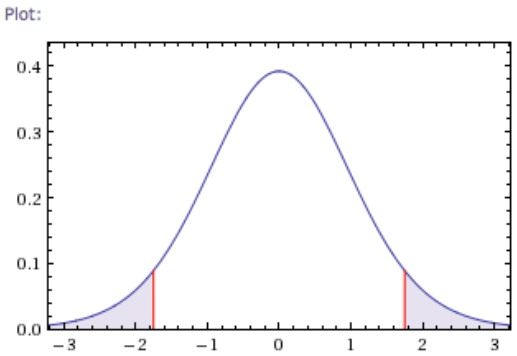
ISLR Chapter 4
Classification

Logistic Regression

(reviewed by B. Haas, 11/2014)

Quantitative vs. Qualitative Response Variable

Quantitative: Y = (some numerical value)



Qualitative: Y = (some categorical variable)

eg. eye color

Classification: predicting qualitative values, or assigning observations to a category or a class.

Most widely used classifiers include:

- Logistic regression (this week)
- Linear discriminant analysis (next week)
- K-nearest neighbors
- Computer intensive methods:
  - Generalized additive models (Chapter 7)
  - Trees, Random forests, Boosting (Chapter 8)
  - Support vector machines (chapter 9)

Examples of Classification Problems

1. 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?

2. 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.

3. 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.

Linear Regression is Not Appropriate for Predicting Qualitative Response Variables

Why Linear Regression Not Appropriate for Qualitative Response?

1. 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?

Alternative potential encodings for response variable:

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

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

Either implies an ordering and relative relationship among the possible responses, which may not make sense. Different encodings lead to different linear models.

There are some cases where linear regression could be appropriate

- A clear numerical relationship exists among the response variable values:

$$Y = \begin{cases} 0 & \text{if mild;} \\ 1 & \text{If moderate;} \\ 2 & \text{If severe.} \end{cases}$$

Ok, if mild < moderate < severe and we agree on their relative differences in values.

There are some cases where linear regression could be appropriate

- The response variable is binary (0|1)

Encode Response for Training

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

Prediction using linear regression:

$$Y = \begin{cases} < 0.5 \quad \text{stroke}; \\ \text{else drug overdose}. \end{cases}$$

Can interpret as Y = **P(drug overdose|X)**

Problem: Not really P-values, as linear regression line can go below zero and above 1.

‘Default’ data set for studying classification

- Response variable:
  - will a person ***default*** on their credit card payment. (Yes/No)
- Explanatory variables:
  - Annual ***income***
  - Monthly credit card ***balance***

- Goal is to predict (default) ~ (income, balance)

Strong Relationship Between Balance and Default. Not so much for Income.

