STAT 480 Statistical Computing Applications

Unit 4. Classification

Lecture 1. Logistic Regression

Department of Statistics, Iowa State University
Spring 2019

Classification

- Classification is a predictive task in which the response takes values across discrete categories (i.e., not continuous), and in the most fundamental case, two categories.
- Examples:
  - Predicting whether a patient will develop breast cancer or remain healthy, given genetic information.
  - Predicting whether or nor a user will like a new product, based on user covariates and a history of his/her previous rating.
  - Predicting the region of Italy in which a brand of olive oil was made, based on its chemical composition.
  - Predicting the next elected president, based on various social, political, and historical measurements.

Classification (Cont.)

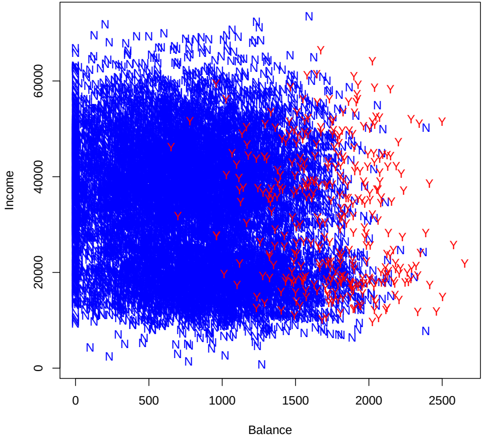
- Similar to our usual setup, we observe pairs (x_i, y_i), $i = 1, \ldots, n$, where y_i gives the class of the ith observation, and x_i is the predictor.
- Though the class labels may actually be $y_i \in \{healthy, sick\}$ or $y_i \in \{Sardinia, Sicily, \ldots\}$, but we can always encode them as

$$y_i \in \{1, 2, \ldots, K\}$$

  where K is the total number of classes.
- Note that there is a big difference between *classification* and *clustering*; in the latter, there is not a pre-defined notion of class membership (and sometimes, not even K), and we are not given labeled examples, but only x_i, $i = 1, \ldots, n$.

Default Data Example

- In this example, 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*.



Binary Classification and Linear Regression

- Supposing that $K = 2$, so that the response is $y_i \in \{1, 2\}$, for $i = 1, \ldots, n$.
- We can use the dummy variable to code the response

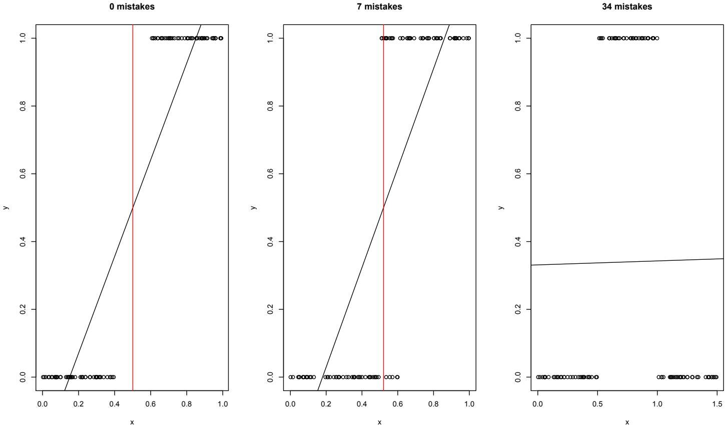
$$Y = \begin{cases} 0, & \text{if default=No;} \\ 1, & \text{if default=Yes.} \end{cases}$$

- You already know a tool that could potentially use in this case for classification: *linear regression*.
- Simply treat the response as if it were continuous, find the linear regression coefficients of y onto the predictors.
- Given a new input x_0, we predict the class to be

$$\hat{f}^{LS}(x_0) = \begin{cases} 0 & \text{if } \hat{\beta}_0 + x_0\hat{\beta}_1 \le 0.5 \\ 1 & \text{if } \hat{\beta}_0 + x_0\hat{\beta}_1 > 0.5 \end{cases}$$

Binary Classification and Linear Regression (Cont.)

