CLASSIFICATION METHODS

Chapter 04 (part 01)

LOGISTIC REGRESSION

Outline

- Cases:
  - Orange Juice Brand Preference
  - Credit Card Default Data
- Why Not Linear Regression?
- Simple Logistic Regression
  - Logistic Function
  - Interpreting the coefficients
  - Making Predictions
  - Adding Qualitative Predictors
- Multiple Logistic Regression

Case 1: Brand Preference for Orange Juice

- We would like to predict what customers prefer to buy: Citrus Hill or Minute Maid orange juice?

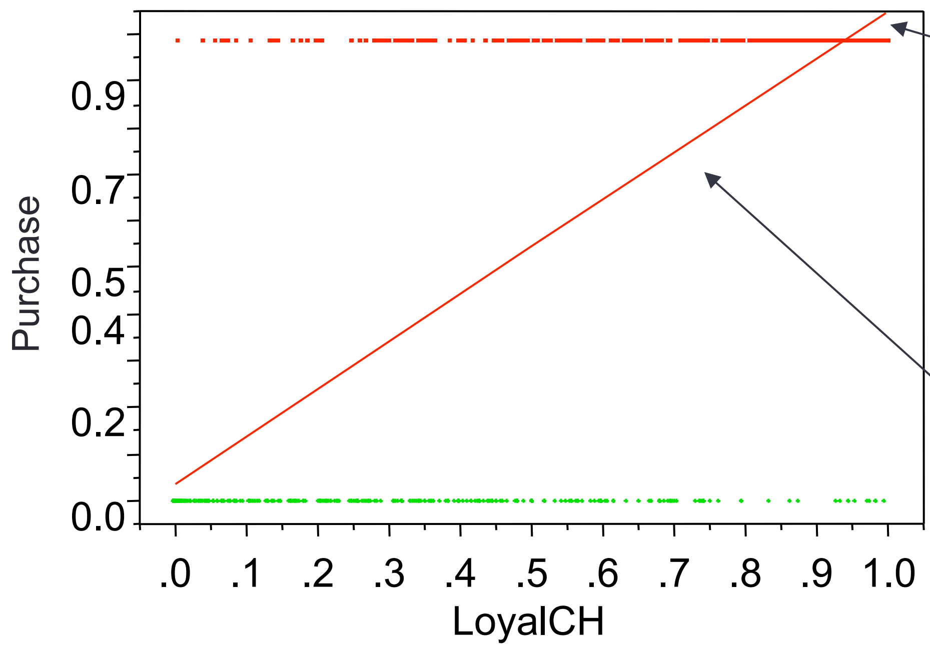
- The Y (Purchase) variable is <u>categorical</u>: 0 or 1

- The X (LoyalCH) variable is a numerical value (between 0 and 1) which specifies the how much the customers are loyal to the Citrus Hill (CH) orange juice

- Can we use Linear Regression when Y is categorical?

Why not Linear Regression?

- When Y only takes on values of 0 and 1, why standard linear regression in inappropriate?



How do we interpret values greater than 1?

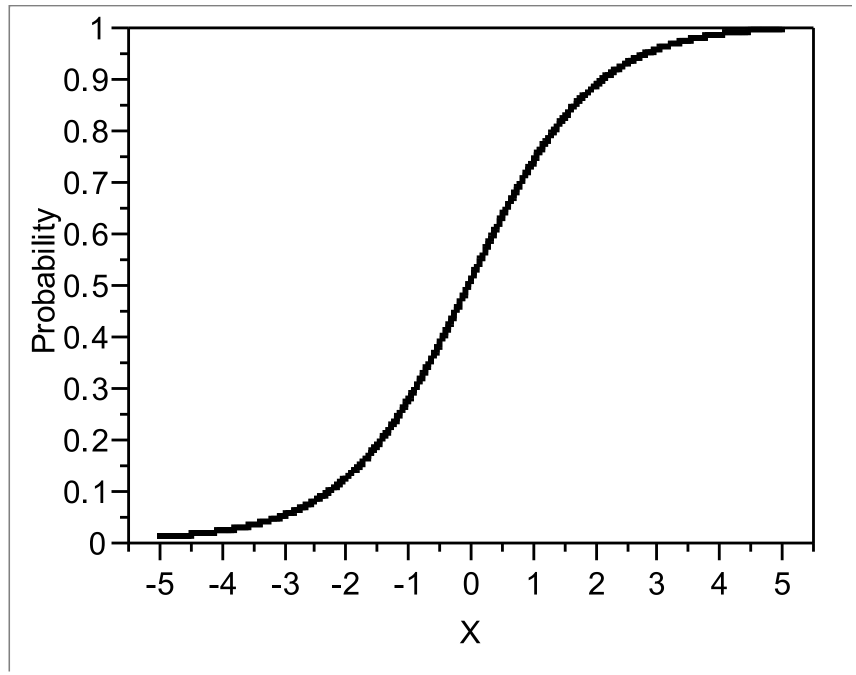
How do we interpret values of Y between 0 and 1?