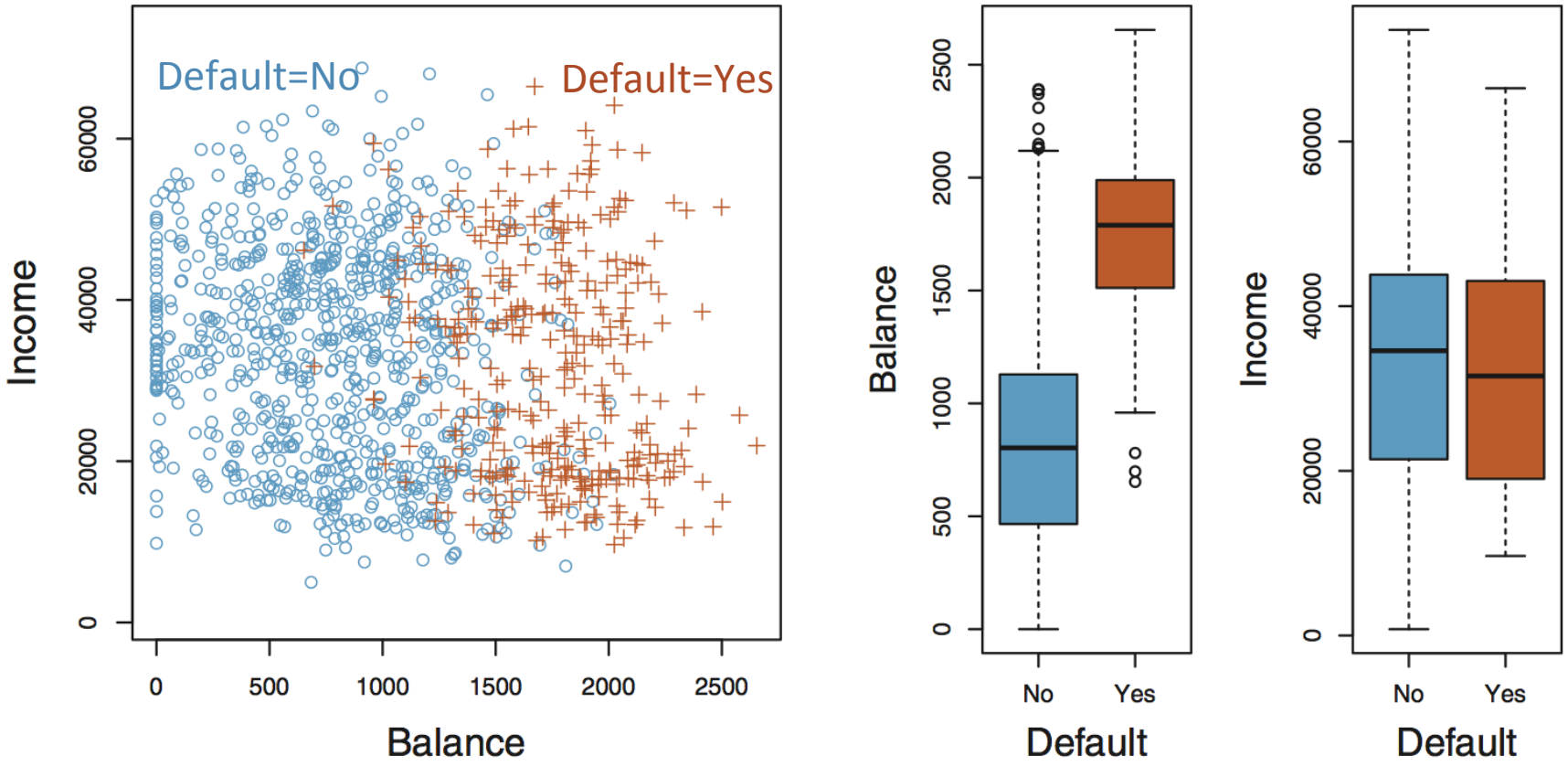


FIGURE 4.1. *The* `Default` *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* `balance` *as a function of*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

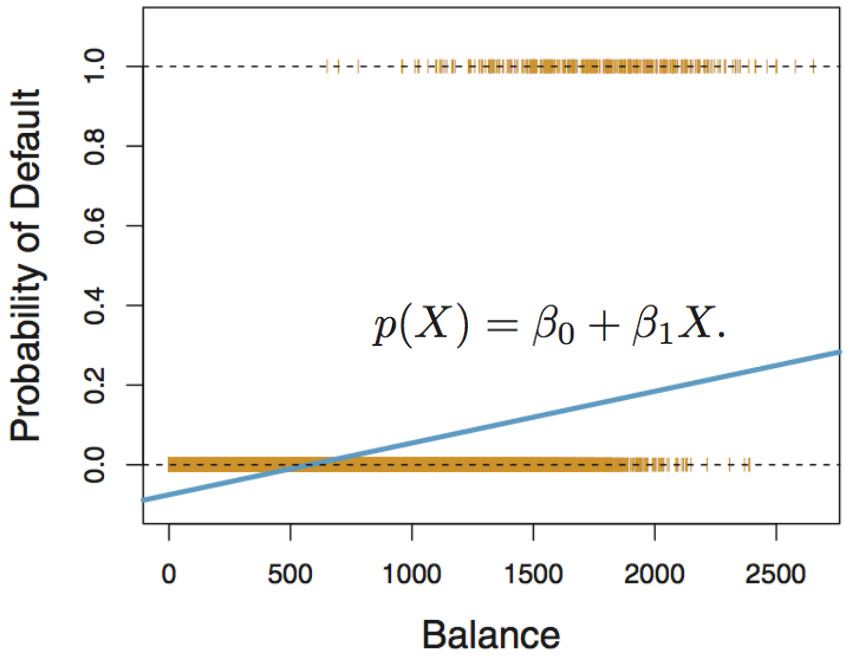
Fit a simple linear model using binary response encoding

Response encoding for lm fit:

default = { 0 if No; 1 If Yes.

