Logistic Regression

Stephen Jun Villejo

School of Statistics

Classification

- The linear regression model we discussed assumes that the response variable is quantitative.
- But in many situations, we deal with *qualitative* or *categorical* response variables.
- The process of predicting a qualitative response for an observation is known as **classification**.
- The most commonly used classification techniques or *classifiers* are: logistic regression, linear discriminant analysis, and K-nearest neighbors.
- More computer-intensive classifiers are: regression trees, classification trees, random forests, boosting, support vector machines.

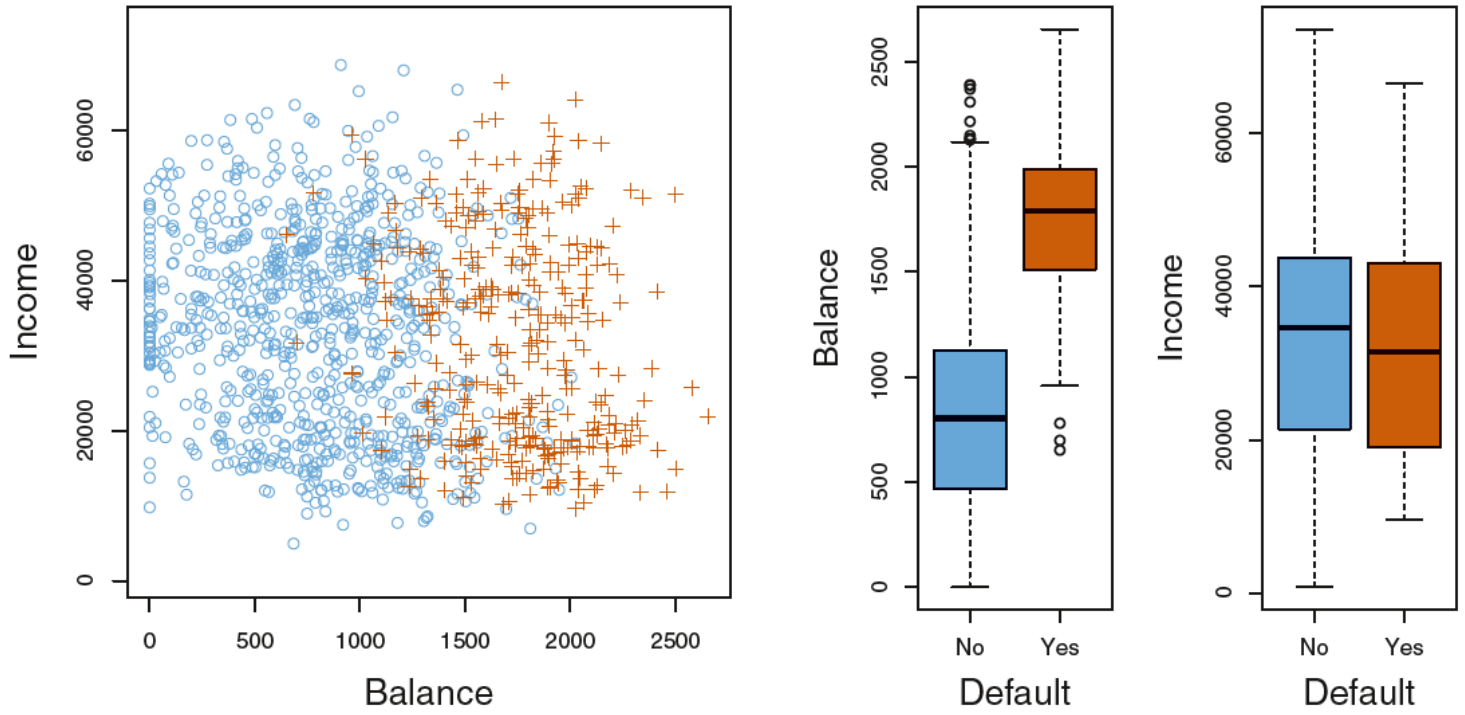
Classification

Examples

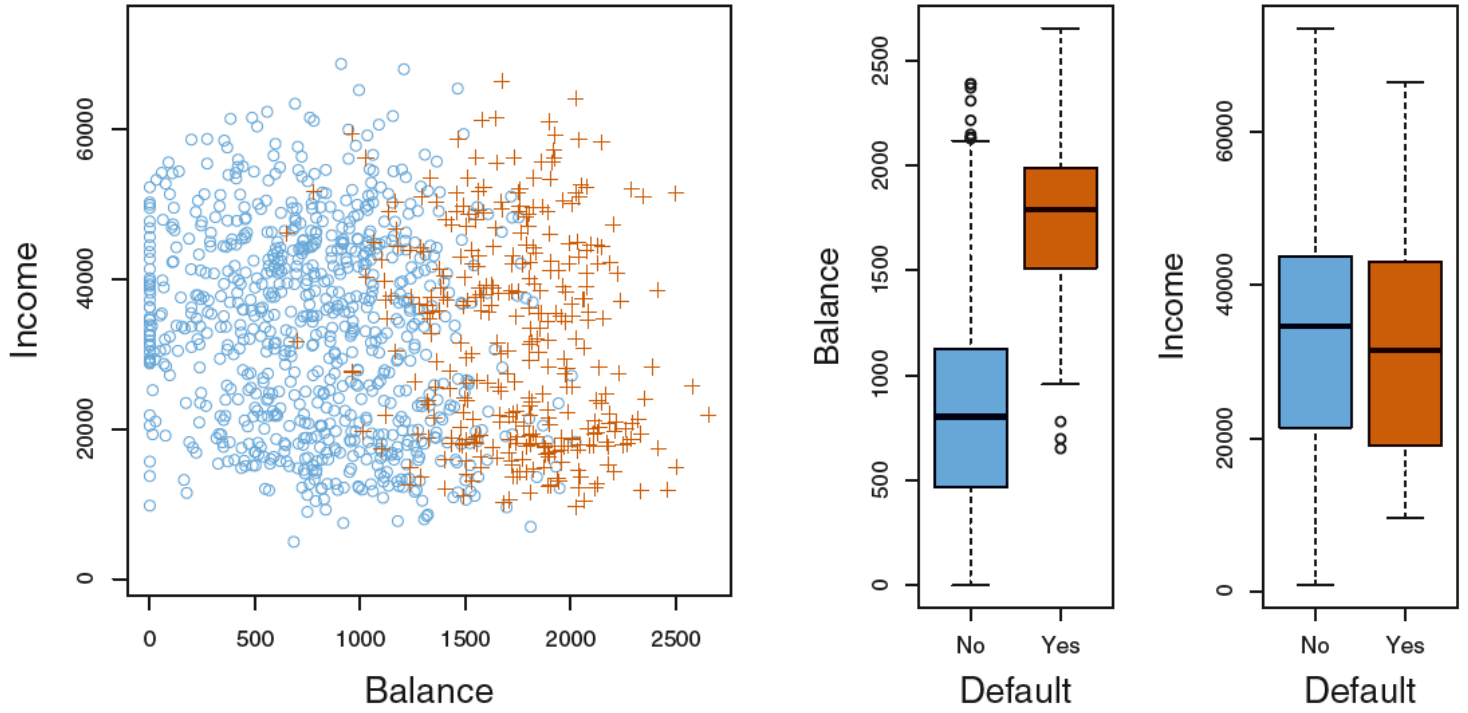
- A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?
- An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.
- On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

Classification

- Suppose 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. Below is the data based on 10,000 individuals.



