# CLASSIFICATION METHODS

Chapter 04 (part 01) – Logistic Regression

# Outline

- Classification problem examples
  - What’s wrong with using linear Regression?
- Simple Logistic Regression
  - Logistic Function
  - Interpreting the coefficients
  - Making Predictions
  - Adding Qualitative Predictors
- Multiple Logistic Regression

# Classification

- Recall that in the regression problem our goal is to estimate the (real) number based on observed features
  - Output Type = Cardinal

- Classification: Estimating the category (class) to which something belongs
- Classes often have no direct underlying numerical relationship but we might use numbers as output values
  - Output Type = Nominal
  - Example: Tank = 1, Non-Tank = 0

# OJ Classification Example

- Goal: predict what customers will buy:
  Citrus Hill orange juice or Minute Maid orange juice
  (based on their brand loyalty to various juice types)

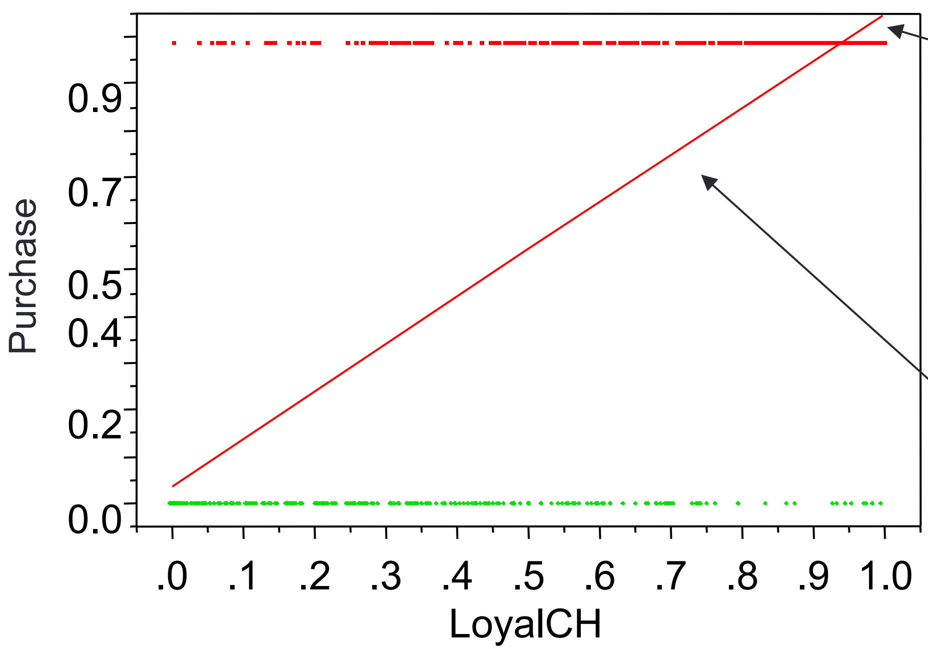
- Y (Purchase CH) is <u>categorical</u>: 0 (no) or 1 (yes)

- X (LoyalCH) <u>numerical</u> (between 0 and 1) which specifies how loyal customers are to the Citrus Hill (CH) brand (0 = not loyal… 1 = completely loyal)

- Could we use Linear Regression when Y is categorical?

# Why not use Linear Regression For category estimation?

- Regression forms a line…




How do we interpret values greater than 1?

