# Classification: Logistic Regression and Linear Discriminant Analysis

(ISLR 4.1 - 4.4)

Yingbo Li

Southern Methodist University

STAT 4399

# Outline

# Classification

One categorical response $Y$:

- Two levels: usually coded as $\{0, 1\}$.
  - $\{$no, yes$\}$,
  - $\{$healthy, diseased$\}$.
- Multiple levels: usually coded as $\{1, 2, \ldots, K\}$.
  - Un-ordered: $\{$orange, purple, other$\}$.
  - Ordered: $\{$mild, moderate, severe$\}$.

  Using numerical coding doesn't necessarily imply ordering, and the gaps between ordered levels may not be the same.

One or more predictors $X = (X_1, X_2, \ldots, X_p)$.

*Classification*: build a function $C(X)$ that takes input $X$ and predicts its value for $Y$

# The `Default` data

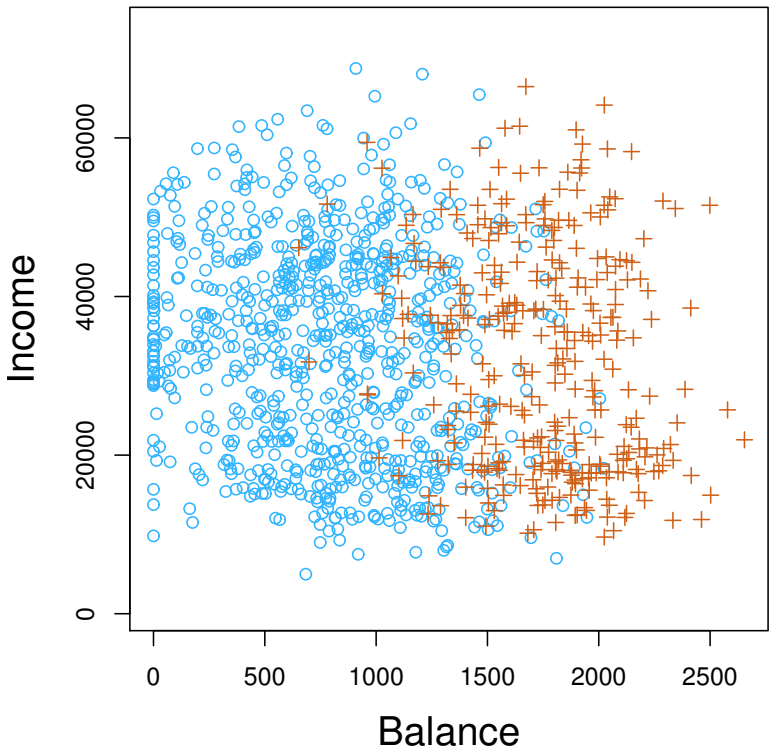
Contains $n = 10000$ observations and 4 variables

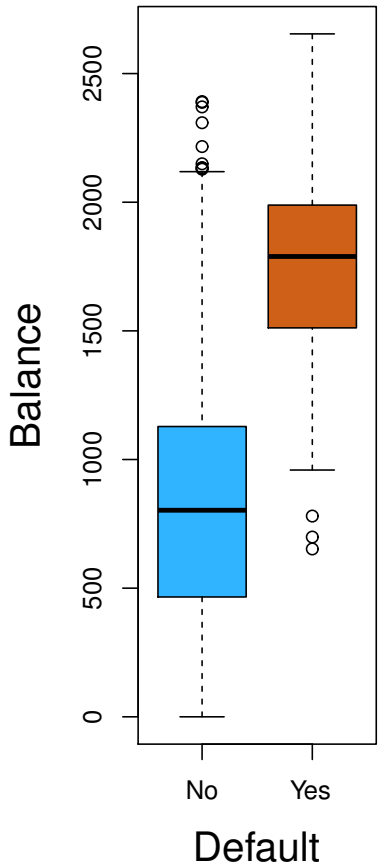
- `default`: $\{\text{yes}, \text{no}\}$.
- `student`: $\{\text{yes}, \text{no}\}$; `balance`, `income`: numerical.

