CEE 616: Probabilistic Machine Learning

M2 Linear Methods: L2a Linear Discriminant Analysis

Jimi Oke

UMass Amherst

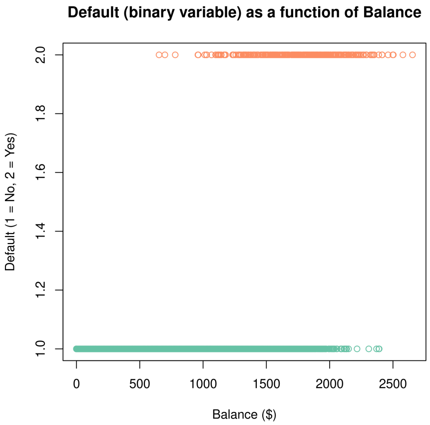
College of Engineering

Tue, Sep 23, 2025

Outline

Why classify?

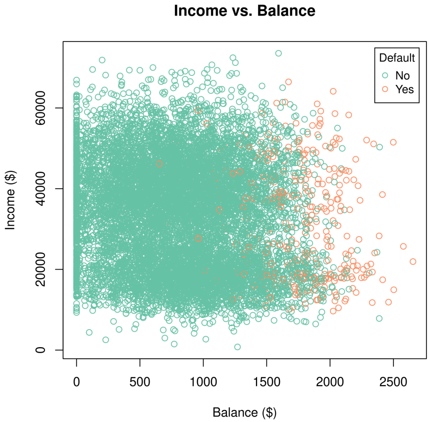
Consider the following plot:



What problems would we face if we tried to fit a linear model to these data?

The classification problem

Consider the following plot



How would you estimate a model to classify a point $(x_{(i,\texttt{balance})}, x_{(i,\texttt{income})})$?

Bayes theorem for classification

Given:

- The conditional density of a variable $\boldsymbol{x}$ in class c: $p_c(\boldsymbol{x})$
- The prior probability of class c: π_c

$$\text{such that: } \sum_{c=1}^{C} \pi_c = 1 \tag{1}$$

Then, using Bayes theorem, we write **class posterior** as:

$$p(y = c|\boldsymbol{x}) = \frac{p_c(\boldsymbol{x})\pi_c}{\sum_{c'=1}^{C} p_{c'}(\boldsymbol{x})\pi_{c'}} \tag{2}$$

With the class posteriors, we can then assign an observation i using the Bayes' classifier:

$$y_i^* = \arg\max_k p(y = c|\boldsymbol{x}_i) \tag{3}$$

(i.e. we assign to the class with the maximum probability)

Generative classifiers

A generative classifier is a model of the form:

$$p(y = c|\boldsymbol{x};\boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|y = c;\boldsymbol{\theta})p(y = c;\boldsymbol{\theta})}{\sum_{c'} p(\boldsymbol{x}|y = c';\boldsymbol{\theta})p(y = c;\boldsymbol{\theta})} \tag{4}$$

where:

- $p(y = c;\boldsymbol{\theta})$ is the class prior
- $p(\boldsymbol{x}|y = c;\boldsymbol{\theta})$ is the class conditional density for class c
- Generative classifiers generate features for each class by sampling from the class conditional density
- Discriminate classifiers model the class posterior $p(y|\boldsymbol{x};\boldsymbol{\theta}$ directly
- Linear discriminant analysis (LDA) arises from the log posterior being a linear function of $\boldsymbol{x}$

Why linear discriminant analysis?

Ideal situation

We know the following for each class:

- True [conditional] probability densities: $p_c(\boldsymbol{x})$
- True parameters: θ_c
- True prior probabilities: π_c

If these were known, the Bayes decision boundary could be computed exactly

In reality

We are not certain!

- Assume Gaussian/normal conditional densities $p_c(\boldsymbol{x})$
- Estimate the parameters $\hat{\mu}_c$ and $\hat{\sigma}^2$ from the sample data
- Estimate the priors from the data

This outlines the **linear discriminant analysis (LDA)** method, which approximates the Bayes classifier

Linear discriminant analysis (LDA)

Let N be the total number of training observations and N_c the number of training observations in Class c.