Problems

- The regression line $\beta_0+\beta_1 X$ can take on any value between negative and positive infinity
- In the orange juice classification problem, Y can only take on two possible values: 0 or 1.
- Therefore the regression line almost always predicts the wrong value for Y in classification problems

Solution: Use Logistic Function

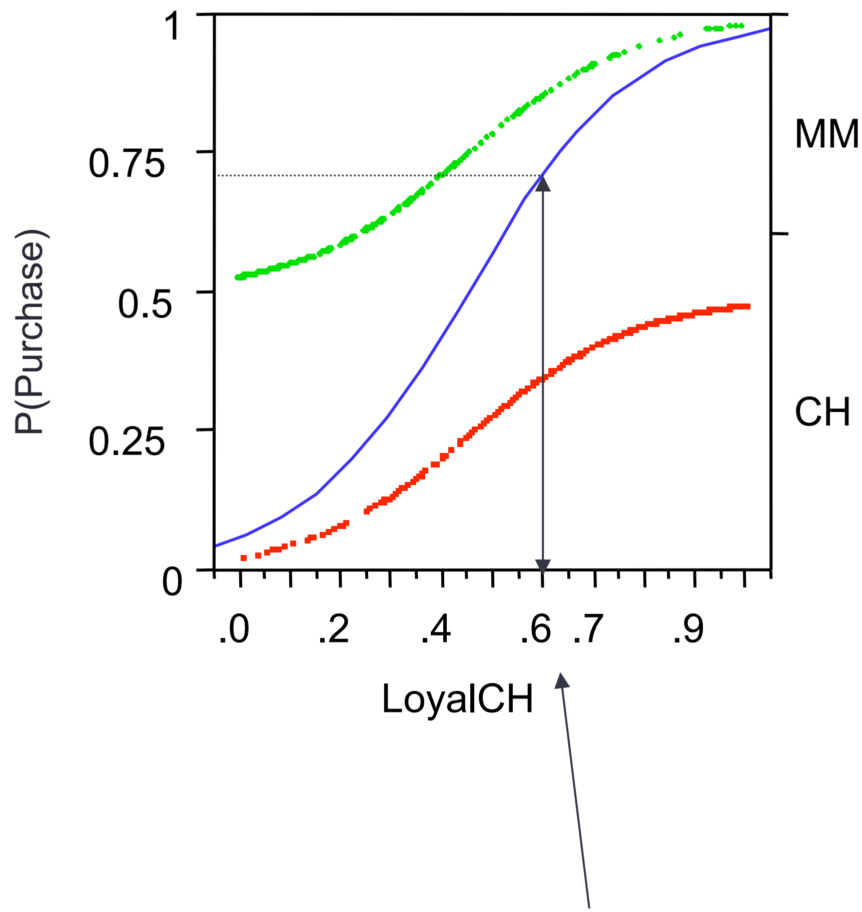
- Instead of trying to predict Y, let's try to predict P(Y = 1), i.e., the probability a customer buys Citrus Hill (CH) juice.
- Thus, we can model P(Y = 1) using a function that gives outputs between 0 and 1.
- We can use the logistic function
- Logistic Regression!



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

Logistic Regression

- Logistic regression is very similar to linear regression
- We come up with b_0 and b_1 to estimate β_0 and β_1.
- We have similar problems and questions as in linear regression
    - e.g. Is β_1 equal to 0? How sure are we about our guesses for β_0 and β_1?



If LoyalCH is about .6 then Pr(CH) ≈ .7.

