Classification Methods: Logistic Regression

Rebecca C. Steorts, Duke University

STA 325, Chapter 4 ISL

Agenda

- What is classification?
- Linear regression
- Logistic regression

Classification

- Recall that linear regression assumes the response is quantitative
- In many cases, the response is qualitative (categorical).

Here, we study approaches for predicting qualitative responses, a process that is known as classification.

Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class.

Classification

We will cover three of the most widely-used classifiers:

1. Logistic Regression
2. Linear Discrimminant Analysis (LDA)
3. Quadratic Discrimminant Analysis (QDA)

More advanced methods in ISL cover methods such as generalized additive models (Chapter 7), trees, random forests, and boosting (Chapter 8), and support vector machines (Chapter 9).

Setup

We have set of training observations $(x_1, y_1), \ldots, (x_n, y_n)$ that we can use to build a classifier.

We want our classifier to perform well on both the training and the test data.

Why not linear regression?

Suppose that we are trying to predict the medical condition of a patient in the emergency room on the basis of her symptoms.

In this simplified example, there are three possible diagnoses: stroke, drug overdose, epileptic seizure.

Why not linear regression?

We could consider encoding these values as a quantitative response variable Y as follows:

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if drug overdose} \\ 3 & \text{if epileptic seizure} \end{cases}$$

This coding implies an **ordering** on the outcomes, putting drug overdose in between stroke and epileptic seizure, and insisting that the difference between stroke and drug overdose is the **same** as the difference between drug overdose and epileptic seizure.

Why not linear regression?

In practice there is no particular reason that this needs to be the case. For instance, one could choose an equally reasonable coding,

$$Y = \begin{cases} 1 & \text{if epileptic seizure} \\ 2 & \text{if stroke} \\ 3 & \text{if drug overdose} \end{cases}$$

which would imply a **totally different relationship** among the three conditions.

Each of these codings would produce fundamentally different linear models that would ultimately lead to different sets of predictions on test observations.

Remark: If the response variable's values did take on a natural ordering, then a 1,2,3 coding would be reasonable.

Why not linear regression?

For a binary (two level) qualitative response, the situation is better.

Consider only two possibilities for the patient's medical condition: stroke and drug overdose:

$$Y = \begin{cases} 0 & \text{if stroke} \\ 1 & \text{if drug overdose} \end{cases}$$

We could then fit a linear regression to this binary response, and predict drug overdose if $\hat{Y} > 0.5$ and stroke otherwise.

However, the dummy variable approach cannot be easily extended to accommodate qualitative responses with more than two levels.

Thus, this is why we consider classification methods.

Default data set

For the rest of the module, we will work with a data set where our goal is to predict the probability a customer defaults on their credit card debt.

We will use this example to

1. introduce logistic regression
2. illustrate the poor quality of linear regression for the above goal
3. and understanding how logistic regression works in practice.

Default data set

Let's consider the Default data set, where the response default falls into one of two categories, Yes or No.

There are four predictors for this data set:

1. default: A factor with levels No and Yes indicating whether the customer defaulted on their debt
2. student: A factor with levels No and Yes indicating whether the customer is a student
3. balance: The average balance that the customer has remaining on their credit card after making their monthly payment
4. income: Income of customer

One goal is to predict which customers will default on their credit card debt.

Default Data set

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

For the Default data, logistic regression models the probability of default (defaulting on credit card debt).

For example, the probability of default given balance can be written as

$$Pr(default = Yes|balance).$$

We will denote the above probability by $p(balance)$ for convenience.

Logistic regression for the default

How should we model the relationship between

$$p(X) = Pr(Y = 1 \mid X) \quad \text{and} \quad X?$$

Recall the linear regression model

$$p(X) = \beta_0 + \beta_1 X$$

where the goal is to use default = Yes to predict balance.

Logistic regression

How does logistic regression work from a modeling perspective?

Logistic regression

We model $p(X)$ using a function that gives outputs between 0 and 1 for all values of X, where

$$p(X) = \beta_0 + \beta_1 X$$

In logistic regression, we use the logistic function,

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}. \tag{1}$$

To fit the model in equation 1 we use **maximum likelihood estimation**

Logistic regression

After some manipulation of equation 1, we find that

$$\frac{p(X)}{1 - p(X)} = e^{\beta_o + \beta_1 X}, \tag{2}$$

where the quantity on the left is called the **odds.**

- The odds can take any value between $[0, \infty]$
- Values of the odds close to 0 and ∞ indicate very low and very high probabilities of default, respectively.

Logistic regression

Taking the log of equation 2 gives us

$$\log(\frac{p(X)}{1-p(X)}) = \beta_o + \beta_1 X, \tag{3}$$

where the left hand side is called the **log-odds** or **logit**.

Note the logistic regression model 1 has a logit linear in X.

Logistic regression versus linear regression

Recall that in a **linear regression model**, β_1 gives the average change in Y associated with a **one-unit increase** in X.