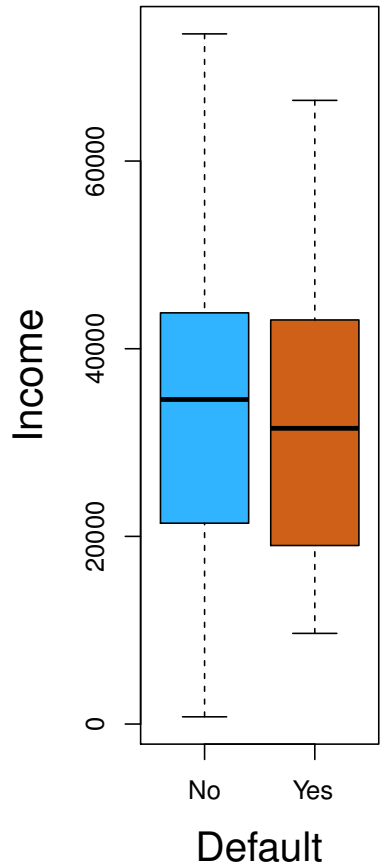
| | default | student |
|---|---|---|
| No | 9667 | 7056 |
| Yes | 333 | 2944 |

| | balance | income |
|---|---|---|
| Min | 0.0 | 772 |
| 1st Qu | 481.7 | 21340 |
| Median | 823.6 | 34553 |
| Mean | 835.4 | 33517 |
| 3rd Qu | 1166.3 | 43808 |
| Max | 2654.3 | 73554 |

We are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit card balance.



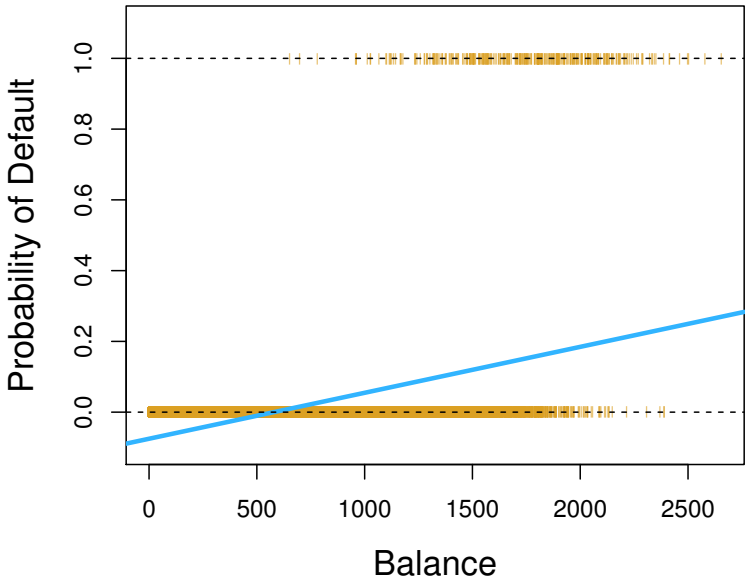
# Model the probability $P(Y = 1 \mid X)$

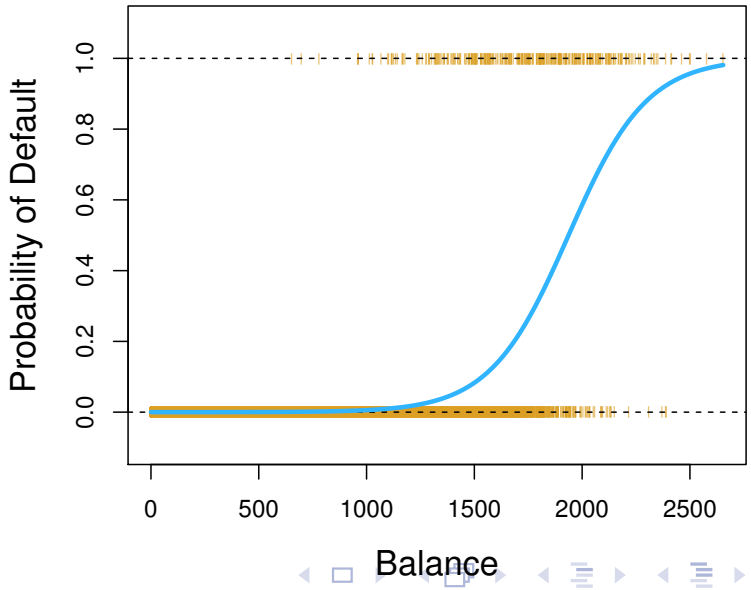
We code the response `default` as

$$Y = \begin{cases} 0 & \text{if no} \\ 1 & \text{if yes} \end{cases}$$

We don't directly regress $P(Y = 1 \mid X)$ on $X$, because

- Normality residuals assumption in linear regression is not satisfied.
- The predicted probability may be greater than 1 or less than 0.



