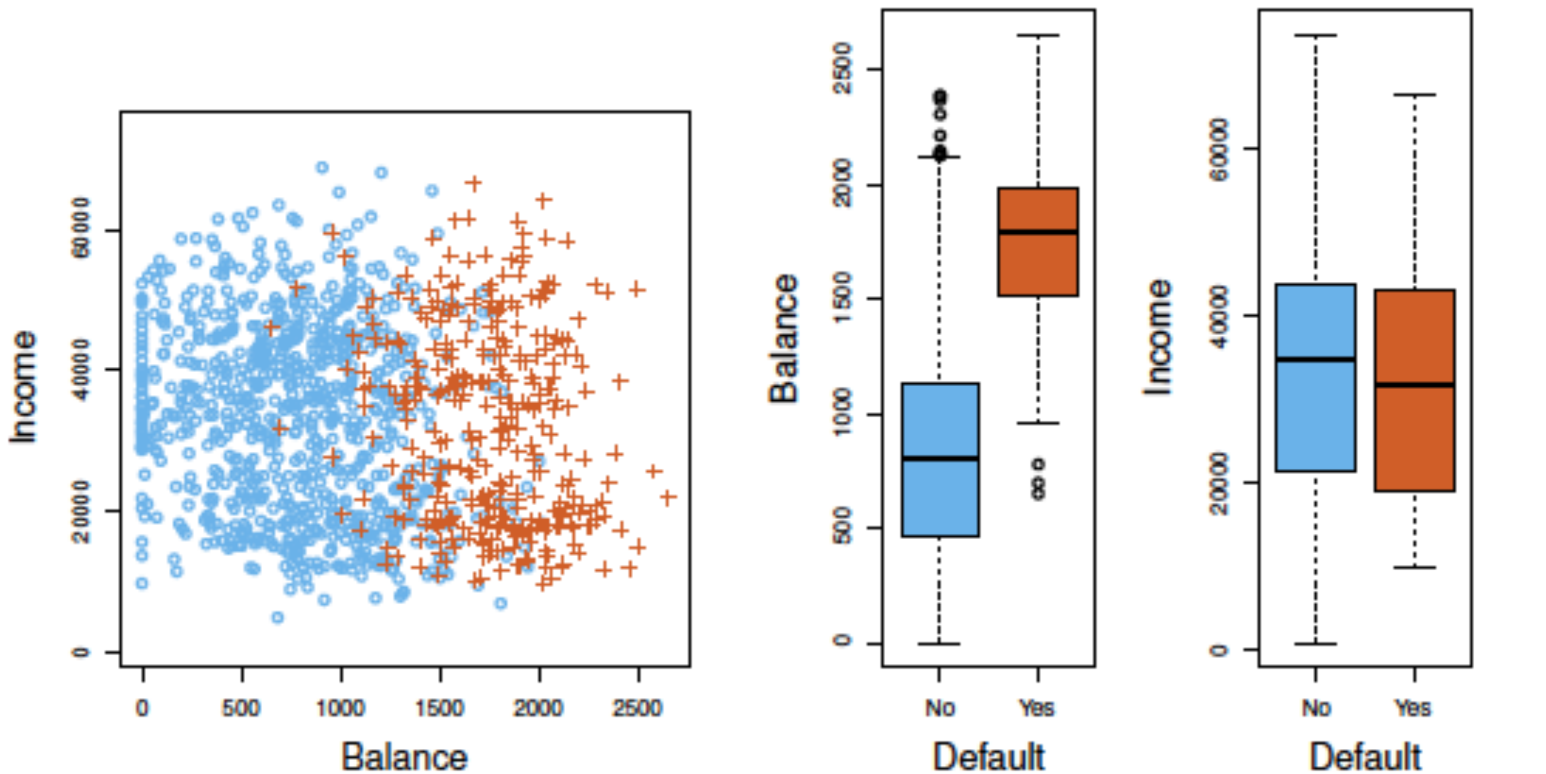
Lecture 8 – Logistic Regression Part 1

Instructor: Hamed Hasheminia

Agenda

- Intro to Logistic Regression
- Odds and Log of Odds
- Using Logistic Regression to make predictions
- How to interpret results of a Logistic Regression model
- How to interpret coefficients of a Logistic Regression Model
- Strengths and weaknesses of Logistic Regression Models

