MATH 189
Classification via Logistic Regression

Wenxin Zhou
UC San Diego

Time: 2:00—3:20 & 3:30—4:50pm TueThur
Location: CENTR 115

Outline

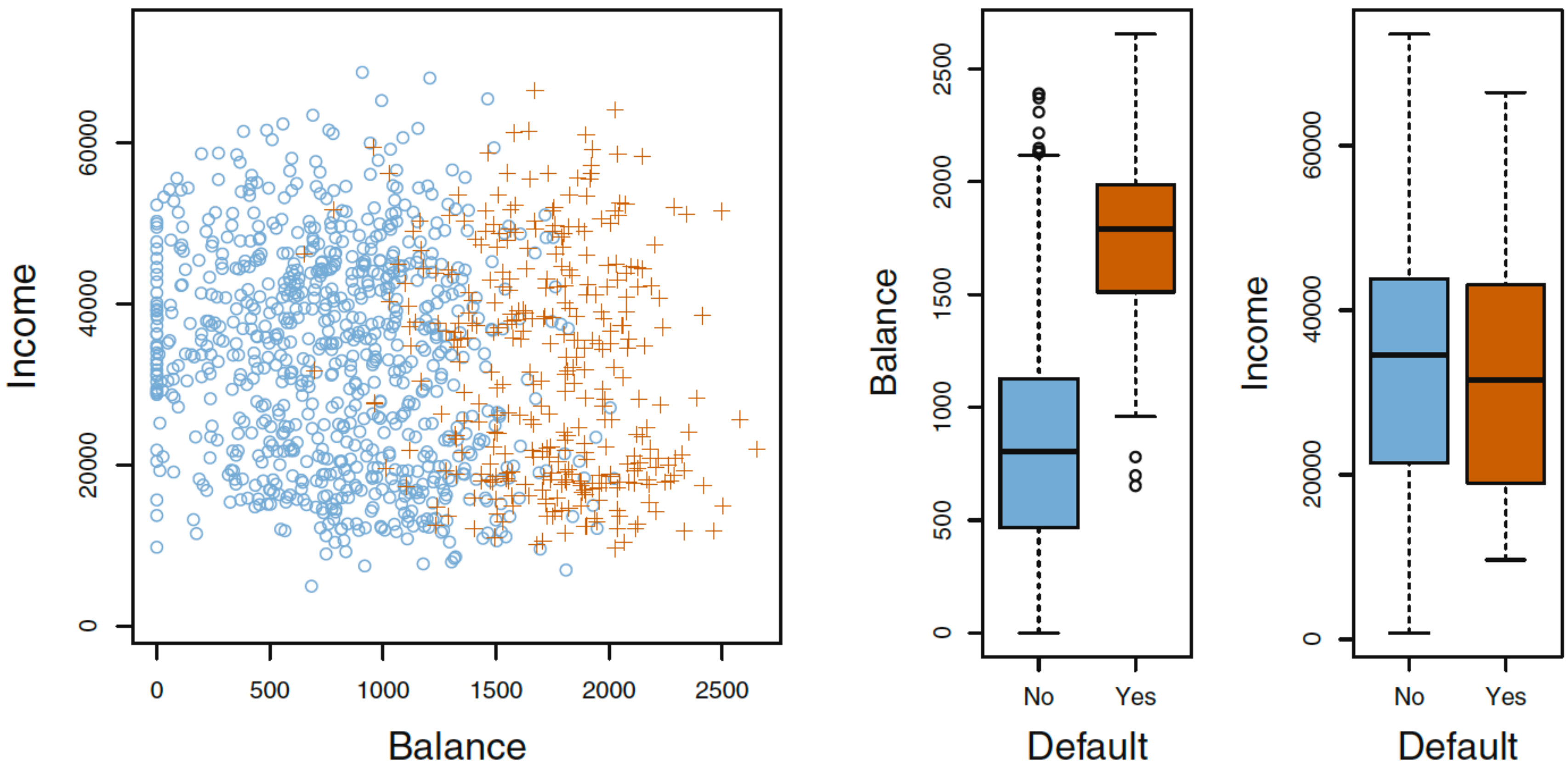
- In the previous lecture, we introduced the method of least squares for linear regression.
  - Simple linear regression
    - Estimation
    - Inference of parameter
    - Assess model accuracy
  - Multivariate linear regression