In contrast, in a **logistic regression model**, increasing X by one unit changes the **log odds** by β_1, or equivalently it multiplies the odds by e^{β_1}.

Estimating the Regression Coefficients

- β_0 and β_1 in equation 1 are unknown and must be estimated using the training data.
- We could use the method of least squares to fit the coefficents, however the method of **maximum likelihood** is preferred since it has better statistical properties.

Maximum Likelihood Estimation (MLE)

The goal is to find the values of β_0 and β_1 that maximize the likelihood function $\ell(\beta_0, \beta_1)$.

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize the likelihood function.

Remark: least squares for linear regression is a special case of maximum likelihood estimation.

Practice exercise

For linear regression and for logistic regression, find the values of β_0 and β_1 that maximize the likelihood function $\ell(\beta_0, \beta_1)$.

Hints: What is the likelihood? Look a the log-likelihood to make life easier. Write down the next steps and then find the solutions!

Logistic regression

First, we install the packages we need.

```r
library(ISLR)
attach(Default)
library(plyr)
names(Default)
```

```
## [1] "default" "student" "balance" "income"
```

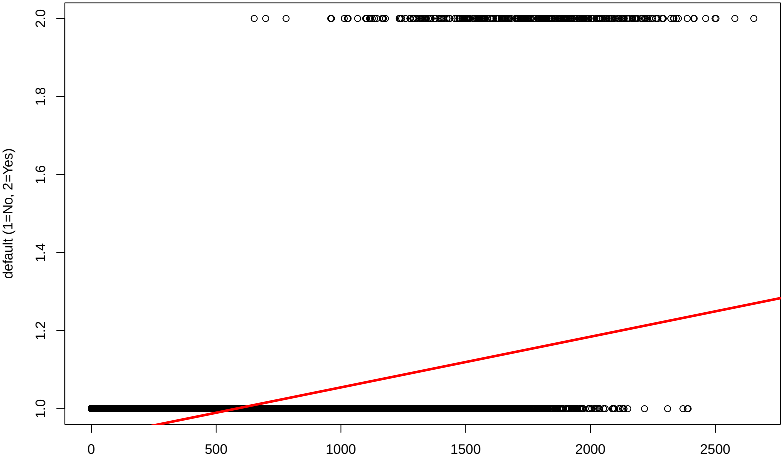
Linear regression

We first investigate linear regression, and then compare it to logistic.

```r
# linear regression
lm.fit <- lm(as.numeric(default)~balance, data=Default)
```

Linear Regression

```
plot(balance, default,xlab="balance",
     ylab= "default (1=No, 2=Yes)")
abline(lm.fit, col="red", lwd=3)
```



The problem with linear regression

For balances close to **zero** we predict a **negative probability of default**

If we were to predict for very large balances, we would get values bigger than 1.

These predictions are not sensible!

The problem with linear regression

This problem is not unique to the credit default data.

Any time a straight line is fit to a binary response that is coded as 0 or 1, in principle we can always predict $p(X) < 0$ for some values of X and $p(X) > 1$ for others (unless the range of X is limited).

Logistic regression

To run a logistic regression, we use the function glm(), giving it an argument of the generalized linear family, which for the logistic regression is the **binomial** family.

Logistic Regression

```r
glm.fit <- glm(default~balance,
               family="binomial", data=Default)
```

Logistic Regression

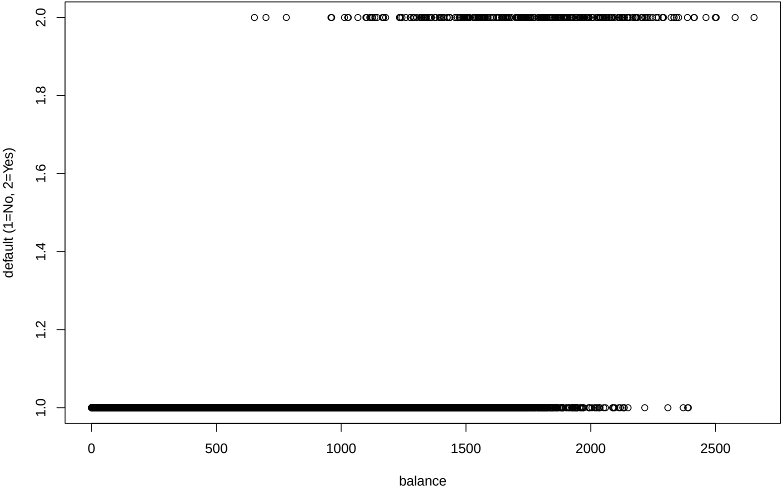
```
summary(glm.fit)
```

```
##
## Call:
## glm(formula = default ~ balance, family = "binomial", data = Default)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
## balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1596.5  on 9998  degrees of freedom
## AIC: 1600.5
##
## Number of Fisher Scoring iterations: 8
```

```
default <- revalue(default, c("Yes"="1", "No"="0"))
# Rescale the fitted values so they
# appear on the plot (not prob anymore)
fitted <- fitted(glm.fit) + 1
```