- In many instances, this actually works reasonably well. For example,

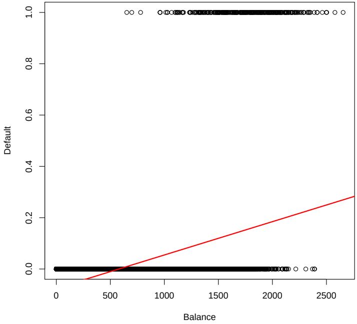


Problem with Linear Regression

We could simplify the plot by drawing a line between the means for the two dependent variable levels, but this is problematic:

(a) the line seems to oversimplify the relationship and

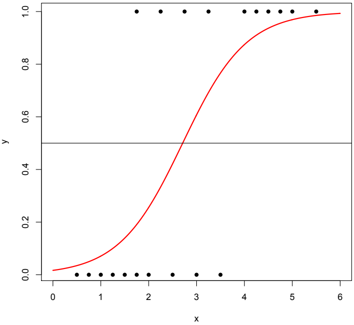
(b) it gives predictions that cannot be observable values of Y for extreme values of X.



- The reason this does't work is because the approach is analogous to fitting a linear model to the probability of the event.
- Probabilities can only take values between 0 and 1, hence, we need a different approach to ensure our model is appropriate for the data.

Problem with Linear Regression (Cont.)

- The shape of this distribution is a cumulative probability distribution.
- We can model the nonlinear relationship between X and Y by transforming one of the variables.
- Two common transformations that result in sigmoid functions are *probit* and *logit* transformations.



Logistic Regression Model

- Rather than modeling the response Y directly, logistic regression models the *probability* that Y belongs to a particular category.

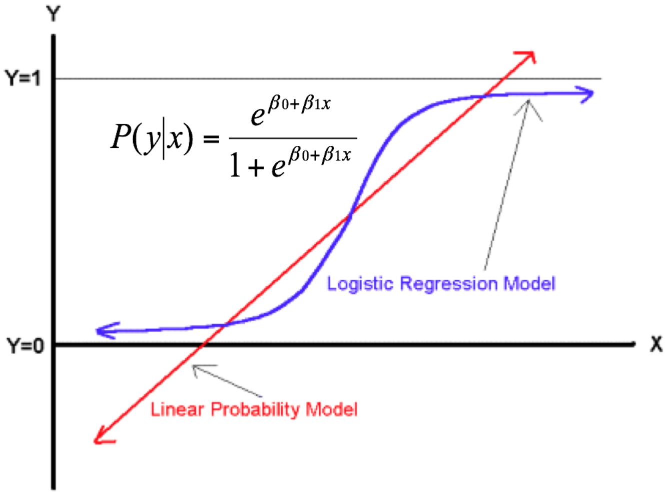
- The *logistic* model solves the following problems:

$$\log\left\{\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right\}=\beta_0+\beta_1 x,$$

  for some unknown β_0 and β_1, which we will estimate directly.

- $P(Y=0|X=x)=1-P(Y=1|X=x)$
  $\Rightarrow\ \log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 x$
  - p is the probability that event Y occurs (range=0 to 1).
  - $p/(1-p)$ is the *odds ratio* (range=0 to ∞).
  - $\log\{p/(1-p)\}$ is *log odds ratio* or *logit* (range=$-\infty$ to ∞).

Comparing Linear Probability & Logistic Models



Odds & Odds Ratios

- The definitions of an *odds*: $odds = \frac{p}{1-p}$.

  The odds has a range from 0 to ∞ with values greater than 1 associated with an event being more likely to occur than not occur and values less than 1 associated with an event that is less likely to occur than not occur.

- The *logit* is defined as the log of the odds:

$$\log(odds) = \log \frac{p}{1-p} = \log(p) - \log(1-p).$$

- This transformation is useful because it creates a variable with a range from $-\infty$ to ∞.

- The interpretation of logits is simple – take the exponential of the logit and you have the odds for the two groups in question.

