S&DS 242/542: Theory of Statistics

Lecture 24: Logistic regression

Classification models

Let (X, Y) be a predictor-response pair. A **classification model** is a model where the response Y is categorical rather quantitative.

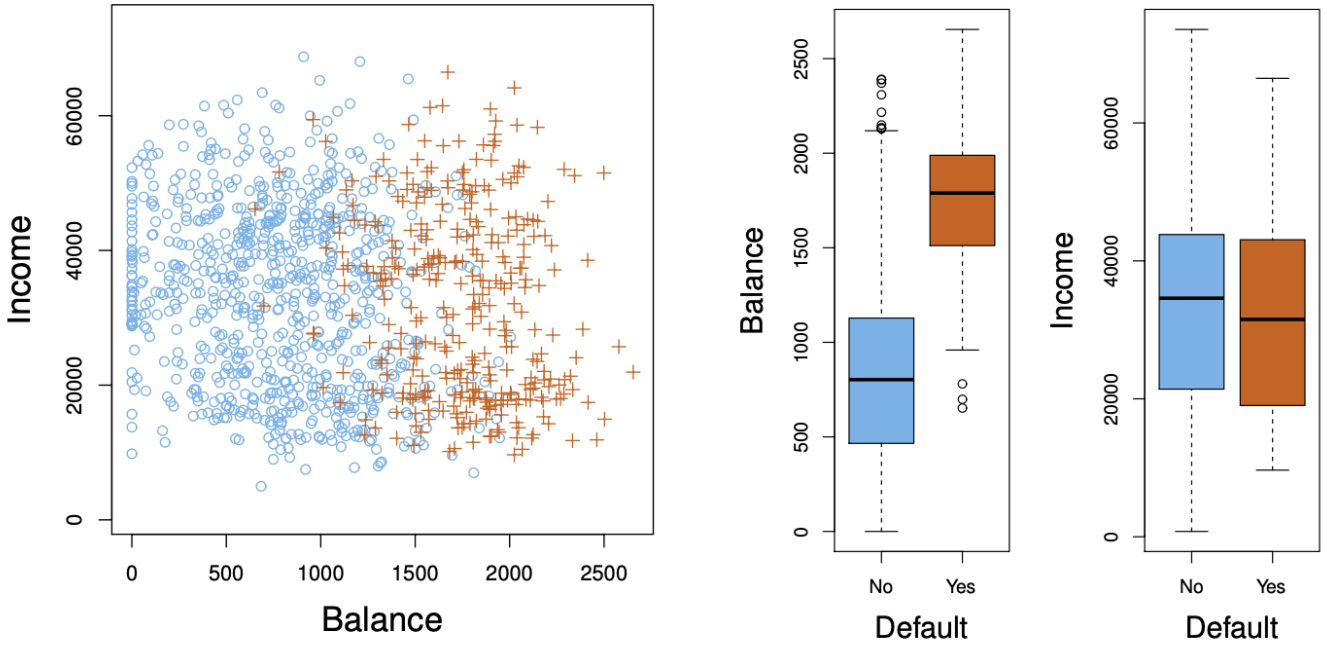
In the simplest setting, Y has two possible classes, which we may encode as 0 and 1. The model may be specified by

$$\mathbb{P}[Y = 1 \mid X] = p(X), \qquad \mathbb{P}[Y = 0 \mid X] = 1 - p(X)$$

where $p(X)$ is an unknown function of the predictors X.

Our goal may be to learn $p(X)$ from data $(X_1, Y_1), \ldots, (X_n, Y_n)$, and to predict a new response $Y_{\text{test}} \in \{0, 1\}$ given X_{test}.

Predicting credit card default (ISLR)



Y: Whether an individual defaults on their credit card payment

X: Monthly credit card balance, annual income, student status

Linear regression?