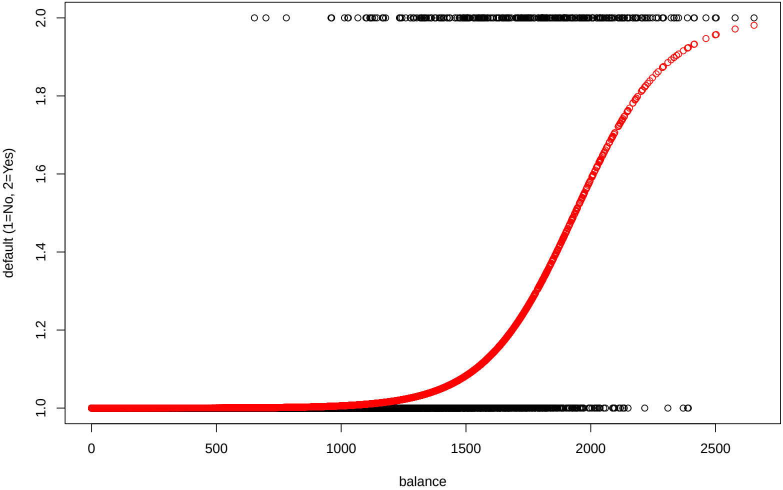
Logistic Regression

```
plot(balance,default, xlab="balance",
     ylab= "default (1=No, 2=Yes)")
```



Logistic Regression

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): '
## graphical parameter
```



Comparison of Linear and Logistic Regression

- Notice that for low balances we now predict the probability of default as close to, but never below, zero.
- Likewise, for high balances we predict a default probability close to, but never above, one.
- The logistic function always produces an S-shaped curve. Regardless of the value of X, we obtain a sensible prediction.
- We also see that the logistic model is better able to capture the range of probabilities compared to the linear regression model.

Estimating the Regression Coefficients

We return to the Default data set to estimate the regression coefficients.

- $\hat{\beta}_1 = 0.0055$, indicating that an increase in balance is associated with an **increase** in the probability of default.
- To be precise, a **one-unit increase** in balance is associated with an increase in the **log odds** of default by 0.0055 units.

```
glm.fit$coeff
```

```
##   (Intercept)       balance
## -10.651330614   0.005498917
```

Making Predictions

This uses all the data to make predictions of default.

```r
predict(glm.fit, type="response")[1:5]
```

```
##            1            2            3            4
## 0.0013056797 0.0021125949 0.0085947405 0.0004344368 0.00
```

For more meaninful predictions, see the lab on page 156 for logistic regression and work through this on your own.

Multiple Logistic Regression

We now consider the problem of predicting a binary response using multiple predictors.

We generalize this as follows:

$$\log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \tag{4}$$

where $X = (X_1, \ldots, X_p)$ are p predictors.

Multiple Logistic Regression

Equation 4 can be re-written as

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}. \tag{5}$$

We estimate $\beta_0, \beta_1, \ldots, \beta_p$ using maximum likelihood estimation.

Multiple Logistic Regression

We return to the Default data set.

Let's estimate a logistic regression model that uses balance, income (in thousands of dollars), and student (status) to predict the probability of default.

Multiple Logistic Regression

```
glm.fit.mlr <- glm(default~balance + income + student, family="binomial", data=Default)
summary(glm.fit.mlr)
```

```
##
## Call:
## glm(formula = default ~ balance + income + student, family = "binomial",
##     data = Default)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
## balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
## income       3.033e-06  8.203e-06   0.370  0.71152
## studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.5  on 9996  degrees of freedom
## AIC: 1579.5
##
## Number of Fisher Scoring iterations: 8
```

Multiple Logistic Regression

There is a surprising result here.

- The p-values associated with balance and student (status) are very small, indicating that each of these variables is associated with the probability of default.
- The coefficient for student (status) is negative, indicating that students are less likely to default than non-students.

Suppose we just look at how student (status) predicts default. In this case, the coefficient for the dummy variable of student is positive, contrasting what we just saw!

Multiple Logistic Regression and Confounding

How is it possible for student status to be associated with an increase in probability of default in one model and a decrease in probability of default in a second model?

Exercise: Read about **confounding** on page 136–137.

- What is confounding?
- Reproduce Figure 4.3 using simple R commands.
- Be able to explain these to others in the class if asked!

Logistic Regression for > 2 Response Classes

We sometimes wish to classify a response variable that has more than two classes.

- Recall that one might have medical conditions that fall into three categories.
- The two-class logistic regression models have multiple-class extensions, but in practice they tend not to be used all that often since they do not work well in practice.
- We turn to other methods that are superior such as linear discrimminant analysis (LDA) instead.