.1334
Constant	-2.0475	0.332

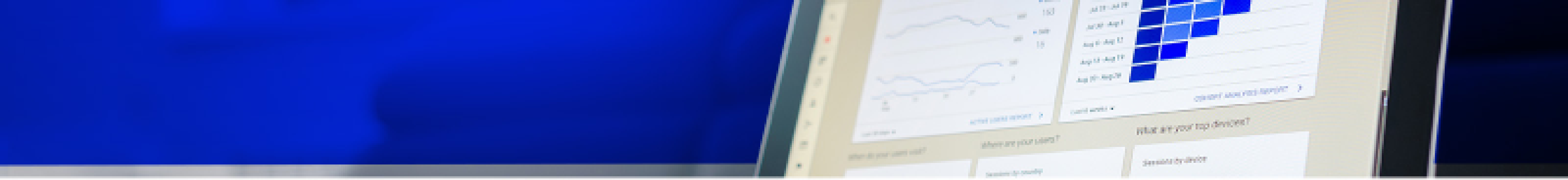
β_1	1.0309
β_0	-2.0475

Interpretation of the parameter β_0 (determines the intercept)

$$\text{intercept} = \frac{e^{\beta_0}}{1+e^{\beta_0}} = \frac{e^{-2.0475}}{1+e^{-2.0475}} = 0.1143$$

Interpretation of the parameter β_1 (determines when p is 0.50 (along with β_0))

$$x = -\frac{\beta_0}{\beta_1} = -\frac{-2.0475}{1.0309} = 1.986$$



Another interpretation of the parameter β_1

$\frac{\beta_1}{4}$ is the rate of increase in p with respect to x when $p = 0.50$

$$\frac{\beta_1}{4} = \frac{1.0309}{4} = 0.258$$

Example Evacuation Data

EVAC	PETS	MOBLHOME	TENURE	EDUC
0	1	0	16	16
0	1	0	26	12
0	1	1	11	13
1	1	1	1	10
1	0	0	5	12
0	0	0	34	12
0	0	0	3	14
0	1	0	3	16
0	1	0	10	12
0	0	0	2	18
0	0	0	2	12
0	1	0	25	16
1	1	1	20	12

The Logistic Regression Model

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"

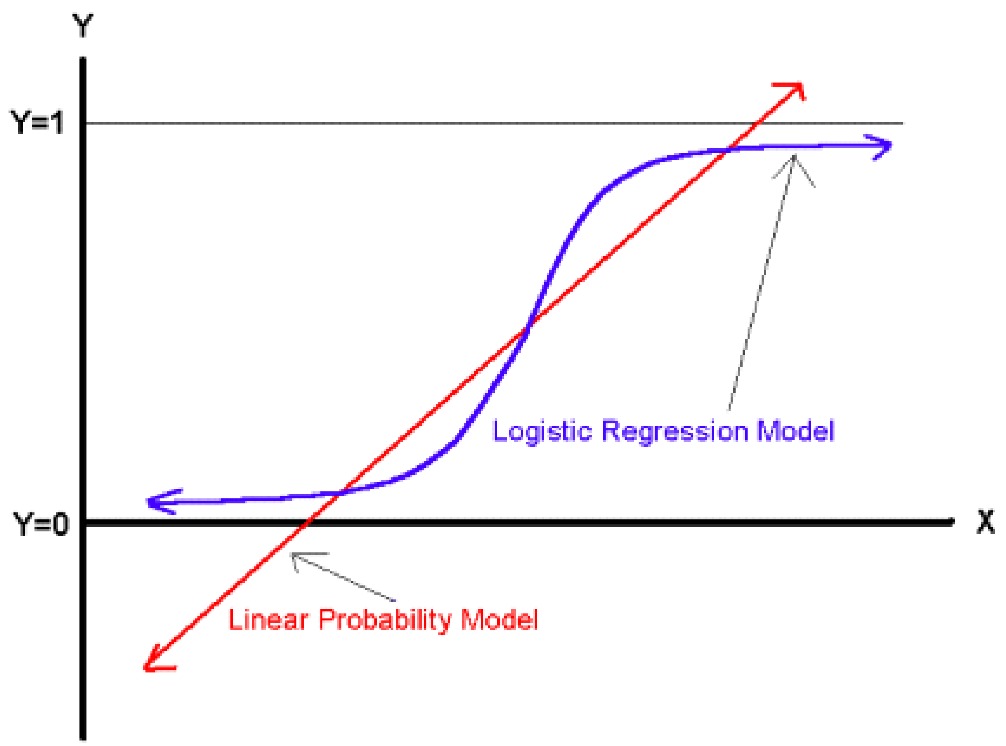
More:

- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
- The estimated probability is:

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

- if you let $\alpha + \beta X = 0$, then $p = .50$
- as $\alpha + \beta X$ gets really big, p approaches 1
- as $\alpha + \beta X$ gets really small, p approaches 0

Comparing the LP and Logit Models



Maximum Likelihood Estimation (MLE)

- 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 (p_1, p_2, ..., p_n) that occur in the sample:

  $$L = \text{Prob}\ (p_1 {*}\ p_2 {*}\ {*}\ {*}\ p_n)$$

- 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
- Or, finds the coefficients that make -2 times the log of the likelihood function (-2LL) as small as possible
- The maximum likelihood estimates solve the following condition:

$$\{Y - p(Y=1)\}X_i = 0$$

summed over all observations, $i = 1,\ldots,n$

Interpreting Coefficients

- Since:

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

  The slope coefficient (β) is interpreted as the rate of change in the "log odds" as X changes … not very useful.