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

# Problems with linear regression for Classification

- The regression line $\beta_0+\beta_1X$ can take on any value between negative and positive infinity
- In the orange juice classification problem, Y should only take on two possible values: 0 or 1.
- Therefore the regression line almost always predicts the wrong value for Y in classification problems

# Classification Example 2: Credit Card Default Data

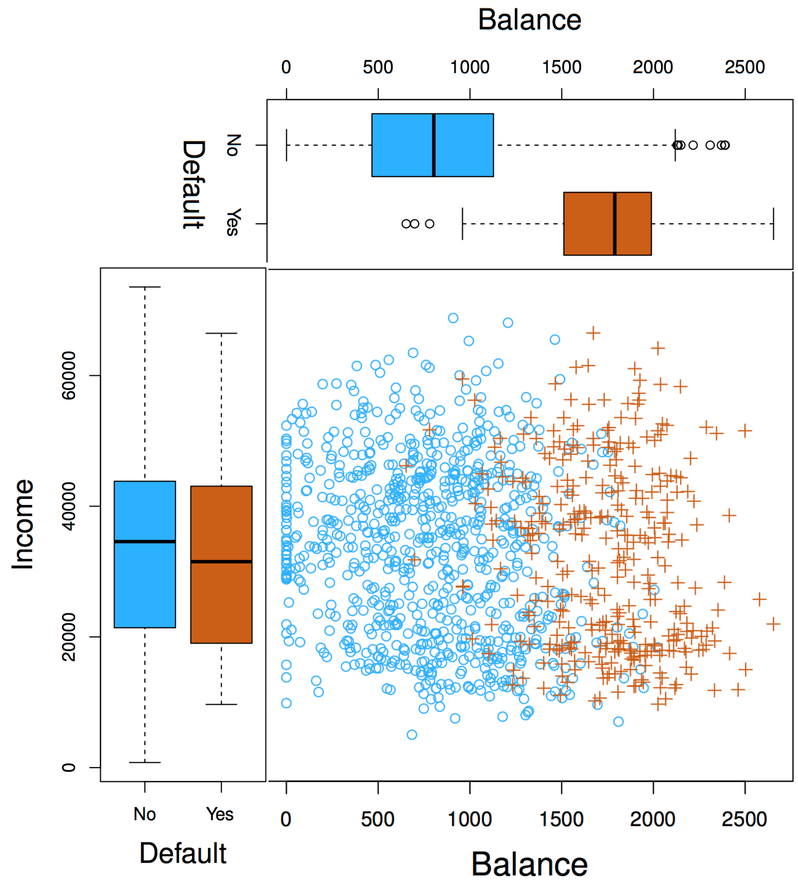
- We would like to be able to predict customers that are likely to default (not pay off their card)

- 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?