Interpretation:

P(***default*** = Yes | ***balance***)



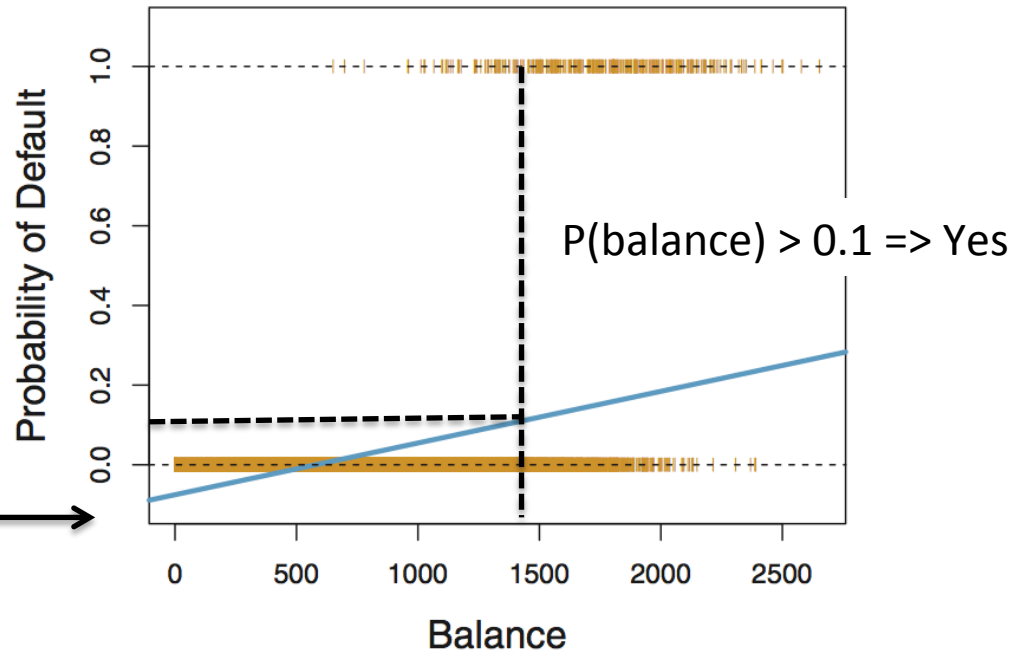
Fit a simple linear model using binary response encoding

Response encoding for lm fit:

default = { 0 if No; 1 If Yes.

Interpretation:

P(***default*** = Yes | ***balance***)



Probability values < 0 !!

Logistic Model to the Rescue

Instead of our linear regression model:

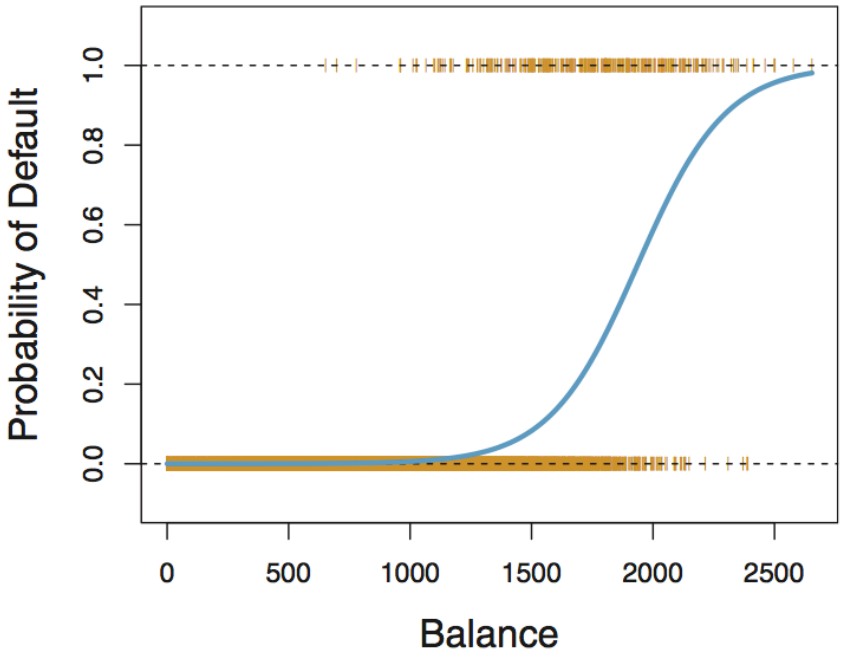
$$p(X) = \beta_0 + \beta_1 X.$$

Use a logistic regression model:

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

Logistic Regression Applied to the *Default* Data

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



Logistic regression:

- Produces and S-shaped curve
- All p(X) values between 0 and 1

Logistic regression model:

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

Odds ratio:

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

Log-odds or logit is linear with respect to X

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

Increasing X by one unit changes the log-odds by B_1.