Suppose $X \in \mathbb{R}^p$ consists of p predictors. One approach for modeling $p(X)$ is to assume a linear model

$$\mathbb{P}[Y = 1 \mid X] = p(X \mid \beta) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

and to fit linear regression to (X, Y), treating $Y \in \{0, 1\}$ as a quantitative response.

Potential downsides of this approach:

- The modeled probability $p(X \mid \beta)$ may fall outside $[0, 1]$
- Difficult to extend to settings where Y has more than 2 classes

This approach is more commonly used in applications with only 2 classes and where responses are harder to predict, i.e. $p(X)$ tends to be far from 0 or 1.

Logistic regression

Logistic regression instead models $p(X)$ by a *logistic function*

$$\mathbb{P}[Y = 1 \mid X] = p(X \mid \beta) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}$$

which always takes values in $(0, 1)$.

Equivalently,

$$\log \frac{p(X \mid \beta)}{1 - p(X \mid \beta)} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Thus the model assumes that $\log \frac{p(X)}{1-p(X)}$ — the *log-odds* of $Y = 1$ to $Y = 0$ — is a linear function of the predictors $X = (x_1, \ldots, x_p)$.

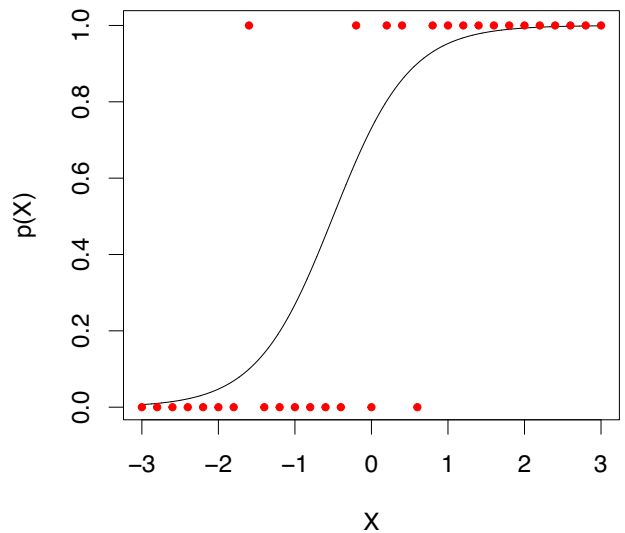
Simple logistic regression

Simple logistic regression model

Consider a single predictor $X \in \mathbb{R}$. The logistic regression model is

$$\mathbb{P}[Y = 1 \mid X] = p(X \mid \beta) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

with parameters $\beta = (\beta_0, \beta_1)$. For $\beta_1 > 0$, $p(X \mid \beta)$ looks like:



- β_0 is the log-odds of $Y = 1$ to $Y = 0$ when $X = 0$
- β_1 is the increase in this log-odds per unit increase of X

Maximum likelihood estimation

The regression coefficients $\beta = (\beta_0, \beta_1)$ are commonly estimated using maximum likelihood.

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are IID with

$$X \sim f_X(x), \qquad \mathbb{P}[Y = 1 \mid X] = p(X \mid \beta) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

The log-likelihood is:

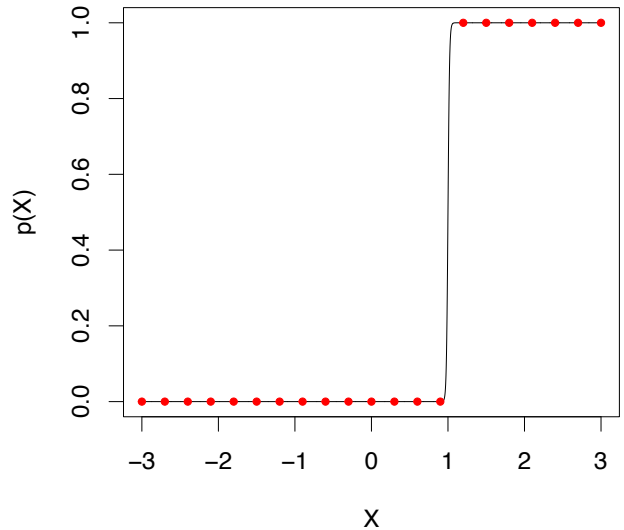
$$\ell_n(\beta) = \sum_{i=1}^{n} \log \left[p(X_i|\beta)^{Y_i} (1 - p(X_i|\beta))^{1-Y_i} \cdot f_X(X_i) \right]$$