# Logistic regression

Denote the success probability $p = P(Y = 1)$, then

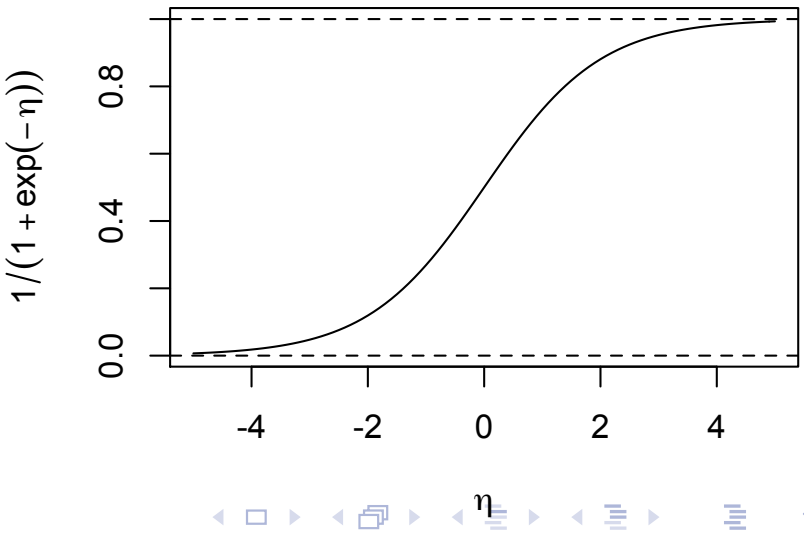
- Odds $\frac{p}{1-p}$: takes value in $(0, \infty)$
- Log-odds $\log\left(\frac{p}{1-p}\right)$: takes value in $(-\infty, \infty)$. Also called logit.

We connect the log-odds with a linear combination of predictors

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

This is equivalent to

$$p = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}}$$



# Maximum likelihood

Consider the case of $p = 1$ predictor. The *likelihood* function is the density of the sampling distribution of $Y$ given the parameters $\beta_0, \beta_1$.

Each $Y_i \sim \text{Bernoulli}(p_i)$, where $p_i$ depends on $X_{i,1}$, its density

$$\begin{aligned} f(Y_i \mid \beta_0, \beta_1) &= p_i^{Y_i}(1-p_i)^{1-Y_i} \\ &= \left(\frac{e^{\beta_0+\beta_1 X_{i,1}}}{1+e^{\beta_0+\beta_1 X_{i,1}}}\right)^{Y_i} \left(\frac{1}{1+e^{\beta_0+\beta_1 X_{i,1}}}\right)^{1-Y_i} \end{aligned}$$

Since observations $Y_1, \ldots, Y_n$ are independent, the likelihood function is

$$\begin{aligned} f(Y \mid \beta_0, \beta_1) &= \prod_{i=1}^{n} f(Y_i \mid \beta_0, \beta_1) \\ &= \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 X_{i,1}}}{1+e^{\beta_0+\beta_1 X_{i,1}}}\right)^{Y_i} \left(\frac{1}{1+e^{\beta_0+\beta_1 X_{i,1}}}\right)^{1-Y_i} \end{aligned}$$

The maximum likelihood estimator (MLE) is

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\max_{\beta_0, \beta_1} f(Y \mid \beta_0, \beta_1)$$

In linear regression (Ch 3), under the normality assumption of residuals $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, the MLE estimators of the regression coefficients are the same as the OLS estimators.

For logistic regression, we can use the `glm` function in `R`:

```
glm(default ~ balance, data = Default, family = binomial);
```

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -10.6513 | 0.3612 | -29.4922 | 0.0000 |
| balance | 0.0055 | 0.0002 | 24.9531 | 0.0000 |

# Interpreting the logistic regression slope

$\hat{\beta}_1 = 0.0055$, and it is significant. What does it mean?

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 \iff \frac{p}{1-p} = e^{\beta_0+\beta_1 X_1}.$$

An unit increase in $X_1$ is associated with

- $\beta_1$ unit increase/decrease in log-odds on average,
- $100(e^{\beta_1} - 1)\%$ increase/decrease in odds.

An increase of $\$10$ in balance is associated with

$$5.65\% = 100 \times (e^{0.0055\times 10} - 1)\%$$

increase in the odds of default.

What does the intercept $\hat{\beta}_0 = -10.6513$ mean?