- Since:

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

  The marginal effect of a change in X on the probability is: $\partial p/\partial X = f(\beta X)\,\beta$

- An interpretation of the logit coefficient which is usually more intuitive is the "odds ratio"
- Since:

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

  $\exp(\beta)$ is the effect of the independent variable on the "odds ratio"

From Program Output:

Variable	B	Exp(B)	1/Exp(B)
PETS	-0.6593	0.5172	1.933
MOBLHOME	1.5583	4.7508	
TENURE	-0.0198	0.9804	1.020
EDUC	0.0501	1.0514	
Constant	-0.916		

“Households with pets are 1.933 times more likely to evacuate than those without pets.”

The logistic transformation (1)

- This type of relationship is described by a special formula.
  - Remember, if the relationship was linear then the equation is just:

$$\boxed{\pi = \alpha + \beta X}$$

  - But the relationship on the graph is actually described by:

$$\boxed{\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X}$$

The logistic transformation (2)

$$\boxed{\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X}$$

This is just the *odds*.

As the probability increases (from zero to 1), the odds increase from 0 to infinity.

The log of the odds then increases from –infinity to +infinity.

So if β is 'large' then as X increases the log of the odds will increase steeply.

The steepness of the curve will therefore increase as β gets bigger.

Fitting this model (1)

- So that’s what we want to do, but how do we do it?
  - With SLR we tried to minimize the squares of the residuals, to get the best fitting line.
  - This doesn’t really make sense here (remember the errors won’t be normally distributed as there’s only two values).
- We use something called *maximum likelihood* to estimate what the β and α are.

Fitting this model (2)

- *Maximum likelihood* is an *iterative process* that estimates the best fitted equation.
  - The iterative bit just means that we try lots of models until we get to a situation where tweaking the equation any further doesn't improve the fit.
  - The maximum likelihood bit is kind of complicated, although the underlying assumptions are simple to understand, and very intuitive. The basic idea is that we find the coefficient value that makes the observed data most likely.

Tiger results

Variable	**Coefficient value**	**Standard error**	***p*-value**
Concentration	-0.07	0.01	0.00
Intercept	3.69	0.72	0.00

- This is how logistic regression results are often reported in articles.
    - It’s clear that concentration span has a negative (and statistically significant) effect on Tiger sightings.
    - But what does the -0.07 actually mean?

Interpreting the coefficients (1)

- What we need to do is think about the equation again, and what an increase in X means.
- So an increase in X of 1 unit will decrease our log (odds) by 0.07.
- If we *antilog* both sides then we could see how the odds change…

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = a + bX$$

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 3.69 - 0.07X$$

Remember the 'hat' sign means the predicted value.

Interpreting the coefficients (2)

- Antilog both sides and we get the odds on the LH side.
- If we enter a value of *X* we can work out what the predicted odds will be.
- Thus the odds of spotting the Tiger (as opposed to not spotting the Tiger) are nearly 5. For every 5 spotters there should be one non-spotter

$$\frac{\hat{\pi}}{1-\hat{\pi}} = e^{a+bX}$$

$$\frac{\hat{\pi}}{1-\hat{\pi}} = e^{3.69-0.07X}$$

$$\frac{\hat{\pi}}{1-\hat{\pi}} = e^{3.69-0.07X}$$

$$\frac{\hat{\pi}}{1-\hat{\pi}} = e^{3.69-0.07\times 30} = 4.90$$

Interpreting the coefficients (3)

- We can also think about what happens to the odds when we increase *X* by a certain amount.
- Another way of writing e^{a+bX} is $e^{a}(e^{b})^{X}$. That means that a one unit increase in *X* multiples the odds by e^{b} (as it's to the power of 1).
- In our case therefore a one unit increase in *X* multiplies the odds by $e^{-0.07}$, or 0.93.
  - When *X* increases from 30 to 31, the odds are 4.90*0.93, or 4.56.
  - When *X* increases from 30 to 40, the odds are $4.90*(0.93)^{10}$, or 2.37.

Yet more coefficient interpretation (1)

- The other way of thinking about things is in terms of probabilities.
- If we rearrange the 'antilogged' equation then we work out what the probability (for a particular value of *X*) would be.
- The probability of a person with CS=30 of Tiger spotting is thus 83%.