$$= \sum_{i=1}^{n} \left[Y_i \underbrace{\log \frac{p(X_i|\beta)}{1 - p(X_i|\beta)}}_{= \beta_0 + \beta_1 X_i} + \log(1 - p(X_i|\beta)) + \log f_X(X_i) \right]$$

$$= \sum_{i=1}^{n} \left[Y_i (\beta_0 + \beta_1 X_i) - \log(1 + e^{\beta_0 + \beta_1 X_i}) + \log f_X(X_i) \right]$$

Maximum likelihood estimation

If the data are perfectly separable by a threshold for X, then the likelihood is maximized when $\beta_1 \to \infty$ or $-\infty$, and the MLE does not exist.



Otherwise the MLE exists and is given by

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1) = \arg\max_{\beta_0, \beta_1} \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \log(1 + e^{\beta_0 + \beta_1 X_i})$$

In general, $\hat{\beta}$ does not have a closed-form expression.

Derivatives of the log-likelihood

$$\frac{\partial l_n}{\partial \beta_0} = \sum_{i=1}^{n} Y_i - \underbrace{\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}}_{=p(X_i|\beta)} = \sum_{i=1}^{n} Y_i - p(X_i|\beta)$$

$$\frac{\partial l_n}{\partial \beta_1} = \sum_{i=1}^{n} Y_i X_i - \frac{X_i e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}} = \sum_{i=1}^{n} (Y_i - p(X_i|\beta)) \cdot X_i$$

$$\Rightarrow \nabla l_n(\beta) = \begin{pmatrix} \frac{\partial l_n}{\partial \beta_0} \\ \frac{\partial l_n}{\partial \beta_1} \end{pmatrix} = \sum_{i=1}^{n} (Y_i - p(X_i|\beta)) \cdot \begin{pmatrix} 1 \\ X_i \end{pmatrix} \in \mathbb{R}^2$$

Wir $Y_i - p(X_i|\beta) = Y_i - 1 + \frac{1}{1+e^{\beta_0+\beta_1 X_i}}$

Derivatives of the log-likelihood

$$\frac{\partial^2 \ell_n}{\partial \beta_0^2} = \sum_{i=1}^{n} - \underbrace{\frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2}}_{= p(x_i|\beta)\cdot(1-p(x_i|\beta))} = - \sum_{i=1}^{n} p(x_i|\beta)(1-p(x_i|\beta))$$

$$\frac{\partial^2 \ell_n}{\partial \beta_0 \partial \beta_1} = \sum_{i=1}^{n} - \frac{x_i e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} = - \sum_{i=1}^{n} p(x_i|\beta)(1-p(x_i|\beta)) \cdot x_i$$

$$\frac{\partial^2 \ell_n}{\partial \beta_1^2} = \sum_{i=1}^{n} - \frac{x_i^2 e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} = - \sum_{i=1}^{n} p(x_i|\beta)(1-p(x_i|\beta)) \cdot x_i^2$$

$$\Rightarrow \nabla^2 \ell_n(\beta) = - \sum_{i=1}^{n} p(x_i|\beta)(1-p(x_i|\beta)) \cdot \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