Parameter estimates

$$\text{Class sample mean:} \quad \hat{\boldsymbol{\mu}}_c = \frac{1}{n_c} \sum_{i:y_i=c} x_i \tag{5}$$

$$\text{Common sample variance:} \quad \hat{\sigma}^2 = \frac{1}{N-C} \sum_{c=1}^{C} \sum_{i:y_i=c} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_c)^2 \tag{6}$$

Class priors

Computed as the fraction of observations belonging to Class c:

$$\hat{\pi}_c = \frac{N_c}{N} \tag{7}$$

LDA classifier

- Like Bayes, the LDA classifier assigns an observation to the class with the maximizing posterior probability:

$$y_i^* = \arg\max_c P(Y = c | X = x_i) \tag{8}$$

- However, this is equivalent to maximizing the RHS of Bayes theorem:

$$y_i^* = \arg\max_c \frac{f_c(\boldsymbol{x}_i)\pi_c}{\sum_{\ell=1}^{C} f_\ell(\boldsymbol{x}_i)\pi_\ell} \tag{9}$$

- Since the denominator is the same for all classes, the LDA rule assigns an observation to the class which maximizes the **discriminant function,** δ_c:

$$\delta_c \sim \log\left[f_c(\boldsymbol{x}_i)\pi_c\right] = \log \pi_c + \log f_c(\boldsymbol{x}_i) \tag{10}$$

In LDA, we assume that f_c is Gaussian. In the univariate case, this is:

$$f(\boldsymbol{x}_c) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{x} - \mu_c)^2\right) \tag{11}$$

LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x\frac{\hat{\mu_c}}{\hat{\sigma}^2} - \frac{\hat{\mu_c}^2}{2\hat{\sigma}^2} \tag{12}$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(\boldsymbol{x})$.

Example: LDA for two classes with equal priors

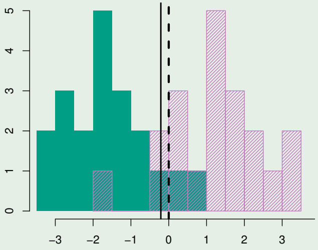


Figure: Bayes decision boundary (dashed) and LDA decision boundary (solid) estimated from the training data

Classes: Class 1 (green); Class 2 (purple)

Sample size and priors:

$$\begin{aligned} n_1 &= n_2 = 20 \\ \hat{\pi}_1 &= \hat{\pi}_2 = \frac{20}{40} = 0.5 \end{aligned}$$

Decision boundary:

$$x = \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$$

Performance (error rate):
Bayes: 10.6%; LDA: 11.1%

Multiple predictors

For a single predictor x, we assumed a *univariate* Gaussian distribution.

If we have multiple predictors, then:

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_D) \tag{13}$$

We then assume sample $\boldsymbol{x}$ has a *multivariate* Gaussian distribution.

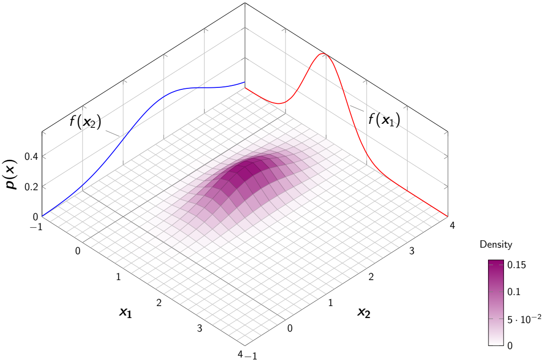


Figure: A multivariate Gaussian density function; $p = 2$ (bivariate Gaussian)

Multivariate Gaussian distributions: notation

If X is a p-dimensional normally distributed random variable, then:

$$X \sim \mathcal{N}(\mu, \mathbf{\Sigma}) \tag{14}$$

where $\mu = \mathbb{E}(\boldsymbol{x})$ and $\mathbf{\Sigma} = Cov(\boldsymbol{x})$ is the $p \times p$ covariance matrix of X.

$$Cov(\boldsymbol{x}_j, X_l) = \mathbb{E}(\boldsymbol{x}_j - \mathbb{E}(\boldsymbol{x}_j)(\boldsymbol{x}_l - \mathbb{E}(\boldsymbol{x}_l)) = \mathbb{E}(\boldsymbol{x}_j X_l) - \mathbb{E}(\boldsymbol{x}_j)\mathbb{E}(\boldsymbol{x}_l) \tag{15}$$

$$Cov(\boldsymbol{x}_j, X_j) = \mathbb{E}(\boldsymbol{x}_j^2) - (\mathbb{E}(\boldsymbol{x}_j))^2 = \mathbb{V}(\boldsymbol{x}_j) \tag{16}$$

