DATA 572: Supervised Learning

2023W2

Shan Du

Classification

- In many situations, the response variable is *qualitative* instead of quantitative.
- Often qualitative variables are referred to as *categorical*; we will use these qualitative terms interchangeably.
- The process for predicting qualitative responses is known as *classification.*
- For classification, we first predict the probability that the observation belongs to each of the categories of a qualitative variable, as the basis for making the classification. In this sense they also behave like regression methods.

Classification

- 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 their symptoms. In this simplified example, there are three possible diagnoses: *stroke*, *drug overdose*, and *epileptic seizure*. We could consider encoding these values as a quantitative response variable, Y, as follows:

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

Classification

We could choose an equally reasonable coding,

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

Order changed so value of 1 can mean anything...not fixed

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

Can use this method if there is a natural order - ex: mild, moderate, and exterme -> 0,1,2

Classification

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

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

- We could then fit a linear regression to this binary response, and predict *drug overdose* if $\hat{Y} > 0.5$ and *stroke* otherwise.
- $X\hat{\beta}$ obtained using linear regression is in fact an estimate of $\Pr(drug\ overdose|X)$.

Classification

- However, if we use linear regression, some of our estimates might be outside the $[0, 1]$ interval making them hard to interpret as probabilities.
- Therefore, it is preferable to use a classification method that is truly suited for qualitative response values.

value might be greater than 1 or less than 0, meaning its an inappropriate probability

Logistic Regression

- Logistic regression models the probability that Y belongs to a particular category.

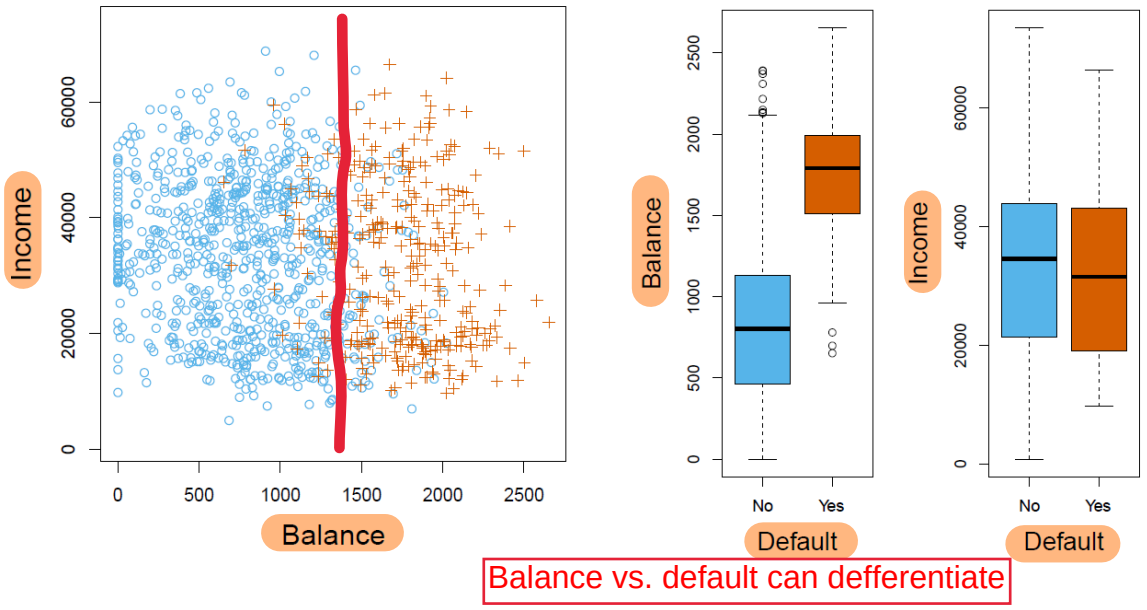

FIGURE 4.1. *The* `Default` *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* `balance` *as a function of* `default` *status.* Right: *Boxplots of* `income` *as a function of* `default` *status.*

Logistic Regression

- The response *default* falls into one of two categories, *Yes* or *No*.
- Rather than modeling this response Y directly, logistic regression models the probability that Y belongs to a particular category, e.g., $\Pr(default = Yes|balance)$. We abbreviate it by $p(balance)$, which will range between 0 and 1.
- Then for any given value of balance, a prediction can be made for *default*.

