WEEK 10

Logistic Regression Model

Lecture outline

- Odds Ratio
- Simple Logistic (Logit) Regression
- Multiple Logistic (Logit) Regression

- ## Categorical Response Variables

Whether or not a person passes the exam

$$Y = \begin{cases} Pass \\ Fail \end{cases}$$

Success of a medical treatment

$$Y = \begin{cases} Survives \\ Dies \end{cases}$$

Whether borrower pays back or defaults on the loan

$$Y = \begin{cases} Pays\ back \\ Defaults \end{cases}$$

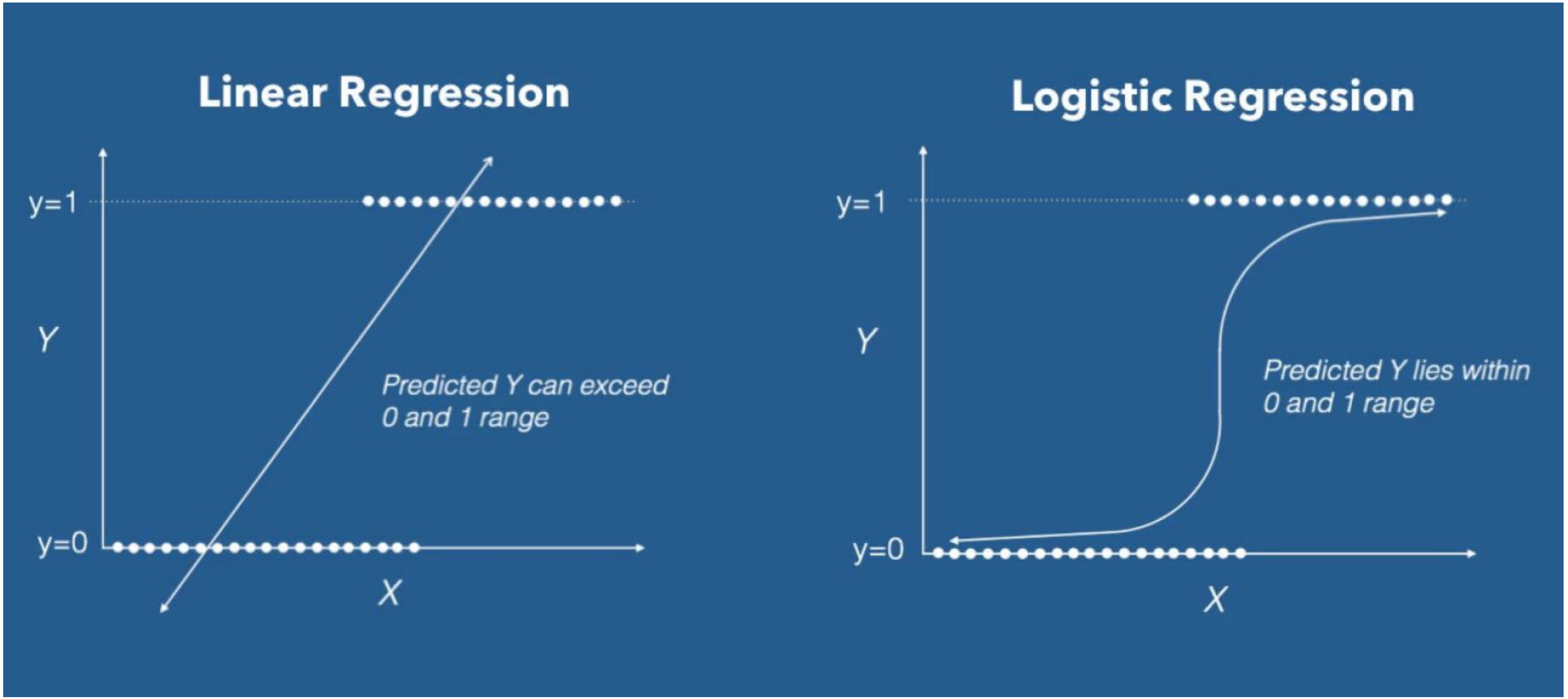
Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

If the OLS (ordinary least squares) model is used when there is a binary dependent variable, the following problems might arise:

- The error terms are heteroskedastic.
- e is not normally distributed because Y takes on only two values;
- The predicted probabilities can be greater than 1 or less than 0.

