Lecture 26 — Logistic regression

26.1 The logistic regression model

Example 26.1. An internet company would like to understand what factors influence whether a visitor to a webpage clicks on an advertisement. Suppose it has available historical data of n ad impressions, each impression corresponding to a single ad being shown to a single visitor. For the ith impression, let $Y_i \in \{0, 1\}$ be such that $Y_i = 1$ if the visitor clicked on the ad, and $Y_i = 0$ otherwise. The internet company also has available various attributes for each impression, such as the position and size of the ad on the webpage, the company or product being advertised, the age and gender of the visitor, the time of day, the month of the year, etc. For each ith impression, suppose that all of these attributes are encoded numerically as p covariates $x_{i1}, \ldots, x_{ip} \in \mathbb{R}$.

The **logistic regression model** assumes each response Y_i is an independent random variable with distribution Bernoulli(p_i), where the log-odds corresponding to p_i is modeled as a linear combination of the covariates plus a possible intercept term:

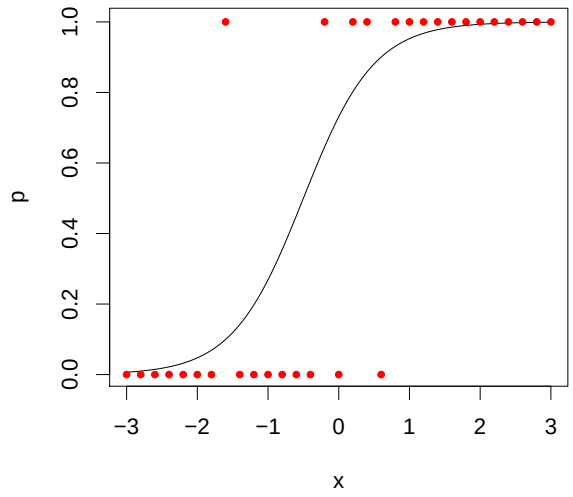
$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

The intercept β_0 represents the "baseline" log-odds of the visitor clicking on the ad, if all of the covariates take value 0. Each coefficient β_j represents the amount of increase or decrease in the log-odds, if the value of the covariate x_{ij} is increased by 1 unit. The above may be equivalently written as

$$\mathbb{P}[Y_i = 1] = p_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}}. \tag{26.1}$$

As in the case of the linear model, we will treat the covariates as fixed and known quantities. The unknown parameters are the regression coefficients $\beta = (\beta_0, \ldots, \beta_p)$.

When there is only one covariate, $p = 1$, we simply write $x_1 = x_{11}, \ldots, x_n = x_{n1}$. The the picture below illustrates the logistic regression model, where the red points correspond to the data values $(x_1, Y_1), \ldots, (x_n, Y_n)$ of the covariate and response, and the black curve shows the probability function $p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$:



26.2 Statistical inference

We will explore the following inferential questions:

- Estimate the regression coefficiets $\beta_0, \beta_1, \ldots, \beta_p$
- Estimate the "conversion" probability that a new impression, with covariate values $(\tilde{x}_1, \ldots, \tilde{x}_p)$, will lead to click on the ad
- Test whether $\beta_j = 0$ for a particular covariate j, say the age of the visitor, and provide a confidence interval for β_j

Since the responses $Y_1, \ldots, Y_n$ are independent Bernoulli random variables, the likelihood for the logistic regression model is given by

$$\text{lik}(\beta_0, \ldots, \beta_p) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i} = \prod_{i=1}^{n}(1-p_i)\left(\frac{p_i}{1-p_i}\right)^{Y_i},$$

where p_i is defined as a function of $\beta_0, \ldots, \beta_p$ and the covariates $x_{i1}, \ldots, x_{ip}$ by equation (26.1). Then, introducing for convenience a covariate $x_{i0} \equiv 1$ for all i that captures the intercept term, the log-likelihood is

$$l(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} Y_i \log\frac{p_i}{1-p_i} + \log(1-p_i) = \sum_{i=1}^{n}\left(Y_i \sum_{j=0}^{p}\beta_j x_{ij} - \log\left(1 + e^{\sum_{j=0}^{p}\beta_j x_{ij}}\right)\right).$$

To estimate the parameters $\beta_0, \ldots, \beta_p$, we may compute the MLE. For the function $f(x) = \log(1+e^x)$, $f'(x) = \frac{e^x}{1+e^x} = 1 - \frac{1}{1+e^x}$. Then setting the partial derivatives of the log-likelihood equal to 0 and applying the chain rule yields the equations (for $m = 0, \ldots, p$)

$$0 = \frac{\partial l}{\partial \beta_m} = \sum_{i=1}^{n} x_{im}\left(Y_i - \frac{e^{\sum_{j=0}^{p}\beta_j x_{ij}}}{1 + e^{\sum_{j=0}^{p}\beta_j x_{ij}}}\right). \tag{26.2}$$

These equations may be solved numerically (e.g. by Newton-Raphson) to obtain the MLEs $\hat{\beta}_0, \ldots, \hat{\beta}_p$. To estimate the conversion probability for a new impression with covariates $\tilde{x}_1, \ldots, \tilde{x}_p$, we may use the plugin estimate

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_1 + \ldots + \hat{\beta}_p \tilde{x}_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_1 + \ldots + \hat{\beta}_p \tilde{x}_p}}. \tag{26.3}$$

