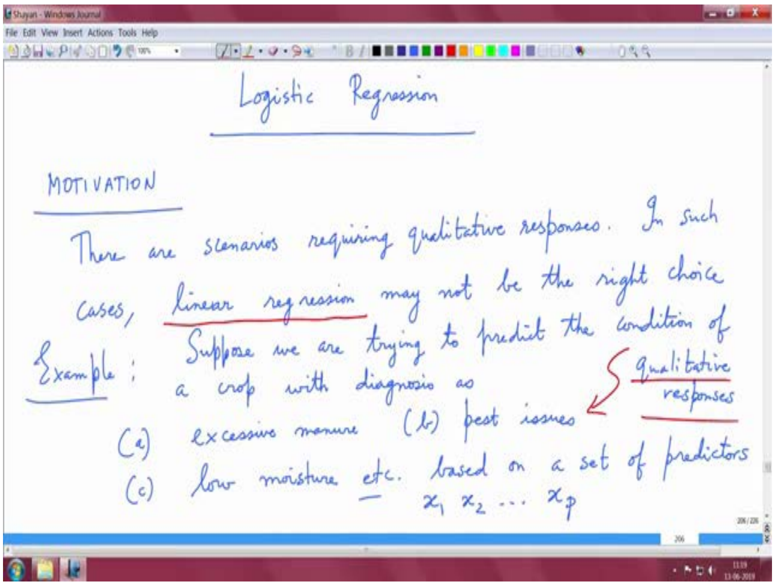
Neural Networks for Signal Processing-I

Prof. Shayan Srinivasa Garani
Department of Electronic System Engineering
Indian Institute of Science, Bengaluru

Lecture – 15
Logistic Regression

Let's dive into the concept of logistic regression. In the previous module, we explored linear regression, and now, in this module, we will shift our focus to logistic regression.

(Refer Slide Time: 00:44)



Logistic Regression

MOTIVATION

There are scenarios requiring qualitative responses. In such cases, linear regression may not be the right choice

Example: Suppose we are trying to predict the condition of a crop with diagnosis as

(a) excessive manure (b) pest issues

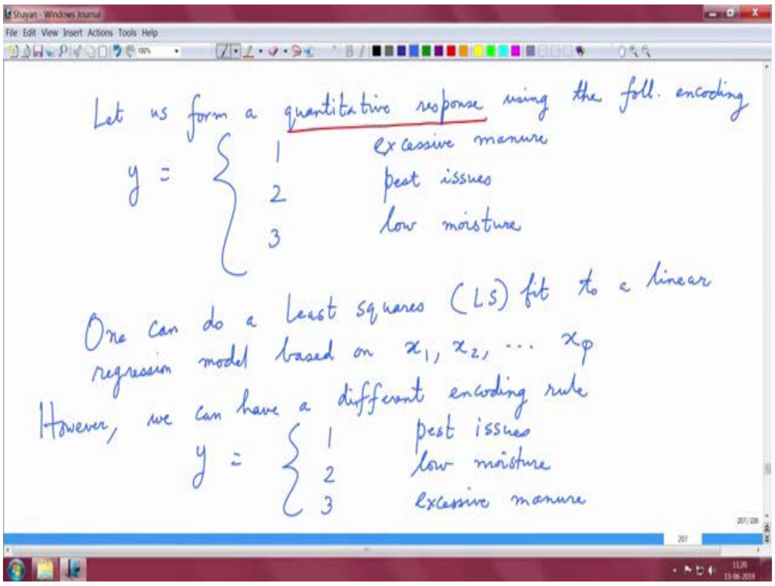
(c) low moisture etc. based on a set of predictors $x_1 \; x_2 \ldots x_p$

qualitative responses

To understand the motivation behind logistic regression, consider that there are situations requiring qualitative responses, where linear regression might not be appropriate. In the previous module, we dealt with linear regression, which is used when the response variable is quantitative. However, when dealing with qualitative responses, linear regression isn't the best fit.

Let's explore an example: imagine we're trying to diagnose the condition of a crop. The diagnosis could involve qualitative outcomes such as excessive manure, pest issues, low moisture content, and so forth, based on a set of predictors $(x_1, x_2, \ldots, x_p)$. In such cases, where the response is qualitative rather than quantitative, logistic regression becomes a more suitable approach.

(Refer Slide Time: 01:47)



Let us form a quantitative response using the foll. encoding

$$y = \begin{cases} 1 & \text{excessive manure} \\ 2 & \text{pest issues} \\ 3 & \text{low moisture} \end{cases}$$

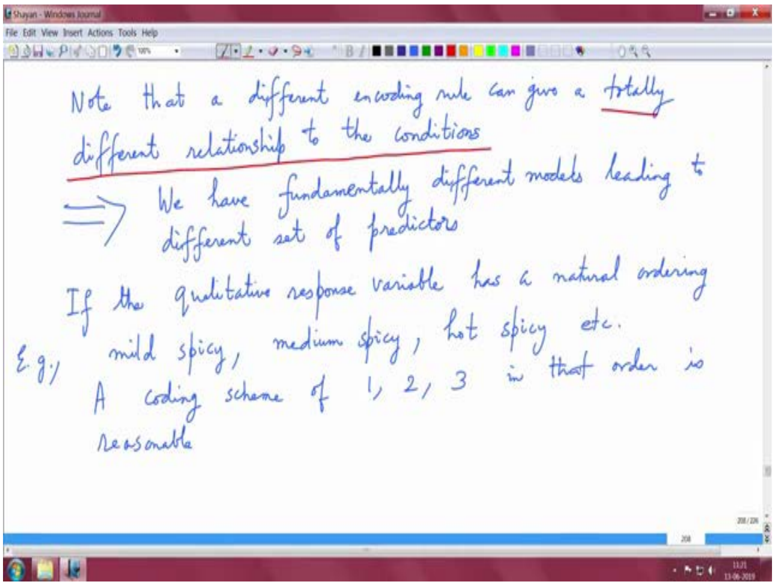
One can do a least squares (LS) fit to a linear regression model based on $x_1, x_2, \dots x_p$

However, we can have a different encoding rule

$$y = \begin{cases} 1 & \text{pest issues} \\ 2 & \text{low moisture} \\ 3 & \text{excessive manure} \end{cases}$$

Let's consider forming a quantitative response for the crop condition problem using the following encoding scheme. We'll assign the variable y three distinct values: 1, 2, and 3. Specifically, y = 1 if the condition is excessive manure, y = 2 if there are pest issues, and y = 3 if there is low moisture content.

(Refer Slide Time: 02:59)



Note that a different encoding rule can give a totally different relationship to the conditions

$\Longrightarrow$ We have fundamentally different models leading to different set of predictors

If the qualitative response variable has a natural ordering

E.g., mild spicy, medium spicy, hot spicy etc.

