Notes on Classification

Yangxinyu Xie

Spring 2019

Last updated February 5, 2019

1. Logistic Regression

❖

1.1 Binary Responses ❖

Consider a simple classification example: we are interested in predicting whether an individual will default on his or her credit card payment, on the basis of monthly credit card balance. The data set is shown in **Figure 1.1**.

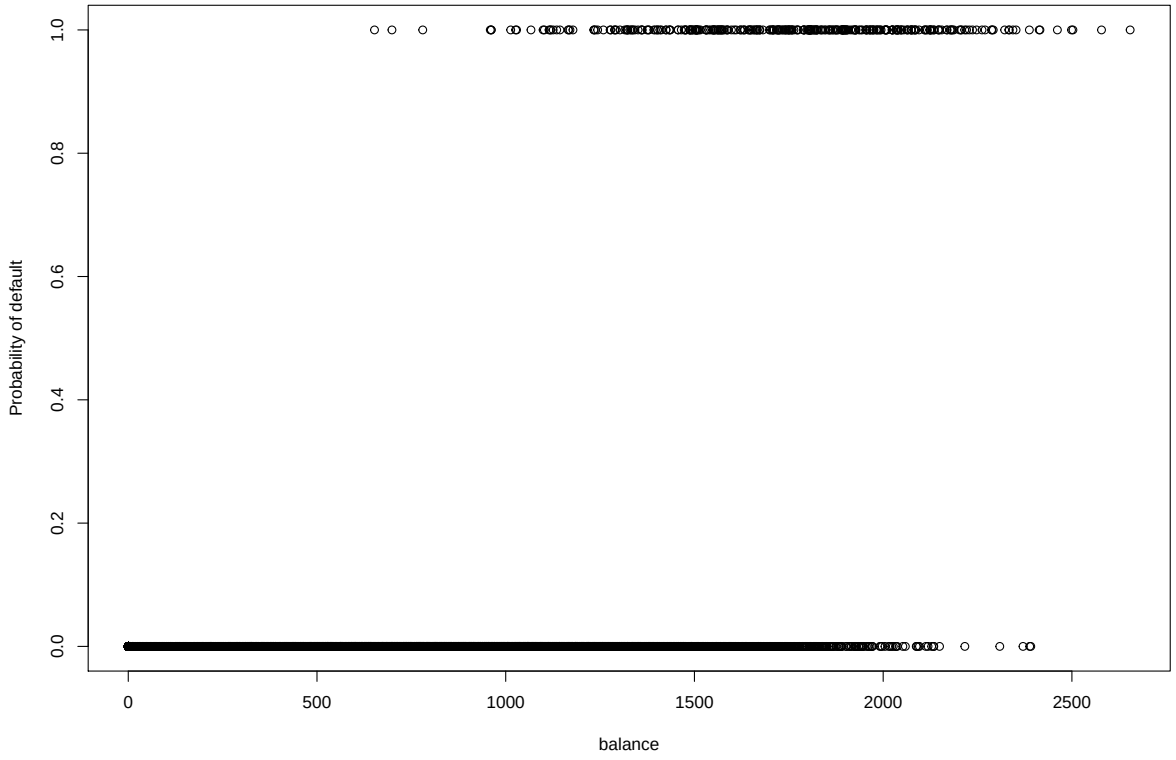

Figure 1.1: Default Dataset in R Package ISLR

For binary responses, we can consider a model

$$Y_i = f(X_i) + \epsilon_i \quad \text{where} \quad Y_i = 0, 1 \tag{1.1.1}$$

Specifically, we can view Y_i as a Bernoulli random variable such that a person tends to default with probability π_i, as shown in **Table 1.1**. From this model, we can obtain $\mathbb{E}(Y_i) = 1 \times \mathbb{P}(Y_i = 1) + 0 \times \mathbb{P}(Y_i = 0) = \pi_i$. That is, in the fitted model, $f(X_i)$ estimates the probability of $Y_i = 1$.