(or multiples the odds by e^{B1})

* The amount p(X) changes due to a one-unit change in X depends on X.

Estimating the Coefficients

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

Find B_0 and B_1 such that

- those that defaulted have p(X) closest to 1
- those that did not default have p(X) closest to 0

Computing the optimal B_0,B_1: find values for B_0,B_1 that maximizes the likelihood:

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

Probability that all those that did default, do default

Probability that all those that did NOT default, do NOT default

Results from Fitting Logistic Regression on the 'default' data

	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.*

$$z\text{-statistic associated with } \beta_1 \text{ is equal to } \hat{\beta}_1 / SE(\hat{\beta}_1)$$

Null hypothesis $H_0 : \beta_1 = 0$ Implies $p(X) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$

(in other words, p(default) does not depend on balance… reject the null due to small P!)

Note: intercept term is not of interest, serves to adjust probabilities for the number of ones and zeros in the data.

Make Predictions Using the Learned Coefficients

Just plug in the coefficients and the value for balance to get P(default)

Ie. For balance = $1000

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

P(default) is less than 1%, so we'll say 'No'

Logistic Regression Using Qualitative Variables

- Earlier, computing P(default | balance) where balance is a numerical (quantitative) value.
- Logistic regression can also be used to fit Qualitative (categorical) variables similarly as in linear regression – use Dummy variables.

Example: P(default | student) where student = {Yes, No}

Encode student status using dummy variable.

Fit $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$. where X { Student = 1, Non-student = 0

P(default)

(all done using a single qualitative variable of student status, entirely ignoring balance)

Logistic Regression Coefficient Fitted for Student Status

	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.*

Coefficient is positive and statistically significant.

So, more likely to default if you're a student.

Can determine exact probabilities of default for student vs. non-student

Just plug in the coefficients and dummy variable student status value

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

P(default) is higher for the student than the non-student.

Multiple Logistic Regression

(can mix quantitative and qualitative explanatory variables, just as in linear regression)

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \qquad (4.6)$$

where $X = (X_1, \ldots, X_p)$ are p predictors. Equation 4.6 can be rewritten as

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}. \qquad (4.7)$$

Fit coefficients for multiple logistic regression

Using

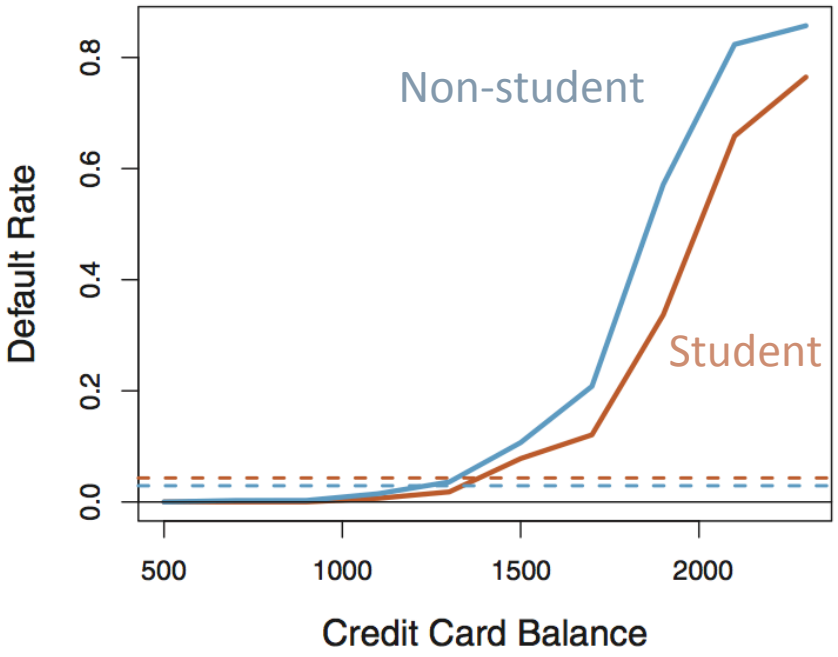
- balance quantitative variable
- Income quantitative variable
- student **qualitative** variable