Concavity of the log-likelihood

The second derivative matrix (Hessian) is

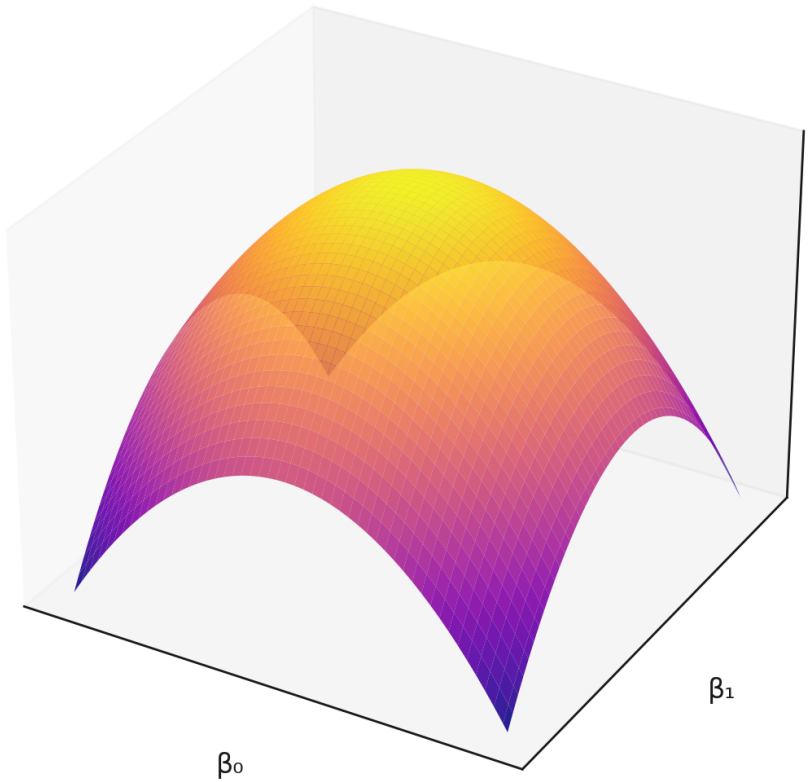
$$\nabla^2 \ell_n(\beta) = -\sum_{i=1}^{n} p(X_i \mid \beta)(1 - p(X_i \mid \beta)) \begin{pmatrix} 1 & X_i \\ X_i & X_i^2 \end{pmatrix}$$

This satisfies, for every non-zero vector $u = (u_0, u_1) \in \mathbb{R}^2$,

$$\begin{aligned} u^\top \nabla^2 \ell_n(\beta) u &= -\sum_{i=1}^{n} p(X_i \mid \beta)(1 - p(X_i \mid \beta))(u_0^2 + 2u_0 u_1 X_i + u_1^2 X_i^2) \\ &= -\sum_{i=1}^{n} p(X_i \mid \beta)(1 - p(X_i \mid \beta))(u_0 + u_1 X_i)^2 < 0 \end{aligned}$$

This means that $\nabla^2 \ell_n(\beta)$ is *negative-definite*, and $\ell_n(\beta)$ is a strictly concave function of $\beta = (\beta_0, \beta_1)$.

Characterization of the MLEs



The MLE $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ (when it exists) is the unique solution of

$$0 = \nabla \ell_n(\beta) = \sum_{i=1}^{n} (Y_i - p(X_i \mid \beta)) \begin{pmatrix} 1 \\ X_i \end{pmatrix}$$

Newton's method and Iterative Reweighted Least Squares

The MLE is often computed numerically using Newton's method:

$$\beta^{(t+1)} = \beta^{(t)} - \left[\nabla^2 \ell_n(\beta^{(t)})\right]^{-1} \nabla \ell_n(\beta^{(t)})$$

(For example, this is implemented in the `glm` function in `R`.)