Probability density function

The multivariate Gaussian/normal probability density function (PDF) is given by:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \mu)^T \mathbf{\Sigma}^{-1}(\boldsymbol{x} - \mu)\right) \tag{17}$$

LDA with multiple predictors

- Assume that observations in cth class are drawn from multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$ with a common covariance matrix
- Estimate distribution parameters $\hat{\boldsymbol{\mu}}_c$, $\hat{\boldsymbol{\Sigma}}$ and priors $\hat{\pi}_c$:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n - C} \sum_{c=1}^{C} \sum_{i: y_i = c} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_c)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_c)^T \tag{18}$$

- Compute the linear discriminant functions:

$$\delta_c(\boldsymbol{x}) = x^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_c - \frac{1}{2} \boldsymbol{\mu}_c{}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_c + \log \pi_c \tag{19}$$

- Assign each observation x_i to the class c which maximizes the linear discriminant functions:

$$\hat{y}_i = \arg\max_c \delta_c(\boldsymbol{x}) \tag{20}$$

LDA with multiple predictors: illustration

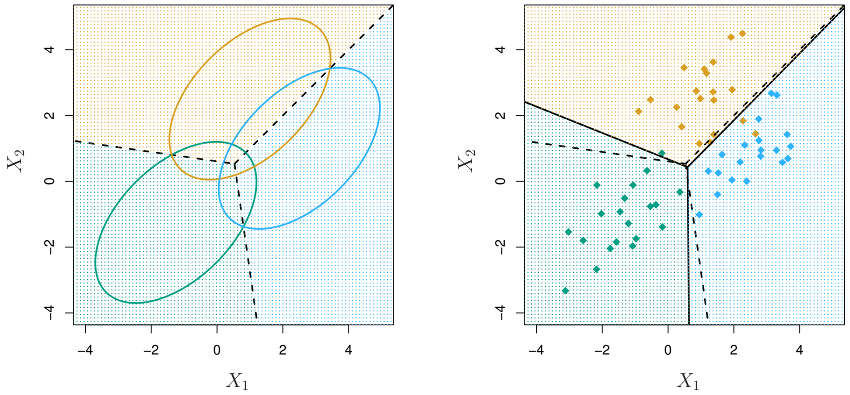


Figure: LDA for $p = 2, C = 3$. (L) Conditional class distributions of X (95% probability ellipses). (R) LDA decision boundaries (solid black) and Bayes decision boundaries (dashed lines).

Bayes' theorem for classification

Define:

$$
\begin{aligned}
C &= \text{number of classes} \\
\pi_c &= \text{prior probability that random observation is in } c\text{th class} \\
p_c(\boldsymbol{x}) &= \Pr(\boldsymbol{x} = x | Y = c) \quad \text{(conditional probability distribution of } X)
\end{aligned}
$$

According to Bayes' theorem, the **posterior probability** that an observation is in class c given x is:

$$
\Pr(Y = c | X = x) = \frac{\pi_c p_c(\boldsymbol{x})}{\sum_{l=1}^{C} \pi_l f_l(\boldsymbol{x})} \tag{21}
$$

We can also express the posterior probability as:

$$
p_c(\boldsymbol{x}) = \Pr(Y = c | X = x) \tag{22}
$$

Bayes' theorem for classification (cont.)

Steps:

1. Determine the probability density function of X: $p_c(\boldsymbol{x})$
2. Determine the prior probability π_c
3. Compute the posterior probability $p_c(\boldsymbol{x})$ using Bayes' theorem
4. Classify the observation based on the maximum probability:

$$\hat{y}_i = c^* = \arg\max_c p_c(\boldsymbol{x}_i) \tag{23}$$

Eq. (23) is called the **decision rule**.

Assumptions

Assumption 1

X is normally distributed:

$$p_c(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(-\frac{1}{2\sigma_c^2}(\boldsymbol{x} - \boldsymbol{\mu}_c)^2\right) \tag{24}$$

where $\boldsymbol{\mu}_c$ and σ_c^2 are the mean and variance of the observations in the cth class.

