ECON3389 Machine Learning in Economics

Module 2
Classification

Alberto Cappello

Department of Economics, Boston College

Fall 2024

Overview

Agenda:

- Qualitative outcomes and classification problem.
- LPM, logit and probit models.
- Estimation, interpretation and accuracy measures.
- Poisson Regression on count data

Readings:

- ISLR sections 4.1, 4.2, 4.3

Qualitative Outcomes

- So far our outcome variable Y has always been assumed to be quantitative, e.g. price, quantity, SAT score, etc.
- Qualitative variables (race, gender, geographical region, type of education, etc.) have only been discussed as predictors.
- But what if we want to quantify a relationship where the outcome Y is a qualitative variable?

Qualitative Outcomes

- So far our outcome variable Y has always been assumed to be quantitative, e.g. price, quantity, SAT score, etc.
- Qualitative variables (race, gender, geographical region, type of education, etc.) have only been discussed as predictors.
- But what if we want to quantify a relationship where the outcome Y is a qualitative variable?
  - A person arrives at an emergency room with symptoms that could be attributed to one of three medical conditions.

Qualitative Outcomes

- So far our outcome variable Y has always been assumed to be quantitative, e.g. price, quantity, SAT score, etc.
- Qualitative variables (race, gender, geographical region, type of education, etc.) have only been discussed as predictors.
- But what if we want to quantify a relationship where the outcome Y is a qualitative variable?
  - A person arrives at an emergency room with symptoms that could be attributed to one of three medical conditions.
  - Online banking service assesses whether a transaction being performed is fraudulent based on user's IP address, past transaction history, transaction amount, etc.

Qualitative Outcomes

- So far our outcome variable Y has always been assumed to be quantitative, e.g. price, quantity, SAT score, etc.
- Qualitative variables (race, gender, geographical region, type of education, etc.) have only been discussed as predictors.
- But what if we want to quantify a relationship where the outcome Y is a qualitative variable?
  - A person arrives at an emergency room with symptoms that could be attributed to one of three medical conditions.
  - Online banking service assesses whether a transaction being performed is fraudulent based on user's IP address, past transaction history, transaction amount, etc.
  - A researchers performs an analysis of socio-economic factors that affect whether a student graduates from college or drops out.

Basic Classification Problem

- Consider a qualitative variable Y that for every observation i takes a single value from a finite set of possible unordered values $\mathcal{C} = \{y_1, y_2, \ldots, y_C\}$.
  - Y = eye color, $\mathcal{C} = \{\text{brown}, \text{blue}, \text{green}\}$
  - Y = medical diagnosis, $\mathcal{C} = \{\text{stroke}, \text{drug overdose}, \text{epileptic seizure}\}$
  - Y = transaction status, $\mathcal{C} = \{\text{fraudulent}, \text{non-fraudulent}\}$

Basic Classification Problem

- Consider a qualitative variable Y that for every observation i takes a single value from a finite set of possible unordered values $\mathcal{C} = \{y_1, y_2, \ldots, y_C\}$.
  - Y = eye color, $\mathcal{C} = \{\text{brown}, \text{blue}, \text{green}\}$
  - Y = medical diagnosis, $\mathcal{C} = \{\text{stroke}, \text{drug overdose}, \text{epileptic seizure}\}$
  - Y = transaction status, $\mathcal{C} = \{\text{fraudulent}, \text{non-fraudulent}\}$
- Given a feature vector X and a qualitative response Y, the classification task is to build a function $C(X)$ that takes as input the feature vector X and predicts its value for Y, i.e. $C(X) \in \mathcal{C}$.
- In most cases we are interested in estimating the probabilities that Y belongs to a category in $\mathcal{C}$ given X, i.e.

$$\Pr(Y = y_c | X) \quad \forall y_c \in \mathcal{C}$$

Linear Probability Model: Problems

- Suppose for our medical condition classification we code

$$Y = \begin{cases} 1, & \text{if } \texttt{Stroke} \\ 2, & \text{if } \texttt{Drug Overdose} \\ 3, & \text{if } \texttt{Epileptic Seizure} \end{cases}$$

Can we use our standard regression model to predict the medical condition of a patient in the emergency room on the basis of her symptoms?

Linear Probability Model: Problems

- Suppose for our medical condition classification we code

$$Y = \begin{cases} 1, & \text{if } \texttt{Stroke} \\ 2, & \text{if } \texttt{Drug Overdose} \\ 3, & \text{if } \texttt{Epileptic Seizure} \end{cases}$$

  Can we use our standard regression model to predict the medical condition of a patient in the emergency room on the basis of her symptoms?
- This coding implies an ordering of outcomes. It also implies that the difference between Drug Overdose and Stroke is the same as the difference between Epileptic Seizure and Drug Overdose.

Linear Probability Model: Problems

- Suppose for our medical condition classification we code

$$Y = \begin{cases} 1, & \text{if } \texttt{Stroke} \\ 2, & \text{if } \texttt{Drug Overdose} \\ 3, & \text{if } \texttt{Epileptic Seizure} \end{cases}$$

  Can we use our standard regression model to predict the medical condition of a patient in the emergency room on the basis of her symptoms?
- This coding implies an ordering of outcomes. It also implies that the difference between Drug Overdose and Stroke is the same as the difference between Epileptic Seizure and Drug Overdose.
- In most cases, it is not possible for us to create a natural ordering in quantitative data

Linear Probability Model: Problems

- Suppose for our medical condition classification we code

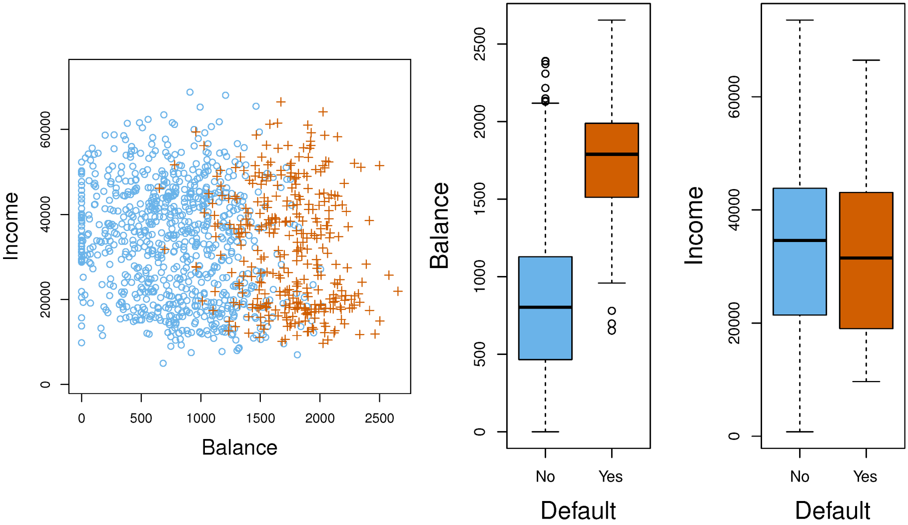
$$Y = \begin{cases} 1, & \text{if } \texttt{Stroke} \\ 2, & \text{if } \texttt{Drug Overdose} \\ 3, & \text{if } \texttt{Epileptic Seizure} \end{cases}$$