- Today we will introduce the logistic regression for classification.
  - Logistic function and logistic model
  - Multiple logistic regression
    - Maximum likelihood estimation
    - Gradient ascend optimization
    - Statistical inference

Classification on a Synthetic Data

- Problem: We are interested in predicting whether an individual will default on his/her credit card payment, on the basis of annual income and monthly credit card balance.
- Data: Default data set from the package "ISLR". It contains 10,000 individuals with 4 variables, default (Yes or No), student (Yes or No), balance and income.
- Visualization:
  - 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



- Goal: We wish to build a model to predict default (Y) for any given value of balance (X_1) and income (X_2).
- Logistic Regression: The response default falls into one of two categories, Yes or No. Logistic regression models the *probability* that Y belongs to a particular category.

For example, the probability of default given balance is

$$p(\text{balance}) := \Pr(\text{default} = \text{Yes} \mid \text{balance}) \ \in [0,1].$$

One might predict default = Yes for any individual for whom $p(\text{balance}) > 0.5$. Alternatively, if a company wishes to be conservative in predicting individuals who are at risk for default, then they may choose to use a lower threshold, such as $p(\text{balance}) > 0.1$.

The Logistic Model (1-dimensional)

Consider the conditional probability

$$p(x) = \mathrm{P}(Y = 1 \,|\, X = x), \quad x \in \mathbb{R}.$$

For convenience we use the generic 0/1 coding for the response (0: no, 1: yes).

The Logistic Model (1-dimensional)

Consider the conditional probability

$$p(x) = \mathrm{P}(Y = 1 \,|\, X = x), \quad x \in \mathbb{R}.$$

For convenience we use the generic 0/1 coding for the response (0: no, 1: yes).

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 this model, we use a method called *maximum likelihood*.

Note that

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

The left-hand side is called the *odds.* Values of the odds close to 0 and ∞ indicate very low and very high probabilities of default, respectively.

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 1/4. Likewise on average 9 out of every 10 people with an odds of 9 will default.

Odds are traditionally used instead of probabilities in horse-racing, since they relate more naturally to the correct betting strategy.

Estimating the Regression Coefficients

- Intuition: We seek estimates for β_0 and β_1 such that the predicted probability $\hat{p}(x)$ of default for each individual corresponds as closely as possible to the individual's observed default status.

- Likelihood function: Given data $\{(y_i, x_i)\}_{i=1}^n$, by model assumption, $\Pr(y_i = 1 \mid x_i) = p(x_i)$, $\Pr(y_i = 0 \mid x_i) = 1 - p(x_i)$. The conditional "density function" of y_i given x_i is

$$p(x_i)^{y_i}\{1 - p(x_i)\}^{1-y_i}.$$

The joint conditional "density" of $y_1, \ldots, y_n$ given $x_1, \ldots, x_n$ is

$$\underbrace{L_n(\beta_0, \beta_1)}_{\text{likelihood function}} = \prod_{i=1}^{n} p(x_i)^{y_i}\{1 - p(x_i)\}^{1-y_i}.$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to *maximize* this likelihood function, a function of (β_0, β_1).

The mathematical properties of maximum likelihood are discussed in MATH 181A.

We will discuss the computation of $(\hat{\beta}_0, \hat{\beta}_1)$ later. In general, logistic regression can be fit using a statistical software package, such as R function glm().

The following table shows the coefficient estimates and related information that result from fitting a logistic regression model on the Default data in order to predict the probability of default=Yes 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

The z-statistic here plays the same role as the t-statistic. For example, the z-statistic associated with β_1 is $\hat{\beta}_1/\mathrm{se}(\hat{\beta}_1)$. Large (absolute) value of the z-statistic indicates evidence against the null hypothesis $H_0 : \beta_1 = 0$.

The null hypothesis $H_0 : \beta_1 = 0$ implies

$$p(X) = \Pr(Y = 1 \,|\, X) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}.$$

In other words, the probability of default does not depend on balance. Since the p-value associated with balance is tiny, we can reject H_0.

We conclude that there is indeed an association between balance and probability of default.

Making Predictions

- Once the coefficients have been estimated, it is a simple matter to compute the probability of default for any given credit card balance.