Classification

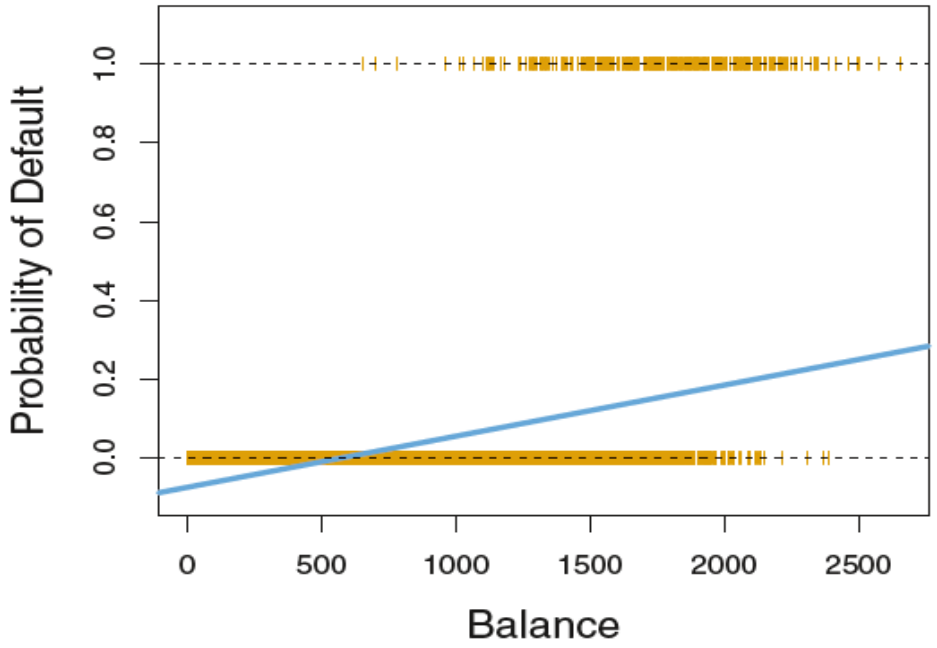


- It appears that individuals who defaulted tended to have higher credit card balances than those who did not.

Classification

Why Not Linear Regression?

- If we use linear regression, some of our estimates might be outside the [0,1] interval.
- Using the default data, some of the estimated probabilities are negative:



Logistic Regression

- Rather than modeling the 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 given `balance` which can be written as

$$\mathrm{P}(\mathrm{default} = \mathrm{Yes}|\mathrm{balance}).$$

- The values of $\mathrm{P}(\mathrm{default} = \mathrm{Yes}|\mathrm{balance})$ will range between 0 and 1. Then for any given of `balance`, a prediction can be made for `Default`.
- For example, one might predict `Default=Yes` if

$$\mathrm{P}(\mathrm{default} = \mathrm{Yes}|\mathrm{balance}) > 0.5.$$

- A lower threshold may be chosen if a company wishes to be conservative.

Logistic Regression

- For a binary qualitative response Y and a single predictor X, the logistic model is given by

$$\mathrm{P}(\mathrm{Y}=1|\mathrm{X})=\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

- The logistic function will always produce an S-shaped curve, and so regardless of X, the estimated probabilities are always between 0 and 1.
- Some manipulation of the model will give

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

- The left-hand side in the equation above is called the **log-odds** or **logit**.