Credit Data



Income
Balance
0
500
1000
1500
2000
2500
20000
40000
60000
Default
No
Yes

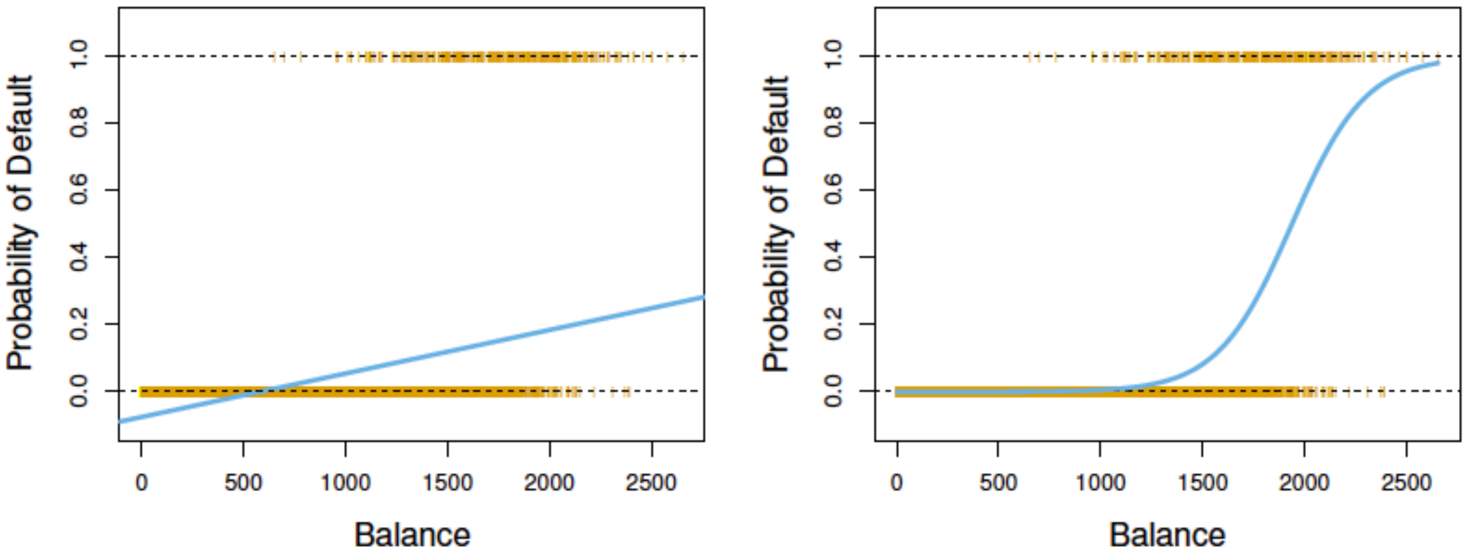
Can we use Linear Regression?

- Suppose that the Default classification task that we code

$$Y = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Yes.} \end{cases}$$

- Can we simply preform a linear regression of Y on X and classify as Yes if Y_hat > 0.5?

Credit data



The orange marks indicate the response Y, either 0 or 1. Linear regression does not estimate Pr(Y = 1|X) well. Logistic regression – the right hand side graph - seems well suited to the task.

Logistic Regression

- Let's write p(x) = Pr(Y =1|X) for short and consider using balance to predict default. Logistic regression uses the form:

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

- (e = 2.51828.. is a mathematical constant [Euler's number.])
- It is easy to see that no matter what values beta_0, beta_1 or X take, P(x) will have values between 0 and 1.