  Can we use our standard regression model to predict the medical condition of a patient in the emergency room on the basis of her symptoms?
- This coding implies an ordering of outcomes. It also implies that the difference between Drug Overdose and Stroke is the same as the difference between Epileptic Seizure and Drug Overdose.
- In most cases, it is not possible for us to create a natural ordering in quantitative data
- A different coding

$$Y = \begin{cases} 1, & \text{if } \texttt{Drug Overdose} \\ 2, & \text{if } \texttt{Stroke} \\ 3, & \text{if } \texttt{Epileptic Seizure} \end{cases}$$

  Will generate a different model and different predictions

Example: Credit Card Defaults



Linear Probability Model

- Suppose for our credit card default classification we code

$$Y = \begin{cases} 0, & \text{if } \texttt{No} \\ 1, & \text{if } \texttt{Yes} \end{cases}$$

Can we use our standard regression model to estimate a regression of Y on X and classify outcome as `Yes` if $\widehat{Y} > 0.5$?

Linear Probability Model

- Suppose for our credit card default classification we code

$$Y = \begin{cases} 0, & \text{if } \texttt{No} \\ 1, & \text{if } \texttt{Yes} \end{cases}$$

  Can we use our standard regression model to estimate a regression of Y on X and classify outcome as `Yes` if $\widehat{Y} > 0.5$?
- Given that Y is binary, we have

$$\mathbb{E}(Y|X) = ?$$

Linear Probability Model

- Suppose for our credit card default classification we code

$$Y = \begin{cases} 0, & \text{if } \texttt{No} \\ 1, & \text{if } \texttt{Yes} \end{cases}$$

  Can we use our standard regression model to estimate a regression of Y on X and classify outcome as `Yes` if $\widehat{Y} > 0.5$?

- Given that Y is binary, we have

$$\mathbb{E}(Y|X) = 1 \cdot \Pr(Y = 1|X) + 0 \cdot \Pr(Y = 0|X) = \Pr(Y = 1|X)$$

  which means that in this case standard linear regression will estimate the probability of outcome $Y = 1$, hence the name *linear probability model* or *LPM*.

Linear Probability Model

- LPM retains all properties of linear regression, but the interpretation of the results is slightly different:

$$\Pr(Y = 1|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

$$\beta_j = \frac{\partial \mathbb{E}[\Pr(Y = 1|X)]}{\partial X_j} = \frac{\mathbb{E}[\Delta \Pr(Y = 1|X)]}{\Delta X_j}$$

so regression coefficients now capture constant *marginal probabilities* of outcome $Y = 1$ given a change in X_j.

Linear Probability Model

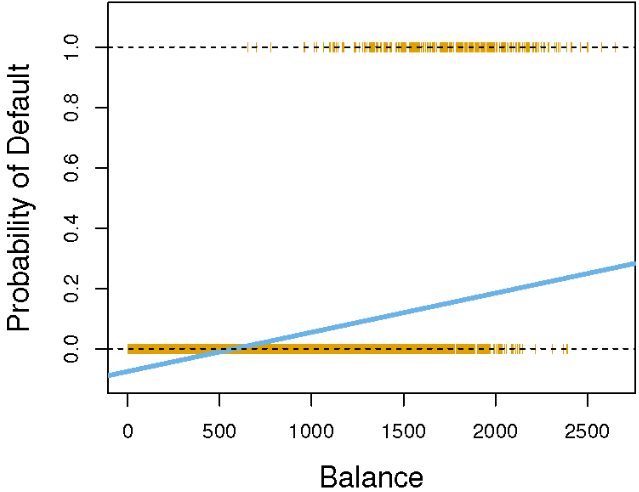
- LPM retains all properties of linear regression, but the interpretation of the results is slightly different:

$$\Pr(Y = 1|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

$$\beta_j = \frac{\partial \mathbb{E}[\Pr(Y = 1|X)]}{\partial X_j} = \frac{\mathbb{E}[\Delta \Pr(Y = 1|X)]}{\Delta X_j}$$

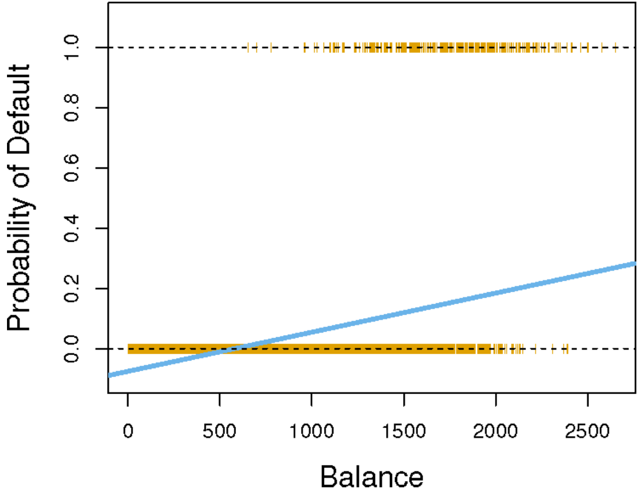
  so regression coefficients now capture constant *marginal probabilities* of outcome $Y = 1$ given a change in X_j.
- LPM works very well in binary classification, especially if sample size is moderate to large.
- But it also has some inherent disadvantages

Linear Probability Model



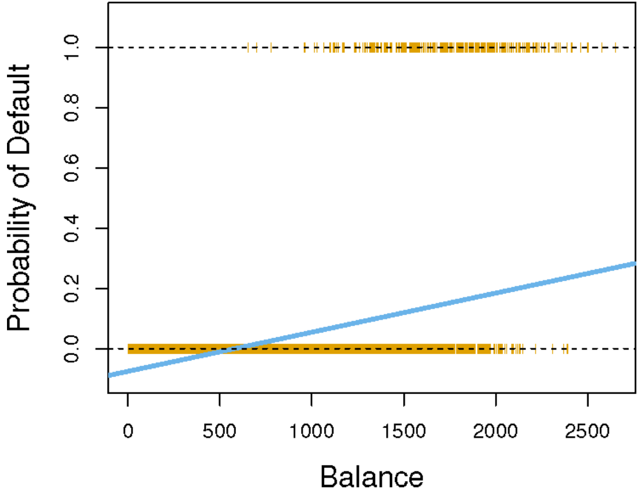
- The orange marks indicate the response Y, either 0 or 1.

