Logistic regression

Suppose we work with binary outputs, i.e., $y_i \in \{0, 1\}$.

Linear regression may not be the best model.
- $x^T \beta \in \mathbb{R}$ not in $\{0, 1\}$.
- Linearity may not be appropriate. Does doubling the predictor doubles the probability of $Y = 1$? (e.g. probability of going to the beach vs outdoors temperature).

**Logistic regression**: Different perspective. Instead of modelling the $\{0, 1\}$ output, we model the probability that $Y = 0, 1$.

**Idea**: We model $P(Y = 1 | X = x)$.
- Now: $P(Y = 1 | X = x) \in [0, 1]$ instead of $\{0, 1\}$.
- We want to relate that probability to $x^T \beta$.

We assume

$$\log \frac{P(Y = 1 | X = x)}{P(Y = 0 | X = x)} = \log \frac{P(Y = 1 | X = x)}{1 - P(Y = 1 | X = x)} = \log \frac{P(Y = 1 | X = x)}{P(Y = 0 | X = x)} = x^T \beta.$$ 

**Logistic regression (cont.)**

Equivalently,

$$P(Y = 1 | X = x) = \frac{e^{x^T \beta}}{1 + e^{x^T \beta}}$$

$$P(Y = 0 | X = x) = 1 - P(Y = 1 | X = x) = \frac{1}{1 + e^{x^T \beta}}$$

The function $f(x) = e^x/(1 + e^x) = 1/(1 + e^{-x})$ is called the logistic function.

$\log \frac{P(Y = 1 | X = x)}{P(Y = 0 | X = x)}$ is the log-odds ratio.

- Larger positive values of $x^T \beta \Rightarrow p \approx 1$.
- Larger negative values of $x^T \beta \Rightarrow p \approx 0$.

In summary, we are assuming:
- $Y | X = x \sim \text{Bernoulli}(p)$.
- $\logit(p) = \logit(E(Y | X = x)) = x^T \beta$.

More generally, one can use a generalized linear model (GLM). A GLM consists of:
- A probability distribution for $Y | X = x$ from the exponential family.
- A linear predictor $\eta = x^T \beta$.
- A link function $g$ such that $g(E(Y | X = x)) = \eta$. 

**Logistic regression (cont.)**
Logistic regression: estimating the parameters

In logistic regression, we are assuming a model for $Y$. We typically estimate the parameter $\beta$ using maximum likelihood.

**Recall:** If $Y \sim \text{Bernoulli}(p)$, then

$$P(Y = y) = p^y(1-p)^{1-y}, \quad y \in \{0,1\}.$$  

Thus, $L(p) = \prod_{i=1}^n p^{y_i}(1-p)^{1-y_i}$.

Here $p = p(x_i, \beta) = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}$. Therefore,

$$L(\beta) = \prod_{i=1}^n p(x_i, \beta)^{y_i}(1-p(x_i, \beta))^{1-y_i}.$$  

Taking the logarithm, we obtain

$$l(\beta) = \sum_{i=1}^n y_i \log p(x_i, \beta) + \log(1-p(x_i, \beta))$$

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

$$= \sum_{i=1}^n [y_i x_i^T \beta - \log(1 + e^{x_i^T \beta})].$$  

Taking the derivative:

$$\frac{\partial}{\partial \beta_j} l(\beta) = \sum_{i=1}^n y_i x_{ij} - x_{ij} e^{x_i^T \beta} \left[ \frac{1}{1 + e^{x_i^T \beta}} \right].$$

Needs to be solved using numerical methods (e.g. Newton-Raphson).

Logistic regression often performs well in applications.

As before, penalties can be added to regularize the problem or induce sparsity. For example,

$$\min_{\beta} -l(\beta) + \alpha \| \beta \|_1$$

$$\min_{\beta} -l(\beta) + \alpha \| \beta \|_2.$$  

---

Logistic regression with more than 2 classes

- Suppose now the response can take any of $\{1, \ldots, K\}$ values.
- Can still use logistic regression.
- We use the categorical distribution instead of the Bernoulli distribution.
- $P(Y = i|X = x) = p_i, 0 \leq p_i \leq 1, \sum_{i=1}^K p_i = 1$.
- Each category has its own set of coefficients:

  $$P(Y = i|X = x) = \frac{e^{x_i^T \beta_i}}{\sum_{i=1}^K e^{x_i^T \beta_i}}.$$  

- Estimation can be done using maximum likelihood as for the binary case.

---

Multiple classes of data

Other popular approaches to classify data from multiple categories.
- **One versus all** (or one versus the rest) Fit the model to separate each class against the remaining classes. Label a new point $x$ according to the model for which $x^T \beta + \beta_0$ is the largest.

Need to fit the model $K$ times.
Multiple classes of data (cont.)

- One versus one:
  1. Train a classifier for each possible pair of classes.
     Note: There are $\binom{K}{2} = K(K-1)/2$ such pairs.
  2. Classify a new point according to a majority vote: count the number of times the new point is assigned to a given class, and pick the class with the largest number.

\[
\begin{align*}
2 & \quad \bullet \\
X & \quad X \\
\cdots & \\
1 & \\
0 & \quad \star \\
\end{align*}
\]

Need to fit the model $\binom{K}{2}$ times (computationally intensive).