To test if a particular coefficient is 0, say $H_0 : \beta_p = 0$, one method is using the generalized likelihood ratio test. This null hypothesis corresponds to a sub-model with one fewer free parameter. We may calculate the sub-model MLEs $\hat{\beta}_{0,0}, \ldots, \hat{\beta}_{0,p-1}$ from the same score equations as (26.2) except with the pth covariate removed, and use the generalized likelihood ratio statistic

$$-2 \log \Lambda = -2 \log \frac{\text{lik}(\hat{\beta}_{0,0}, \ldots, \hat{\beta}_{0,p-1}, 0)}{\text{lik}(\hat{\beta}_0, \ldots, \hat{\beta}_p)}.$$

When the number of impressions n is large, we may perform an approximate level-α test of H_0 by rejecting H_0 when $D > \chi^2_1(\alpha)$, since the difference between model dimensionalities here is 1.

We may obtain a confidence interval for β_j from the MLE estimate $\hat{\beta}_j$ and an estimate of its standard error: We compute the Fisher information $I_{\mathbf{Y}}(\beta) = -\mathbb{E}_\beta[\nabla^2 l(\beta)]$ by calculating the second partial derivatives of l: For $f(x) = \log(1 + e^x)$, $f''(x) = \frac{e^x}{(1+e^x)^2}$. Then

$$\frac{\partial^2 l}{\partial \beta_m \partial \beta_l} = -\sum_{i=1}^{n} x_{im} x_{il} \frac{e^{\sum_{j=0}^p \beta_j x_{ij}}}{(1 + e^{\sum_{j=0}^p \beta_j x_{ij}})^2} = -X_m^T W X_l,$$

where we have set $X_j = (x_{1j}, \ldots, x_{nj})$ as the jth column of the matrix of covariates as in Lecture 25, and defined the $n \times n$ diagonal matrix

$$W := W(\beta) = \text{diag}\left(\frac{e^{\sum_{j=0}^p \beta_j x_{1j}}}{(1 + e^{\sum_{j=0}^p \beta_j x_{1j}})^2}, \ldots, \frac{e^{\sum_{j=0}^p \beta_j x_{nj}}}{(1 + e^{\sum_{j=0}^p \beta_j x_{nj}})^2} \right).$$

So $\nabla^2 l(\beta) = -X^T W X$, $I_{\mathbf{Y}}(\beta) = X^T W X$, and the approximate sampling distribution of $\hat{\beta}$ for large n is $\mathcal{N}(\beta, (X^T W X)^{-1})$. Letting $\hat{W} = W(\hat{\beta})$ be the plugin estimate of the diagonal matrix W, we may estimate the standard error of $\hat{\beta}_j$ by $\hat{\text{se}}_j = \sqrt{((X^T \hat{W} X)^{-1})_{jj}}$, and obtain a 95% confidence interval for β_j as $\hat{\beta}_j \pm z(0.025)\hat{\text{se}}_j$.

Remark 26.2. A word of caution regarding model misspecification: The above standard error estimates $\hat{\text{se}}_j$ (which are the standard errors reported by most logistic regression software) are only expected to be accurate when the logistic regression model is correctly specified—that is, when the Y_i's are truly independent random variables with distribution Bernoulli(p_i), where the log-odds for each p_i is the same linear combination of the covariates. This is because, as in the case of n IID observations, the covariance of $\hat{\beta}$ is given by the inverse Fisher information only in a correctly specified model.

Logistic regression is still oftentimes used as a tool for binary classification problems even if the model does not yield an extremely accurate fit to the data, as long as the model has

good classification accuracy. In such settings, the MLE $\hat{\beta}$ represents the "closest" logistic regression model (in the given covariates) to the true distribution of $Y_1, \ldots, Y_n$, in the sense of KL-divergence as in Lecture 16. The standard error for $\hat{\beta}_j$ may be robustly estimated using either a sandwich estimator or the non-parametric bootstrap. For the logistic regression model, the sandwich estimate of the covariance matrix of $\hat{\beta}$ is given by[1]

$$(X^T \hat{W} X)^{-1} (X^T \tilde{W} X)(X^T \hat{W} X)^{-1},$$

where $\tilde{W} = \text{diag}((Y_1 - \hat{p}_1)^2, \ldots, (Y_n - \hat{p}_n)^2)$ and $\hat{p}_i$ is the fit probability for the ith observation, defined by the right side of equation (26.1) with $\hat{\beta}$ in place of β. The (j, j) element of this matrix gives a sandwich estimate for the variance of $\hat{\beta}_j$. Alternatively, one may use the **pairs bootstrap**, which pairs the covariates and response for each ith observation into a single data vector $(x_{i1}, \ldots, x_{ip}, Y_i)$, and then draws bootstrap samples by randomly selecting, with replacement, n of these vectors. The logistic regression model is fit to each such bootstrap sample to yield an MLE $\hat{\beta}^*$, and the standard error of $\hat{\beta}_j$ is estimated by the empirical standard deviation of $\hat{\beta}_j^*$ across bootstrap samples.

[1] See Liang and Zeger, "Longitudinal data analysis using generalized linear models" or Agresti, "Categorical Data Analysis" Section 12.3.3.