# Exploring the (credit card) Default Dataset



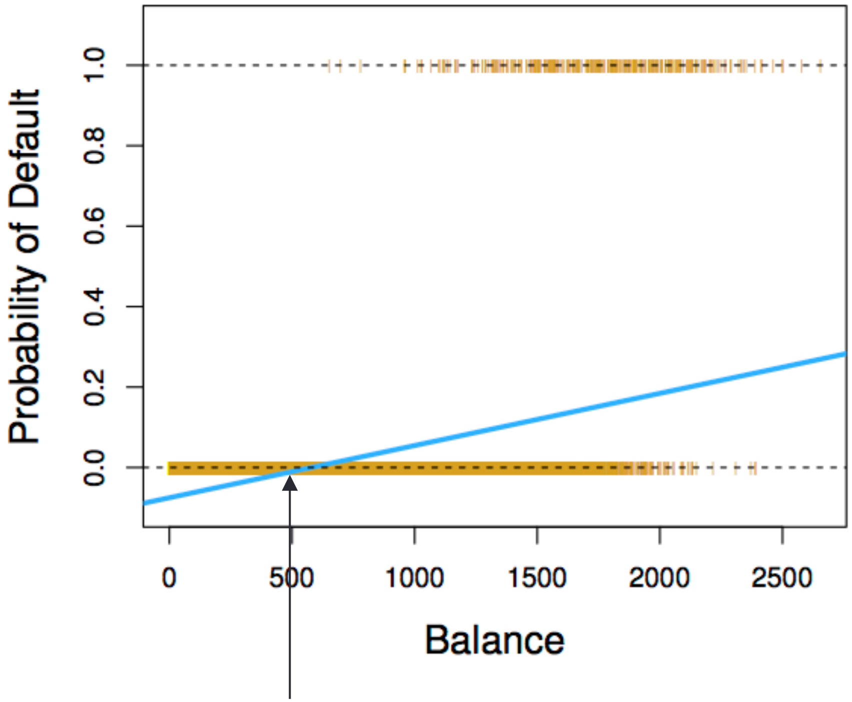

**Concept Check:**

**Is there a meaningful relationship between Balance and Defaulting?**

**Is there a meaningful relationship between Income and defaulting?**

# Why not Linear Regression?

- For very low balances we predict a negative probability
- For high balances we predict a probability above 1

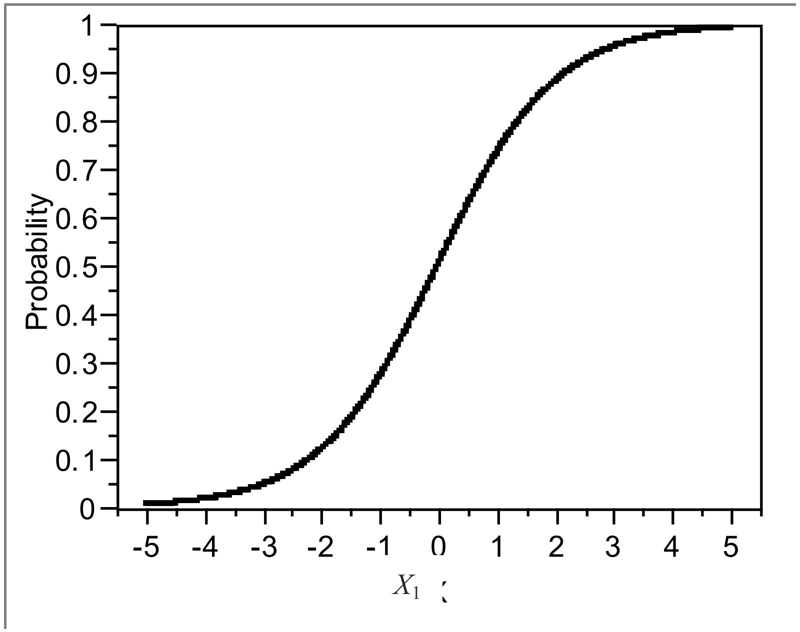



When Balance < 500, Pr(default) is negative!

# 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.
  - Model P(Y = 1) with a function that gives outputs between 0 and 1.
  - Determine the Boolean answer by thresholding p
- Logistic function: Logistic Regression

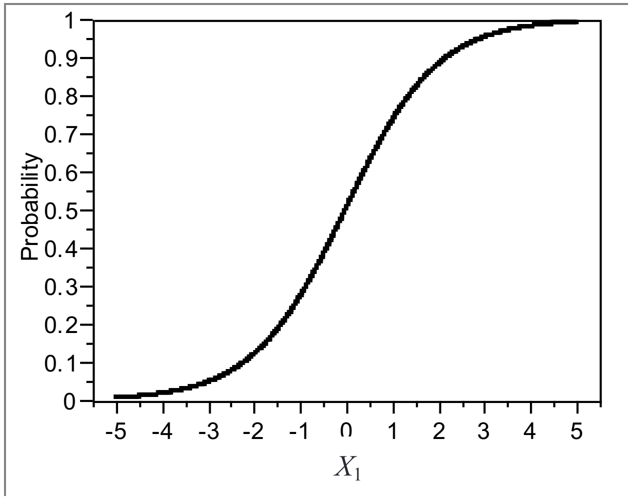
$$p = P(y=1) = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$$



# Logistic Function: Thinking & Coding Practice

- What do you think happens to the shape of the curve as you alter the size and sign of $\beta_0$? $\beta_1$?