Assumption 2

There is a common variance across all C classes:

$$\sigma_1^2 = \cdots = \sigma_C^2 \tag{25}$$

Bayes classifier

Given the assumptions of normally distributed X and constant variance, then the **posterior probability** according to Bayes is:

$$p_c(\boldsymbol{x}) = \frac{\pi_c \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(-\frac{1}{2\sigma_c^2}(\boldsymbol{x} - \boldsymbol{\mu}_c)^2\right)}{\sum_{l=1}^{C} \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(\boldsymbol{x} - \mu_l)^2\right)} \tag{26}$$

The Bayes classifier assigns an observation to the class for which $p_c(\boldsymbol{x})$ is the largest.

For the ith observation, this assignment can be written as:

$$y_i = c^* = \arg\max_c p_c(\boldsymbol{x}) = \arg\max_c \log(p_c(\boldsymbol{x})) = \arg\max_c \delta_c(\boldsymbol{x}) \tag{27}$$

The term $\delta_c(\boldsymbol{x})$ denotes the **linear discriminant functions**. .

Linear discriminant functions

To derive, take the log of the posterior probability:

$$\log(p_c(\boldsymbol{x})) = \log\left(\frac{\pi_c}{\sqrt{2\pi}\sigma_c}\right) - \frac{(\boldsymbol{x}-\boldsymbol{\mu}_c)^2}{2\sigma_c^2} - \sum_{l=1}^{C}\left[\log\left(\frac{\pi_l}{\sqrt{2\pi}\sigma_l}\right) - \frac{(\boldsymbol{x}-\mu_l)^2}{2\sigma_l^2}\right]$$

$$= \log\left(\frac{\pi_c}{\sqrt{2\pi}\sigma}\right) - \frac{(\boldsymbol{x}-\boldsymbol{\mu}_c)^2}{2\sigma^2} - \sum_{l=1}^{C}\left[\log\left(\frac{\pi_l}{\sqrt{2\pi}\sigma}\right) - \frac{(\boldsymbol{x}-\mu_l)^2}{2\sigma^2}\right]$$

$$= \log(\pi_c) - \log\left(\sqrt{2\pi}\sigma\right) - \frac{x^2}{2\sigma^2} + \frac{x\boldsymbol{\mu}_c}{2\sigma^2} - \frac{\boldsymbol{\mu}_c{}^2}{2\sigma^2}$$

$$- \sum_{l=1}^{C}\left[\log\left(\frac{\pi_l}{\sqrt{2\pi}\sigma}\right) - \frac{(\boldsymbol{x}-\mu_l)^2}{2\sigma^2}\right]$$

To find the maximizing value of $\log(p_c(\boldsymbol{x}))$, we discard the terms that are constant in c and thus obtain the **linear discriminant functions**:

$$\delta_c(\boldsymbol{x}) = \log(\pi_c) + x\frac{\boldsymbol{\mu}_c}{\sigma^2} - \frac{\boldsymbol{\mu}_c{}^2}{2\sigma^2} \tag{28}$$

Linear discriminant functions (cont.)

Given C classes each with priors π_c, means $\boldsymbol{\mu}_c$ and common variance σ, we express the discriminant function as:

$$
\begin{aligned}
\delta_1(\boldsymbol{x}) &= \log(\pi_1) + x\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} \\
\delta_2(\boldsymbol{x}) &= \log(\pi_2) + x\frac{\boldsymbol{\mu}_2}{\sigma^2} - \frac{\boldsymbol{\mu}_2{}^2}{2\sigma^2} \\
&\vdots \qquad \vdots \\
\delta_C(\boldsymbol{x}) &= \log(\pi_C) + x\frac{\boldsymbol{\mu}_c}{\sigma^2} - \frac{\boldsymbol{\mu}_c{}^2}{2\sigma^2}
\end{aligned}
$$

For an observation with $X = x$, the *decision rule* assigns Y to the class c for which $\delta_c(\boldsymbol{x})$ is the greatest.

Bayes decision boundary: binary case

In the binary case, $C = 2$. The discriminant functions are:

$$\delta_1(\boldsymbol{x}) = \log(\pi_1) + x\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2}$$
$$\delta_2(\boldsymbol{x}) = \log(\pi_2) + x\frac{\boldsymbol{\mu}_2}{\sigma^2} - \frac{\boldsymbol{\mu}_2{}^2}{2\sigma^2}$$

The decision rule assigns an observation to Class 1 if $\delta_1(\boldsymbol{x}) > \delta_2(\boldsymbol{x})$.
Now, consider the situation where the *priors are equal*:

$$\pi_1 = \pi_2$$

implying that an observation is *equally likely* to come from Class 1 or Class 2. The decision rule would then **assign to Class 1** if:

$$x\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} > x\frac{\boldsymbol{\mu}_2}{\sigma^2} - \frac{\boldsymbol{\mu}_2{}^2}{2\sigma^2}$$
$$\implies \quad x\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) > \frac{\boldsymbol{\mu}_1{}^2 - \boldsymbol{\mu}_2{}^2}{2}$$

Bayes decision boundary (cont.)

