Co-financed by the Connecting Europe
Facility of the European Union

xAIM

Logistic Regression Models

Trustworthy AI - Lecturer: Emanuela Raffinetti; Python instructor: Alex Gramegna
E-mail: emanuela.raffinetti@unipv.it; alex.gramegna01@universitadipavia.it

1. The Logistic Model
   - Bernoulli Distribution
   - The linear probabilistic model
   - The Logit Model

Co-financed by the Connecting Europe Facility of the European Union

Bernouilli distribution- I

- We consider a case in which the response variable, or the variable that we want to predict, is binary.
- Hence the variable y_i can have only two values (typically 0 and 1).
- We view y_i as a realization of a random variable Y_i that can take the values 1 and 0 with probabilities π_i and $1 - \pi_i$, respectively. The distribution of Y_i is called a Bernoulli distribution with parameter π_i, and can be written in compact form as

$$Pr\{Y_i = y_i\} = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

for $y_i = 0, 1$. Note that if $y_i = 1$ we obtain π_i, and if $y_i = 0$ we obtain 1 - π_i.

Bernouilli distribution- II

- The expected value and variance of Y_i are:

$$E(Y_i) = \mu_i = \pi_i$$
$$Var(Y_i) = \sigma_i^2 = \pi_i(1 - \pi_i)$$

- It is important to mention that the mean and variance depend on the underlying probability π_i.
- Any factor that affects π_i will affect not only the mean but also the variance of the observations.

The linear probabilistic model

- We are defining a model in which the probability of a particular event is dependent on a set of covariates X_i.
- The simplest way to achieve this would be to let π_i be a linear function of the covariates.
- More formally,

$$\pi_i = x_i'\beta$$

- where β is a vector of regression coefficients.

Issues with LPM

- The main issue with this linear probabilistic model is that the probability π_i on the left-hand side has to be a value between 0 and 1, but the linear predictor $x_i'\beta$ on the right-hand-side can take any real value.
- Hence, there is no guarantee that the predicted values will be in the correct range unless complex restrictions are imposed on the coefficients.
- Solution: we can transform the probability to remove the range restrictions, and model the transformation as a linear function of the covariates.

The logit transformation

- Two steps of the logistic transformation:
- First, we move from the probability π_i to the odds:

$$odds_i = \frac{\pi_i}{1 - \pi_i}$$

  - If the probability of an event is 1/2, the odds are one-to-one or even.
  - If the probability is 1/3, the odds are one-to-two.
- Second, we take logarithms, calculating the logit or log-odds