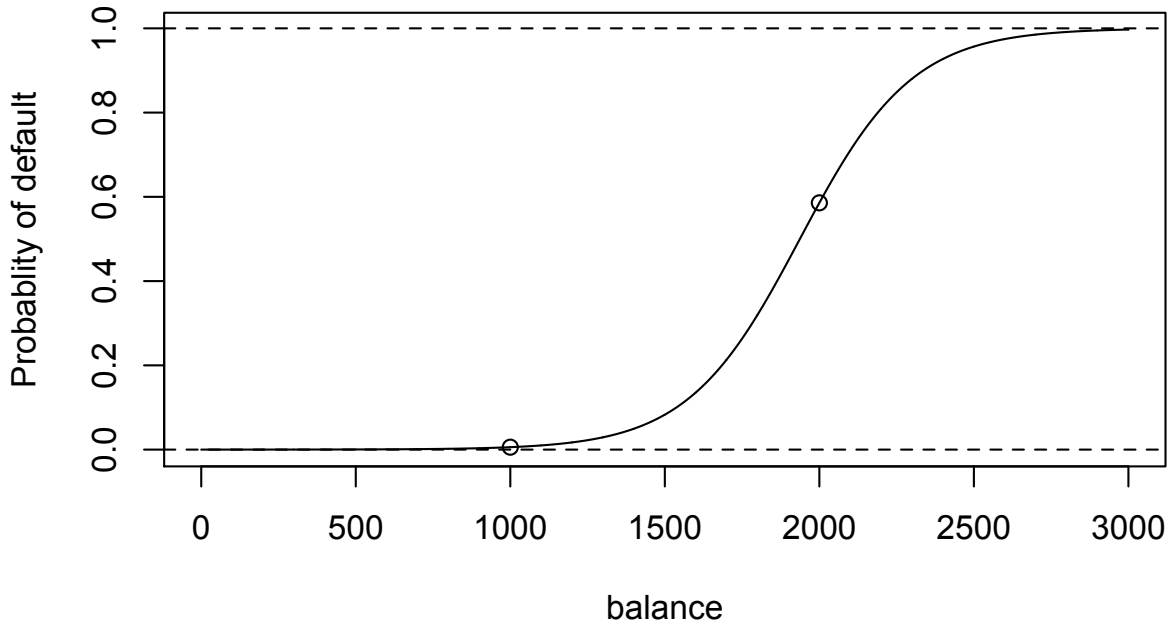
# Prediction

Predict the probability of default of someone of a balance $X$:

$$\hat{p} = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 X)}}$$

For an individual

- with a balance of $\$1000$, $\hat{p} = 0.006$.
- with a balance of $\$2000$, $\hat{p} = 0.586$.



# A categorical predictor

Fit a logistic regression, using `student` as the predictor: code it as a dummy variable: 0 - no, 1 - yes.

`glm(default ~ student, data = Default, family = binomial);`

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -3.5041 | 0.0707 | -49.5542 | 0.0000 |
| student[Yes] | 0.4049 | 0.1150 | 3.5202 | 0.0004 |

Predicted probability of default

- for a student is $\hat{p} = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1)}} = 0.0431.$
- for a non-student is $\hat{p} = \frac{1}{1+e^{-\hat{\beta}_0}} = 0.0292.$

Students tend to have higher default probabilities than non-students.

# Multiple logistic regression

```
glm(default ~ ., data = Default, family = binomial);
```

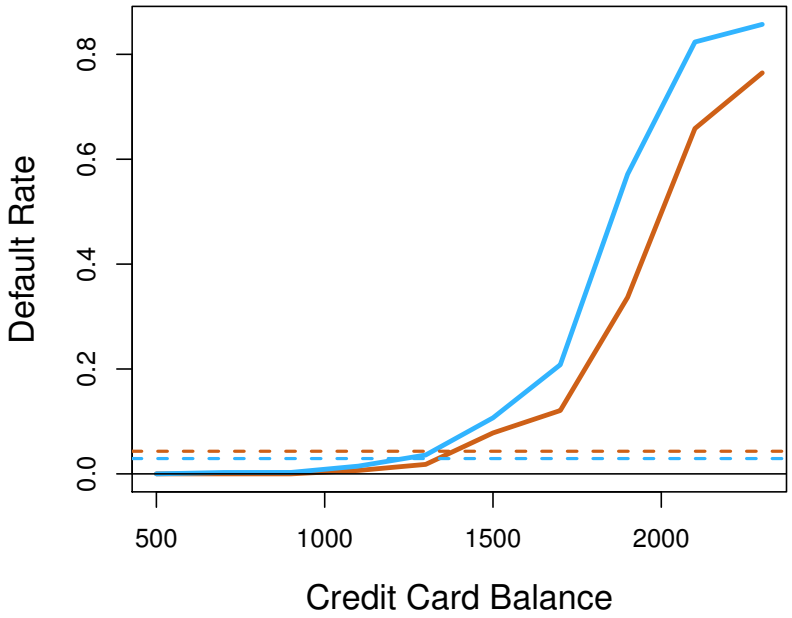
| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -10.8690 | 0.4923 | -22.0801 | 0.0000 |
| student[Yes] | -0.6468 | 0.2363 | -2.7376 | 0.0062 |
| balance | 0.0057 | 0.0002 | 24.7376 | 0.0000 |
| income | 0.0000 | 0.0000 | 0.3698 | 0.7115 |