Proposition

Given parameters $\beta^{(t)} = (\beta_0^{(t)}, \beta_1^{(t)})$, define sample weights and adjusted responses, for $i = 1, \ldots, n$, by

$$w_i^{(t)} = p(X_i \mid \beta^{(t)})(1 - p(X_i \mid \beta^{(t)}))$$
$$Z_i^{(t)} = \beta_0^{(t)} + \beta_1^{(t)} X_i + \left(Y_i - p(X_i \mid \beta^{(t)})\right)/w_i^{(t)}$$

Then $\beta^{(t+1)}$ solves the weighted least-squares problem

$$\beta^{(t+1)} = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} w_i^{(t)} (\beta_0 + \beta_1 X_i - Z_i^{(t)})^2$$

Newton's method and Iterative Reweighted Least Squares

Proof: Weighted least-squares solutions solve:

$$0 = \sum_{i=1}^{n} w_i^{(t)} \left(\beta_0 + \beta_1 X_i - \underbrace{z_i^{(t)}}_{\beta_0^{(t)} + \beta_1^{(t)} X_i + \frac{Y_i - \rho(X_i|\beta^{(t)})}{w_i^{(t)}}} \right)$$

$$= (\beta_0 - \beta_0^{(t)}) \cdot \sum_{i=1}^{n} w_i^{(t)} + (\beta_1 - \beta_1^{(t)}) \cdot \sum_{i=1}^{n} w_i^{(t)} X_i - \sum_{i=1}^{n} (Y_i - \rho(X_i|\beta^{(t)}))$$

$$0 = \sum_{i=1}^{n} w_i^{(t)} \cdot X_i \left(\beta_0 + \beta_1 X_i - z_i^{(t)} \right)$$

$$= (\beta_0 - \beta_0^{(t)}) \cdot \sum_{i=1}^{n} w_i^{(t)} X_i + (\beta_1 - \beta_1^{(t)}) \sum_{i=1}^{n} w_i^{(t)} X_i^2 - \sum_{i=1}^{n} (Y_i - \rho(X_i|\beta^{(t)})) \cdot X_i$$

$$\Rightarrow \underbrace{\sum_{i=1}^{n} (Y_i - \rho(X_i|\beta^{(t)})) \cdot \begin{pmatrix} 1 \\ X_i \end{pmatrix}}_{= \nabla \ell_n(\beta^{(t)})} = \underbrace{\sum_{i=1}^{n} w_i^{(t)} \begin{pmatrix} 1 & X_i \\ X_i & X_i^2 \end{pmatrix}}_{= -\nabla^2 \ell_n(\beta^{(t)})} \begin{pmatrix} \beta_0 - \beta_0^{(t)} \\ \beta_1 - \beta_1^{(t)} \end{pmatrix} \Rightarrow \beta^{(t+1)} = \beta^{(t)} - (\nabla^2 \ell_n)^{-1} \nabla \ell_n$$

Fisher information matrix

The standard errors of $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ may be derived from the Fisher information matrix.

Let $f(X, Y \mid \beta)$ denote the likelihood for a single observation. Then the Fisher information matrix for $\beta = (\beta_0, \beta_1)$ is

$$\begin{aligned} I(\beta) &= -\mathbb{E}\big[\nabla^2 \log f(X, Y \mid \beta)\big] \\ &= \mathbb{E}\left[p(X \mid \beta)(1 - p(X \mid \beta)) \begin{pmatrix} 1 & X \\ X & X^2 \end{pmatrix}\right] \in \mathbb{R}^{2\times 2} \end{aligned}$$

where $\mathbb{E}$ denotes the expectation over the distribution of X. This distribution is usually unknown, and $I(\beta)$ is estimated using

$$\widehat{I}(\beta) = \frac{1}{n}\sum_{i=1}^{n} p(X_i \mid \beta)(1 - p(X_i \mid \beta)) \begin{pmatrix} 1 & X_i \\ X_i & X_i^2 \end{pmatrix}$$

The matrix $n \cdot \widehat{I}(\beta)$ is often called the **observed information**.