Table 1: Y_i as a Bernoulli Random Variable

$\mathbb{P}(Y_i = 1)$	π_i
$\mathbb{P}(Y_i = 0)$	$1 - \pi_i$

1.2 Problems with Linear Regression ❖

Back to the default example, to predict the probability of whether an individual will default given his or her monthly credit card balance, we can try a linear model where $f(X_i) = \beta_0 + \beta_1 X_i$.

$$\hat{y}_i = -0.0751920 + 0.0001299 x_i \tag{1.2.1}$$

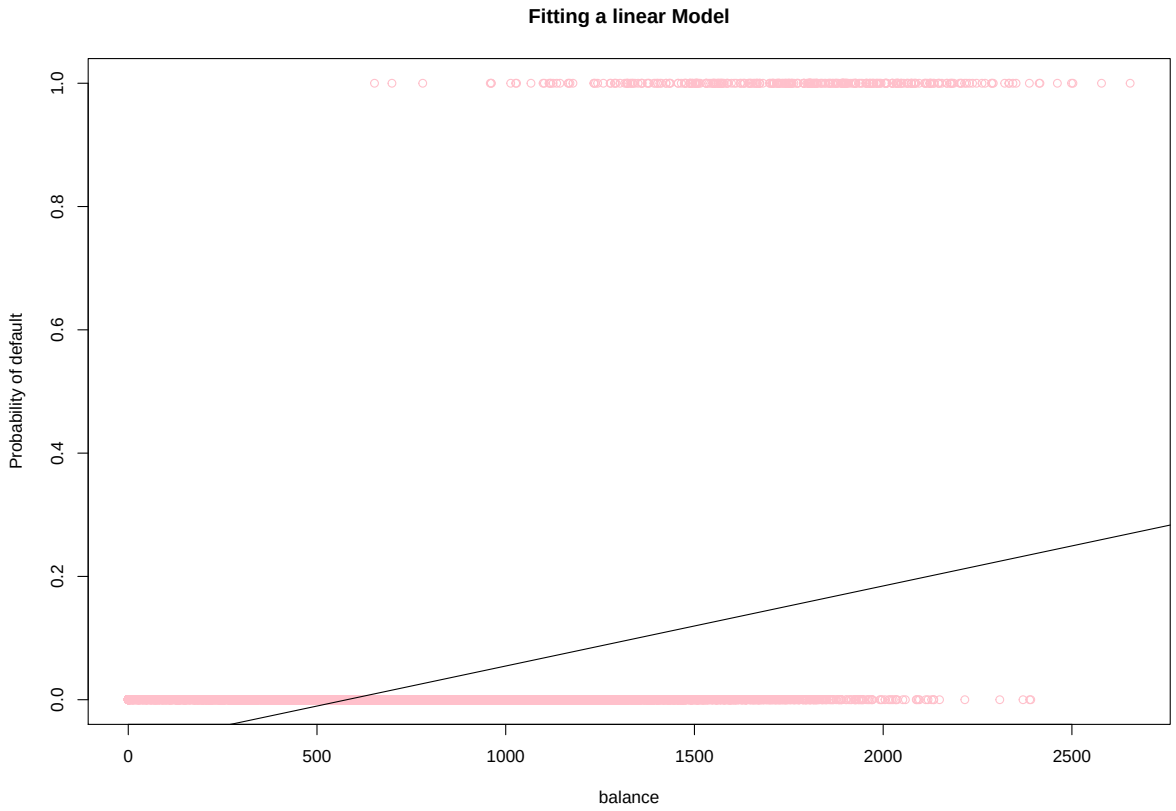

Figure 1.2: Linear Model

As we observe from **Figure 1.2**, the linear model have done quite poorly. In fact, there are three major problems with linear regression model in this context:

1. **Constraints on Regression Function** Based on our model in (1.1.1), we obtained that $0 \leq \mathbb{E}(Y_i) = \pi_i \leq 1$. As we have noticed, the linear model can easily fall outside these limits.

2. **Normal Error Terms** Simple linear model assume that the error term stems from a normal distribution; however, in model (1.2.1), we have

$$\begin{cases} Y_i = 1 : \epsilon_i = 1 - f(X_i) = 1 - (\beta_0 + \beta_1 X_i) \\ Y_i = 0 : \epsilon_i = -f(X_i) = -(\beta_0 + \beta_1 X_i) \end{cases} \tag{1.2.2}$$

   which breaks the assumption.

3. **Non-constant Variance** From model (1.1.1), we assume that $f(X_i)$ is fixed. In other words, the variability of Y_i comes from the ϵ_i. Hence,

$$\sigma^2(\epsilon_i) = \sigma^2(Y_i) = \pi_i(1 - \pi_i) = (\mathbb{E}(Y_i))(1 - \mathbb{E}(Y_i)) = (\beta_0 + \beta_1 X_i)[1 - (\beta_0 + \beta_1 X_i)]. \tag{1.2.3}$$

Since $\sigma^2(\epsilon_i)$ depends on X_i, it breaks the linear model assumption of constant variance.