Logistic Regression

- The quantity

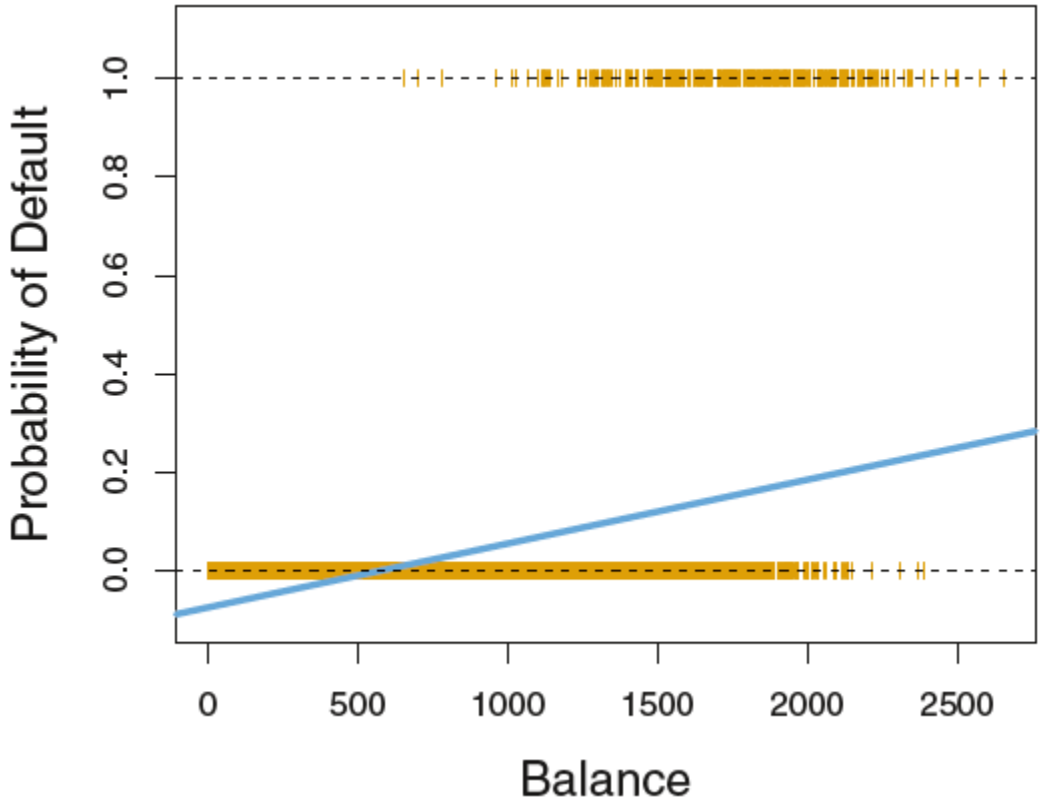
$$\frac{P(Y = 1|X)}{P(Y = 0|X)}$$

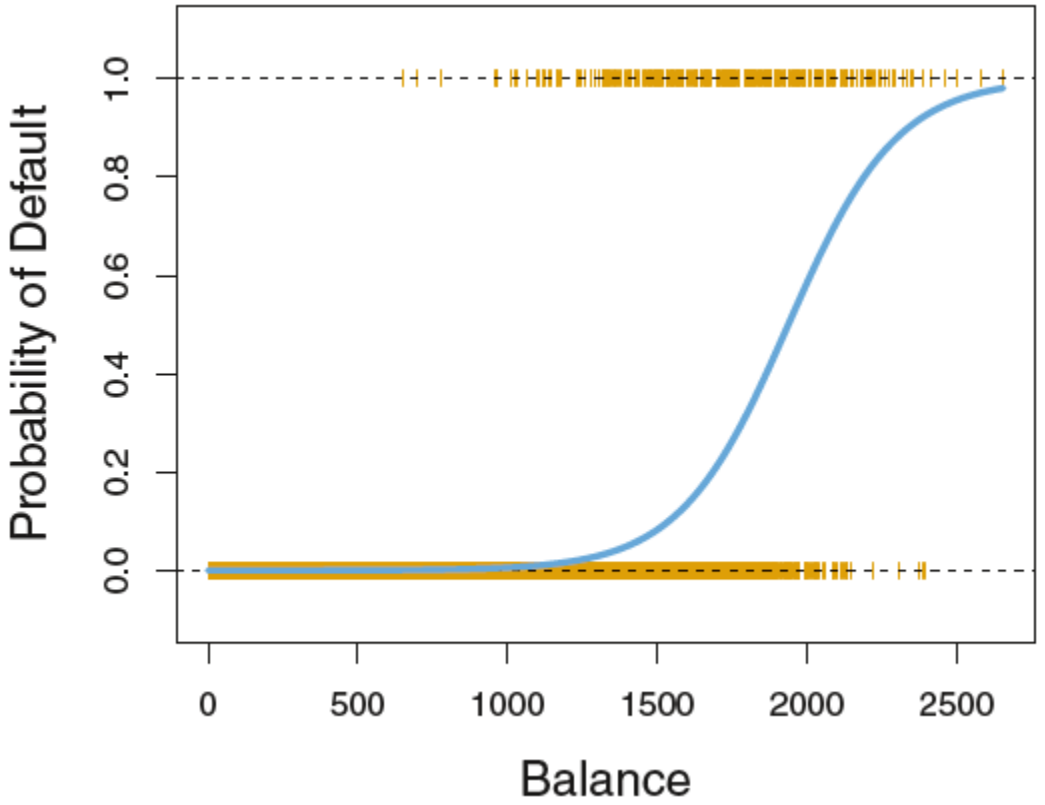
  is called the **odds**, and can take any value from 0 to ∞.

- Value of the odds close to 0 means that the probability of success is very low, while odds close to 1 means that the probability of success is high.

Logistic Regression

- Figure on the left are the estimated probabilities using linear regression, while figure on the right are the estimated probabilities using logistic regression.





Logistic Regression

Interpretation of β_1

- Increasing X by one unit changes the log-odds by β_1, or equivalently, it multiplies the odds by e^{β_1}.
- The relationship between $\mathrm{P}(\mathrm{Y} = 1|\mathrm{X})$ and X is not linear, so β_1 does not correspond to the change in $\mathrm{P}(\mathrm{Y} = 1|\mathrm{X})$ associated with a one-unit increase in X.
- The amount that $\mathrm{P}(\mathrm{Y} = 1|\mathrm{X})$ changes due to a one-unit change in X will depend on the current value of X.
- But regardless of the value of X, if β_1 is positive, then increasing X will be associated with increasing $\mathrm{P}(\mathrm{Y} = 1|\mathrm{X})$.