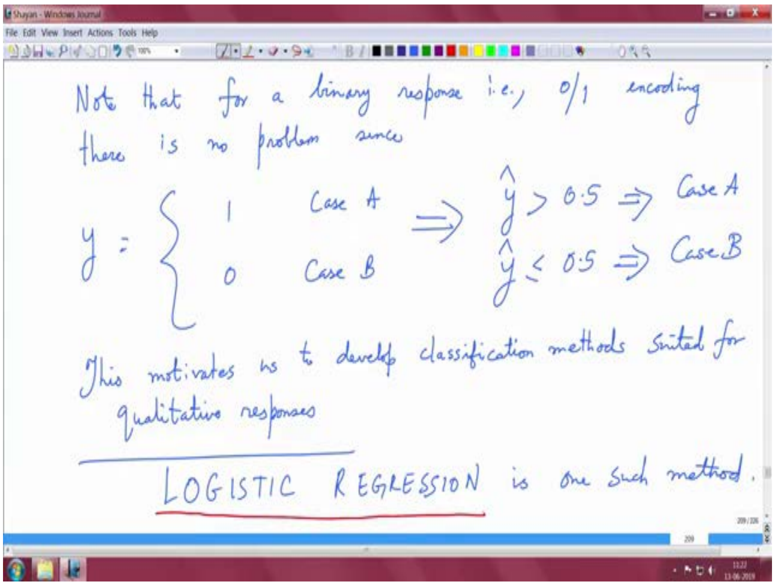
A coding scheme of 1, 2, 3 in that order is reasonable

One can use least squares fitting to apply a linear regression model based on predictors $x_1, x_2, \dots, x_p$. However, it's also possible to use an alternative encoding scheme. For instance, you could assign y = 1 for pest issues, y = 2 for low moisture content, and y = 3 for excessive manure in the soil. Different encoding rules can be employed depending on how you want to represent the qualitative responses.

Why do we use different encoding rules? The choice of encoding can significantly alter the relationships between the predictors and the conditions. This is crucial because it can lead to fundamentally different models and, consequently, different sets of predictors.

This issue is less problematic when the qualitative response variable has a natural order. For instance, consider a restaurant where the food is categorized by spice level: mild, medium, and hot. Using an encoding scheme where 1 represents mild spice, 2 represents medium spice, and 3 represents hot spice is reasonable because there is an inherent natural order to these categories. However, in the crop production problem, there was no such natural ordering for the response variables, which makes the choice of encoding even more critical.

(Refer Slide Time: 04:05)



Note that for a binary response i.e., 0/1 encoding there is no problem since

$$y = \begin{cases} 1 & \text{Case A} \\ 0 & \text{Case B} \end{cases} \Rightarrow \begin{array}{l} \hat{y} > 0.5 \Rightarrow \text{Case A} \\ \hat{y} \leq 0.5 \Rightarrow \text{Case B} \end{array}$$

This motivates us to develop classification methods suited for qualitative responses

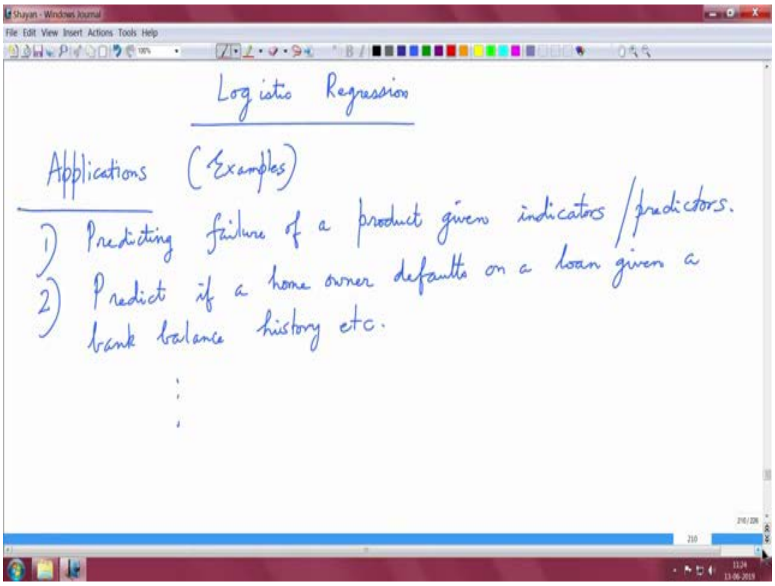
LOGISTIC REGRESSION is one such method.

When dealing with binary responses using a 0-1 encoding scheme, the process is straightforward. For instance, if we encode the response variable y as 1 for case A and 0 for case B, then if our predicted value $\hat{y}$ is greater than 0.5, we interpret it as case A. Conversely, if $\hat{y}$ is less than or equal to 0.5, we interpret it as case B. This method works well for binary responses.

However, the situation becomes more complex when dealing with qualitative responses that involve more than two categories, such as ternary or quaternary responses. In these cases, we need to develop classification methods that are suited to handle multiple categories.

One such method is logistic regression. This technique becomes particularly useful in these scenarios and motivates us to explore it further. We will first examine logistic regression in the context of binary responses, and then extend our understanding to cases involving multiple predictors and multi-class problems.

(Refer Slide Time: 05:31)



Logistic Regression

Applications (Examples)

1) Predicting failure of a product given indicators/predictors.

2) Predict if a home owner defaults on a loan given a bank balance history etc.

:

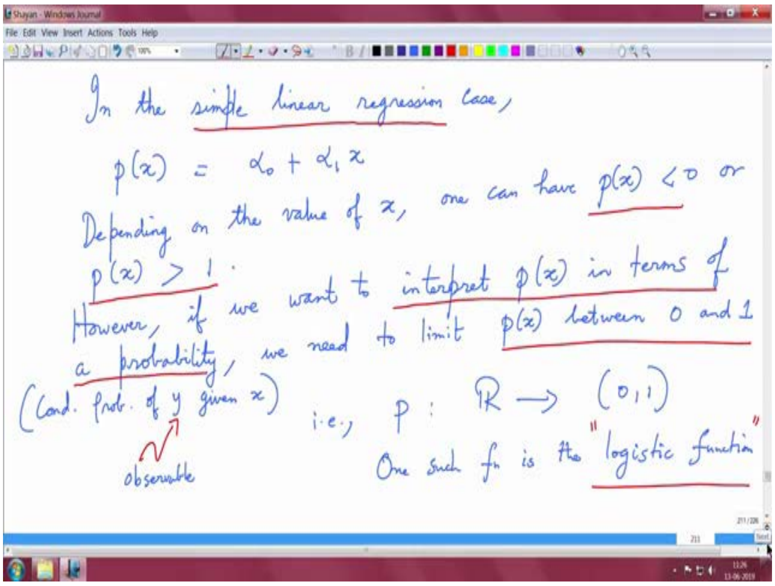
Before we dive into the details of the logistic regression model, let’s explore some practical applications where it can be effectively utilized. Here are a couple of notable examples:

1. Predicting Product Failure: Logistic regression can be employed to predict whether a product will fail based on various indicators. These indicators could include a range of predictive variables that help assess the likelihood of failure.

2. Bank Loan Default Prediction: A practical application is in the realm of banking, where logistic regression can be used to predict whether a homeowner might default on a loan based on their bank balance history. This tool can be instrumental for banks in making informed decisions about issuing loans and evaluating the financial health of loan applicants.

These examples illustrate the versatility and usefulness of logistic regression in addressing real-world problems.

(Refer Slide Time: 06:36)



In the simple linear regression case,

$p(x) = \alpha_0 + \alpha_1 x$

Depending on the value of x, one can have $p(x) < 0$ or $p(x) > 1$.

