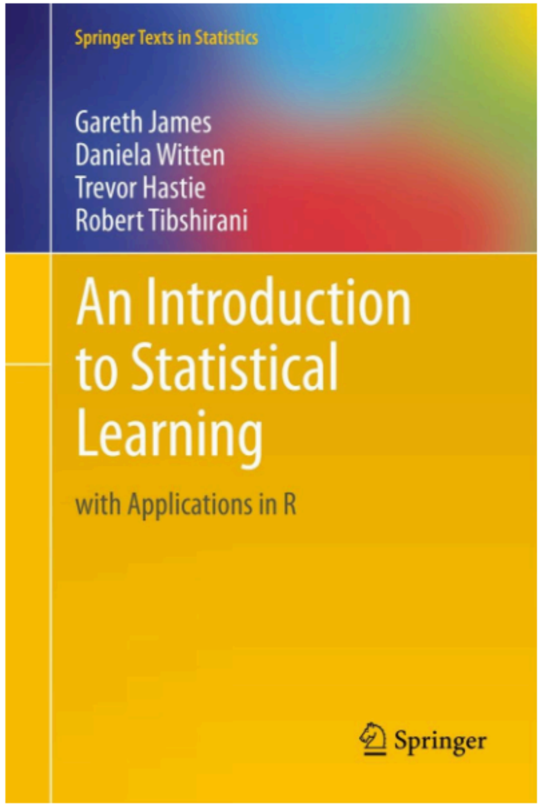
4. Logistic Regression

ESC Spring 2018 – Data Mining and Analysis

SeoHyeong Jeong



Textbook:

An Introduction to Statistical Learning

Lecture Slides:

Stanford Stats 202: Data Mining and Analysis

Spring 17’ ESC Statistical Data Analysis

REFERENCE

Reading:

An Introduction to Statistical Learning

chapter 4.1 An Overview of Classification
chapter 4.2 Why Not Linear Regression
chapter 4.3 Logistic Regression

Table of Contents

Why Not Linear Regression?

- If we have a good estimate for the conditional probability $\hat{p}(Y|X)$, we can use the classifier:

$$\hat{y}_0 = \text{argmax}_y \ \hat{P}(Y = y \mid X = x_0).$$

- Suppose Y is a binary variable. Could we use a linear model?

$$P(Y = 1|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Problems:
  - This would allow probabilities <0 and >1.
  - Difficult to extend to more than 2 categories.

Linear Regression vs. Logistic Regression

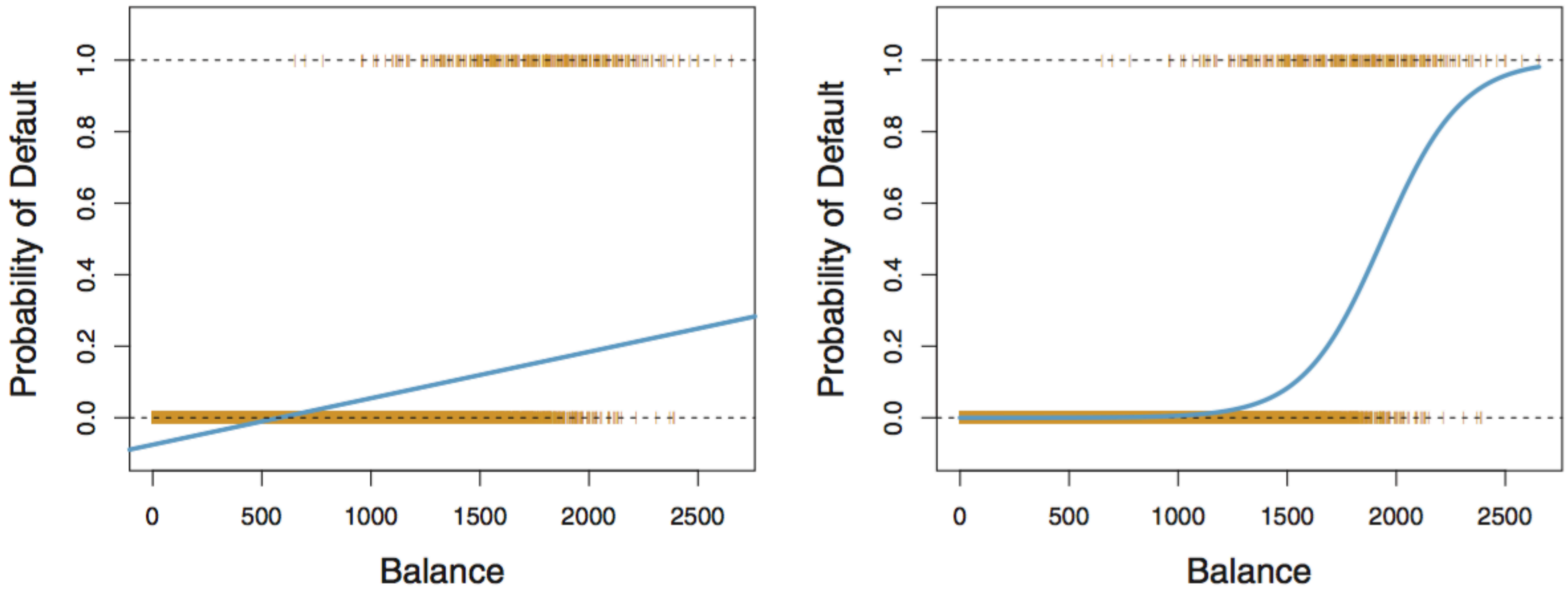


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default` *(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Logistic Regression