Linear discriminant analysis (LDA)

- Categorical data $Y$. Predictors $X_1, \ldots, X_p$.
  - We saw how logistic regression can be used to predict $Y$ by modelling the log-odds
    \[
    \log \frac{P(Y = 1 | X = x)}{P(Y = 0 | X = x)} = x^T \beta.
    \]
  - More now examine other models for $P(Y = i | X = x)$.

Recall: Bayes' theorem (Rev. Thomas Bayes, 1701–1761). Given two events $A, B$:

\[
P(A|B) = \frac{P(B|A)P(A)}{P(B)}
\]

Using Bayes' theorem

- $P(Y = i | X = x)$ harder to model.
- $P(X = x | Y = i)$ easier to model.

Going back to our prediction using Bayes' theorem:

\[
P(Y = i | X = x) = \frac{P(X = x | Y = i)P(Y = i)}{P(X = x)}
\]

Using Bayes' theorem

More precisely, suppose

- $Y \in \{1, \ldots, k\}$,
- $P(Y = i) = \pi_i \quad (i = 1, \ldots, k)$,
- $P(X = x | Y = i) \sim f_i(x) \quad (i = 1, \ldots, k)$.

Then

\[
P(Y = i | X = x) = \frac{P(X = x | Y = i)P(Y = i)}{P(X = x)}
\]

= \frac{P(X = x | Y = i)P(Y = i)}{\sum_{j=1}^{k} P(X = x | Y = j)P(Y = j)}

= \frac{f_i(x)\pi_i}{\sum_{j=1}^{k} f_j(x)\pi_j}.

- We can easily estimate $\pi_i$ using the proportion of observations in category $i$.
- We need a model for $f_i(x)$. 
Using a Gaussian model: LDA and QDA

A natural model for the $f_j$s is the multivariate Gaussian distribution:

$$f_j(x) = \frac{1}{\sqrt{(2\pi)^p \det \Sigma_j}} e^{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)}.$$

**Linear discriminant analysis (LDA):** We assume $\Sigma_j = \Sigma$ for all $j = 1, \ldots, k$.

**Quadratic discriminant analysis (QDA):** general case, i.e., $\Sigma_j$ can be distinct.

Note: When $p$ is large, using QDA instead of LDA can dramatically increase the number of parameters to estimate.

In order to use LDA or QDA, we need:
- An estimate of the class probabilities $\pi_j$.
- An estimate of the mean vectors $\mu_j$.
- An estimate of the covariance matrices $\Sigma_j$ (or $\Sigma$ for LDA).

**Estimating the parameters**

**LDA:** Suppose we have $N$ observations, and $N_j$ of these observations belong to the $j$ category ($j = 1, \ldots, k$). We use

- $\hat{\pi}_j = N_j/N$.
- $\hat{\mu}_j = \frac{1}{N_j} \sum_{y_i=j} x_i$ (average of $x$ over each category).
- $\hat{\Sigma} = \frac{1}{N-k} \sum_{j=1}^k \sum_{y_i=j} (x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T$. (Pooled variance.)

**QDA:** quadratic decision boundary

Let us now examining the log-odds for QDA: in that case no simplification occurs as before

$$\log \frac{P(Y = l | X = x)}{P(Y = m | X = x)} = \log \frac{f_l(x)}{f_m(x)} + \log \frac{\pi_l}{\pi_m} + 1 \frac{1}{2} \log \frac{\det \Sigma_m}{\det \Sigma_l} - \frac{1}{2} (x - \mu_l)^T \Sigma_l^{-1}(x - \mu_l) - \frac{1}{2} (x - \mu_m)^T \Sigma_m^{-1}(x - \mu_m).$$

**LDA: linearity of the decision boundary**

In the previous figure, we saw that the decision boundary is linear. Indeed, examining the log-odds:

$$\log \frac{P(Y = l | X = x)}{P(Y = m | X = x)} = \log \frac{f_l(x)}{f_m(x)} + \log \frac{\pi_l}{\pi_m} + \frac{1}{2} (\mu_l + \mu_m)^T \Sigma^{-1}(\mu_l - \mu_m) + x^T \Sigma^{-1}(\mu_l - \mu_m)$$

$$= \beta_0 + x^T \beta.$$

Note that the previous expression is linear in $x$.

Recall that for logistic regression, we model

$$\log \frac{P(Y = l | X = x)}{P(Y = m | X = x)} = \beta_0 + x^T \beta.$$

How is this different from LDA?
- In LDA, the parameters are more constrained and are not estimated the same way.
- Can lead to smaller variance if the Gaussian model is correct.
- In practice, logistic regression is considered safer and more robust.
- LDA and logistic regression often return similar results.
Despite their simplicity, LDA and QDA often perform very well.

- Both techniques are widely used.

Problems when \( n < p \):

- Estimating covariance matrices when \( n \) is small compared to \( p \) is challenging.

  - The sample covariance (MLE for Gaussian)
  
  \[
  S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T
  \]
  
  has rank at most \( \min(n, p) \) so is singular when \( n < p \).

  - This is a problem since \( \Sigma \) needs to be inverted in LDA and QDA.

Many strategies exist to obtain better estimates of \( \Sigma \) (or \( \Sigma_j \)). Among them:

- Regularization methods. E.g. \( \hat{\Sigma}(\lambda) = \hat{\Sigma} + \lambda I \).
- Graphical modelling (discussed later during the course).