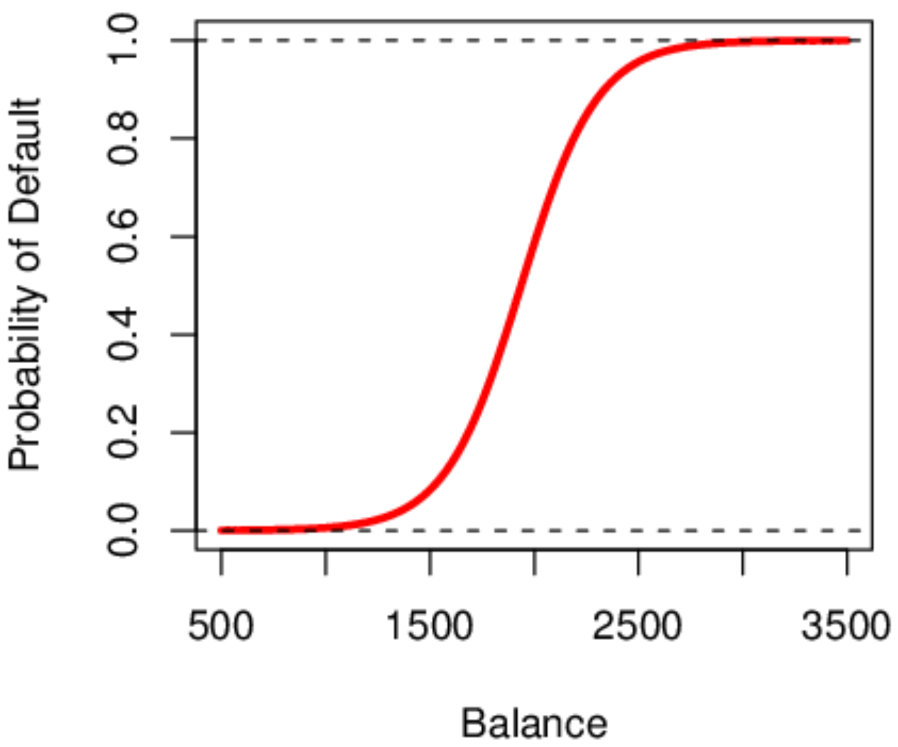
$$p = P(y=1) = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$$




- Write a function which accepts $\beta_0$, $\beta_1$, and X and returns P(Y=1)
  - $\beta_0$ and $\beta_1$ are scalars
  - X is a (n x 1) matrix.
  - P(Y=1) is a (n x 1) matrix
- Plot the results and see what happens when you alter the betas. Does it match your intuition?

# Logistic Function on Bank Default Data

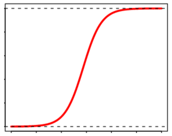
- The probability of default is close to, but not less than zero for low balances.
- … and close to, but not above 1 for high balances



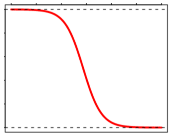
# Interpreting $\beta_1$

$$p = P(y=1) = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$$

- Interpreting what $\beta_1$ means is not very easy with logistic regression, simply because we are predicting P(Y) and not Y
- If $\beta_1$ =0, no relationship between Y and X
- If $\beta_1$ >0, when X gets larger Y approaches 1



- If $\beta_1$ <0, when X gets larger Y approaches 0



- But how much bigger or smaller depends on where we are on the slope
- **Concept Check:**
  - **How is the logistic line altered by changing $\beta_0$ ?**

# Logistic Regression Assessment: Are the coefficients significant?

- We still want to perform a hypothesis test to see whether we can be sure that are $\beta_0$ and $\beta_1$ significantly different from zero.
- We use a Z test instead of a T test (due to the process used to compute the coefficients), but that doesn't change the way we interpret the *p*-value
- Here the *p*-value for balance is very small, and $\hat{\beta}_1$ is positive, so we are sure that if the credit balance increases, 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 | $< 0.0001$ |

# Logistic Regression
# Making Predictions

- 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%).