Linear Probability Model



- The orange marks indicate the response Y, either 0 or 1.
- Blue line is estimated linear regression, which produces negative predictions for $\Pr(Y = 1|X)$ when `Balance` is less than 500.

Linear Probability Model



- The orange marks indicate the response Y, either 0 or 1.
- Blue line is estimated linear regression, which produces negative predictions for $\Pr(Y = 1|X)$ when `Balance` is less than 500.
- One way is to simply ignore the problem and bound any predictions from above and from below.
- But can we do better?

Moving away from LPM

- The problem with prediction in LPM stems from the fact that we use linear combination of features X as the estimated probability itself.

Moving away from LPM

- The problem with prediction in LPM stems from the fact that we use linear combination of features X as the estimated probability itself.
- To avoid this problem, we can instead use a one-to-one mapping $F(\cdot)$ from $\mathbb{R}$ to $[0, 1]$ interval:

$$\Pr(Y = 1|X) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$$

Moving away from LPM

- The problem with prediction in LPM stems from the fact that we use linear combination of features X as the estimated probability itself.
- To avoid this problem, we can instead use a one-to-one mapping $F(\cdot)$ from $\mathbb{R}$ to $[0, 1]$ interval:

$$\Pr(Y = 1|X) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$$

- What could serve as a function $F(\cdot)$? Infinitely many possibilities exist, but because this question was first addressed by statisticians, a natural choice was a *cumulative distribution function* (cdf) from some distribution.

Moving away from LPM

- The problem with prediction in LPM stems from the fact that we use linear combination of features X as the estimated probability itself.
- To avoid this problem, we can instead use a one-to-one mapping $F(\cdot)$ from $\mathbb{R}$ to $[0,1]$ interval:

$$\Pr(Y = 1|X) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$$

- What could serve as a function $F(\cdot)$? Infinitely many possibilities exist, but because this question was first addressed by statisticians, a natural choice was a *cumulative distribution function* (cdf) from some distribution.
- For any random variable Z its cdf $F_Z(\cdot)$ is by definition:

$$F_Z(a) = \Pr(Z \leq a)$$

- In classical statistical learning the two most common choice for $F(\cdot)$ are standard normal and logistic cdf.

Probit and Logit Models

- For simplicity, let's use the matrix notation:

$$\boldsymbol{X\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Probit and Logit Models

- For simplicity, let's use the matrix notation:

$$\boldsymbol{X}\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

- In *probit* model $F(\cdot)$ is assumed to be the cdf of a standard normal distribution:

$$F(\boldsymbol{X}\beta) = \Phi(\boldsymbol{X}\beta) = \int_{-\infty}^{X\beta} \frac{1}{\sqrt{2\pi}} \exp^{-\frac{z^2}{2}} dz$$

Probit and Logit Models

- For simplicity, let's use the matrix notation:

$$\boldsymbol{X\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

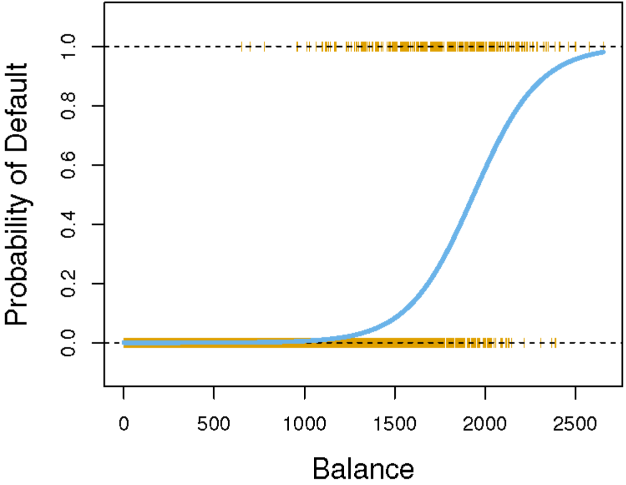
- In *probit* model $F(\cdot)$ is assumed to be the cdf of a standard normal distribution:

$$F(\boldsymbol{X\beta}) = \Phi(\boldsymbol{X\beta}) = \int_{-\infty}^{X\beta} \frac{1}{\sqrt{2\pi}} \exp^{-\frac{z^2}{2}} dz$$

- In *logit* model $F(\cdot)$ is assumed to be the cdf of a logistic distribution:

$$F(\boldsymbol{X\beta}) = \Lambda(\boldsymbol{X\beta}) = \int_{-\infty}^{X\beta} \frac{\exp(-z)}{(1+\exp(-z))^2} dz = \frac{\exp^{X\beta}}{1+\exp^{X\beta}}$$

Probit and Logit Models



With either normal or logistic cdf as our $F(\cdot)$ function the estimated probability will, by definition, always lie in $[0, 1]$ interval.

Generalized Linear Models

- Both logit and probit are special cases of the so-called *generalized linear models* (GLMs). Every GLM consists of three parts: the structural component, the link function and the response distribution.

Generalized Linear Models

- Both logit and probit are special cases of the so-called *generalized linear models* (GLMs). Every GLM consists of three parts: the structural component, the link function and the response distribution.
- *Structural component* is simply the linear combination of our predictors: $\boldsymbol{X}\beta$.

Generalized Linear Models

- Both logit and probit are special cases of the so-called *generalized linear models* (GLMs). Every GLM consists of three parts: the structural component, the link function and the response distribution.
- *Structural component* is simply the linear combination of our predictors: $\boldsymbol{X}\beta$.
- The *link function* $g(\mu)$ is such that its inverse gives us the (conditional) mean of our outcome Y as a function of the structural component:

$$g(\mu) = \boldsymbol{X}\beta \quad \text{or} \quad \mathbb{E}(Y|\boldsymbol{X}) = \mu = g^{-1}(\boldsymbol{X}\beta)$$

Generalized Linear Models

- Both logit and probit are special cases of the so-called *generalized linear models* (GLMs). Every GLM consists of three parts: the structural component, the link function and the response distribution.
- *Structural component* is simply the linear combination of our predictors: $\boldsymbol{X}\boldsymbol{\beta}$.
- The *link function* $g(\mu)$ is such that its inverse gives us the (conditional) mean of our outcome Y as a function of the structural component:

$$g(\mu) = \boldsymbol{X}\boldsymbol{\beta} \quad \text{or} \quad \mathbb{E}(Y|\boldsymbol{X}) = \mu = g^{-1}(\boldsymbol{X}\boldsymbol{\beta})$$

- The link function is the key to GLMs: since the distribution of the *response variable* Y is non-normal (in our simple example it is binomial), it's what lets us connect the structural component $\boldsymbol{X}\boldsymbol{\beta}$ to the response Y — it 'links' them (hence the name).