- We model the joint probability as:

$$P(Y = 1 \mid X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}},$$

$$P(Y = 0 \mid X) = \frac{1}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

- This is the same as using a linear model for the log odds:

$$\log\left[\frac{P(Y = 1 \mid X)}{P(Y = 0 \mid X)}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

Fitting Logistic Regression

- The training data is a list of pairs $(x_1, y_1), \ldots, (x_n, y_n)$. In the linear model,

$$\log\left[\frac{P(Y=1 \mid X)}{P(Y=0 \mid X)}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

  we don't observe the left hand side.

- We cannot use a least squares fit as we did for the linear regression.

- **Solution**: **MLE (Maximum Likelihood Estimation)**
  The likelihood is the probability of the training data, for a fixed set of coefficients $\beta_0, \ldots, \beta_p$:

$$\prod_{i=1}^{n} P(Y = y_i \mid X = x_i)$$

$$= \underbrace{\prod_{i;y_i=1} \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}}_{\text{Probability of responses } = 1} \underbrace{\prod_{j;y_j=0} \frac{1}{1+e^{\beta_0+\beta_1 x_{j1}+\cdots+\beta_p x_{jp}}}}_{\text{Probability of responses } = 0}$$

- Choose estimates $\hat{\beta}_0, \ldots, \hat{\beta}_p$ which maximize the likelihood.
- Solved with numerical methods (e.g. Newton's algorithm).

Interpreting β_1

- Let $p(X) = Pr(Y = 1|X)$
- Assume logistic regression of the form

$$log \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

- For binary response Y, let $Y = 1$ imply success and $Y = 0$ imply failure.
- The *odds* of success are defined to be

$$\text{odds} = \frac{Pr(Y = 1)}{1 - Pr(Y = 1)}$$

- In logistic regression:

$$Pr(Y = 1 \mid X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- Now, the odds:

$$\text{odds} = \frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{\frac{1}{1+e^{\beta_0+\beta_1 X}}} = e^{\beta_0+\beta_1 X}$$

- For a continuous explanatory variable, the *odds ratio* is defined as:

$$\text{odds ratio} = \frac{\text{odds}(X+1)}{\text{odds}(X)} = \frac{e^{\beta_0+\beta_1 (X+1)}}{e^{\beta_0+\beta_1 X}} = e^{\beta_1}$$

- Thus e^{β_1} represents the change in the odds of the outcome of Y by increasing X by 1 unit.
  - If $\beta_1 = 0$, $e^{\beta_1} = 1$, the odds are the same at all X levels. This implies X does not have influence on the outcome of Y.
  - If $\beta_1 > 0$, $e^{\beta_1} > 1$, the odds increase as X increases
  - If $\beta_1 < 0$, $e^{\beta_1} < 1$, the odds decrease as X increases
- There are many cases that several explanatory variables are needed to classify Y.
- In this case, β coefficients can be interpreted the same way as logistic regression with a single variable.

Logistic Regression in R

```
> glm.fit=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,
    data=Smarket,family=binomial)
> summary(glm.fit)

Call:
glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5
    + Volume, family = binomial, data = Smarket)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
 -1.45    -1.20     1.07     1.15     1.33

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.12600    0.24074   -0.52     0.60
Lag1        -0.07307    0.05017   -1.46     0.15
Lag2        -0.04230    0.05009   -0.84     0.40
Lag3         0.01109    0.04994    0.22     0.82
Lag4         0.00936    0.04997    0.19     0.85
Lag5         0.01031    0.04951    0.21     0.83
Volume       0.13544    0.15836    0.86     0.39
```

- We can estimate the Standard Error of each coefficient.
- The $z - statistic$ is the equivalent of the $t - statistic$ in linear regression:

$$z = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)}.$$

- The $p - values$ are test of the null hypothesis $\beta_j = 0$ (Wald's test).
- Other possible hypothesis test: likelihood ratio test (chi-square distribution) is useful for testing whether groups of variables have coefficients equal to 0.

Wald test:

Wald test statistic $z^2 = \left(\frac{\hat{\beta}_1}{\widehat{se}(\hat{\beta}_1)}\right)^2 \sim X^2(1)$

Likelihood ratio test:

$2(\log \text{likelihood}(\hat{\beta}_0, \hat{\beta}_1) - \log \text{likelihood}(\hat{\beta}_0)) \sim X^2(1)$