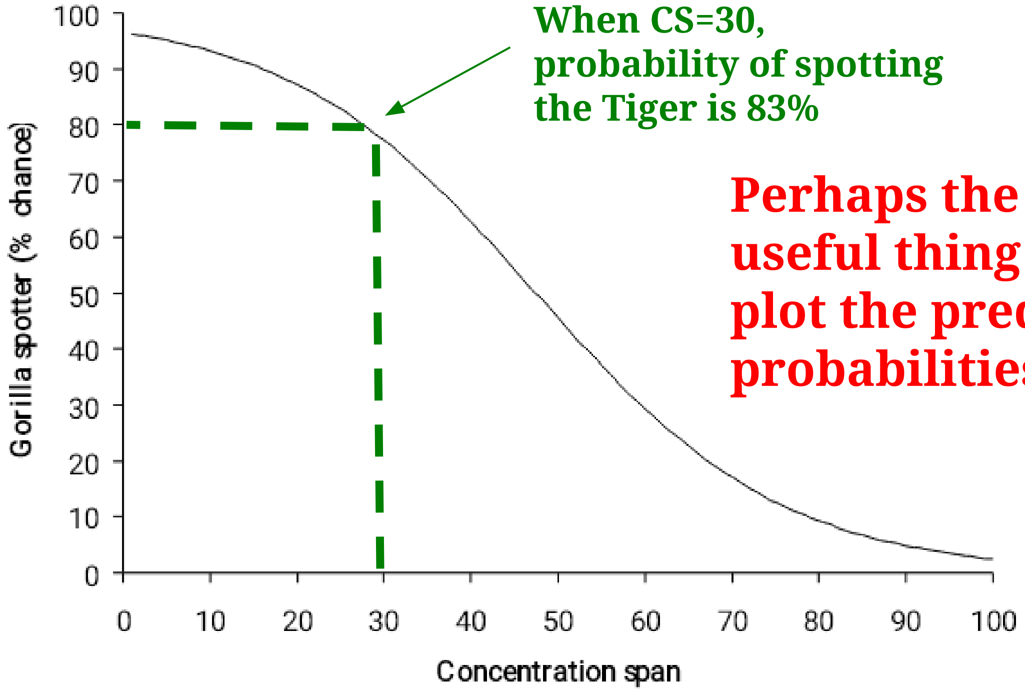
$$\frac{\hat{\pi}}{1-\hat{\pi}} = e^{3.69-0.07X}$$

$$\hat{\pi} = \frac{e^{3.69-0.07X}}{1+e^{3.69-0.07X}}$$

$$\hat{\pi} = \frac{e^{3.69-0.07X}}{1+e^{3.69-0.07X}}$$

$$\hat{\pi} = \frac{e^{3.69-0.07\times30}}{1+e^{3.69-0.07\times30}} = 0.83$$

Yet more coefficient interpretation (2)



Perhaps the most useful thing to do is to plot the predicted probabilities

Comparing models (1)

- One of the most important differences between logistic regression and linear regression is in how we compare models.
  - Remember for linear regression we looked at how the adjusted R^2 changed. If there was a significant increase when we added another variable (or interaction) then we thought the model had improved.
- For logistic regression there are a variety of ways of looking model improvement.

Comparing models (2)

- The best way of comparing models is to use something called the *likelihood-ratio test*.
  - When we were using OLS regression, we were trying to minimize the sum of squares, for logistic regression we are trying to maximize something called the *likelihood function* (normally called L).
  - To see whether our model has improved by adding a variable (or interaction, or squared term), we can compare the maximum of the likelihood function for each model (just like we compared the R^2 before for OLS regressions).

Comparing models (3)

- In fact, just to complicate matters we actually compare the maximised values of -2*log *L*.

$$\boxed{LR = (-2 \times \log L_o) - (-2 \times \log L_1)}$$

First model's maximised value

Second model's maximised value

- By logging the *L*s and multiplying them by -2, this statistic conveniently ends up with a chi-square distribution. This means we test whether there is a statistically significant improvement with reference to the χ^2 distribution.

Linear Regression	Logistic Regression
Used to predict the continuous dependent variable using a given set of independent variables.	Used to predict the categorical dependent variable using a given set of independent variables.
The outputs produced must be a continuous value, such as price and age.	The outputs produced must be Categorical values such as 0 or 1, Yes or No.
The relationship between the dependent variable and independent variable must be linear.	The relationship DOES NOT need to be linear between the dependent and independent variables.
Used for solving Regression problems.	Used for solving Classification problems.
We are finding and using the line of best fit to help us easily predict outputs.	We are using the S-curve (Sigmoid) to help us classify predicted outputs.
Least square estimation method is used for the estimation of accuracy.	Maximum likelihood estimation method is used for the estimation of accuracy.
There is a possibility of collinearity between the independent variables.	There should not be any collinearity between the independent variable.