1.3 Simple Logistic Regression Model ❖

We introduce a commonly used generalized linear model called logistic regression model. Let Y_i be a Bernoulli random variable with $\mathbb{E}(Y_i) = \pi_i$, where

$$\pi_i = \frac{exp(\beta_0 + \beta_1 X_i)}{1 + exp(\beta_0 + \beta_1 X_i)} \tag{1.3.1}$$

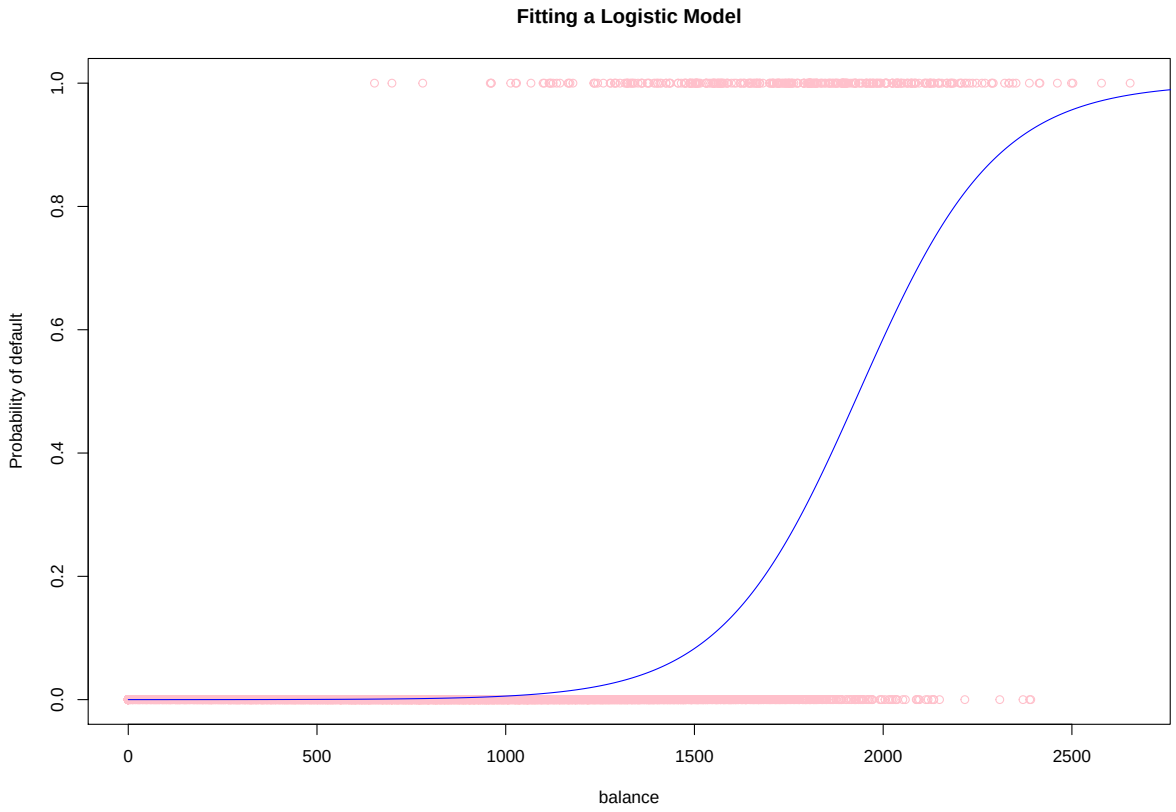

Figure 1.3: Simple Logistic Model

Exercise 1.1. Rewrite the simple logistic model 1.3.1 as $\log(\pi_i/(1-\pi_i)) = \beta_0 + \beta_1 X_i$

Solution. Let $c = \beta_0 + \beta_1 X_i$, we get

$$\begin{aligned} \pi_i &= \frac{e^c}{1+e^c} \\ \pi_i + \pi_i e^c &= e^c \\ \pi_i &= e^c(1-\pi_i) \\ \frac{\pi_i}{(1-\pi_i)} &= e^c \end{aligned}$$