Logistic Regression

Example

- Below are the estimates from fitting a logistic regression model on `Default` data, in order to predict the probability of default using `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

- We see that an increase in balance is associated with an increase in the probability of default.
- A one-unit increase in the balance is associated with an increase in the log odds of default by 0.0055 units.
- The p-value is small, so we conclude that there is indeed an association between balance and the probability of default.

Logistic Regression

Example

- It is a simple matter to compute the probability of default for any given credit card balance. For example, to predict the default probability for an individual with a balance of 1000, we have

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1,000}}{1+e^{-10.6513+0.0055\times 1,000}} = 0.00576,$$

  which is below 1%.

- In contrast, the predicted probability of default for an individual with a balance of 2000 is much higher, equal to 0.586 or 58.6%.

Multiple Logistic Regression

- When we have multiple predictors, the logistic model is now given by

$$\log\left(\frac{\mathrm{P}(\mathrm{Y}=1|\mathrm{X})}{\mathrm{P}(\mathrm{Y}=0|\mathrm{X})}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p.$$

which may also be rewritten as

$$\mathrm{P}(\mathrm{Y}=1|\mathrm{X})=\frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p}}.$$

Logistic Regression

Example

- Below are the estimates from fitting a logistic regression model on `Default` data, in order to predict the probability of default using `balance`, `income`, and `student`.

	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

- The variables `balance` and `student` are significant because the p-values are small.

Logistic Regression

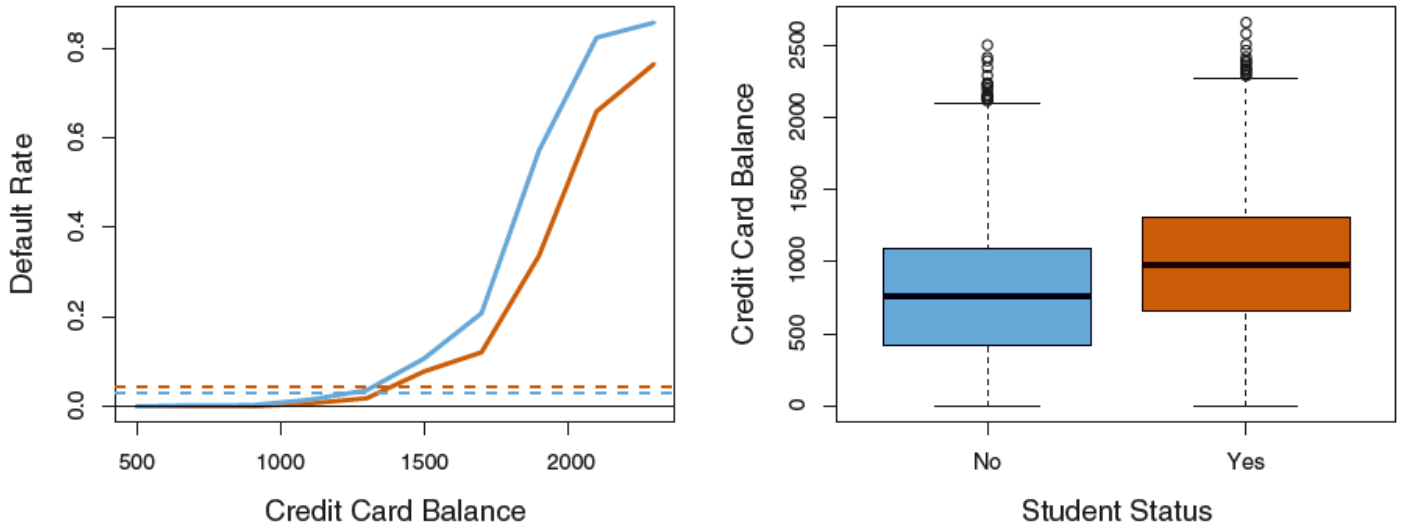
Example

	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

- Higher levels of debt (balance) is associated with higher probability of default.
- For a fixed value of balance and income, a student is less likely to default than a non-student.

Logistic Regression

Example



- Students tend to hold high levels of debt. In other words, students are more likely to have large credit card balances, which is in turn associated with higher probabilities of default.
- A student is riskier than a non-student if no information about the student's credit card balance is available. However, that student is less risky than a non-student with the same credit card balance.

R Exercise