Case 2: Credit Card Default Data

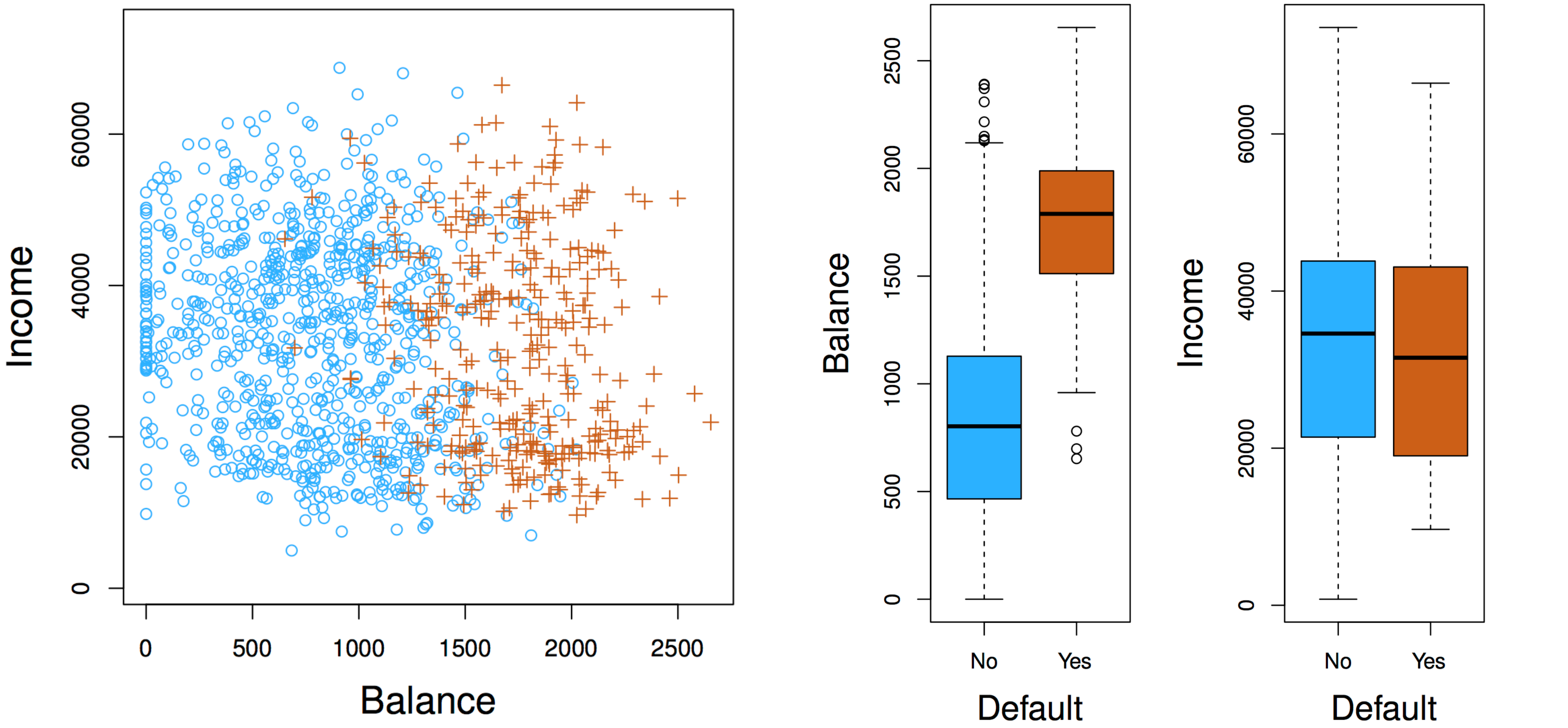
- We would like to be able to predict customers that are likely to default