Logistic Regression

- A bit of rearrangement gives

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

- This monotone transformation is called the log odds or logit transformation of p(X).

Maximum Likelihood Estimation

- Most statistical packages can fit linear logistic regression models by maximum likelihood.

	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

- What is our estimated probability of default for someone with a balance of $1000?
- How about with a balance of $2000?

Making predictions

- Balance = \$1000

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

- Balance = \$2000

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586$$

What if we have categorical inputs?

- Let's do it again, using student as the predictor.

	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

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

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

Logistic regression with several variables

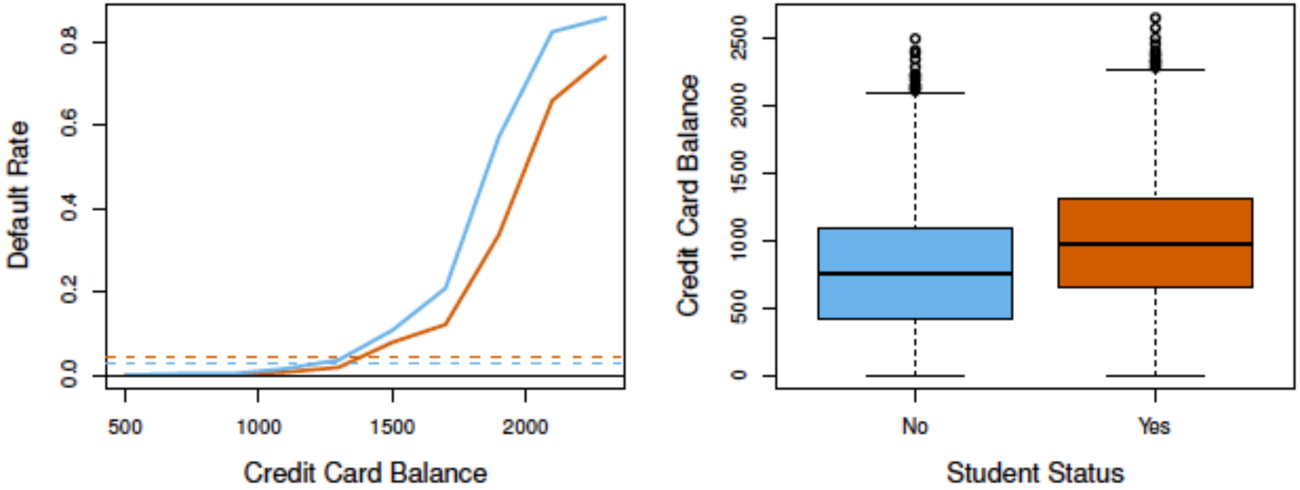
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

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

	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

Why is coefficient for `student` negative, while it was positive before?

Confounding



- Students tend to have higher balances than non-students, so their marginal default rate is higher than for non-students.
- But for each level of balance, students default less than non-students

How do we interpret coefficients?

- One can explain beta_i as an approximation for percentage change in odds for a unit change of X_i keeping all other variables constant.
- Slightly more accurate way of measuring the precentage change of odds for a unit change of X_i would be to use
  - exp(beta_i) – 1
- For small values of beta_i, these two measures are very close.

Let's interpret coefficients of Logistic Regression

	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

	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

A few notes about Logistic regression lines

- For categorical data, these models are among the most interpretable models.
- If you have more than 2 categories, then you shall use another class of logistic regression models called multinomial regression line. (you can use such models for prediction but model's interpretability decreases.)
- If we you have well-separated classes, then logistic regression model give you unstable results
- You do not need to standardize your inputs to train a logistic regression model
- Almost every concept discussed for linear regression lines can be used in logistic regression lines – i.e. (p-values, interaction effects, etc)

Summary

- We discussed the use of Logistic Regression
- The definition of odds were introduced
- We used Logistic Regression to make predictions
- We learned hot to interpret the results of a Logistic Regression model
- Lastly we discussed strengths and weaknesses of Logistic Regression Models