The Logistic Model

- How should we model the relationship between $p(X) = \Pr(Y = 1|X)$ and X?
- If we use a linear regression model to represent these probabilities:

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

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 (shown below).

The Logistic Model

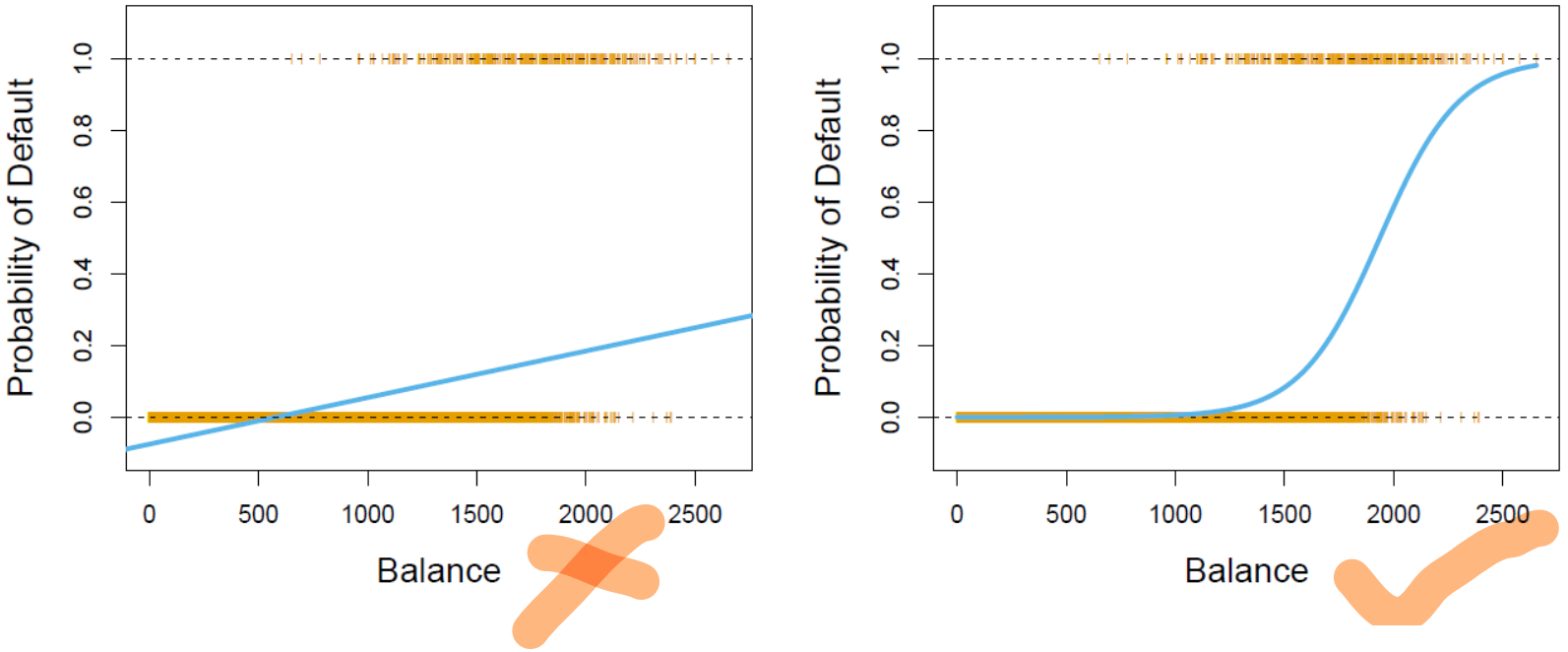

FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default` *(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using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

The Logistic Model

- To avoid this problem, we must model $p(X)$ using a function that gives outputs between 0 and 1 for all values of X. Many functions meet this description. In logistic regression, we use the *logistic function*,

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

To fit the model, we use a method called *maximum likelihood*.

The Logistic Model

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

is called the *odds*, and can take on any value between odds 0 and ∞.

- Values of the odds close to 0 and ∞ indicate very low and very high probabilities of default, respectively.
- Odds are traditionally used instead of probabilities in horse-racing, since they relate more naturally to the correct betting strategy.