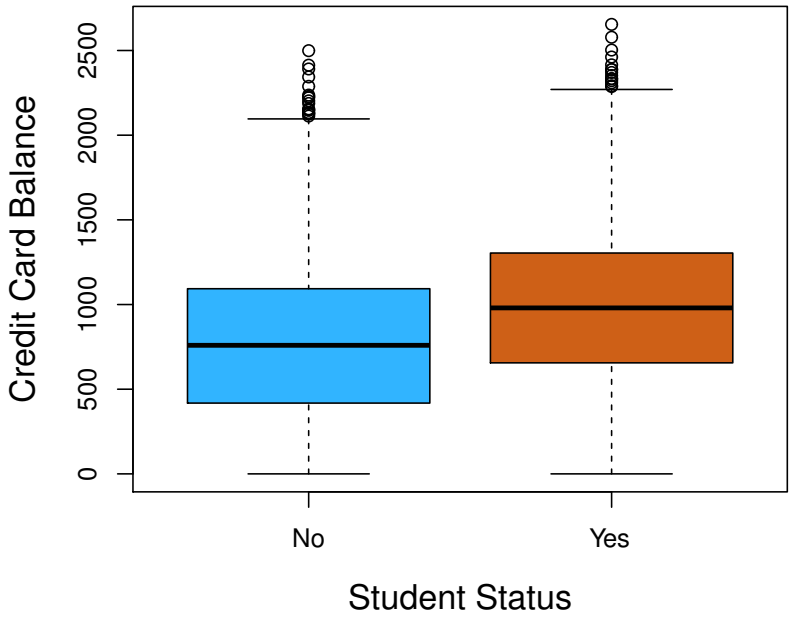
Why is the coefficient for `student` negative here, but positive in the previous logistic regression where $p = 1$?

Given the rest the predictors, ...

For a fixed value of `balance`, a student tends to have a lower default probability than a non-student.

# Students (orange) vs. non-students (blue)




To whom should credit be offered?

- Students tend to have higher balances than non-students, so their marginal default rate is higher than for non-students. *A student is risker than non students if no information about the credit card balance is available*
- But for each level of balance, students default less than non-students. *However, that student is less risky than a non student with the same credit card balance!*

# Discriminant analysis

Model the distribution of $X$ in each class $k$ separately, and then use *Bayes theorem* to flip things around and obtain $P(Y = k \mid X)$.

- Why discriminant analysis?
  - More stable than the logistic regression when the two classes are well-separated.
  - Can handle more than two classes.
- Linear discriminant analysis:

$$D = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  - $D$: discriminant function
  - $\beta_0$: constant
  - $\beta_1, \ldots, \beta_p$: discriminant coefficients
- We want to discriminate between the different categories
- Good predictors tend to have large $\beta_j$'s (weight)

# Bayes theorem

- For events: let $\{A_1, \ldots, A_K\}$ be a partition of the sample space, i.e., $A_i \cap A_j = \emptyset$ for any $1 < i, j < K$, and $\sum_{i=1}^{K} P(A_i) = 1$, then

$$P(A_k \mid B) = \frac{P(B \mid A_k)P(A_k)}{\sum_{j=1}^{K} P(B \mid A_j)P(A_j)}$$

- When $Y \in \{1, 2, \ldots, K\}$ is discrete, then

$$P(Y = k \mid X) = \frac{f(X \mid Y = k)P(Y = k)}{\sum_{j=1}^{K} f(X \mid Y = j)P(Y = j)}$$

  - Denote $\pi_k = P(Y = k)$, is the *prior* probability of class $k$.
  - $P(Y = k \mid X)$ is the *posterior* probability of class $k$, given the observation $X$.
  - Denote $f_k(X) = f(X \mid Y = k)$ is the density of $X$ in class $k$.