	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.*

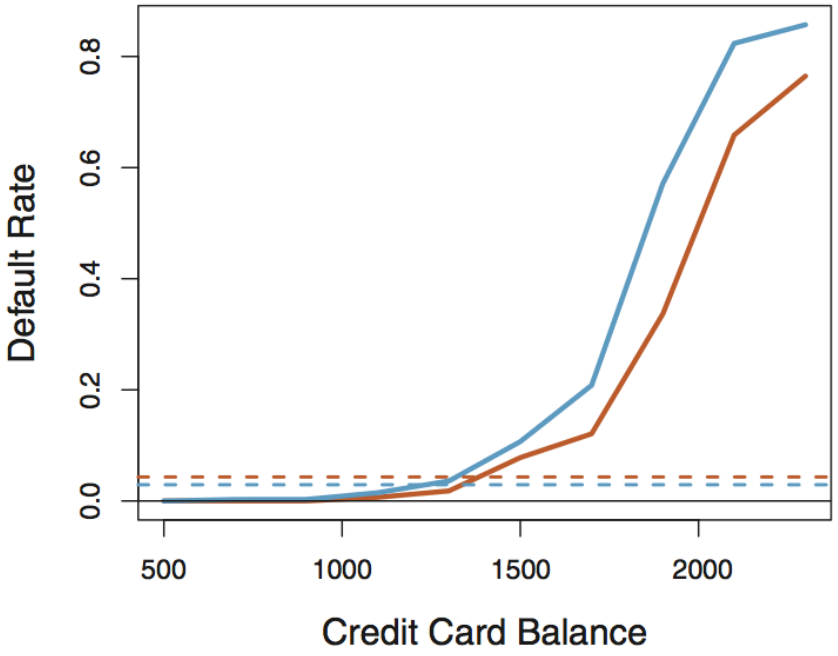
Balance and student status both statistically significant.

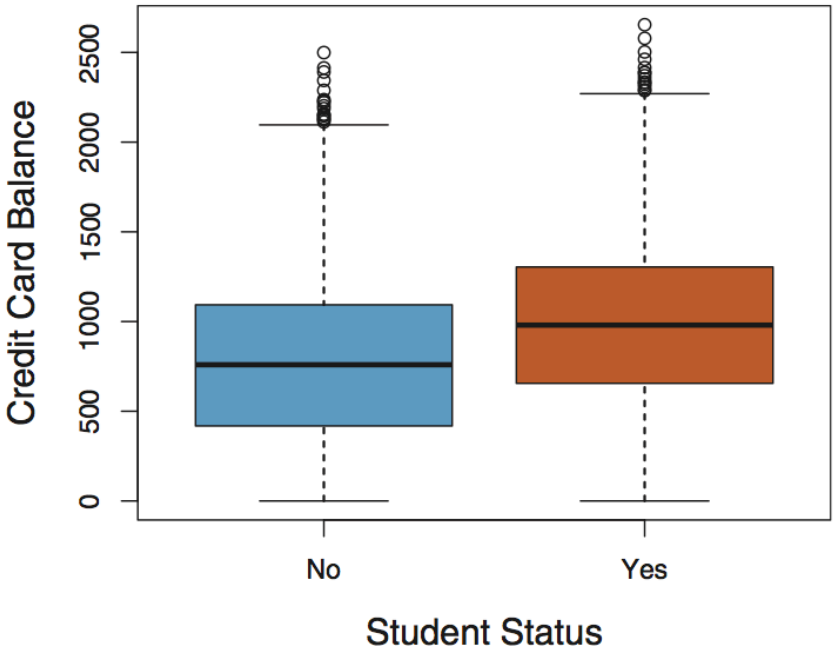
But… now a negative coefficient for student as compared to a positive coef earlier?



At a given balance, student default rate is lower.

Cumulatively (all balances and incomes), student default rate is slightly higher.





Can see that balances are higher for students than non-students, and this is a confounding correlative effect.

Beware: results from using one predictor alone may be different from models using multiple predictors! (especially when predictors are correlated)

Again, compute probabilities for the response variable by plugging in coefficients and variable values.

By substituting estimates for the regression coefficients from Table 4.3 into (4.7), we can make predictions. For example, a student with a credit card balance of \$1,500 and an income of \$40,000 has an estimated probability of default of

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}} = 0.058. \qquad (4.8)$$

A non-student with the same balance and income has an estimated probability of default of

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}} = 0.105. \qquad (4.9)$$

Determine response classes based on these probability values.

Logistic Regression for > 2 response classes?

- Don't bother…. Just use LDA

Next week: Linear Discriminant Analysis (LDA)