Generalized Linear Models

- Because our outcome Y is binary, we have $\mathbb{E}(Y|\boldsymbol{X}) = \Pr(Y = 1|\boldsymbol{X})$, and thus our inverse link function g^{-1} is simply the function that defines conditional probability of $Y = 1$ given $\boldsymbol{X}$.
- For probit and logit models the corresponding cumulative distribution functions act as inverse link functions:

$$\begin{aligned} \text{Probit} &: \mathbb{E}(Y|\boldsymbol{X}) = \mu_{Y|X} = \Pr(Y = 1|\boldsymbol{X}) = \Phi(\boldsymbol{X}\beta) \\ \text{Logit} &: \mathbb{E}(Y|\boldsymbol{X}) = \mu_{Y|X} = \Pr(Y = 1|\boldsymbol{X}) = \Lambda(\boldsymbol{X}\beta) \end{aligned}$$

Generalized Linear Models

- Because our outcome Y is binary, we have $\mathbb{E}(Y|\boldsymbol{X}) = \Pr(Y = 1|\boldsymbol{X})$, and thus our inverse link function g^{-1} is simply the function that defines conditional probability of $Y = 1$ given $\boldsymbol{X}$.
- For probit and logit models the corresponding cumulative distribution functions act as inverse link functions:

$$\begin{aligned} \text{Probit} &: \mathbb{E}(Y|\boldsymbol{X}) = \mu_{Y|X} = \Pr(Y = 1|\boldsymbol{X}) = \Phi(\boldsymbol{X}\beta) \\ \text{Logit} &: \mathbb{E}(Y|\boldsymbol{X}) = \mu_{Y|X} = \Pr(Y = 1|\boldsymbol{X}) = \Lambda(\boldsymbol{X}\beta) \end{aligned}$$

- Note: standard MLR is also a special case of GLM with $g(\mu) = \mu = \boldsymbol{X}\beta$.

Generalized Linear Models

- Because our outcome Y is binary, we have $\mathbb{E}(Y|\boldsymbol{X}) = \Pr(Y = 1|\boldsymbol{X})$, and thus our inverse link function g^{-1} is simply the function that defines conditional probability of $Y = 1$ given $\boldsymbol{X}$.
- For probit and logit models the corresponding cumulative distribution functions act as inverse link functions:

$$\begin{aligned}\text{Probit} &: \mathbb{E}(Y|\boldsymbol{X}) = \mu_{Y|X} = \Pr(Y = 1|\boldsymbol{X}) = \Phi(\boldsymbol{X}\beta)\\ \text{Logit} &: \mathbb{E}(Y|\boldsymbol{X}) = \mu_{Y|X} = \Pr(Y = 1|\boldsymbol{X}) = \Lambda(\boldsymbol{X}\beta)\end{aligned}$$

- Note: standard MLR is also a special case of GLM with $g(\mu) = \mu = \boldsymbol{X}\beta$.
- The two key differences of probit/logit models and usual MLR are estimation method and marginal effects calculation/interpretation.

Maximum Likelihood Estimation: Idea

- Suppose $X_1, X_2, ..., X_n$ form a random sample from a distribution for which the pdf is $f(x|\theta)$

Maximum Likelihood Estimation: Idea

- Suppose $X_1, X_2, ..., X_n$ form a random sample from a distribution for which the pdf is $f(x|\theta)$
- For every observed vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, we can define the joint pdf as follows

$$L(\theta) = f_n(\boldsymbol{x}|\theta) = f(x_1|\theta).f(x_2|\theta).f(x_3|\theta)...f(x_n|\theta) \tag{1}$$

Maximum Likelihood Estimation: Idea

- Suppose $X_1, X_2, ..., X_n$ form a random sample from a distribution for which the pdf is $f(x|\theta)$
- For every observed vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, we can define the joint pdf as follows

$$L(\theta) = f_n(\boldsymbol{x}|\theta) = f(x_1|\theta).f(x_2|\theta).f(x_3|\theta)...f(x_n|\theta) \tag{1}$$

- $L(\theta)$ is called the *Likelihood Function*

Maximum Likelihood Estimation: Idea

- Suppose $X_1, X_2, ..., X_n$ form a random sample from a distribution for which the pdf is $f(x|\theta)$
- For every observed vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, we can define the joint pdf as follows

$$L(\theta) = f_n(\boldsymbol{x}|\theta) = f(x_1|\theta).f(x_2|\theta).f(x_3|\theta)...f(x_n|\theta) \tag{1}$$

- $L(\theta)$ is called the *Likelihood Function*
- The MLE estimator of θ will find the parameter values that maximize $L(\theta)$

Maximum Likelihood Estimation: Idea

- Suppose $X_1, X_2, ..., X_n$ form a random sample from a distribution for which the pdf is $f(x|\theta)$
- For every observed vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, we can define the joint pdf as follows

$$L(\theta) = f_n(\boldsymbol{x}|\theta) = f(x_1|\theta).f(x_2|\theta).f(x_3|\theta)...f(x_n|\theta) \tag{1}$$

- $L(\theta)$ is called the *Likelihood Function*
- The MLE estimator of θ will find the parameter values that maximize $L(\theta)$
- In other words, it will find the parameter value that maximize the likelihood of the observed data being drawn from $f(x|\theta)$

Maximum Likelihood Estimation: Idea

- Suppose $X_1, X_2, ..., X_n$ form a random sample from a distribution for which the pdf is $f(x|\theta)$
- For every observed vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, we can define the joint pdf as follows

$$L(\theta) = f_n(\boldsymbol{x}|\theta) = f(x_1|\theta).f(x_2|\theta).f(x_3|\theta)...f(x_n|\theta) \tag{1}$$

- $L(\theta)$ is called the *Likelihood Function*
- The MLE estimator of θ will find the parameter values that maximize $L(\theta)$
- In other words, it will find the parameter value that maximize the likelihood of the observed data being drawn from $f(x|\theta)$
- Suppose $x_1, x_2, ..., x_n$ form a random sample from a normal distribution for which the mean μ is unknown and variance σ^2 is known

Maximum Likelihood Estimation: Idea

- Suppose $X_1, X_2, ..., X_n$ form a random sample from a distribution for which the pdf is $f(x|\theta)$
- For every observed vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, we can define the joint pdf as follows

$$L(\theta) = f_n(\boldsymbol{x}|\theta) = f(x_1|\theta).f(x_2|\theta).f(x_3|\theta)...f(x_n|\theta) \tag{1}$$

- $L(\theta)$ is called the *Likelihood Function*
- The MLE estimator of θ will find the parameter values that maximize $L(\theta)$
- In other words, it will find the parameter value that maximize the likelihood of the observed data being drawn from $f(x|\theta)$
- Suppose $x_1, x_2, ..., x_n$ form a random sample from a normal distribution for which the mean μ is unknown and variance σ^2 is known
- The likelihood function of μ is

$$L(\mu) = f_n(\boldsymbol{x}|\mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{2}$$

Maximum Likelihood Estimation: Example 1

- **Example**: Suppose I toss a coin 100 times and get 56 heads. What is the MLE of the probability of heads in a single toss?