$$logit(\pi_i) = \log \frac{\pi_i}{1 - \pi_i}$$

Graphical representation of the logit transformation

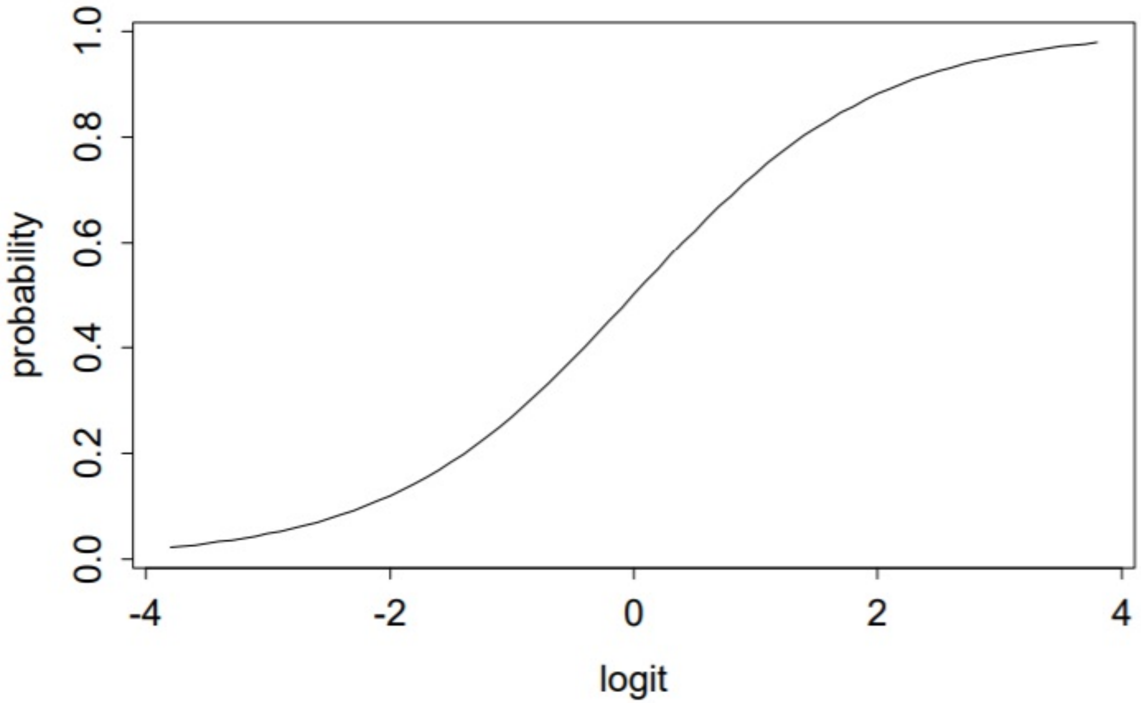

Figure: The Logit Transformation

Specification - I

- Suppose that we have k independent observations $y_1, \ldots, y_k$, and that the i-th observation can be treated as a realization of a random variable Y_i.
- We assume that Y_i has a binomial distribution.

$$Y_i \sim Bin(n_i, \pi_i)$$

  with binomial denominator n_i and probability π_i.
- Bernoulli is a special case in which $n_i = 1$.

Co-financed by the Connecting Europe Facility of the European Union

Specification - II

- Suppose further that the logit of the underlying probability π_i is a linear function of the predictors.

$$logit(\pi_i) = x_i' \beta$$

  where x_i is a vector of covariates and β is a vector of regression coefficients.
- The model above is a generalized linear model with binomial response and logit link.

Interpretation

- The regression coefficients β_j shows the change in the logit of the probability associated with a unit change in the j-th predictor holding all other predictors constant.
- Exponentiating the equation from the previous slide, we find that the odds for the i-th unit are given by
$$\frac{\pi_i}{1-\pi_i} = \exp(x_i'\beta)$$
- When x_i is a binary variable, the regression coefficient is the odds ratio: $\exp(\beta) = \dfrac{ODDS(x=1)}{ODDS(x=0)}$
- When x_i is a continuous variable, the regression coefficient is the change in logit(π_i) for a unit increase in the predictor.

Estimation

The unknown parameters of a logistic regression can be estimated by the maximum likelihood method, using an iterative numerical optimisation algorithm.
This allows to check the hypothesis that a single parameter is equal to zero using a statistical test:

$$Z = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}$$

which converges to a standard normal distribution.

Co-financed by the Connecting Europe Facility of the European Union

Deviance test for logistic regression - I

- The residual variance of a model, which is based on the distance of each point from its estimate, is replaced by the residual deviance (G^2), which is based on the distance of the probability of each point from its estimate.

Source	Formula	DF
$G^2(M)$	$2\sum_{i=1}^{C} n_i \log(n_i/\hat{n}_i)$	$C-p-1$

- The deviance expresses the distance between the observed (n_i) and the fitted ($\hat{n}_i$) frequencies in the i-th cell of the contingency table. It can be shown that

$$G^2 \sim \chi^2(C-p-1)$$

so a test can be applied to evaluate the goodness of fit of a model.

Co-financed by the Connecting Europe Facility of the European Union

Deviance test for logistic regression - II

Alternatively, the goodness of fit can be estimated with the χ^2 statistics:

$$\chi^2 = \sum_{i=1}^{C} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$$

which is also asymptotically chi-squared.

Co-financed by the Connecting Europe Facility of the European Union

Deviance test for nested model comparison

- When we compare two models, for example a simpler model M_1 with a more complex model M_2, we can use a test based on the difference between the corresponding deviances:

$$G^2(M_1) - G^2(M_2) \sim \chi^2(k)$$

  where k is the difference in number of parameters (complexity) between model M_2 and M_1.
- When the number of additional parameters $k = 1$ the deviance difference is distributed as a $\chi^2(1)$.
- We will come back to this when studying model selection.

Co-financed by the Connecting Europe Facility of the European Union

Reference

- Kutner M.H., Nachstsheim C.J., Neter J.: Applied Linear Regression Models, 4° edition, McGraw Hill Irwin (2004), available at: https://www.academia.edu/32804953/Applied_Linear_Regression_Models_4th_edition