- Possible X variables are:
  - Annual Income
  - Monthly credit card balance

- The Y variable (Default) is <u>categorical</u>: Yes or No

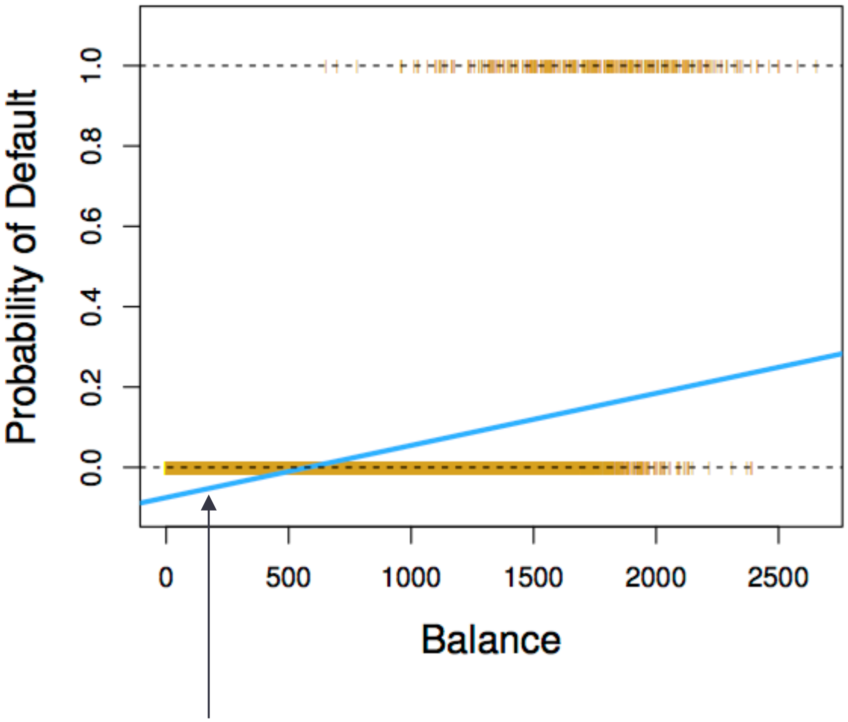
- How do we check the relationship between Y and X?

The Default Dataset



Why not Linear Regression?

- If we fit a linear regression to the Default data, then for very low balances we predict a negative probability, and for high balances we predict a probability above 1!

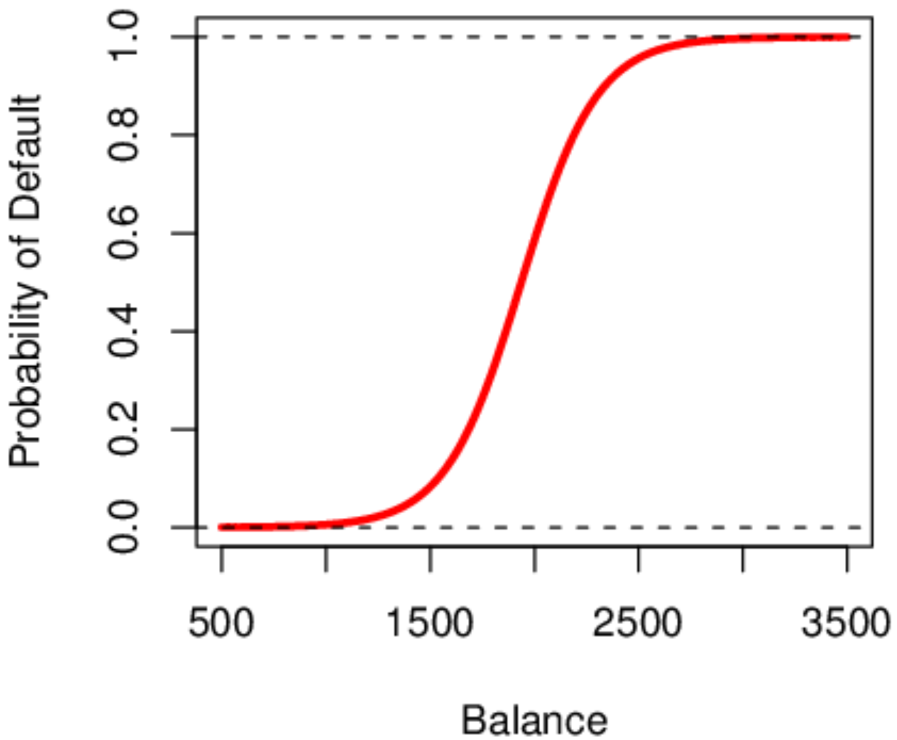


When Balance < 500,

Pr(default) is negative!

Logistic Function on Default Data

- Now the probability of default is close to, but not less than zero for low balances. And close to but not above 1 for high balances



Interpreting β_1

- Interpreting what β_1 means is not very easy with logistic regression, simply because we are predicting P(Y) and not Y.
- If $\beta_1 = 0$, this means that there is no relationship between Y and X.
- If $\beta_1 > 0$, this means that when X gets larger so does the probability that Y = 1.
- If $\beta_1 < 0$, this means that when X gets larger, the probability that Y = 1 gets smaller.
- But how much bigger or smaller depends on where we are on the slope

Are the coefficients significant?

- We still want to perform a hypothesis test to see whether we can be sure that are β_0 and β_1 significantly different from zero.
- We use a Z test instead of a T test, but of course that doesn't change the way we interpret the p-value
- Here the p-value for balance is very small, and b_1 is positive, so we are sure that if the balance increase, then the probability of default will increase as well.