- Using the coefficient estimates, we predict that the default probability for an individual with a balance of \$1000 is

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

  which is below 1%.

- The predicted probability of default for an individual with a balance of \$2000 is much higher, approximately 58.6%.

- Alternatively, one may predict the probability of default from student status. To see this, create a dummy variable that takes on a value of 1 for students and 0 for non-students.

	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

The coefficient for the dummy variable is positive, and the associated p-value is statistically significant. This indicates that students tend to have higher default probabilities than non-students.

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

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

Multiple Logistic Regression

Goal: predicting a binary response using multiple predictors.

Generalize the previous models as follows:

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

where $X = (X_1, \ldots, X_p)^\intercal$ is p-vector of *covariates*. Equivalently,

$$p(X) = \frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}.$$

Let $\beta = (\beta_1, \ldots, \beta_p)^\intercal$ be the vector of *regression coefficients*;

β_0 is referred to as the *intercept*.

Define the *logistic function* as

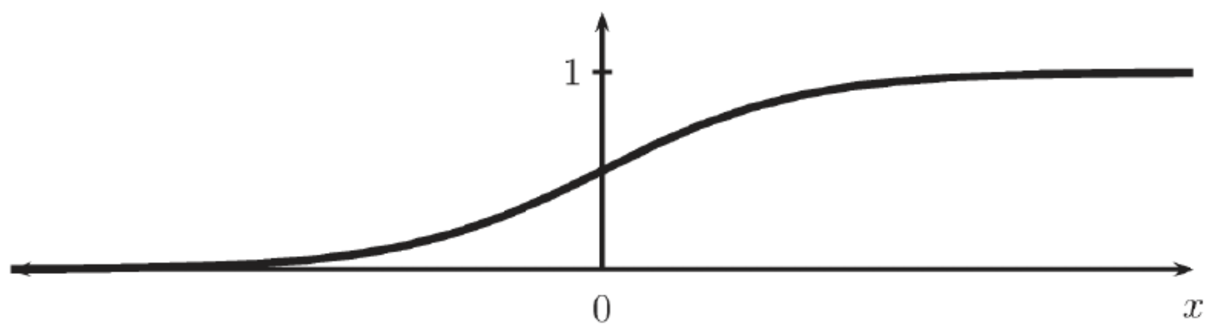
$$\phi(t) = \frac{1}{1+e^{-t}}, \ \ t \in \mathbb{R}.$$

Then, the *logistic regression model* for $(Y, X) \in \{0,1\} \times \mathbb{R}^p$ is

$$p(X) = \Pr(Y = 1 \,|\, X) = \phi(\beta_0 + X^\intercal \beta).$$

On the other hand,

$$\Pr(Y = 0 \,|\, X) = 1 - \phi(\beta_0 + X^\intercal \beta).$$



- Data: $(Y_1, X_1), \ldots, (Y_n, X_n)$ are independent from (Y, X).

- Likelihood function: The conditional density of Y_i given $X_i = (X_{i1}, \ldots, X_{ip})^\intercal$ is

$$\phi(\beta_0 + X_i^\intercal \beta)^{Y_i}\{1 - \phi(\beta_0 + X_i^\intercal \beta)\}^{1-Y_i}.$$

Hence, the joint density of $(Y_1, \ldots, Y_n)$ given $X_1, \ldots, X_n$ is

$$L_n(\beta_0, \beta) = \prod_{i=1}^{n} \phi(\beta_0 + X_i^\intercal \beta)^{Y_i}\{1 - \phi(\beta_0 + X_i^\intercal \beta)\}^{1-Y_i}.$$

This $L_n : \mathbb{R} \times \mathbb{R}^p \to [0,\infty)$ is called the *likelihood function*.

Maximum Likelihood Estimator

- The *maximum likelihood estimator* $(\hat{\beta}_0, \hat{\beta})$ is defined as the maximizer of the likelihood function:

$$(\hat{\beta}_0, \hat{\beta}) \in \arg \max_{(\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^p} L_n(\beta_0, \beta).$$

- Log-likelihood: In both theory and practice, it is easier to deal with the logarithm of likelihood function

$$\ell_n(\beta_0, \beta) = \log L_n(\beta_0, \beta)$$

$$= \sum_{i=1}^{n} Y_i \log \phi(\beta_0 + X_i^\intercal \beta) + (1 - Y_i) \log\{1 - \phi(\beta_0 + X_i^\intercal \beta)\} .$$

Default Data

- Back to default data, we take X_1: balance, X_2: income and $X_3 \in \{0,1\}$: student status.