However, if we want to interpret $p(x)$ in terms of a probability, we need to limit $p(x)$ between 0 and 1

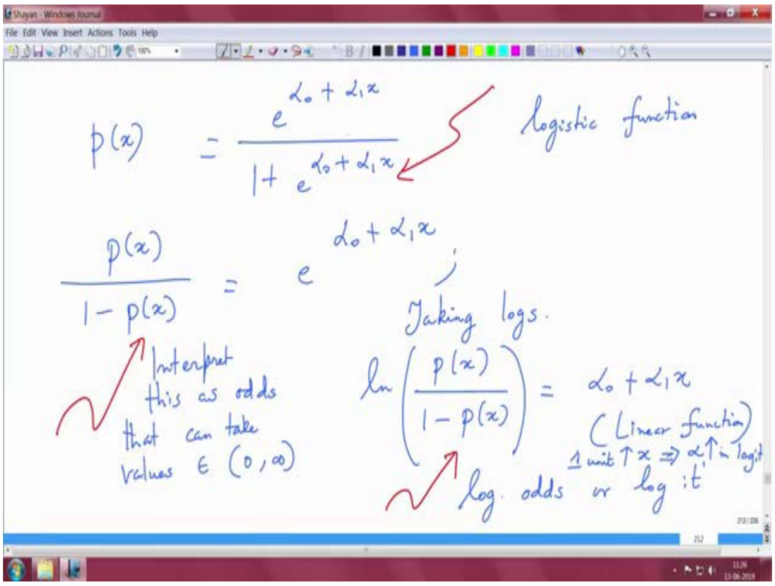
(Cond. Prob. of y given x) i.e., $p : \mathbb{R} \rightarrow (0,1)$

observable

One such fn is the "logistic function"

Let’s start with a basic example of linear regression. Consider a simple linear regression model where p(x) = α0 + α1 x, with x being the predictor variable. In this model, depending on the value of x and the coefficients α0 and α1, p(x) can potentially fall outside the range of 0 to 1. This can be problematic if we wish to interpret p(x) as a probability, since probabilities must lie between 0 and 1.

(Refer Slide Time: 08:18)



$$p(x) = \frac{e^{\alpha_0 + \alpha_1 x}}{1 + e^{\alpha_0 + \alpha_1 x}}$$

logistic function

$$\frac{p(x)}{1 - p(x)} = e^{\alpha_0 + \alpha_1 x};$$

Taking logs.

$$\ln\left(\frac{p(x)}{1 - p(x)}\right) = \alpha_0 + \alpha_1 x$$

(Linear function)

1 unit ↑ $x \Rightarrow \alpha_1 \uparrow$ in logit

Interpret this as odds that can take values $\in (0, \infty)$

log odds or logit

To address this issue, we need p(x) to be constrained within the interval [0, 1] because it must represent a valid probability function. One function that naturally satisfies this constraint is the logistic function. The logistic function maps any real-valued number into a value between 0 and 1, making it an ideal choice for ensuring that our predictions are interpreted as valid probabilities.

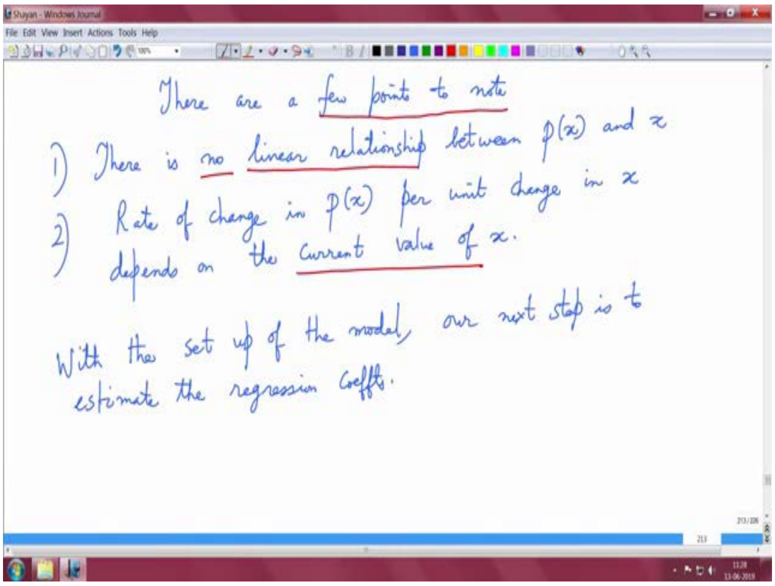
Let’s delve into the details of the logistic function. The logistic function is expressed as a ratio of two terms. The numerator is $e^{\alpha_0+\alpha_1 x}$, and the denominator is $1 + e^{\alpha_0+\alpha_1 x}$. This function can be interpreted in terms of odds. For instance, if p(x) represents the probability of getting a head, then 1 - p(x) represents the probability of getting a tail.

In this context, the odds ratio $\frac{p(x)}{1-p(x)}$ can be expressed as $e^{\alpha_0+\alpha_1 x}$. This ratio ranges from 0 to infinity because p(x) lies between 0 and 1. Taking the natural logarithm of this odds ratio, we get:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \alpha_0 + \alpha_1 x.$$

This equation is linear in x, meaning that a one-unit increase in x results in an increase of α_1 units in the logarithm of the odds. This log-odds is a key component in interpreting the likelihood function in logistic regression.

(Refer Slide Time: 10:10)



There are a few point to note

1) There is no linear relationship between $p(x)$ and x

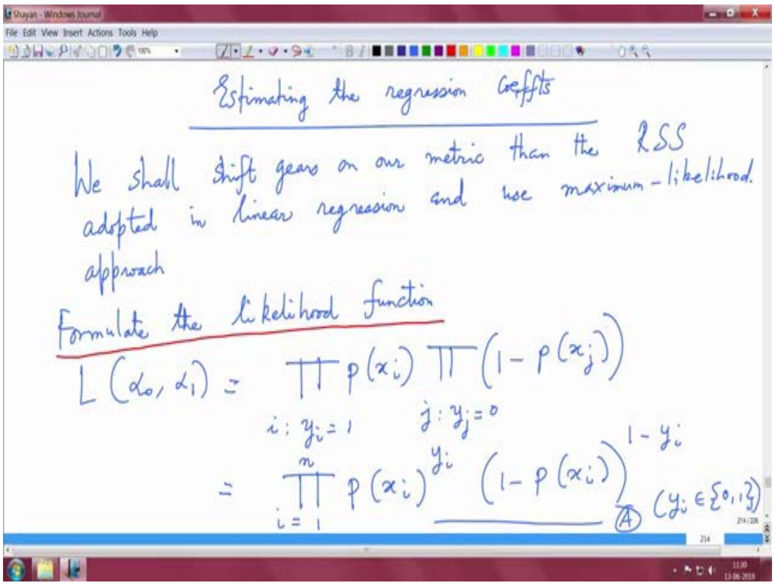
2) Rate of change in $p(x)$ per unit change in x depends on the current value of x.

With the set up of the model, our next step is to estimate the regression coeffts.

Here are a few key points to consider. First, there is no direct linear relationship between p(x) and x. Instead, the rate at which p(x) changes with each unit change in x depends on the current value of x. This means that the effect of x on p(x) is not constant but varies depending on x.

With this model setup in mind, our next task is to estimate the regression coefficients, namely α_0 and α_1.

(Refer Slide Time: 10:46)



Estimating the regression Coeffts

We shall shift gears on our metric than the RSS adopted in linear regression and use maximum-likelihood. approach

Formulate the likelihood function

$$L(\alpha_0, \alpha_1) = \prod_{i:\, y_i=1} p(x_i) \prod_{j:\, y_j=0} \left(1 - p(x_j)\right)$$

$$= \prod_{i=1}^{n} p(x_i)^{y_i} \underline{\left(1 - p(x_i)\right)^{1-y_i}} \quad \textcircled{A} \quad (y_i \in \{0, 1\})$$

Now, let's delve into how we address this problem by changing our approach. In linear regression, we used the residual sum of squares as our metric, which led us to optimize a quadratic cost function. However, in the case of logistic regression, we turn our attention to the maximum likelihood function, which offers superior statistical properties.