At the Bayes decision boundary, the $\delta_1 = \delta_2$, thus:

$$x\,(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{{\boldsymbol{\mu}_1}^2 - {\boldsymbol{\mu}_2}^2}{2} = \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)}{2}$$

$$x = \frac{(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)}{2}$$

Thus, given a univariate predictor X whose two classes have equal priors $\pi_1 = \pi_2$, the decision boundary lies midway between the means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

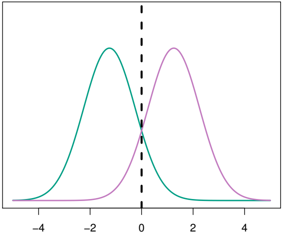


Figure: Bayes decision boundary (dashed line) shown for the priors of the distributions of 2 classes: Class 1 (green) and Class 2 (purple)

Quadratic discriminant analysis (QDA)

- A key assumption in LDA is that all classes share a common covariance structure: $\mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_C = \mathbf{\Sigma}$.
- If we discard this assumption, then:

$$X \sim \mathcal{N}(\boldsymbol{\mu}_c, \mathbf{\Sigma}_{c}) \tag{29}$$

and we can no longer ignore the $\hat{\mathbf{\Sigma}}_c$ terms in posterior probabilities.
- This results in discriminant functions that are **quadratic** in x:

$$\begin{aligned}
\delta_c(\boldsymbol{x}) &= -\frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{\mu}_c})^T \mathbf{\Sigma}_c^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}_c}) - \frac{1}{2}\log|\mathbf{\Sigma}_c| + \log \pi_c & (30)\\
&= -\frac{1}{2}x^T \mathbf{\Sigma}_c^{-1}\boldsymbol{x} + x^T \mathbf{\Sigma}_c^{-1}\boldsymbol{\mu}_c - \frac{1}{2}\hat{\boldsymbol{\mu}_c}^T \mathbf{\Sigma}_c^{-1}\hat{\boldsymbol{\mu}_c} - \frac{1}{2}\log|\mathbf{\Sigma}_c| + \log \pi_c & (31)
\end{aligned}$$

- Under this assumption of class-specific covariance, we perform **quadratic discriminant analysis**.

QDA considerations: bias-variance trade-off

Number of parameters:

- To find $\hat{\mathbf{\Sigma}}$ in LDA, $D(D+1)/2$ parameters must be estimated
- In QDA, since there are C covariance matrices, $CD(D+1)/2$ must be estimated

Bias-variance:

- Thus QDA is more flexible (lower bias, but potentially higher variance)
- LDA might be more stable (lower variance, but potentially higher bias) if the constant $\mathbf{\Sigma}$ assumption does not hold for the data.
- Generally, with fewer observations, LDA is preferred

QDA vs. LDA

QDA produces a quadratic decision boundary which performs better in classifying the observations if the covariance matrices are different for each class.

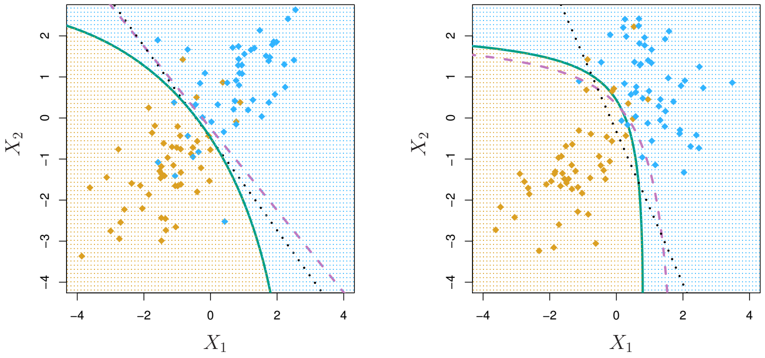


Figure: (L): shared covariance across classes (linear boundary). (R): different covariance in each class (quadratic boundary). Bayes (purple dashed); LDA (black dotted); QDA (green solid)

Approximating a quadratic decision boundary

QDA can be approximated by LDA by including second-order terms in the predictor space, i.e.

$$(X_1, X_2) \rightarrow (X_1, X_2, X_1X_2, X_1^2, X_2^2)$$

The LDA approximation is may be less accurate but more stable, as fewer parameters are required.