Interpretation

- The logit distribution constrains the estimated probabilities to lie between 0 and 1.
- the estimated probability is

$$p = P(Y = 1|X = x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

- If $\beta_0 + \beta_1 x = 0$ then $p = 0.5$.
- As $\beta_0 + \beta_1 x$ gets really big, p approaches 1.
- As $\beta_0 + \beta_1 x$ gets really small, p approaches 0.

Estimating Logistic Regression Coefficients

- Suppose that we are given a sample (x_i, y_i), $i = 1, \ldots, n$. Here y_i denotes the class $\in \{0, 1\}$ of the ith observation.
- Assume that the classes are conditionally independent given $x_1, \ldots, x_n$, then

$$\mathbb{L}(\beta_0, \beta_1) = \prod_{i=1}^{n} P(Y = y_i | X = x_i)$$

  the likelihood of these n observations, so the log likelihood is

$$l(\beta_0, \beta_1) = \sum_{i=1}^{n} \log P(Y = y_i | X = x_i).$$

- For convenience, we define the indicator $u_i = \begin{cases} 1 & \text{if } y_i = 1 \\ 0 & \text{if } y_i = 0 \end{cases}$

Estimating Logistic Regression Coefficients

- The log-likelihood can be written as

$$
\begin{aligned}
l(\beta_0, \beta_1) &= \sum_{i=1}^{n} \log P(Y = y_i | x = x_i) \\
&= \sum_{i=1}^{n} [u_i(\beta_0 + \beta_1 x_i) - \log \{1 + \exp(\beta_0 + \beta_1 x_i)\}]
\end{aligned}
$$

- The coefficients are estimated by maximizing the likelihood,

$$
\sum_{i=1}^{n} [u_i(\beta_0 + \beta_1 x_i) - \log \{1 + \exp(\beta_0 + \beta_1 x_i)\}]
$$

Estimation for Default Data

- Default: Customer default records for a credit card company; available in the ISLR library.
- For Default data, estimated coefficients of the logistic regression model that predicts the probability of `default` using `balance`.

	Coefficient	Std.Error	Z statistic	p-value
Intercept	-10.6513	0.3612	-29.5	$< 2 \times 10^{-16}$
balance	0.0055	0.0002	24.9	$< 2 \times 10^{-16}$

- One unit increase in `balance` is associated with an increase in the log odds of `default` by 0.0055 units.

Making Predictions

- For example, we predict that the `default` probability for an individual with a `balance` of \$1,000

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}} = \frac{e^{-10.6513 + 0.0055 * 1000}}{1 + e^{-10.6513 + 0.0055 * 1000}} = 0.00576$$

  which is below 1%.

- The predicted probability of `default` for an individual with a `balance` of \$2,000 is much higher, and equals 0.586.

Multiple Logistic Regression

- Model:

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

- Estimated coefficients for Default Data

	Coefficient	Std.Error	Z statistic	p-value
Intercept	-11.54047	0.434756	-26.5	$< 2 \times 10^{-16}$
balance	0.00565	0.000227	24.8	$< 2 \times 10^{-16}$
income	0.00002	0.000005	4.2	3×10^{-5}

- For an individual with a credit card `balance` of \$1,500 and an `income` of \$4,000 has an estimated probability of `default` of

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2}} = \frac{e^{-11.54 + 0.0056 * 1500 + 0.00002 * 4000}}{1 + e^{-11.54 + 0.0056 * 1500 + 0.00002 * 4000}} = 0.048$$

Classification Tables

- Choose a cutoff value on the probability scale, say 50%, and classify all predicted values above that as predicting an event, and all below that cutoff value as not predicting the event.
- Construct a 2-by-2 table of data, since we have dichotomous observed outcomes, and have now created dichotomous "fitted values", when we used the cutoff

	Observed positive	Observed negative
Predicted positive (above cutoff)	a	b
Predicted negative (below cutoff)	c	d

- Consider: sensitivity $= a/(a+c)$, specificity $= d/(b+d)$. Higher values indicate a better fit of the model.