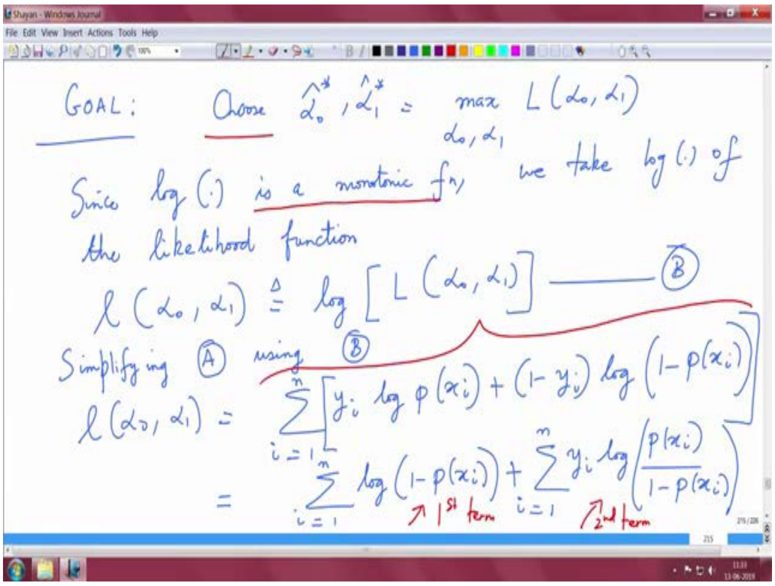
To formulate the likelihood function, denoted as $\mathcal{L}$, we express it as follows:

$$\mathcal{L}(\alpha_0, \alpha_1) = \prod_{y_i=1} p(x_i) \times \prod_{y_j=0} \left(1 - p(x_j)\right)$$

Here, $\prod_{y_i=1} p(x_i)$ represents the product of probabilities for observations where y_i is 1, and $\prod_{y_j=0} \left(1 - p(x_j)\right)$ represents the product of the complements of probabilities for observations where y_j is 0. This can be compactly written in the given form, where y_i takes on the values 0 or 1, and the corresponding probability for each observation is $p(x_i)$.

The goal is to maximize this likelihood function, $\mathcal{L}(\alpha_0, \alpha_1)$, with respect to the unknown parameters α0 and α1. By doing so, we aim to find the parameters that best fit our model and accurately describe the data.

(Refer Slide Time: 12:40)



GOAL: Choose $\hat{\alpha}_0^*, \hat{\alpha}_1^* = \max_{\alpha_0, \alpha_1} L(\alpha_0, \alpha_1)$

Since $\log(\cdot)$ is a monotonic fn, we take $\log(\cdot)$ of the likelihood function

$$l(\alpha_0, \alpha_1) \triangleq \log\left[L(\alpha_0, \alpha_1)\right] \text{ ——— } Ⓑ$$

Simplifying Ⓐ using Ⓑ

$$l(\alpha_0, \alpha_1) = \sum_{i=1}^{n}\left[y_i \log p(x_i) + (1-y_i)\log\left(1-p(x_i)\right)\right]$$

$$= \underbrace{\sum_{i=1}^{n} \log\left(1-p(x_i)\right)}_{\nearrow \text{1st term}} + \underbrace{\sum_{i=1}^{n} y_i \log\left(\frac{p(x_i)}{1-p(x_i)}\right)}_{\nearrow \text{2nd term}}$$

Thus, we need to determine the optimal estimates of the parameters α0 and α1. The asterisk (*) denotes optimality, while the hat (^) indicates that these are estimates maximizing the likelihood function.

Looking at the likelihood function itself, it appears quite complex due to the product of terms, making direct optimization challenging. To simplify, we use a well-known technique: taking the logarithm of the likelihood function. The advantage of this approach is that the logarithm is a monotonic function, meaning it preserves the order of the values and does not change the fundamental properties of the function.

We define $l(\alpha_0, \alpha_1)$ as the logarithm of the likelihood function, which allows us to simplify the expression. After performing the mathematical operations, we obtain the following form:

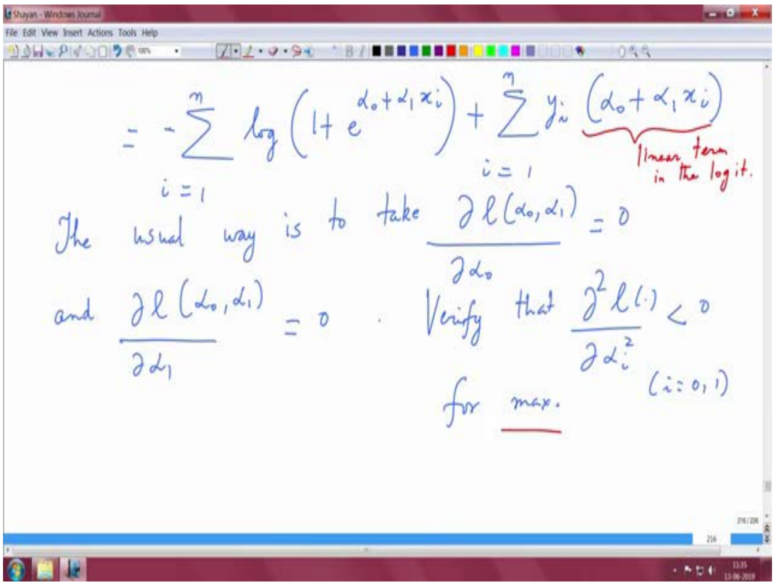
$$l(\alpha_0, \alpha_1) = \sum_{i=1}^{n}\left[y_i \log p\,(x_i) + (1-y_i)\log\left(1-p(x_i)\right)\right]$$

This form is more manageable. By rearranging the equation, we can isolate yi in one term and handle the rest separately. Thus, we simplify it to:

$$\sum_{i=1}^{n} \log\big(1 - p(x_i)\big) + \sum_{i=1}^{n} y_i \log\left(\frac{p(x_i)}{1 - p(x_i)}\right)$$

This simplification is straightforward. The second term, $\log\left(\frac{p(x_i)}{1-p(x_i)}\right)$, represents the log of the odds, which ties back to the linear relationship we discussed earlier. We will use this property to further simplify our analysis.

(Refer Slide Time: 15:57)



$$= -\sum_{i=1}^{n} \log\left(1 + e^{\alpha_0 + \alpha_1 x_i}\right) + \sum_{i=1}^{n} y_i \underbrace{(\alpha_0 + \alpha_1 x_i)}_{\text{linear term in the logit.}}$$

The usual way is to take $\frac{\partial \ell(\alpha_0, \alpha_1)}{\partial \alpha_0} = 0$ and $\frac{\partial \ell(\alpha_0, \alpha_1)}{\partial \alpha_1} = 0$. Verify that $\frac{\partial^2 \ell(\cdot)}{\partial \alpha_i^2} < 0$ $(i = 0, 1)$ for max.