Maximum Likelihood Estimation: Example 1

- **Example**: Suppose I toss a coin 100 times and get 56 heads. What is the MLE of the probability of heads in a single toss?
- **Model**: $L(p) = L(p; n, x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{100}{56} p^{56} (1-p)^{44}$

Maximum Likelihood Estimation: Example 1

- **Example**: Suppose I toss a coin 100 times and get 56 heads. What is the MLE of the probability of heads in a single toss?
- **Model**: $L(p) = L(p; n, x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{100}{56} p^{56} (1-p)^{44}$
- $L(0.5) = 0.038$
- $L(0.52) = 0.058$
- $L(0.54) = 0.073$
- $L(0.56) = 0.081$
- $L(0.58) = 0.073$

Maximum Likelihood Estimation: Method

- **Method 1**
  In the previous example, it is easy for us to write a simple equation that describes the likelihood surface that can be differentiated to find the MLE estimate

Maximum Likelihood Estimation: Method

- **Method 1**
  In the previous example, it is easy for us to write a simple equation that describes the likelihood surface that can be differentiated to find the MLE estimate
- Step 1: Take the log

$$\ln L = \ln \binom{100}{56} + \ln(p^{56}(1-p)^{44}) \tag{3}$$

$$\ln L = 56 \ln(p) + 44 \ln(1-p) \tag{4}$$

$$\tag{5}$$

Maximum Likelihood Estimation: Method

- **Method 1**
  In the previous example, it is easy for us to write a simple equation that describes the likelihood surface that can be differentiated to find the MLE estimate
- Step 1: Take the log

$$
\begin{aligned}
\ln L &= \ln \binom{100}{56} + \ln(p^{56}(1-p)^{44}) && (3)\\
\ln L &= 56 \ln(p) + 44 \ln(1-p) && (4)\\
& && (5)
\end{aligned}
$$

- Step 2: Differentiate the log likelihood to find the optimal parameter

$$
\begin{aligned}
\frac{56}{p} - \frac{44}{(1-p)} &= 0 && (6)\\
56(1-p) - 44p &= 0 && (7)\\
p &= \frac{56}{100} && (8)
\end{aligned}
$$

Maximum Likelihood Estimation: Method

- **Method 2**
  In the previous example, we used the information that sum of bernoullis follows a binomial distribution to construct the overall likelihood surface

Maximum Likelihood Estimation: Method

- **Method 2**
  In the previous example, we used the information that sum of bernoullis follows a binomial distribution to construct the overall likelihood surface
- In many cases this might not be possible because we are not working with such distributions and the model is at the level of individual coin tosses

Maximum Likelihood Estimation: Method

- **Method 2**
  In the previous example, we used the information that sum of bernoullis follows a binomial distribution to construct the overall likelihood surface
- In many cases this might not be possible because we are not working with such distributions and the model is at the level of individual coin tosses
- In such cases we need to construct the overall likelihood surface using the individual likelihoods

Maximum Likelihood Estimation: Method

- **Method 2**
  In the previous example, we used the information that sum of bernoullis follows a binomial distribution to construct the overall likelihood surface
- In many cases this might not be possible because we are not working with such distributions and the model is at the level of individual coin tosses
- In such cases we need to construct the overall likelihood surface using the individual likelihoods
- Each individual coin toss follows a Bernoulli distribution. Suppose $X = 1$ when heads and 0 otherwise

$$L(p; x) = p^x (1 - p)^{1-x}$$

Maximum Likelihood Estimation: Method

- **Method 2**: Remember the data is given to us

Observation	Outcome (x)	Likelihood of outcome
1	1	p
2	0	1-p
3	1	p
...		
99	0	1-p
100	1	p
Total	56	?

- What is the overall/joint likelihood of entries in the second column?

Maximum Likelihood Estimation: Method

- **Method 2**: Remember the data is given to us

Observation	Outcome (x)	Likelihood of outcome
1	1	p
2	0	1-p
3	1	p
...		
99	0	1-p
100	1	p
Total	56	?

- What is the overall/joint likelihood of entries in the second column?
- Each coin toss is independent

$$L(p) = p.p.p.p..(1-p)(1-p)...(1-p) \tag{9}$$

$$L(p) = p^{56}(1-p)^{44} \tag{10}$$

$$\ln L(p) = \ln(p^{56}(1-p)^{44}) \tag{11}$$

OLS as a special case of MLE

- Main assumption: The errors follow a normal distribution with mean 0 and variance σ^2

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$
$$\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

OLS as a special case of MLE

- Main assumption: The errors follow a normal distribution with mean 0 and variance σ^2

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$
$$\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

- The likelihood function of (β) is

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$
$$\ln(L(\beta)) = \sum_{i=1}^{n} \ln(\frac{1}{\sigma\sqrt{2\pi}}) - \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

OLS as a special case of MLE

- Main assumption: The errors follow a normal distribution with mean 0 and variance σ^2

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$
$$\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

- The likelihood function of (β) is

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$
$$\ln(L(\beta)) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

- The first term does not depend on (β) and the second term has a constant σ^2 that we can bring outside the summation

$$\ln(L(\beta)) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Maximum Likelihood Estimation

- Because we no longer have a direct connection between Y and our structural component $\boldsymbol{X}\beta$, we need to specify our loss function in a different way. Using our link function, we can for every observation i wright down the probability of observing a certain value of Y_i given values of $\boldsymbol{X}_i$

Maximum Likelihood Estimation

- Because we no longer have a direct connection between Y and our structural component $\boldsymbol{X}\beta$, we need to specify our loss function in a different way. Using our link function, we can for every observation i wright down the probability of observing a certain value of Y_i given values of $\boldsymbol{X}_i$
- For example, for a logit model we have:

$$\Pr(Y = Y_i|\boldsymbol{X}_i) = \left(\frac{\exp^{X_i\beta}}{1+\exp^{X_i\beta}}\right)^{Y_i}\left(1-\frac{\exp^{X_i\beta}}{1+\exp^{X_i\beta}}\right)^{1-Y_i}$$