The Logistic Model

- For example, on average 1 in 5 people with an odds of 1/4 will default, since $p(X) = 0.2$ implies an odds of $\frac{0.2}{1-0.2} = 1/4$.

1/5 = 0.2

- On average 9 out of every 10 people with an odds of 9 will default, since $p(X) = 0.9$ implies an odds of $\frac{0.9}{1-0.9} = 9$.
- By taking the logarithm of both sides

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

We have *log odds* or *logit*. We see that the logistic regression model has a logit that is linear in X.

Estimating the Regression Coefficients

- The coefficients β_0 and β_1 are unknown. We must use data to estimate the coefficients.
- Although we could use (non-linear) least squares to fit the model, the more general method of *maximum likelihood* is preferred, since it has better statistical properties.
- The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows:

Estimating the Regression Coefficients

- We seek estimates for β_0 and β_1 such that the predicted probability $\hat{p}(x_i)$ of default for each individual corresponds as closely as possible to the individual's observed default status.
- That is, we try to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that plugging these estimates into the model for $p(X)$, yields a number close to one for all individuals who defaulted, and a number close to zero for all individuals who did not.

Estimating the Regression Coefficients

- This intuition can be formalized using a mathematical equation called a *likelihood function*:

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

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

Making Predictions

- Once the coefficients have been estimated, we can compute the probability of *default* for any given credit card balance.

	Coefficient	Std. error	z-statistic	p-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

TABLE 4.1. *For the* Default *data, estimated coefficients of the logistic regression model that predicts the probability of* default *using* balance*. A one-unit increase in* balance *is associated with an increase in the log odds of* default *by* 0.0055 *units.*

- The *default* probability for an individual with a balance of $1,000 is 0.00576, which is below 1%. In contrast, the predicted probability of default for an individual with a balance of $2,000 is much higher, and equals 0.586 or 58.6 %.

Making Predictions

- One can use qualitative predictors with the logistic regression model.

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	−3.5041	0.0707	−49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

TABLE 4.2. *For the* `Default` *data, estimated coefficients of the logistic regression model that predicts the probability of* `default` *using student status. Student status is encoded as a dummy variable, with a value of* 1 *for a student and a value of* 0 *for a non-student, and represented by the variable* `student[Yes]` *in the table.*

- As an example, the *Default* data set contains the qualitative variable *student*. To fit a model that uses student status as a predictor variable, we simply create a dummy variable that takes on a value of 1 for students and 0 for non-students.

Multiple Logistic Regression

- We can predict a binary response using multiple predictors.

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

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

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}$$

Multiple Logistic Regression

- We use the maximum likelihood method to estimate $\beta_0, \beta_1, \ldots, \beta_p$.

	Coefficient	Std. error	z-statistic	p-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

TABLE 4.3. *For the* Default *data, estimated coefficients of the logistic regression model that predicts the probability of* default *using* balance, income, *and student status. Student status is encoded as a dummy variable* student[Yes], *with a value of* 1 *for a student and a value of* 0 *for a non-student. In fitting this model,* income *was measured in thousands of dollars.*