Hence, we obtain

$$\beta_0 + \beta_1 X_i = \log_e e^c = \log_e \left(\frac{\pi_i}{(1-\pi_i)} \right) \tag{1.3.2}$$

□

1.4 Likelihood Function ❖

Typically, we choose maximum likelihood method to estimate parameters. For example, in simple linear regression, least square estimates of β_0 and β_1 are maximum likelihood estimates. The likelihood function of simple logistic model 1.3.1 is

$$g(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \tag{1.4.1}$$

By taking the logarithm,

$$\begin{aligned} \log_e[g(Y_1, \ldots, Y_n)] &= \sum_{i=1}^{n} Y_i \log_e(\pi_i) + (1-Y_i)\log_e(1-\pi_i) \\ &= \sum_{i=1}^{n} Y_i \log_e(\frac{\pi_i}{1-\pi_i}) + \sum_{i=1}^{n} \log_e(1-\pi_i) \end{aligned} \tag{1.4.2}$$

From 1.3.1 and 1.3.2, we obtain,

$$\begin{aligned} \log_e[g(Y_1, \ldots, Y_n)] &= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) + \sum_{i=1}^{n} \log_e\{[1 + e^(\beta_0 + \beta_1 X_i)]^{-1}\} \\ &= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \log_e[1 + exp(\beta_0 + \beta_1 X_i)] \end{aligned} \tag{1.4.3}$$

In fact, no simple formula for estimating β_0 and β_1 is available. Usually, statistical packages will utilize numerical search procedure to obtain maximum likelihood estimates for β_0 and β_1. By running `glm` package in R, we obtain a model

$$\hat{\pi}_i' = \log_e(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) = -10.651331 + 0.005499 x_i \tag{1.4.4}$$

and its result shown in **Figure 1.3**

1.5 Multiple Logistic Regression & Polynomial Logistic Regression ❖

The extension from simple logistic regression to multiple logistic regression and polynomial logistic regression is intuitive. In fact, we can express general logistic regression model as

$$\mathbb{E}[Y_i] = \pi_i = \frac{exp(\boldsymbol{X_i^T \beta})}{1 + exp(\boldsymbol{X_i^T \beta})} \tag{1.5.1}$$

Remark 1.2. Convergence difficulties in the numerical search procedure for finding the maximum likelihood estimates of the multiple logistic regression function may be encountered when the predictor variable are highly correlated or when there is a large number of predictor variables.

2. Linear Discriminant Analysis

❖

2.1 Bayes' Theorem ❖

The classical Bayes' Theorem states that

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(X,Y)}{\mathbb{P}(X)} = \frac{\mathbb{P}(X|Y)\mathbb{P}(Y)}{\mathbb{P}(X)} \tag{2.1.1}$$

We have seen this in logistic regression, where given X_i, we compute the probability of $Y_i = 1$ and classify Y_i into group 1 if $\mathbb{P}(Y|X) > 0.5$. Following the same idea, we consider a scenario where there are more than 2 classes.

Proposition 2.1. *Suppose there are K classes in total. Let π_k be the **prior probability** that a randomly chosen observation comes from the kth class and let the **density function** $f_k(x) \equiv \mathbb{P}(X = x|Y = k)$, then the Bayes' Theorem states that*

$$p_k(x) = \mathbb{P}(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \tag{2.1.2}$$

*where $p_k(x)$ is the posterior probability that an observation X_i belongs to the kth class, **given** the predictor value for that observation.*

2.2 Linear Discriminant Analysis ❖

The assumption of Discriminant Analysis is that in 2.1.2, $X(k) \sim \mathcal{N}(\mu_k, \Sigma_k)$, or equivalently,

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$$

Typically, the **Linear Discriminant Analysis** assumes that for all $k, \Sigma_k = \Sigma$ where Σ is the pooled covariance matrix. When deciding the boundary between any two classes, we consider the log ratio:

$$\begin{aligned}
\log \frac{\mathbb{P}(Y = k|X = x)}{\mathbb{P}(Y = l|X = x)} =& \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l} \\
=& \log \frac{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)]}{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(x-\mu_l)^T\Sigma^{-1}(x-\mu_l)]} + \log \frac{\pi_k}{\pi_l} \\
=& -\frac{1}{2}[(x-\mu_k)^T\Sigma^{-1}(x-\mu_k) - (x-\mu_l)^T\Sigma^{-1}(x-\mu_l)] + \log \frac{\pi_k}{\pi_l} \\
=& -\frac{1}{2}[x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_k - \mu_k^T\Sigma^{-1}x + \mu_k^T\Sigma^{-1}\mu_k \\
& - x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_l + \mu_l^T\Sigma^{-1}x - \mu_l^T\Sigma^{-1}\mu_l] + \log \frac{\pi_k}{\pi_l} \\
=& -\frac{1}{2}[-x^T\Sigma^{-1}(\mu_k - \mu_l) - (\mu_k - \mu_l)^T\Sigma^{-1}x + \mu_k^T\Sigma^{-1}\mu_k - \mu_l^T\Sigma^{-1}\mu_l] + \log \frac{\pi_k}{\pi_l} \\
=& x^T\Sigma^{-1}(\mu_k - \mu_l) - \frac{1}{2}(\mu_k + \mu_l)^T\Sigma^{-1}(\mu_k - \mu_l) + \log \frac{\pi_k}{\pi_l}
\end{aligned} \tag{2.2.1}$$

When the log ratio is greater than zero, we classify X as k and l otherwise.

Exercise 2.2. Show that classifying observations based on the boundary set by log ratio in 2.2.1 is equivalent to maximizing

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log \pi_k \tag{2.2.2}$$

Solution. Set $\log \frac{\mathbb{P}(Y=k|X=x)}{\mathbb{P}(Y=l|X=x)} = 0$. We obtain,

$$x^T\Sigma^{-1}(\mu_k - \mu_l) - \frac{1}{2}(\mu_k + \mu_l)^T\Sigma^{-1}(\mu_k - \mu_l) + \log \frac{\pi_k}{\pi_l} = 0$$

which is

$$-\frac{1}{2}[-x^T\Sigma^{-1}\mu_k - \mu_k^T\Sigma^{-1}x + \mu_k^T\Sigma^{-1}\mu_k + x^T\Sigma^{-1}\mu_l + \mu_l^T\Sigma^{-1}x - \mu_l^T\Sigma^{-1}\mu_l] + \log \frac{\pi_k}{\pi_l} = 0$$

Rearranging the parameters,

$$-\frac{1}{2}[-x^T\Sigma^{-1}\mu_k - \mu_k^T\Sigma^{-1}x + \mu_k^T\Sigma^{-1}\mu_k] + \log\pi_k = \frac{1}{2}[x^T\Sigma^{-1}\mu_l + \mu_l^T\Sigma^{-1}x - \mu_l^T\Sigma^{-1}\mu_l] + \log\pi_l$$

$$x^T\Sigma^{-1}\mu_k - -\frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k = x^T\Sigma^{-1}\mu_l - \frac{1}{2}\mu_l^T\Sigma^{-1}\mu_l + \log\pi_l$$

where the left hand side is $\delta_k(x)$ and the right hand side is $\delta_l(x)$. Hence, we choose the class k' that maximizes $\delta_{k'}(x)$. □

As the decision boundary $\delta_k(x)$ is only a linear combination of x, we see that in p dimension, a decision boundary will simply be a hyperplane.

2.3 Parameter Estimation ❖

The estimates of parameters in $\delta_k(x)$ 2.2.2 are

$$\begin{aligned}\hat{\pi}_k &= N_k/N \\ \hat{\mu}_k &= \sum_{g_i=k} x_i/N_k \\ \hat{\Sigma} &= \sum_{k=1}^{K}\sum_{g_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T/(N-k)\end{aligned} \tag{2.3.1}$$

Since x is a $p \times 1$ vector where p is the number of predictor variables, we see that there are $(K-1)\times(p+1)$ parameters to estimate for LDA.

2.4 Quadratic Discriminant Analysis ❖

Quadratic Discriminant Analysis does not follow the assumption that $\Sigma_k = \Sigma \forall k$, instead, it assumes that each category should have its own Σ_k, hence a greater level of flexibility than LDA. Following similar algebraic moves, we will get

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k \tag{2.4.1}$$

There will be $(K-1)\times[p(p+3)/2+1]$ parameters to estimate, meaning a greater scale of variance of the model.

References

[1] . M. James et al, An Introduction to Statistical Learning: With Applications in R. New York: Springer, 2013

[2] . H. Kutner, Applied Linear Statistical Models. (5th ed.) Boston: McGraw-Hill Irwin, 2005

[3] . Hastie, R. Tibshirani and Friedman, J. H. (Jerome H.), The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer, 2001.