Asymptotic normality of the MLE

Theorem

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ *be IID, where*

$$\mathbb{P}[Y = 1 \mid X] = p(X \mid \beta) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

(i.e. the logistic regression model is correctly specified). Then, under regularity conditions,

$$\sqrt{n}(\hat{\beta} - \beta) \to \mathcal{N}\left(0, I(\beta)^{-1}\right)$$

in distribution as $n \to \infty$.

Let $[\cdot]_{00}, [\cdot]_{01}, [\cdot]_{10}, [\cdot]_{11}$ denote the entries of a 2×2 matrix. Then for large n, the standard errors of $\hat{\beta}_0, \hat{\beta}_1$ are approximately

$$\mathrm{se}_0 = \sqrt{[(n \cdot I(\beta))^{-1}]_{00}}, \quad \mathrm{se}_1 = \sqrt{[(n \cdot I(\beta))^{-1}]_{11}}$$

Standard error estimates and confidence intervals

A model-based estimate for se_1 is given by

$$\hat{\mathrm{se}}_1 = \sqrt{[(n \cdot \widehat{I}(\hat{\beta}))^{-1}]_{11}}$$

where $n \cdot \widehat{I}(\hat{\beta})$ is the observed information evaluated at the MLE $\hat{\beta}$.

Assuming that the logistic regression model for $\mathbb{P}[Y = 1 \mid X]$ is correct with true parameter β, we expect $\hat{\beta} \to \beta$ and $\widehat{I}(\beta) \to I(\beta)$ in probability as $n \to \infty$, so $\hat{\mathrm{se}}_1 / \mathrm{se}_1 \to 1$.

A confidence interval for β_1 with asymptotic coverage $1 - \alpha$ is then given by

$$\hat{\beta}_1 \pm z^{(\alpha/2)} \hat{\mathrm{se}}_1$$

Standard error estimates and confidence intervals for β_0 may be obtained analogously.

Hypothesis tests

A test of

$$H_0 : \beta_1 = 0 \qquad \text{vs.} \qquad H_1 : \beta_1 > 0$$

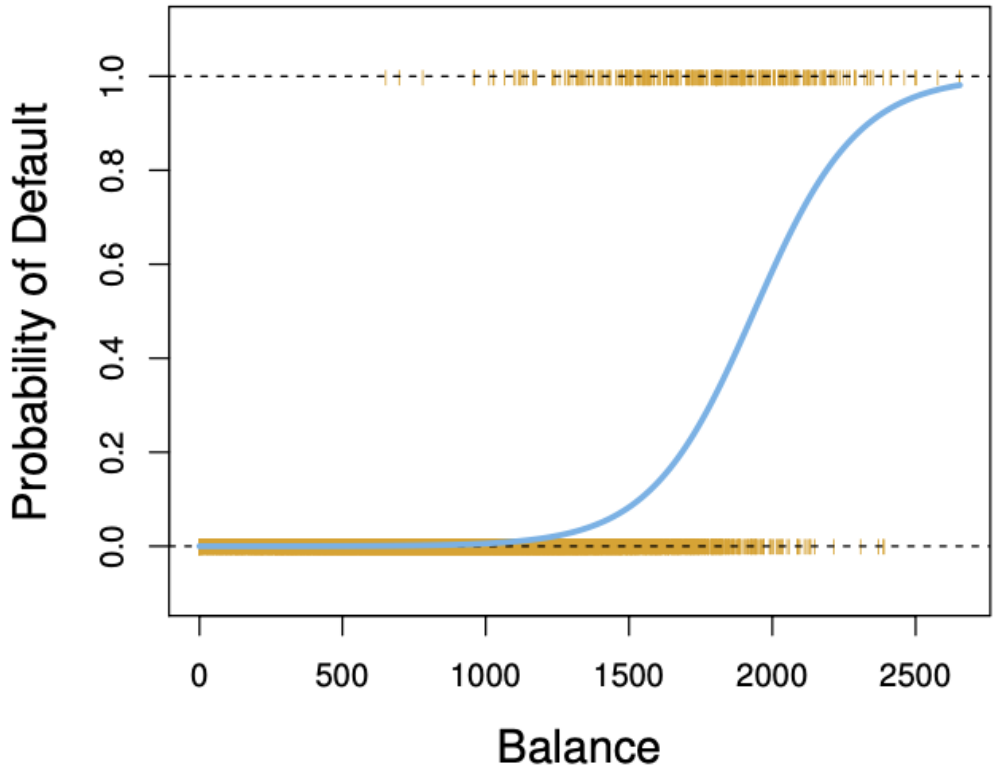
may be based on the **z-statistic** $\frac{\hat{\beta}_1}{\hat{\mathrm{se}}_1}$. Under H_0, $\hat{\beta}_1/\mathrm{se}_1 \to \mathcal{N}(0,1)$ in distribution as $n \to \infty$. Then by Slutsky's Lemma,