# Get yourself a 10 Deutsche Mark bill

# Linear discriminant analysis, $p = 1$

- Under the *assumptions*:
  - In the class $Y = k$,

$$X \sim \mathsf{N}(\mu_k, \sigma^2) \iff f_k(X) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(X-\mu_k)^2}{2\sigma^2}}.$$

  - The variance $\sigma^2$ is the same across classes.
- Assign the observation with predictor $X$ to the class $C(x)$

$$C(X) = \arg\max_k \; P(Y = k \mid X) = \arg\max_k \; f_k(X)\pi_k$$

- Under the assumptions of normality and equal variance, we assign the observation to the class for which

$$\delta_k(X) = \log(\pi_k) - \frac{\mu_k^2}{2\sigma^2} + \frac{\mu_k}{\sigma^2} X$$

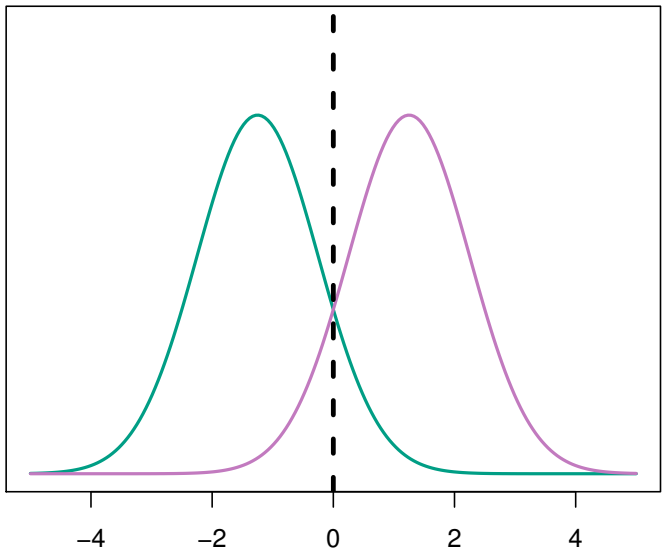
is the largest. Note: $\delta_k(X)$ is a linear function in $X$.

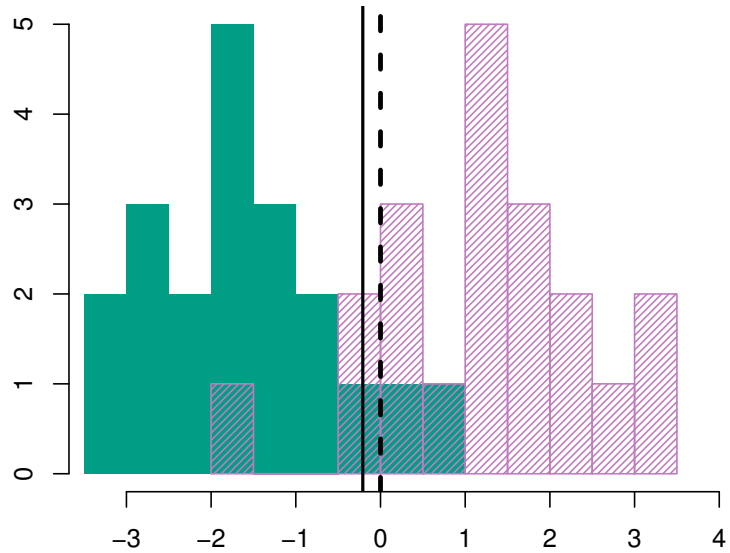
# An example of $K = 2$

When $\pi_1 = \pi_2$, the decision boundary is:

$$\delta_1(X) = \delta_2(X) \Longrightarrow X = \frac{\mu_1 + \mu_2}{2}$$

In the training data, $n_1 = n_2 = 20$




- LDA decision boundary (solid line): error rate 11.1%
- Bayesian decision boundary (dashed line): error rate: 10.6%

# Parameter estimation

Estimators of the unknown parameters: $\mu_1, \ldots, \mu_K, \sigma^2, \pi_1, \ldots, \pi_K$.

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:Y_i=k} X_i,$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:Y_i=k} (X_i - \hat{\mu}_k)^2$$

$$\hat{\pi}_k = \frac{n_k}{n},$$

Estimated LDA discriminant function:

$$\hat{\delta}_k(X) = \log(\hat{\pi}_k) - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \frac{\hat{\mu}_k}{\hat{\sigma}^2} X$$