	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

Making Prediction

- Suppose an individual has an average balance of $1000. What is their probability of default?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.00576$$

- The predicted probability of default for an individual with a balance of $1000 is less than 1%.

- For a balance of $2000, the probability is much higher, and equals to 0.586 (58.6%).

Qualitative Predictors in Logistic Regression

- We can predict if an individual default by checking if she is a student or not. Thus we can use a qualitative variable "Student" coded as (Student = 1, Non-student =0).
- b_1 is positive: This indicates students tend to have higher default probabilities than 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

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

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

Multiple Logistic Regression

- We can fit multiple logistic just like regular regression

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

Multiple Logistic Regression- Default Data

- Predict Default using:
  - Balance (quantitative)
  - Income (quantitative)
  - Student (qualitative)

	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

Predictions

- A student with a credit card balance of $1,500 and an income of $40,000 has an estimated probability of default

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 1}} = 0.058.$$

An Apparent Contradiction!

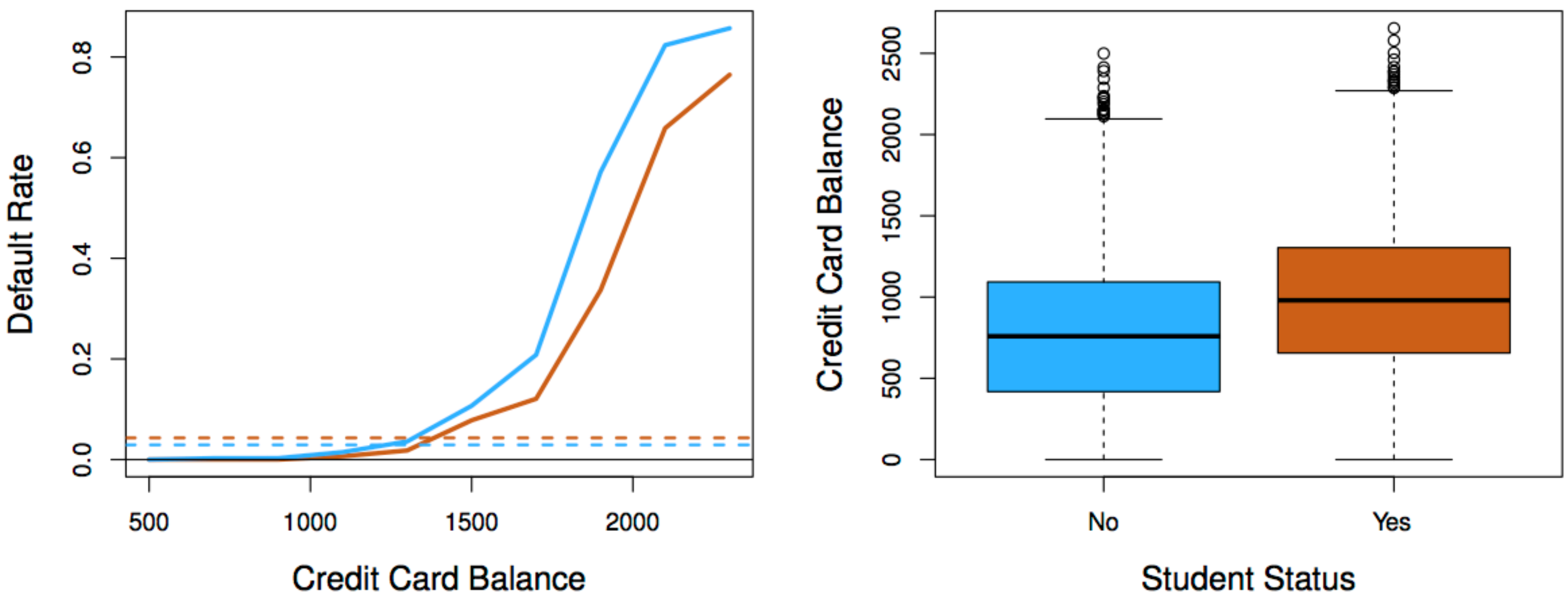
	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

Positive

	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

Negative

Students (Orange) vs. Non-students (Blue)



To whom should credit be offered?

- A student is risker than non students if no information about the credit card balance is available

- However, that student is less risky than a non student with the same credit card balance!