$$\hat{\beta}_1/\hat{\mathrm{se}}_1 \to \mathcal{N}(0,1)$$

so an asymptotic level-α test may reject H_0 when $\frac{\hat{\beta}_1}{\hat{\mathrm{se}}_1} > z^{(\alpha)}$, the upper-$\alpha$ point of the standard normal distribution.

Under misspecification of the logistic regression model for $\mathbb{P}[Y = 1 \mid X]$, this estimate $\hat{\mathrm{se}}_1 = \sqrt{[(n \cdot \widehat{I}(\hat{\beta}))^{-1}]_{11}}$ may not be accurate for the standard error of $\hat{\beta}_1$ even for large n, and robust estimators (e.g. using nonparametric bootstrap) may be preferred.

Predicting credit card default (ISLR)



	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

Prediction on test data

Given a new observation $X_{\text{test}} = x$, the probability that $Y_{\text{test}} = 1$ under the logistic regression model is

$$\mathbb{P}[Y_{\text{test}} = 1 \mid X_{\text{test}} = x] = p(x \mid \beta)$$

which we may estimate by

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

For large n, the standard error of $p(x \mid \hat{\beta})$ may be computed from the approximate normal distribution for $\hat{\beta}_0 + \hat{\beta}_1 x$ and the delta method. This may then be used to construct asymptotic confidence intervals for $p(x \mid \beta)$.

Multiple predictors and multiple classes

Logistic regression with multiple predictors

For p predictors $X = (x_1, \ldots, x_p) \in \mathbb{R}^p$ and a binary response $Y \in \{0, 1\}$, the logistic regression model takes the form

$$\mathbb{P}[Y = 1 \mid X] = p(X \mid \beta) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}$$

with parameters $\beta = (\beta_0, \ldots, \beta_p) \in \mathbb{R}^{p+1}$.

These parameters are commonly estimated by the MLE

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\max} \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) - \log(1 + e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}})$$

which exists as long as the data are not perfectly separable by thresholding a linear function of X. There is in general no closed-form expression for $\hat{\beta}$, and $\hat{\beta}$ is often computed by Newton's method for small p or gradient ascent for large p.

Fisher information matrix

The Fisher information matrix for $\beta = (\beta_0, \ldots, \beta_p)$ is given by

$$I(\beta) = -\mathbb{E}\big[\nabla^2 \log f(X, Y \mid \beta)\big]$$

$$= \mathbb{E}\left[p(X \mid \beta)(1 - p(X \mid \beta)) \cdot \begin{pmatrix} 1 & x_1 & x_2 & \cdots & x_p \\ x_1 & x_1^2 & x_1 x_2 & \cdots & x_1 x_p \\ x_2 & x_2 x_1 & x_2^2 & \cdots & x_2 x_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_p & x_p x_1 & x_p x_2 & \cdots & x_p^2 \end{pmatrix} \right]$$

with expectation $\mathbb{E}$ over the distribution of $X = (x_1, \ldots, x_p)$.