Maximum Likelihood Estimation

- Because we no longer have a direct connection between Y and our structural component $\boldsymbol{X}\beta$, we need to specify our loss function in a different way. Using our link function, we can for every observation i wright down the probability of observing a certain value of Y_i given values of $\boldsymbol{X}_i$
- For example, for a logit model we have:

$$\Pr(Y = Y_i|\boldsymbol{X}_i) = \left(\frac{\exp^{X_i\beta}}{1+\exp^{X_i\beta}}\right)^{Y_i}\left(1-\frac{\exp^{X_i\beta}}{1+\exp^{X_i\beta}}\right)^{1-Y_i}$$

- With the default assumption of i.i.d. observations we can wright down the joint probability or *likelihood function* of seeing our sample:

$$\ell(\beta) = \prod_{i=1}^{n} \Pr(Y = Y_i|\boldsymbol{X}_i)$$

Maximum Likelihood Estimation

- *Maximum likelihood* estimation (ML) is a method that chooses parameters $\boldsymbol{\beta}$ so as to minimize the loss function in form of the negative of the log likelihood function:

$$\widehat{\boldsymbol{\beta}}_{ML} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} - \ln \ell(\boldsymbol{\beta})$$

Maximum Likelihood Estimation

- *Maximum likelihood* estimation (ML) is a method that chooses parameters $\boldsymbol{\beta}$ so as to minimize the loss function in form of the negative of the log likelihood function:

$$\widehat{\boldsymbol{\beta}}_{ML} = \underset{\boldsymbol{\beta}}{\text{argmin}} - \ln \ell(\boldsymbol{\beta})$$

- Under some general conditions $\widehat{\boldsymbol{\beta}}_{ML}$ is efficient, consistent and asymptotically normal, just like $\widehat{\boldsymbol{\beta}}_{OLS}$

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	−10.6513	0.3612	−29.5	<0.0001
`balance`	0.0055	0.0002	24.9	<0.0001

- But unlike OLS, ML is a more general estimation procedure and allows one to recover structural parameters such as $\boldsymbol{\beta}$ in models that are far more flexible than standard MLR.

Logistic Regression

- A key difference of logit/probit models from LPM is the fact that margial effects are now calculated and interpreted in a different way:

$$\Pr(Y = 1|\boldsymbol{X}) = p(\boldsymbol{X}) = \frac{\exp^{X\beta}}{1 + \exp^{X\beta}}$$
$$\frac{p(\boldsymbol{X})}{1 - p(\boldsymbol{X})} = \exp^{X\beta}$$
$$\ln \frac{p(\boldsymbol{X})}{1 - p(\boldsymbol{X})} = X\beta$$

- The quantity on the left is called *log-odds* or *logit*
- The logistic regression model has a logit that is linear in X
- In a logistic regression model, increasing X by one unit changes the log odds by its corresponding β

Marginal Effects in Probit/Logit

- The other key difference of logit/probit models from LPM is the fact that margial effects are now calculated and interpreted in a different way:

$$\text{Probit}: \frac{\partial p(\boldsymbol{X})}{\partial X_j} = \frac{\partial \Phi(\boldsymbol{X}\boldsymbol{\beta})}{\partial X_j}\beta_j \neq \beta_j \qquad \text{Logit}\ : \frac{\partial p(\boldsymbol{X})}{\partial X_j} = \frac{\partial \Lambda(\boldsymbol{X}\boldsymbol{\beta})}{\partial X_j}\beta_j \neq \beta_j$$

where $p(\boldsymbol{X}) = \Pr(Y = 1|\boldsymbol{X})$ for simplicity

Marginal Effects in Probit/Logit

- The other key difference of logit/probit models from LPM is the fact that margial effects are now calculated and interpreted in a different way:

$$\text{Probit}: \frac{\partial p(\boldsymbol{X})}{\partial X_j} = \frac{\partial \Phi(\boldsymbol{X}\boldsymbol{\beta})}{\partial X_j}\beta_j \neq \beta_j \qquad \text{Logit}: \frac{\partial p(\boldsymbol{X})}{\partial X_j} = \frac{\partial \Lambda(\boldsymbol{X}\boldsymbol{\beta})}{\partial X_j}\beta_j \neq \beta_j$$

  where $p(\boldsymbol{X}) = \Pr(Y = 1|\boldsymbol{X})$ for simplicity
- The marginal effects now depend on values of all variables in $\boldsymbol{X}$, so we need to either estimate the marginal effects at a specific value of all our predictors (typically means or medians) or calculate their average over all values of $\boldsymbol{X}$ in our sample.

Multinomial Logistic Regression

- It is possible to extend the two-class logistic regression approach to the setting of $K > 2$ classes
- Suppose we have K classes. First we need to define a base class (K^{th} one)

$$\Pr(Y = k|\boldsymbol{X}) = \frac{\exp^{X\beta_k}}{1 + \sum_{l=1}^{K-1} \exp^{X\beta_l}}$$
$$\Pr(Y = K|\boldsymbol{X}) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp^{X\beta_l}}$$

- It can be shown that the above model implies

$$\frac{\Pr(Y = k|\boldsymbol{X})}{\Pr(Y = K|\boldsymbol{X})} = \exp^{X\beta_k}$$
$$\ln \frac{\Pr(Y = k|\boldsymbol{X})}{\Pr(Y = K|\boldsymbol{X})} = X\beta_k$$

- The interpretation of the β_k is with respect to the base category K

Goodness-of-fit Measures

- While LPM can use the standard R^2 as a well-defined goodness-of-fit measure, it is not an option for logit/probit models due to the different loss function.
- In standard MLR an single R^2 value of 0.95 is an evidence of an excellent fit, but in classification problems we are often more interested in *class-specific* performance, especially in areas such as medicine or biology.