Now, let's examine the quantity we have. We obtain:

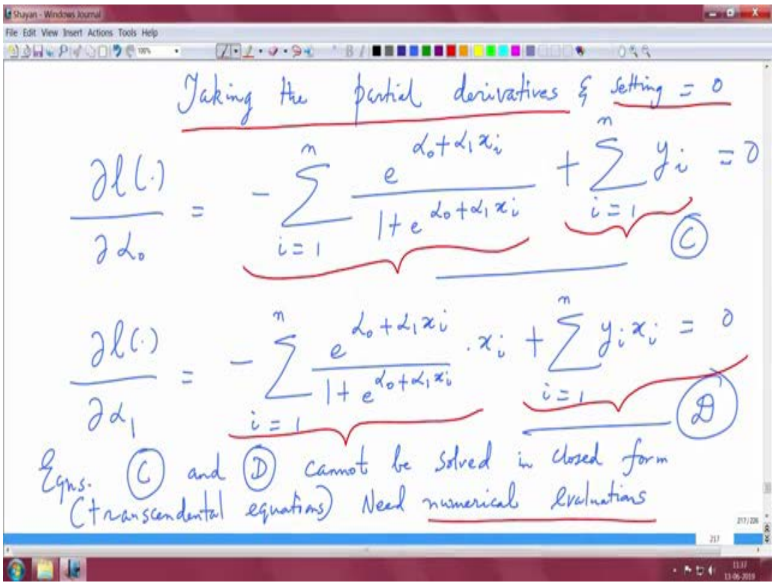
$$-\sum_{i=1}^{n} \log(1 + e^{\alpha_0 + \alpha_1 x_i}) + \sum_{i=1}^{n} y_i(\alpha_0 + \alpha_1 x_i)$$

Here, the first term comes from the logistic regression model, reflecting the log of the odds, and the second term arises from the linear component within the logistic function. With this simplified expression in hand, we can proceed to calculus to find the optimal values of α_0 and α_1.

To do this, we compute the partial derivatives of this expression with respect to α_0 and α_1. We then set these partial derivatives to zero to find the critical points. To confirm that these critical points correspond to a maximum, we check the second derivatives to ensure they are less than zero, which is the condition for a maximum.

This process involves standard calculus techniques.

(Refer Slide Time: 17:06)



Taking the partial derivatives & setting = 0

$$\frac{\partial \ell(\cdot)}{\partial \alpha_0} = \underbrace{-\sum_{i=1}^{n} \frac{e^{\alpha_0+\alpha_1 x_i}}{1+e^{\alpha_0+\alpha_1 x_i}}} + \underbrace{\sum_{i=1}^{n} y_i} = 0 \quad \textcircled{C}$$

$$\frac{\partial \ell(\cdot)}{\partial \alpha_1} = \underbrace{-\sum_{i=1}^{n} \frac{e^{\alpha_0+\alpha_1 x_i}}{1+e^{\alpha_0+\alpha_1 x_i}} \cdot x_i} + \underbrace{\sum_{i=1}^{n} y_i x_i} = 0 \quad \textcircled{D}$$

Eqns. $\textcircled{C}$ and $\textcircled{D}$ cannot be solved in closed form (transcendental equations) Need numerical evaluations

To determine the optimal values for α0 and α1, we begin by taking the partial derivatives of the log-likelihood function with respect to each parameter and setting these derivatives equal to zero.

First, let's find the partial derivative with respect to α0. The derivative of the log-likelihood function involves the term:

$$\frac{1}{1+e^{\alpha_0+\alpha_1 x_i}}$$

To find the derivative, we apply the chain rule. The derivative of the denominator $1+e^{\alpha_0+\alpha_1 x_i}$ is:

$$e^{\alpha_0+\alpha_1 x_i}$$

The coefficient in front of α0 is 1. Thus, after simplifying, we obtain:

$$-\sum_{i=1}^{n}\left(\frac{e^{\alpha_0+\alpha_1 x_i}}{1+e^{\alpha_0+\alpha_1 x_i}}\right)+\sum_{i=1}^{n} y_i$$

Next, we compute the partial derivative with respect to α1. Here, the derivative involves:

$$\frac{1}{1+e^{\alpha_0+\alpha_1 x_i}}$$

Taking the derivative of this term with respect to α1 gives:

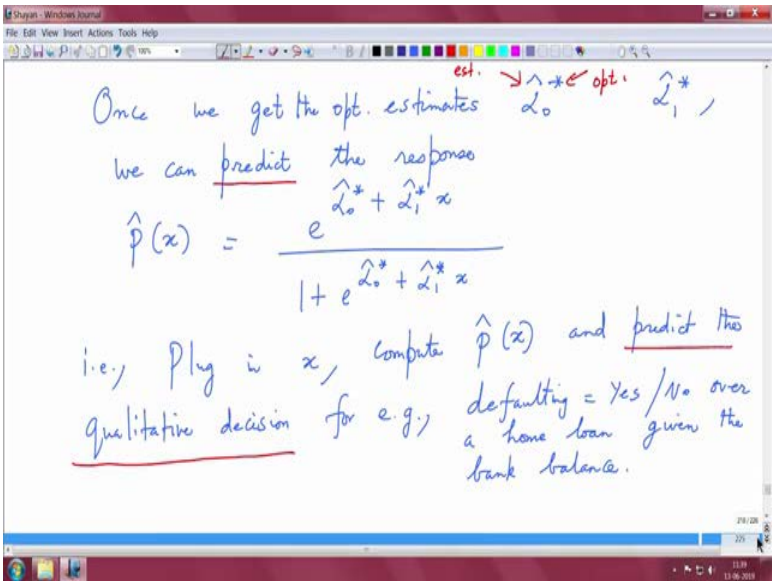
$$e^{\alpha_0+\alpha_1 x_i} \cdot x_i$$

Thus, the partial derivative with respect to α1 is:

$$-\sum_{i=1}^{n}\left(\frac{e^{\alpha_0+\alpha_1 x_i} \cdot x_i}{1+e^{\alpha_0+\alpha_1 x_i}}\right)+\sum_{i=1}^{n} y_i x_i$$

Now, we have derived the necessary equations, but solving these equations analytically is challenging because they are transcendental equations. This complexity arises from the presence of the exponential function and summation terms. Therefore, finding the exact solutions for α0 and α1 typically requires numerical methods, which are beyond the scope of our current discussion. Once the partial derivatives are set up, you would use numerical solvers to find the optimal parameter estimates.

(Refer Slide Time: 19:59)



Once we get the opt. estimates $\overset{\text{est.} \searrow}{\hat{\alpha}_0^{* \swarrow \text{opt.}}}$ $\hat{\alpha}_1^*$,

we can predict the response

$$\hat{p}(x) = \frac{e^{\hat{\alpha}_0^* + \hat{\alpha}_1^* x}}{1+e^{\hat{\alpha}_0^* + \hat{\alpha}_1^* x}}$$

i.e., Plug in x, compute $\hat{p}(x)$ and predict the qualitative decision for e.g., defaulting = Yes / No over a home loan given the bank balance.

Once we obtain the optimal estimates for the parameters, our goal is to predict the responses. The predicted probability, $\hat{p}(x)$, is given by the logistic function:

$$\hat{p}(x) = \frac{e^{\widehat{\alpha_0^*}+\widehat{\alpha_1^*}x}}{1+e^{\widehat{\alpha_0^*}+\widehat{\alpha_1^*}x}}$$

Here, $\widehat{\alpha_0^*}$ and $\widehat{\alpha_1^*}$ denote the optimal estimates for the parameters, with the hat symbol indicating an estimate and the star denoting optimal conditions.

With these optimal estimates, you can predict the qualitative outcome by interpreting the result as a probability. For instance, given information like a bank balance, you can determine whether a loan default is likely or not.