Linear vs Logit Model



Source: https://www.datacamp.com/

Simple Logistic Regression Model

- Logistic regression works with **odds ratio**, the ratio of "success" (1) to "failure" (0).

$$\text{odds} = \frac{\pi}{1-\pi} = \frac{probability\ of\ success}{probability\ of\ failure}$$

Logit form

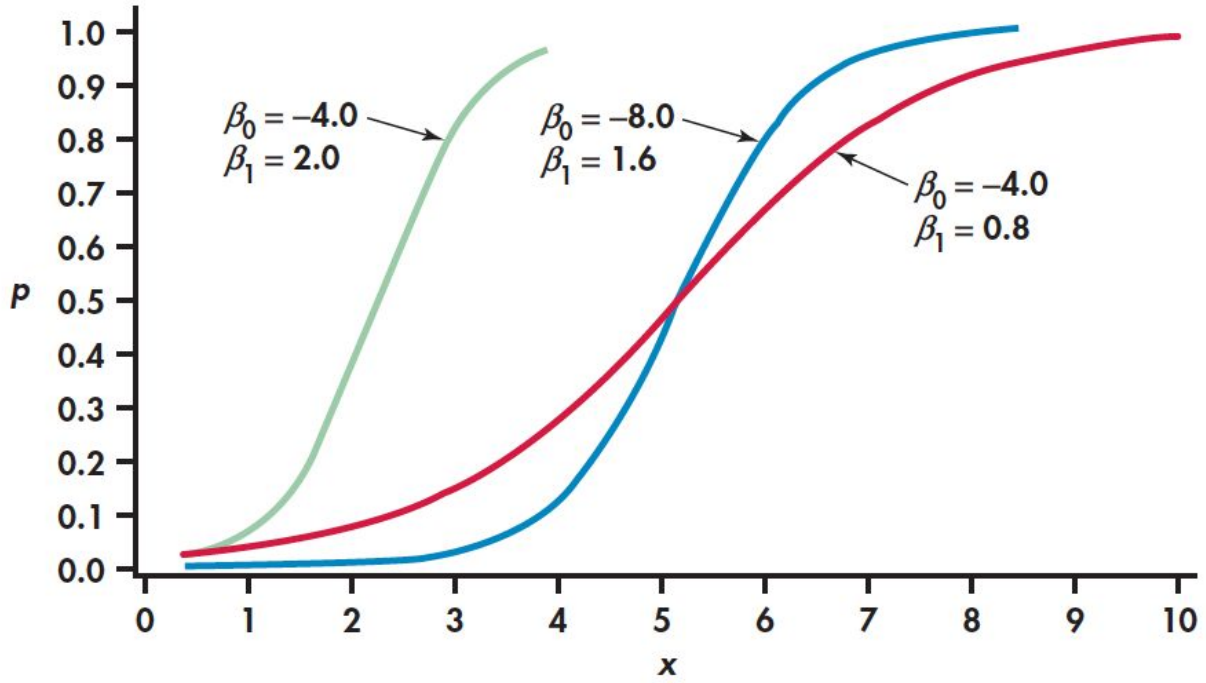
$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

Note: This is a natural log (aka "ln")

Probability form

$$\pi = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Figure 10.2 Plot of p (estimate of π) versus x for selected values of β_0 and β_1.



Estimation of coefficients

Maximum Likelihood Estimation (MLE) is a statistical method for estimating the coefficients of a logistic model.

Likelihood function:

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

The estimates b_0 and b_1 are chosen to maximize this likelihood function.

Learn more about MLE:

https://medium.com/codex/logistic-regression-and-maximum-likelihood-estimation-function-5d8d998245f9

Example: Titanic Dataset

- Let's use Titanic dataset and predict the probability of survival based on Fare of the passengers.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.931730   0.095215  -9.786  < 2e-16 ***
Fare         0.015084   0.002228   6.771 1.28e-11 ***
```

There is a positive relationship with the odds of survival and passenger fares.

£1 increase in fare is associated with an increase in the log odds $[\ln(\frac{p}{1-p})]$ of survival by 0.015 units.

R Code:

```
titanic <- read.csv("titanic.csv")
logit.m <- glm(Survived ~ Fare, data = titanic, family = "binomial")
summary(logit.m)
```

Fitted simple logit model

$$Ln\left(\frac{p}{1-p}\right) \text{ or } Log\left(\frac{p}{1-p}\right) = -0.9317 + 0.0151 * Fare$$

Question: What is the probability to survive if the passenger paid £300 for the trip? P(Y = survived | x = 300) = ?

Solution:

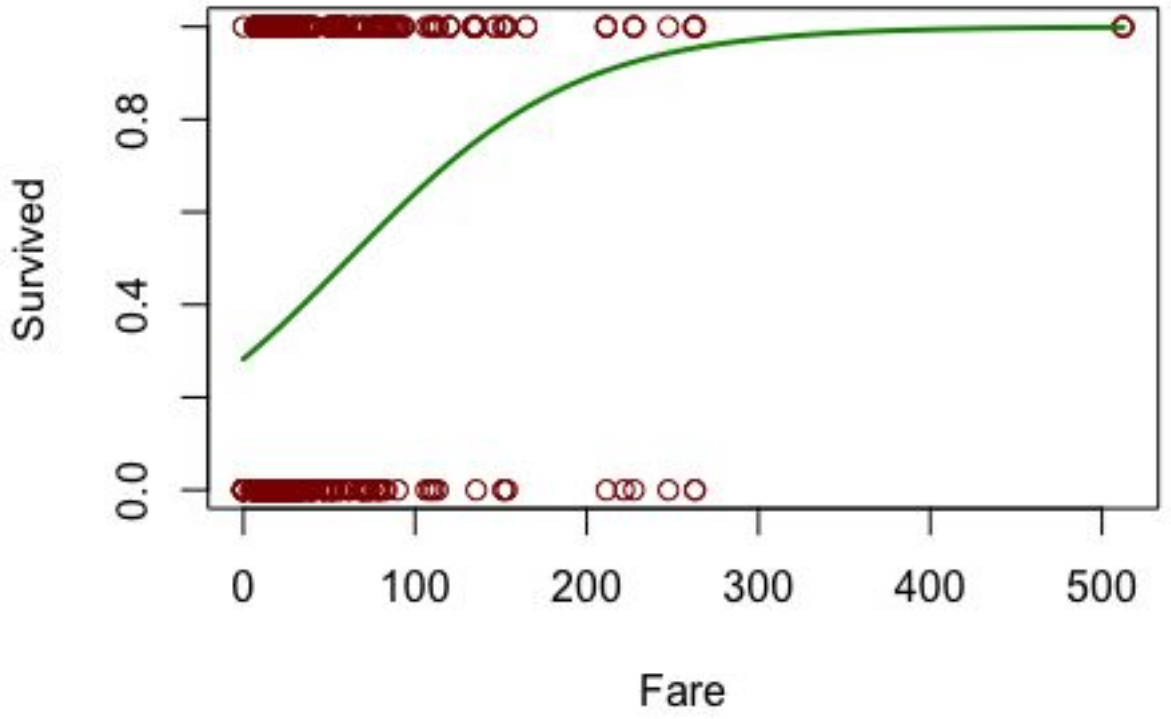
$$ln\left(\frac{p}{1-p}\right) = -0.9317 + 0.0151 * 300$$

$$\frac{p}{1-p} = e^{-0.9317 + 0.0151*300}$$

$$p = \frac{e^{-0.9317+0.0151*300}}{1+e^{-0.9317+0.0151*300}}$$

P(Y = 1| x = 300) = 0.9734

There is about 97% chance to survive if the fare amount paid is £300.



Multiple Logistic Regression

General Logit Model: $\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 \ldots + \beta_k * x_k + \varepsilon$

Fitted (estimated model): $\log(\frac{p}{1-p}) = b_0 + b_1 * x_1 + b_2 * x_2 \ldots + b_k * x_k$

$$p = \frac{e^{(b_0 + b_1 * x_1 + b_2 * x_2 \ldots + b_k * x_k)}}{1 + e^{(b_0 + b_1 * x_1 + b_2 * x_2 \ldots + b_k * x_k)}}$$

R Output:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.4137743  0.4786357   7.132 9.87e-13 ***
Pclass2     -1.0557724  0.3303871  -3.196 0.001396 **
Pclass3     -2.3438225  0.3337248  -7.023 2.17e-12 ***
Sexmale     -2.5773356  0.2100574 -12.270  < 2e-16 ***
Age         -0.0284005  0.0079198  -3.586 0.000336 ***
Fare        -0.0009845  0.0024805  -0.397 0.691433
```

$$\ln\left(\frac{p}{1-p}\right) = 3.414 - 1.056 * Pclass2 - 2.344*\text{Pclass3} - 2.577*\text{Male} - 0.028*\text{Age} - 0.001*\text{Fare}$$

Interpretations: If the passenger is female, ceteris paribus, the odds of surviving is $e^{2.577} = 13.16$ times higher than male passengers.

If the passenger is in First Class, ceteris paribus, the odds of surviving is $e^{1.056} = 2.87$ times higher than those who are in second class and $e^{2.344} = 10.42$ times higher than those who are in third class .

Exercises

The Global Findex database is the world's most comprehensive data set on how adults save, borrow, make payments, and manage risk. The data are collected in partnership with Gallup, Inc., through nationally representative surveys of more than 150,000 adults in over 140 economies. "findex_uzb.csv" provides some of the variables from Findex dataset of Uzbekistan for 2017. Responses are recorded as 1 if "yes", 0 if "no".

There are total of 1000 observations in the dataset.

a) Fit a simple logit model to predict whether the person saved (1) or not saved (0) based on age.

Comment on the significance of Age.

b) Predict the probability of saving at a financial institution for a person aged 45 years old.

c) Fit a multiple logit model to predict whether the person saved (1) or not saved (0) based on age, gender, education level, and employment status.

d) Compute the odds ratio and probability for a female, 30 years old, with secondary education, and employed.

Exercises

- Using ISLR package, load the “Default” dataset.

a) Check whether there is a problem of Multicollinearity.
b) Fit a logit model to predict the probability of default using remaining ones as independent variables. Provide interpretations about the coefficients.
c) Comment on the significance of independent variables. Use α = 0.01.
d) What is the probability of default for a student whose credit balance is $2000 and income of $8000? How about for a non-student?

References

1. James G. et al. An Introduction to Statistical Learning (with Applications in R). ISBN: 978-1-4614-7137-0. p.p. 156 – 160.

2. https://www.datacamp.com/community/tutorials/logistic-regression-R

3. https://machinelearningmastery.com/logistic-regression-for-machine-learning/

THE END

Thank you for your attention!