Goodness-of-fit Measures

- While LPM can use the standard R^2 as a well-defined goodness-of-fit measure, it is not an option for logit/probit models due to the different loss function.
- In standard MLR an single R^2 value of 0.95 is an evidence of an excellent fit, but in classification problems we are often more interested in *class-specific* performance, especially in areas such as medicine or biology.
- Consider the case where $Y = 1$ means a positive test result. Then our model's predictions fall into one of the 4 possible cases:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$		
$\widehat{Y} = 1$		

Goodness-of-fit Measures

- While LPM can use the standard R^2 as a well-defined goodness-of-fit measure, it is not an option for logit/probit models due to the different loss function.
- In standard MLR an single R^2 value of 0.95 is an evidence of an excellent fit, but in classification problems we are often more interested in *class-specific* performance, especially in areas such as medicine or biology.
- Consider the case where $Y = 1$ means a positive test result. Then our model's predictions fall into one of the 4 possible cases:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	True Negative	
$\widehat{Y} = 1$		

Goodness-of-fit Measures

- While LPM can use the standard R^2 as a well-defined goodness-of-fit measure, it is not an option for logit/probit models due to the different loss function.
- In standard MLR an single R^2 value of 0.95 is an evidence of an excellent fit, but in classification problems we are often more interested in *class-specific* performance, especially in areas such as medicine or biology.
- Consider the case where $Y = 1$ means a positive test result. Then our model's predictions fall into one of the 4 possible cases:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	True Negative	False Negative (Type II Error)
$\widehat{Y} = 1$		

Goodness-of-fit Measures

- While LPM can use the standard R^2 as a well-defined goodness-of-fit measure, it is not an option for logit/probit models due to the different loss function.
- In standard MLR an single R^2 value of 0.95 is an evidence of an excellent fit, but in classification problems we are often more interested in *class-specific* performance, especially in areas such as medicine or biology.
- Consider the case where $Y = 1$ means a positive test result. Then our model's predictions fall into one of the 4 possible cases:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	True Negative	False Negative (Type II Error)
$\widehat{Y} = 1$	False Positive (Type I Error)	

Goodness-of-fit Measures

- While LPM can use the standard R^2 as a well-defined goodness-of-fit measure, it is not an option for logit/probit models due to the different loss function.
- In standard MLR an single R^2 value of 0.95 is an evidence of an excellent fit, but in classification problems we are often more interested in *class-specific* performance, especially in areas such as medicine or biology.
- Consider the case where $Y = 1$ means a positive test result. Then our model's predictions fall into one of the 4 possible cases:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	True Negative	False Negative (Type II Error)
$\widehat{Y} = 1$	False Positive (Type I Error)	True Positive

Goodness-of-fit Measures

- While LPM can use the standard R^2 as a well-defined goodness-of-fit measure, it is not an option for logit/probit models due to the different loss function.
- In standard MLR an single R^2 value of 0.95 is an evidence of an excellent fit, but in classification problems we are often more interested in *class-specific* performance, especially in areas such as medicine or biology.
- Consider the case where $Y = 1$ means a positive test result. Then our model's predictions fall into one of the 4 possible cases:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	True Negative	False Negative (Type II Error)
$\widehat{Y} = 1$	False Positive (Type I Error)	True Positive

Goodness-of-fit Measures

- While LPM can use the standard R^2 as a well-defined goodness-of-fit measure, it is not an option for logit/probit models due to the different loss function.
- In standard MLR an single R^2 value of 0.95 is an evidence of an excellent fit, but in classification problems we are often more interested in *class-specific* performance, especially in areas such as medicine or biology.
- Consider the case where $Y = 1$ means a positive test result. Then our model's predictions fall into one of the 4 possible cases:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	True Negative	False Negative (Type II Error)
$\widehat{Y} = 1$	False Positive (Type I Error)	True Positive

  - For a COVID-19 test or cancer screening, we care more FN then about FP.

Goodness-of-fit Measures

- While LPM can use the standard R^2 as a well-defined goodness-of-fit measure, it is not an option for logit/probit models due to the different loss function.
- In standard MLR an single R^2 value of 0.95 is an evidence of an excellent fit, but in classification problems we are often more interested in *class-specific* performance, especially in areas such as medicine or biology.
- Consider the case where $Y = 1$ means a positive test result. Then our model's predictions fall into one of the 4 possible cases:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	True Negative	False Negative (Type II Error)
$\widehat{Y} = 1$	False Positive (Type I Error)	True Positive

  - For a COVID-19 test or cancer screening, we care more FN then about FP.
  - For city administration FPs in traffic cameras and speeding tickets are more important.

Goodness-of-fit Measures

- While LPM can use the standard R^2 as a well-defined goodness-of-fit measure, it is not an option for logit/probit models due to the different loss function.
- In standard MLR an single R^2 value of 0.95 is an evidence of an excellent fit, but in classification problems we are often more interested in *class-specific* performance, especially in areas such as medicine or biology.
- Consider the case where $Y = 1$ means a positive test result. Then our model's predictions fall into one of the 4 possible cases:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	True Negative	False Negative (Type II Error)
$\widehat{Y} = 1$	False Positive (Type I Error)	True Positive

  - For a COVID-19 test or cancer screening, we care more FN then about FP.
  - For city administration FPs in traffic cameras and speeding tickets are more important.
  - In judicial system both FP and FN are equally important.

Goodness-of-fit Measures

- Consider the following *confusion matrix*, depicting the prediction results for the `Default` dataset, using 50% as a threshold for probability of default:

		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

Goodness-of-fit Measures

- Consider the following *confusion matrix*, depicting the prediction results for the `Default` dataset, using 50% as a threshold for probability of default:

		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

- If we simply look at pure prediction precision, then:
  - Our total error rate is $(23 + 252)/10000 = 2.75\%$, which seems low enough.

Goodness-of-fit Measures

- Consider the following *confusion matrix*, depicting the prediction results for the `Default` dataset, using 50% as a threshold for probability of default:

		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

- If we simply look at pure prediction precision, then:
  - Our total error rate is $(23 + 252)/10000 = 2.75\%$, which seems low enough.
  - Out of 104 predicted defaults 81 ended up being classified correctly, which means only $23/9667 = 0.24\%$ of all non-defaults were classified incorrectly.

Goodness-of-fit Measures

- Consider the following *confusion matrix*, depicting the prediction results for the `Default` dataset, using 50% as a threshold for probability of default:

		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

- If we simply look at pure prediction precision, then:
  - Our total error rate is $(23 + 252)/10000 = 2.75\%$, which seems low enough.
  - Out of 104 predicted defaults 81 ended up being classified correctly, which means only $23/9667 = 0.24\%$ of all non-defaults were classified incorrectly.
  - However, out of 333 true defaults we managed to miss $252/333 = 75.67\%$, which could be an unacceptably high error rate for this class.