	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

- Prediction: A student with a credit card balance of $\$1500$ and an income of $\$40{,}000$ has an estimated probability of default of

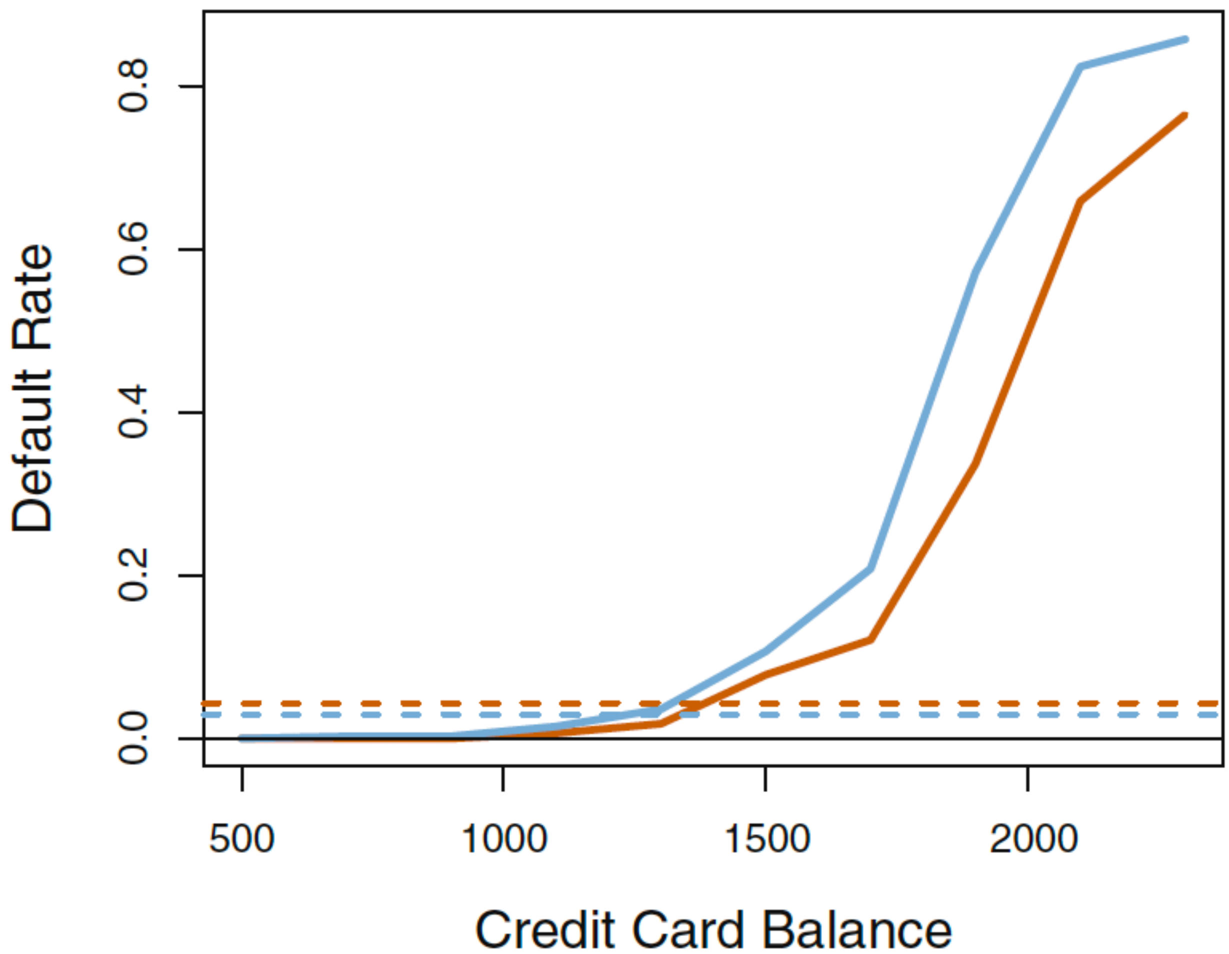
$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}} = 0.058.$$