We have thus formulated a simple model for binary logistic regression with a single predictor variable. The natural extension of this model involves predicting binary outcomes given multiple predictors.

Before diving into the details of extending the model, let's consider some motivating examples where logistic regression is particularly useful. Suppose I am an undergraduate student deciding between pursuing a degree in pure science or engineering. My decision might be influenced by my grades, interests, and preferences. Using these inputs, logistic regression can help determine whether I should opt for pure science or engineering.

Another example is in an election scenario where an individual must choose between two political parties. Given demographic characteristics and personal preferences, logistic regression can predict whether the individual is likely to vote for Party A or Party B. Additionally, logistic regression can be used to forecast election results based on exit polls and other relevant factors.

With these applications in mind, let's explore how to formulate a logistic regression model for multiple predictors. As discussed earlier, the logarithm of the odds ratio:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \alpha_0 + \alpha_1 x$$

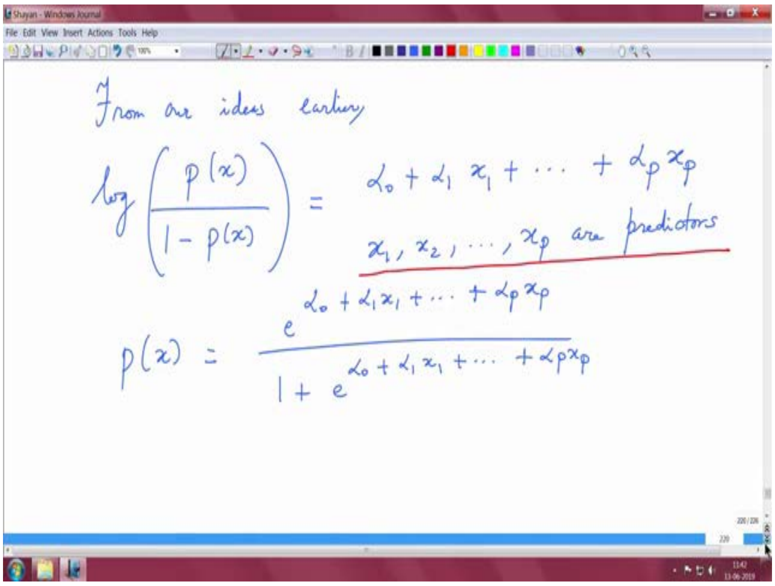
was shown to be a linear function.

This concept can be extended to accommodate multiple predictor variables $x_1, x_2, \ldots, x_p$. We can express the log-odds ratio as:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p$$

Here, $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_p$ are the parameters we need to estimate, and $x_1, x_2, \ldots, x_p$ are the predictor variables.

(Refer Slide Time: 23:17)



From our ideas earlier

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p$$

$x_1, x_2, \dots, x_p$ are predictors

$$p(x) = \frac{e^{\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p}}{1 + e^{\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p}}$$

Through straightforward algebraic manipulation, we derive the expression for the predicted probability p(x). It is given by:

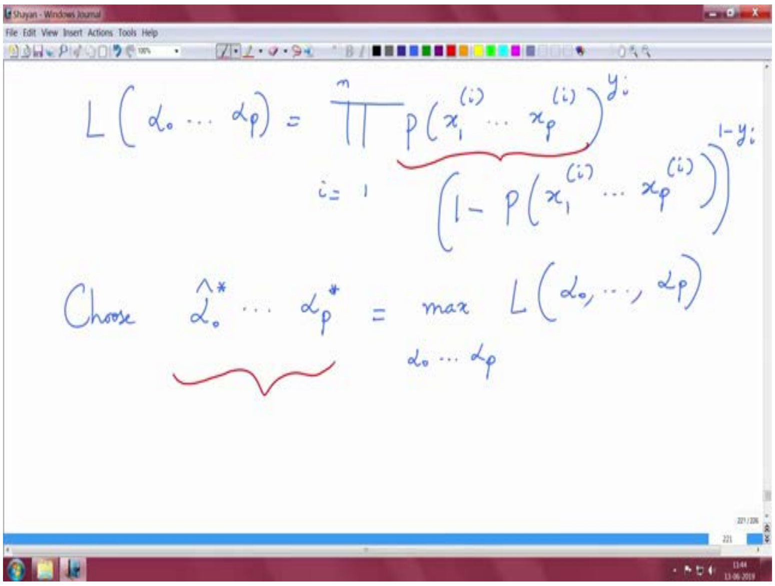
$$p(x) = \frac{e^{\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\cdots+\alpha_p x_p}}{1+e^{\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\cdots+\alpha_p x_p}}$$

This represents the ratio of two terms, where the numerator is the exponential function of the linear combination of the predictor variables, and the denominator is one plus the same exponential function.

To set up the likelihood function, we start by considering the binary response variable y_i, which can be either 0 or 1. We need to look at this in terms of the prediction variables $x_{1i}, x_{2i}, \dots, x_{pi}$ for each observation i. So, effectively, we have two types of variables: i representing the observation, and y_i representing the binary response, with x_1, x_2, …, x_p as the predictor variables.

The likelihood function for this scenario can be formulated in a manner similar to how we approached the binary case with a single predictor variable. The goal is to estimate the parameters α_0, α_1, …, α_p that maximize this likelihood function. The process should be fairly intuitive if you're familiar with the binary case; it follows a similar logic.

(Refer Slide Time: 24:34)

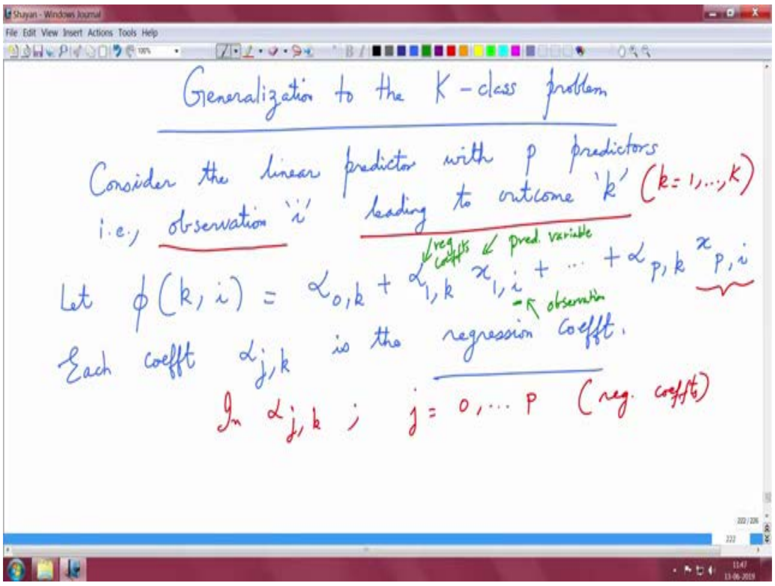


$$L(\alpha_0 \dots \alpha_p) = \prod_{i=1}^{m} \underbrace{p(x_1^{(i)} \dots x_p^{(i)})}^{y_i} \left(1 - p(x_1^{(i)} \dots x_p^{(i)})\right)^{1-y_i}$$

$$\text{Choose } \underbrace{\hat{\alpha}_0^* \dots \alpha_p^*} = \max_{\alpha_0 \dots \alpha_p} L(\alpha_0, \dots, \alpha_p)$$

To simplify the problem, you might make certain assumptions that allow you to interpret the likelihood function more easily. By breaking down the joint probability term with these assumptions, you can simplify the likelihood equation and proceed to perform maximum likelihood estimation (MLE) for the parameters.