Goodness-of-fit Measures

- This why in classification problems it is important to evaluate class-specific precision via the following four measures:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	True Negative Rate (TNR) or *specificity*: $TNR = TN/N = 9644/9667 = 99.76\%$	
$\widehat{Y} = 1$		

Goodness-of-fit Measures

- This why in classification problems it is important to evaluate class-specific precision via the following four measures:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	True Negative Rate (TNR) or *specificity*: $TNR = TN/N = 9644/9667 = 99.76\%$	False Negative Rate (FNR): $FNR = FN/P = 252/333 = 75.67\%$
$\widehat{Y} = 1$		

Goodness-of-fit Measures

- This why in classification problems it is important to evaluate class-specific precision via the following four measures:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	True Negative Rate (TNR) or *specificity*: $TNR = TN/N = 9644/9667 = 99.76\%$	False Negative Rate (FNR): $FNR = FN/P = 252/333 = 75.67\%$
$\widehat{Y} = 1$	False Positive Rate (FPR): $FPR = FP/N = 23/9667 = 0.24\%$	

Goodness-of-fit Measures

- This why in classification problems it is important to evaluate class-specific precision via the following four measures:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	True Negative Rate (TNR) or *specificity*: $TNR = TN/N = 9644/9667 = 99.76\%$	False Negative Rate (FNR): $FNR = FN/P = 252/333 = 75.67\%$
$\widehat{Y} = 1$	False Positive Rate (FPR): $FPR = FP/N = 23/9667 = 0.24\%$	True Positive Rate (TPR) or *sensitivity*: $TPR = TP/P = 81/333 = 24.33\%$

Goodness-of-fit Measures

- This why in classification problems it is important to evaluate class-specific precision via the following four measures:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	True Negative Rate (TNR) or *specificity*: $TNR = TN/N = 9644/9667 = 99.76\%$	False Negative Rate (FNR): $FNR = FN/P = 252/333 = 75.67\%$
$\widehat{Y} = 1$	False Positive Rate (FPR): $FPR = FP/N = 23/9667 = 0.24\%$	True Positive Rate (TPR) or *sensitivity*: $TPR = TP/P = 81/333 = 24.33\%$

Goodness-of-fit Measures

- This why in classification problems it is important to evaluate class-specific precision via the following four measures:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	True Negative Rate (TNR) or *specificity*: $TNR = TN/N = 9644/9667 = 99.76\%$	False Negative Rate (FNR): $FNR = FN/P = 252/333 = 75.67\%$
$\widehat{Y} = 1$	False Positive Rate (FPR): $FPR = FP/N = 23/9667 = 0.24\%$	True Positive Rate (TPR) or *sensitivity*: $TPR = TP/P = 81/333 = 24.33\%$

- As one can easily see, all four measures are related to each other. In particular, the following two identities must always hold:

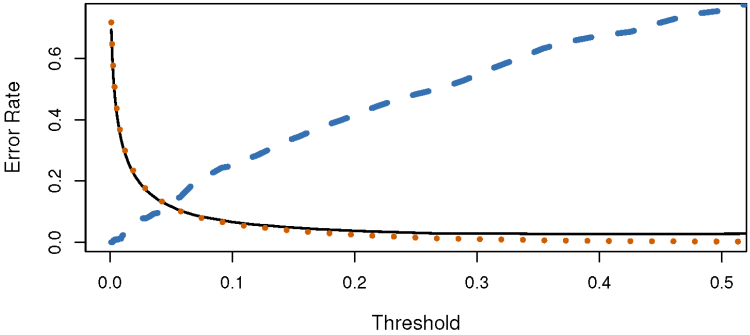
$$\text{TNR} + \text{FPR} = 100\% \quad \text{and} \quad \text{TPR} + \text{FNR} = 100\%$$

Goodness-of-fit Measures

The table on the previous slide was constructed using the rule $\widehat{Y} = 1$ if $\widehat{\Pr}(Y = 1|\boldsymbol{X}) > 0.5$, because 0.5 is the most common probability threshold used for classification predictions. However, the values of all 4 goodness-of-fit metrics will change if we change this threshold.

Goodness-of-fit Measures

The table on the previous slide was constructed using the rule $\widehat{Y} = 1$ if $\widehat{\Pr}(Y = 1|\boldsymbol{X}) > 0.5$, because 0.5 is the most common probability threshold used for classification predictions. However, the values of all 4 goodness-of-fit metrics will change if we change this threshold.



Black line: total error rate

Blue dashes: FNR

Orange dots: FPR

Based on this chart, we might want to set our threshold to 0.05 to achieve better error rate composition.

Logit vs Probit

While logit and probit models usually deliver very similar estimation results (especially on large datasets), modern statistical learning overwhelmingly prefers to use logistic regression. Why?

Logit vs Probit

While logit and probit models usually deliver very similar estimation results (especially on large datasets), modern statistical learning overwhelmingly prefers to use logistic regression. Why?

- **Coefficient interpretation.** In Economics we are interested in calculating and interpreting marginal effects, but in logit model one can also interpret the actual values of $\widehat{\beta}_j$ themselves. This is because in logit model coefficient $\widehat{\beta}_j$ shows how the log of odds ratio changes with changes in X_j:

$$ln\left(\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right) = \boldsymbol{X}\boldsymbol{\beta} \quad \Rightarrow \quad \beta_j = \frac{\partial ln\left(\frac{p(X)}{1-p(X)}\right)}{\partial X_j}$$

Logit vs Probit

While logit and probit models usually deliver very similar estimation results (especially on large datasets), modern statistical learning overwhelmingly prefers to use logistic regression. Why?

- **Coefficient interpretation.** In Economics we are interested in calculating and interpreting marginal effects, but in logit model one can also interpret the actual values of $\widehat{\beta}_j$ themselves. This is because in logit model coefficient $\widehat{\beta}_j$ shows how the log of odds ratio changes with changes in X_j:

$$ln\left(\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right) = \boldsymbol{X}\beta \quad \Rightarrow \quad \beta_j = \frac{\partial ln\left(\frac{p(X)}{1-p(X)}\right)}{\partial X_j}$$

- **Random utility models.** Suppose consumer is choosing between two alternatives based on utility that is a function of observable product attributes $\boldsymbol{X}$ and a random utility shock ϵ. Then if ϵ follows Type I EV distribution, consumer's choice probabilities will take logit form (McFadden, D. (1973)).

Logit vs Probit

While logit and probit models usually deliver very similar estimation results (especially on large datasets), modern statistical learning overwhelmingly prefers to use logistic regression. Why?

- **Coefficient interpretation.** In Economics we are interested in calculating and interpreting marginal effects, but in logit model one can also interpret the actual values of $\widehat{\beta}_j$ themselves. This is because in logit model coefficient $\widehat{\beta}_j$ shows how the log of odds ratio changes with changes in X_j:

$$ln\left(\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right) = \boldsymbol{X}\beta \quad \Rightarrow \quad \beta_j = \frac{\partial ln\left(\frac{p(X)}{1-p(X)}\right)}{\partial X_j}$$

- **Random utility models.** Suppose consumer is choosing between two alternatives based on utility that is a function of observable product attributes $\boldsymbol{X}$ and a random utility shock ϵ. Then if ϵ follows Type I EV distribution, consumer's choice probabilities will take logit form (McFadden, D. (1973)).
- **Generalized choice models and information theory** (Matejka, F. and McKay, A. (2015)).