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

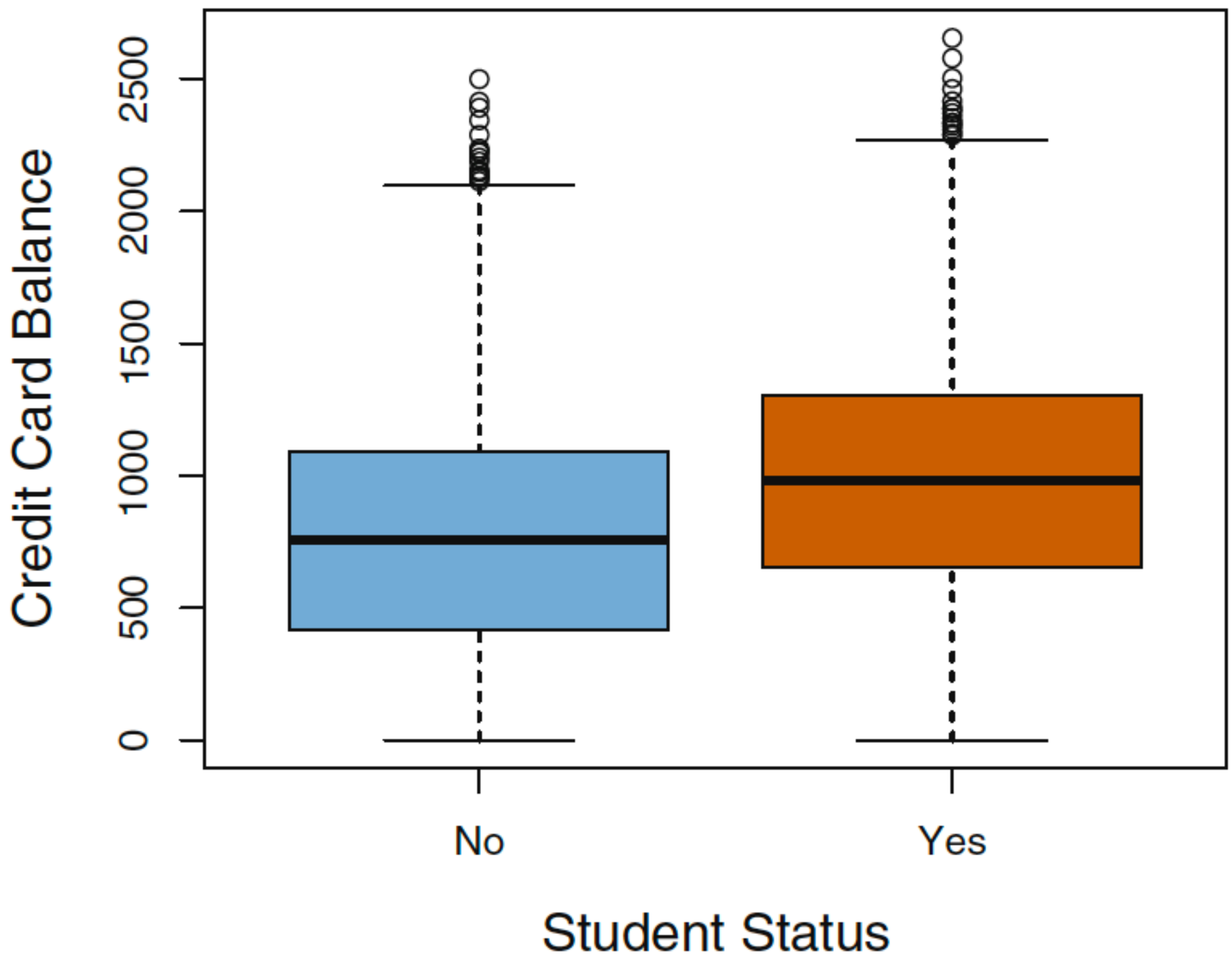
$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}} = 0.105.$$

Here we multiply the income coefficient estimate from Table by 40, rather than by 40,000, because in that table the model was fit with income measured in units of $\$1000$.

The two-class logistic regression models have multiple-class extensions (*multiple-class logistic regression*), but in practice they tend not to be used all that often. The software for it is available in R.



Default rates for students (orange) and non-students (blue)



Boxplots of balance for students (orange) and non-students (blue)