(Refer Slide Time: 26:28)



Generalization to the K-class problem

Consider the linear predictor with p predictors leading to outcome 'k' $(k = 1, \dots, K)$ i.e., observation 'i'

$$\text{Let } \phi(k, i) = \alpha_{0,k} + \overset{\text{reg coeffts}}{\alpha_{1,k}} \overset{\text{pred. variable}}{x_{1,\underset{\text{observation}}{i}}} + \dots + \alpha_{p,k} \underbrace{x_{p,i}}$$

Each coeff $\alpha_{j,k}$ is the regression coeff.

$$\text{In } \alpha_{j,k} \; ; \quad j = 0, \dots p \quad (\text{reg. coeffts})$$

Our goal doesn't end with the binary case; we can extend this approach to a K-class problem. Let's consider a scenario with p predictors and an observation i that can lead to one of K possible outcomes. In this K-class problem, the outcome, denoted as k, can range from 1 to K. Essentially, we need to determine which of these classes an observation belongs to.

To generalize our approach for this K-class scenario, we start by formulating a linear predictor. This linear predictor, which I will denote as $\phi_{k,i}$, is given by the following equation:

$$\phi_{k,i} = \alpha_{0,k} + \alpha_{1,k}x_{1,i} + \alpha_{2,k}x_{2,i} + \cdots + \alpha_{p,k}x_{p,i}.$$

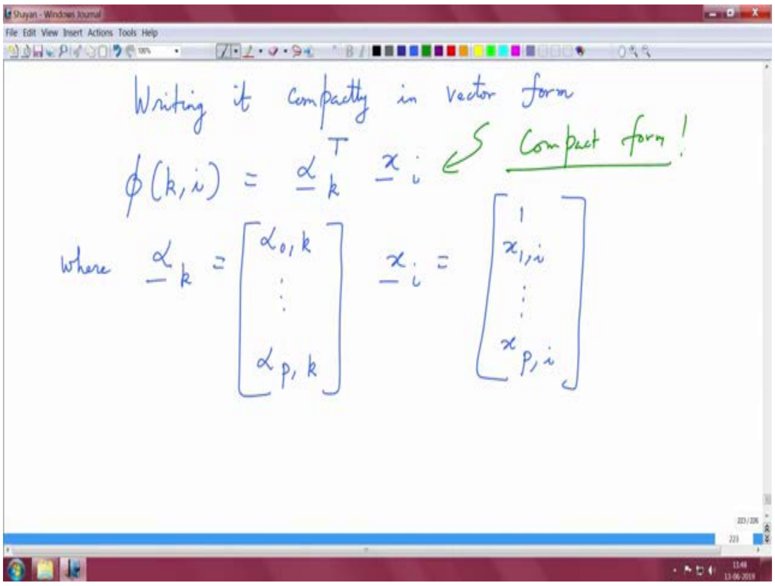
In this equation:

- $x_{1,i}, x_{2,i}, \ldots, x_{p,i}$ are the predictor variables, where the first index indicates the predictor variable, and the second index indicates the observation.
- $\alpha_{j,k}$ are the regression coefficients, where j ranges from 0 to p, and k denotes the outcome class. These coefficients are the parameters we need to estimate.

- i represents the observation index, and k represents the class label.

At first glance, this formulation might seem quite complex due to the multiple indices and coefficients involved.

(Refer Slide Time: 29:37)



Writing it compactly in vector form

$$\phi(k,i) = \underline{\alpha}_k^T \, \underline{x}_i \quad \text{Compact form!}$$

$$\text{where } \underline{\alpha}_k = \begin{bmatrix} \alpha_{0,k} \\ \vdots \\ \alpha_{p,k} \end{bmatrix} \quad \underline{x}_i = \begin{bmatrix} 1 \\ x_{1,i} \\ \vdots \\ x_{p,i} \end{bmatrix}$$

We can simplify this significantly by using vector notation. The linear predictor $\phi_{k,i}$ can be interpreted as the dot product between two vectors: α_k and x_i, represented as follows:

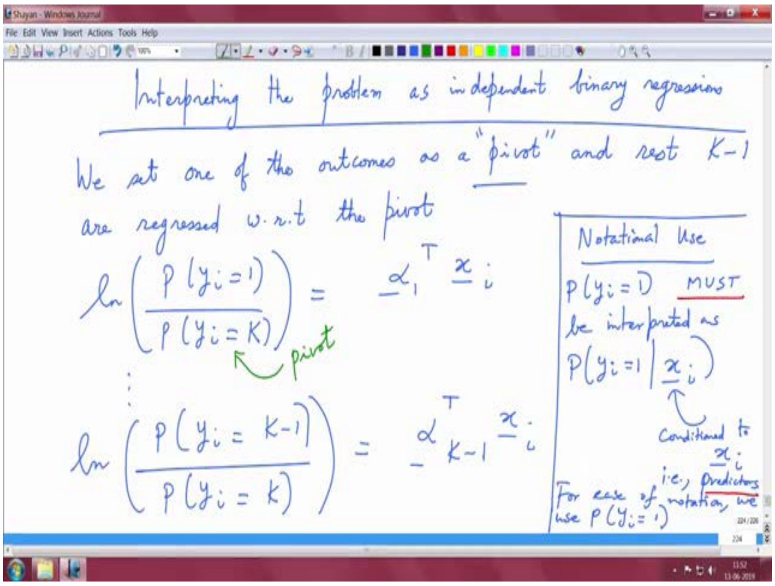
$$\phi_{k,i} = \alpha_k^T x_i.$$

Here’s what this means:

- α_k is a vector that stacks all the regression coefficients corresponding to the p predictors for outcome class k, including the intercept term.
- x_i is a vector that stacks all p predictors $x_{1,i}$ through $x_{p,i}$ for observation i. The first coordinate of x_i is typically set to 1 to account for the intercept term, consistent with how we define the linear predictor.

This compact notation makes it much easier to handle and work with the model.

(Refer Slide Time: 30:42)



Interpreting the problem as independent binary regressions

We set one of the outcomes as a "pivot" and rest K-1 are regressed w.r.t the pivot

$$\ln\left(\frac{P(y_i=1)}{P(y_i=K)}\right) = \underline{\alpha}_1^T \underline{x}_i$$

pivot

$$\vdots$$

$$\ln\left(\frac{P(y_i=K-1)}{P(y_i=K)}\right) = \underline{\alpha}_{K-1}^T \underline{x}_i$$

Notational Use

$P(y_i=1)$ MUST be interpreted as $P(y_i=1 \mid \underline{x}_i)$

Conditioned to $\underline{x}_i$ i.e., predictors

For ease of notation, we use $P(y_i=1)$

To address the K-class classification problem using independent binary regressions, we take a different approach compared to the binary regression case. In binary regression, we could easily interpret odds with two outcomes (0 versus 1). However, for a K-class problem, we need to adapt our approach since we can't directly apply the odds concept.

The strategy involves fixing one of the K outcomes as a reference or pivot class. We then compare the remaining K-1 outcomes relative to this pivot. This allows us to perform regressions with respect to this fixed pivot class.

Let’s break this down:

1. Interpretation of Probabilities:

We denote p(y_i = k) as the probability that observation i belongs to class k, given the predictor vector x_i (comprising p predictors). To simplify notation, we omit explicit conditioning on x_i for readability and write p(y_i = k) directly.