Multiple Logistic Regression

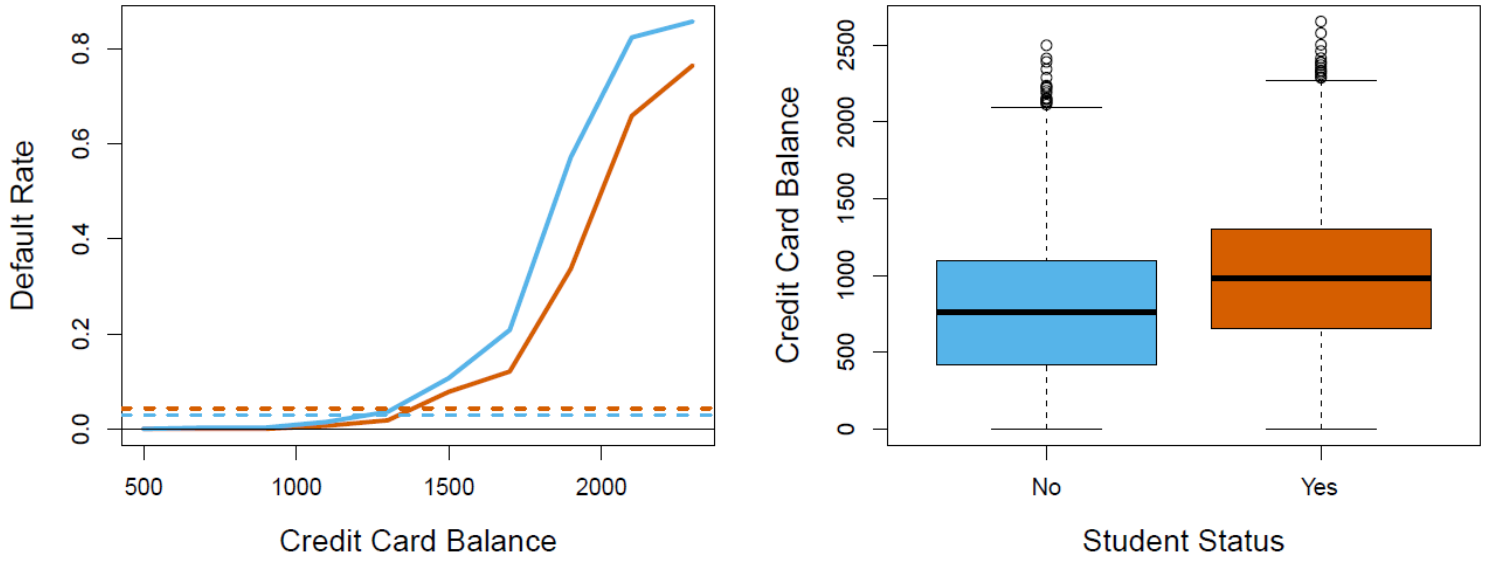

FIGURE 4.3. *Confounding in the* `Default` *data.* Left: *Default rates are shown for students (orange) and non-students (blue). The solid lines display default rate as a function of* `balance`*, while the horizontal broken lines display the overall default rates.* Right: *Boxplots of* `balance` *for students (orange) and non-students (blue) are shown.*

Multiple Logistic Regression

- A student with a credit card balance of $1,500 and an income of $40, 000 has an estimated probability of default of 0.058.
- A non-student with the same balance and income has an estimated probability of default of 0.105.

Multinomial Logistic Regression

- We can classify a response variable that has more than two classes. This extension is sometimes known as multinomial logistic regression.
- To do this, we first select a single class to serve as the *baseline*; without loss of generality, we select the Kth class for this role. Then we replace the model with

Multinomial Logistic Regression

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\ldots+\beta_{kp}x_p}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\ldots+\beta_{lp}x_p}}$$

for $k = 1, .., K-1$, and

$$\Pr(Y = K|X = x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\ldots+\beta_{lp}x_p}}$$

Multinomial Logistic Regression

- Therefore for $k = 1, .., K-1$,

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = K|X = x)}\right)$$
$$= \beta_{k0} + \beta_{k1}x_1 + \ldots + \beta_{kp}\,x_p$$

- The log odds between any pair of classes is linear in the features.

Multinomial Logistic Regression

- The decision to treat the Kth class as the baseline is unimportant. For example, when classifying emergency room visits into *stroke*, *drug overdose*, and *epileptic seizure*, suppose that we fit two multinomial logistic regression models: one treating *stroke* as the baseline, another treating *drug overdose* as the baseline.
- The coefficient estimates will differ between the two fitted models due to the differing choice of baseline, but the fitted values (predictions), the log odds between any pair of classes, and the other key model outputs will remain the same.

Multinomial Logistic Regression

- We now briefly present an alternative coding for multinomial logistic regression, known as the *softmax* coding.
- The softmax coding is used extensively in some areas of the machine learning literature.
- In the softmax coding, rather than selecting a baseline class, we treat all K classes symmetrically, and assume that for $k = 1, .., K,$

Multinomial Logistic Regression

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\ldots+\beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\ldots+\beta_{lp}x_p}}$$

- Thus, rather than estimating coefficients for $K-1$ classes, we actually estimate coefficients for all K classes.
- The log odds ratio between the kth and k'th classes equals

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = k'|X = x)}\right)$$
$$= (\beta_{k0} - \beta_{k'0}) + (\beta_{k1} - \beta_{k'1})x_1 + \ldots + (\beta_{kp} - \beta_{k'p})x_p$$