This may be estimated by $\widehat{I}(\beta)$ which replaces this expectation by the sample average over the training data $X_1, \ldots, X_n$.

Asymptotic normality of the MLE

Theorem

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ *be IID, where* $\mathbb{P}[Y = 1 \mid X] = p(X \mid \beta)$ *(i.e. the logistic regression model is correctly specified). Then, under regularity conditions,*

$$\sqrt{n}(\hat{\beta} - \beta) \to \mathcal{N}\left(0, I(\beta)^{-1}\right)$$

in distribution as $n \to \infty$.

Thus the standard error of $\hat{\beta}_j$ for the j^{th} predictor is approximately

$$\text{se}_j = \sqrt{[(n \cdot I(\beta))^{-1}]_{jj}}$$

We may estimate this by $\hat{\text{se}}_j = \sqrt{[(n \cdot \widehat{I}(\hat{\beta}))^{-1}]_{jj}}$, and use this estimate in confidence intervals and hypothesis tests for β_j.

Predicting credit card default (ISLR)

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	−10.8690	0.4923	−22.08	<0.0001
`balance`	0.0057	0.0002	24.74	<0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	−0.6468	0.2362	−2.74	0.0062

As in our discussion of linear regression, the meaning of β_j depends on the other predictors in the model, and represents the effect of the j^{th} predictor on $p(X)$ *after controlling for* the other predictors.

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	−3.5041	0.0707	−49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

In this example, being a student decreases the predicted probability of default ($Y = 1$) when controlling for credit card balance and income, but increases this probability in a univariate regression model. This occurs because students tend to carry higher balances.

Multinomial logistic regression

Suppose Y has K classes, encoded as $\{0, 1, \ldots, K-1\}$ where is no meaning to the ordering of these classes.

We may extend logistic regression to the **multinomial logistic regression** model,

$$\mathbb{P}[Y = 0 \mid X] = \frac{1}{1 + \sum_{\ell=1}^{K-1} e^{\beta_{\ell 0} + \beta_{\ell 1} x_1 + \ldots + \beta_{\ell p} x_p}}$$

$$\mathbb{P}[Y = k \mid X] = \frac{e^{\beta_{k0} + \beta_{k1} x_1 + \ldots + \beta_{kp} x_p}}{1 + \sum_{\ell=1}^{K-1} e^{\beta_{\ell 0} + \beta_{\ell 1} x_1 + \ldots + \beta_{\ell p} x_p}} \text{ for } k = 1, \ldots, K-1$$

This has parameters $(\beta_{k0}, \ldots, \beta_{kp})$ for each class $k = 1, \ldots, K-1$, where the log-odds of $Y = k$ to $Y = 0$ is modeled linearly as

$$\log \frac{\mathbb{P}[Y = k \mid X]}{\mathbb{P}[Y = 0 \mid X]} = \beta_{k0} + \beta_{k1} x_1 + \ldots + \beta_{kp} x_p$$

Log-odds and parameter differences

One may think of Class 0 as a baseline class, where we fix its parameters by convention

$$(\beta_{00}, \beta_{01}, \ldots, \beta_{0p}) = (0, 0, \ldots, 0)$$

Then for each pair of classes $k, \ell \in \{0, 1, \ldots, K-1\}$,

$$\log \frac{\mathbb{P}[Y = k \mid X]}{\mathbb{P}[Y = \ell \mid X]} = (\beta_{k0} - \beta_{\ell 0}) + (\beta_{k1} - \beta_{\ell 1})x_1 + \ldots + (\beta_{kp} - \beta_{\ell p})x_p$$

so the parameter differences $\beta_{kj} - \beta_{\ell j}$ model the log-odds of $Y = k$ to $Y = \ell$.

The choice of baseline class is unimportant, in the sense that choosing a different baseline class will shift $(\beta_{0j}, \ldots, \beta_{k-1,j})$ by a constant while keeping the differences $\beta_{kj} - \beta_{\ell j}$ unchanged.