# Logistic Regression
# Encoding Qualitative Predictors

- We can predict if an individual will 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).
- $\hat{\beta}_1$ 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 Feature Logistic Regression

- We can fit multiple logistic regression coefficients

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

# Multiple Feature Logistic Regression Credit Card 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 |

# Making Predictions with multiple-feature Logistic Regression

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

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

# Interpreting multiple-feature Logistic Regression Explain what happened here…

- The sign of the student coefficient changes when adding more features – Why?

| | 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 |

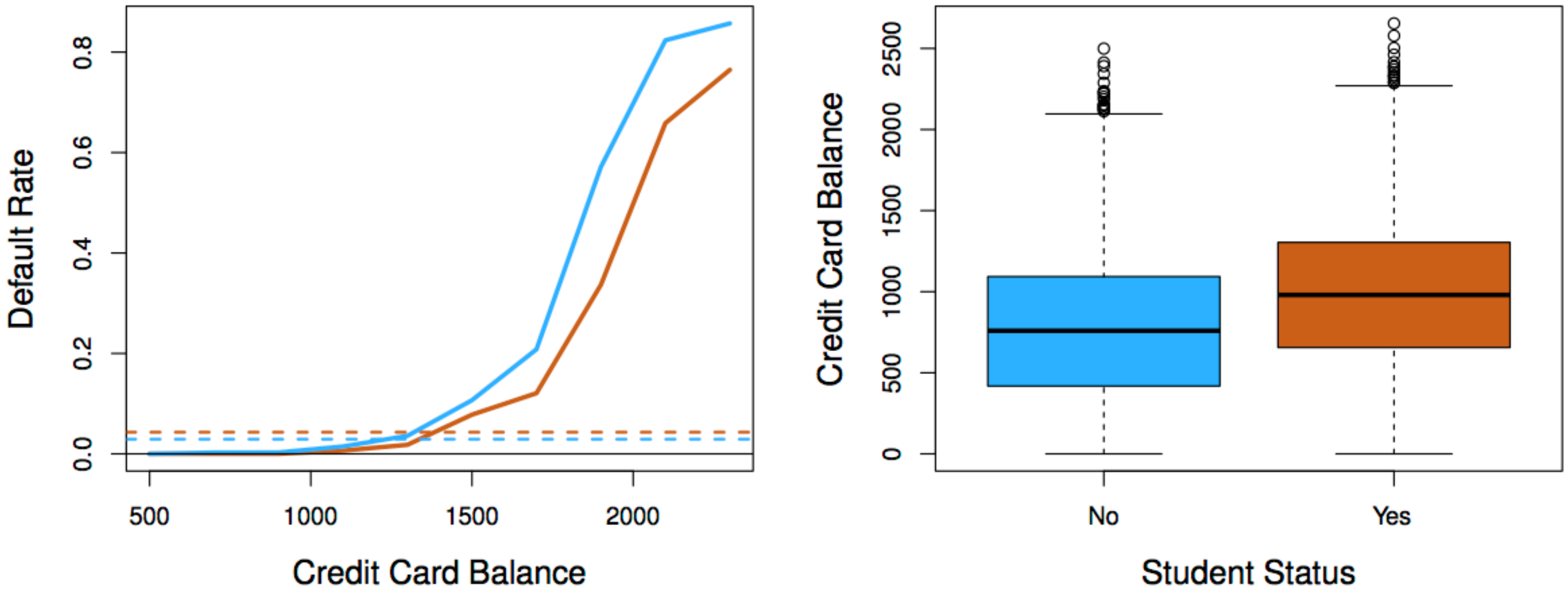
Positive

| | 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

# Interpreting multiple-feature Logistic Regression
# To whom should credit be offered?

- A student (orange) is risker than non students (blue) *if no information about the credit card balance is available*




- However, for two individuals *with the same credit card balance*, *the student is less risky* than a non student