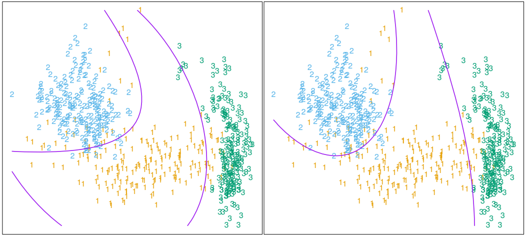


Figure: (L) LDA approximation of decision boundary. (R) QDA decision boundary. $D = 2, C = 3$.

Outlook

- Next lecture: L2b: Logistic regression
- Reading for today's lecture: **PMLI** 9.1-2; **ESL** 4.3
- Optional: Naive Bayes classifier (NBC) **PMLI** 9.3

The generalized linear model (GLM)

- Conventional linear models have the form:

$$y_i \sim N(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2) \tag{32}$$

  where
  - y_i is a continuous response
  - $\boldsymbol{x}_i$ is a vector of quantitative and/or qualitative explanatory variables
- Generalized linear models (GLMs) were introduced to extend this framework to allow y_i to be modeled by other exponential family distributions besides the normal/Gaussian, e.g.
  - exponential
  - binomial/multinomial (with fixed number of trials)
  - Poisson
- In the GLM framework:
  - The mean of y_i is given by μ_i
  - μ_i can be specified by a nonlinear function of $\boldsymbol{x}_i^T \boldsymbol{\beta}$
  - Note that the simple linear regression is a special case of GLM in which $\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and y_i follows a Gaussian distribution

Back to Recap

GLM components

A GLM consists of three parts:

- **Random component:** this is the probability distribution of the response variable
- **Systematic component:** specifies the explanatory variables within the linear combination of their coefficients ($\boldsymbol{X}\beta$)
- **Link function $\boldsymbol{g(\mu)}$:** defines the relationship between the random and systematic components:
  - Simple linear regression (identity link function):

$$g(\mu_i) = g(\mathbb{E}(y_i)) = \boldsymbol{x}_i^T \boldsymbol{\beta} \tag{33}$$

  - Binary logistic regression (logit link function):

$$g(\mu_i) = g(p(\boldsymbol{x}_i)) = \text{logit}(p(\boldsymbol{x}_i)) = \ln\left(\frac{p(\boldsymbol{x}_i)}{1 - p(\boldsymbol{x}_i)}\right) = \boldsymbol{x}_i^T \boldsymbol{\beta} \tag{34}$$

Assumptions of GLM

- The observations of the response variable $\boldsymbol{y}$ are i.i.d.
- Response variable y_i is typically exponentially distributed (not restricted to being normally distributed)
  - Implies that errors need not be normally distributed (but should be independent)
- Link function is linear with respect to the coefficients (β_j)
  - Relationship between response and explanatory variables does not have to be linear
  - Explanatory variables can be nonlinear transformations of original values (as in simple linear regression)
- Variance may not homogeneous (i.e. homoscedasticity is not a requirement)
- Parameters are estimated via MLE

Commonly used GLM models and their components

Model	Random component	Link function
Linear regression	Gaussian	Identity: $g(\mu_i) = \mu_i = \beta^T x_i$
Binary logistic regression	Bernoulli	Logit: $g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$
Probit regression	Bernoulli	Probit: $g(\mu_i) = \Phi^{-1}(\mu_i)$
Multinomial logit/logistic	Categorical	Multinomial logit: $g(\mu_{ic}) = \ln\left(\frac{\mu_{ic}}{\mu_{iC}}\right)$
Poisson regression	Poisson	Log: $g(\mu_i) = \ln(\mu_i)$

Note that in all cases, the link function always results in:

$$g(\mu_i) = \beta^T x_i \tag{35}$$

Its job is to "link" the response to the systematic component via a suitable transformation that results in a linear function of the β's.

Further reading on GLMs

- German Rodriguez's lecture notes on GLMs: `https://data.princeton.edu/wws509/notes/`
- Penn State: `https://online.stat.psu.edu/stat504/lesson/6/6.1` (Including more on logistic and multinomial logistic)