2. Formulation:

To use the logarithm of the likelihood function, we select a class k as the reference class. The log-odds of an outcome being in class k versus the reference class are given by:

$$\log \frac{p(y_i = k)}{p(y_i = K)} = \alpha_k^T x_i$$

where α_k represents the coefficients for class k and x_i is the vector of predictors for observation i. For each class k, this gives us a series of equations.

3. Simplified Formulation:

The log-probability ratio for the outcomes can be expressed as:

$$\log \frac{p(y_i = k)}{p(y_i = K)} = \alpha_k^T x_i$$

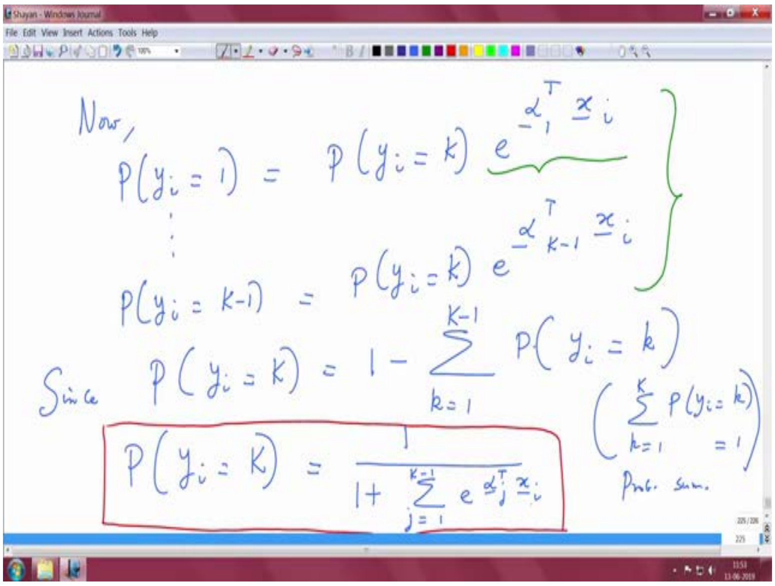
By doing this for all K-1 classes relative to the pivot class, we obtain a set of equations for each class comparison.

4. Simplification:

To predict the outcome, we use the exponential of the linear combinations. The prediction for a given outcome can be scaled by this exponential term relative to the pivot class.

By applying this approach, we effectively handle the K-class problem using multiple binary regressions, simplifying the process while accommodating the complexities introduced by multiple classes.

(Refer Slide Time: 33:56)



Now,

$$P(y_i = 1) = P(y_i = K)\, \underbrace{e^{\underline{\alpha}_1^T \underline{x}_i}}$$

$$\vdots$$

$$P(y_i = K-1) = P(y_i = K)\, e^{\underline{\alpha}_{K-1}^T \underline{x}_i}$$

$$\text{Since } P(y_i = K) = 1 - \sum_{k=1}^{K-1} P(y_i = k) \qquad \left(\sum_{k=1}^{K} P(y_i = k) = 1\right) \text{ Prob. sum.}$$

$$\boxed{P(y_i = K) = \frac{1}{1 + \sum_{j=1}^{K-1} e^{\underline{\alpha}_j^T \underline{x}_i}}}$$

To solve this, we take a straightforward approach: we apply the exponential function to both sides of the equation and simplify accordingly. Here's how it works:

1. Exponential Application:

By taking the exponential of both sides of the equation, we simplify the expression. For instance, when you apply the exponential function, you obtain:

$$p(y_i = k) = \frac{e^{\alpha_k^T x_i}}{1 + \sum_{j=1}^{K-1} e^{\alpha_j^T x_i}}$$

This formula applies for each class k, and you repeat this process for all K-1 possibilities to get a complete set of equations.

2. Probability Summation:

According to probability rules, the sum of probabilities for all possible outcomes must equal 1. Therefore, the probability that the outcome is class K is:

$$p(y_i = K) = \frac{1}{1 + \sum_{j=1}^{K-1} e^{\alpha_j^T x_i}}$$

Here, α_j represents the regression vectors for each of the K-1 classes, and x_i is the predictor vector for observation i.

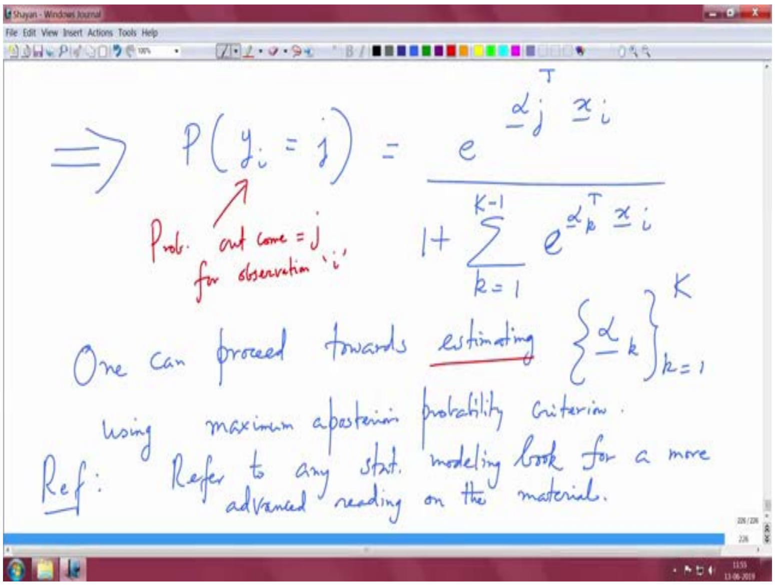
3. Resulting Formula:

Combining these equations, the probability that the outcome is class K can be expressed as:

$$p(y_i = K) = \frac{1}{1 + \sum_{j=1}^{K-1} e^{\alpha_j^T x_i}}$$

In summary, we simplify the model by applying the exponential function and ensuring that the sum of probabilities is 1. This provides us with a clear formula for predicting the probability of each class outcome based on the given predictors.

(Refer Slide Time: 35:44)



$$\Rightarrow P(y_i = j) = \frac{e^{\underline{\alpha}_j^T \underline{x}_i}}{1 + \sum_{k=1}^{K-1} e^{\underline{\alpha}_k^T \underline{x}_i}}$$

Prob. 'out come = j for observation 'i'

One can proceed towards estimating $\{\underline{\alpha}_k\}_{k=1}^{K}$ using maximum aposteriori probability criterion.

Ref: Refer to any stat. modeling book for a more advanced reading on the material.

To compute the probability p(y$_i$ = j) for a specific outcome j, we use the formula:

$$p(y_i = j) = \frac{e^{\alpha_j^T x_i}}{1 + \sum_{k=1}^{K-1} e^{\alpha_k^T x_i}}$$

This formula represents the probability of outcome j for observation i, where α$_j$ are the regression vectors corresponding to the j-th class, and x$_i$ is the vector of predictors for observation i. The term in the denominator accounts for the normalization across all possible outcomes.

To estimate these regression vectors α$_k$ for k = 1, 2, …, K, we typically use the maximum a posteriori (MAP) estimation criterion. For a detailed derivation of these methods, consulting a statistical modeling textbook would be beneficial, especially for understanding the nuances of multi-class problems.

For our purposes, we've covered the core concepts of both linear and logistic regression. These techniques are particularly valuable for classification problems within the framework of supervised learning. We’ll conclude our discussion here, but feel free to explore further in a pattern recognition course, where these topics might be expanded upon in greater depth.

Thank you.