LOGISTIC REGRESSION

Presented by:
Bikram Pun, 191707
Arjun Bhandari, 191733

Table Of Contents

What is Logistic Regression?

- A statistical analysis used to examine relationships between
  - Independent *variables (predictors) and a dependant variable (criterion)*
- A linear model for classification and probability estimation.
- The main difference is in logistic regression
  - *the criterion is nominal (predicting group membership).*
- For example, do age and gender predict whether one signs up for swimming lessons (yes/no)

Types of Logistic Regression

- There are primarily 2 types of logistic regression:
  - *(1) Binary and (2) Multinomial models.*
- The difference lies in the types of the criterion variable
- Binary logistic regression is for a dichotomous criterion
  - *(i.e., 2-level variable)*
- Multinomial logistic regression is for a multicategorical criterion
  - *(i.e., a variable with more than 2 levels)*
- This set of slides focuses on <u>binary logistic regression</u>

Why use logistic regression?

- There are many important research topics for which the dependent variable is "limited."
- For example: voting, morbidity or mortality, and participation data
  - *is not continuous or distributed normally.*
- Binary logistic regression is a type of regression analysis
  - *where the dependent variable is a dummy variable:*
    - coded 0 (did not vote) or 1(did vote)

The Linear Probability Model

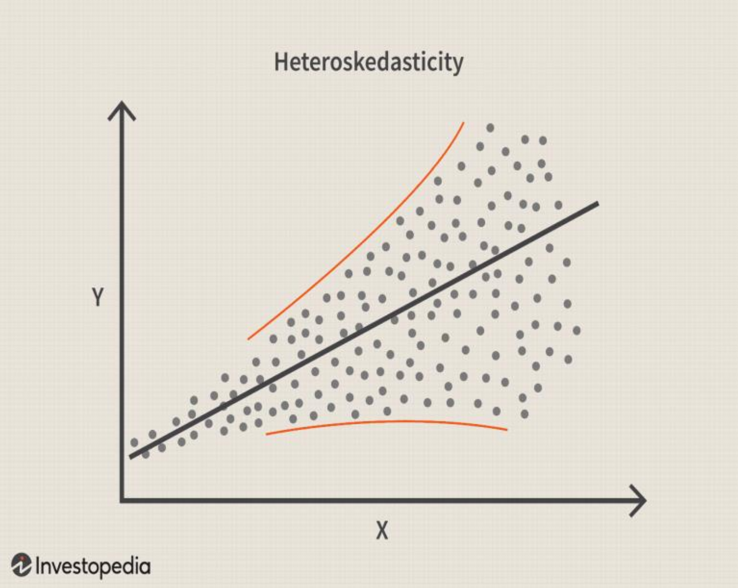
Consider the linear probability (LP) model:

$$Y = a + BX + e$$

where

- Y is a dummy dependent variable, =1 if event happens, =0 if event doesn't happen,
- a is the coefficient on the constant term,
- B is the coefficient(s) on the independent variable(s),
- X is the independent variable(s), and
- e is the error term.

Problems with using the LP model



- The error terms are heteroskedastic
  - *heteroskedasticity occurs when the variance of the dependent variable is different with different values of the independent variable*

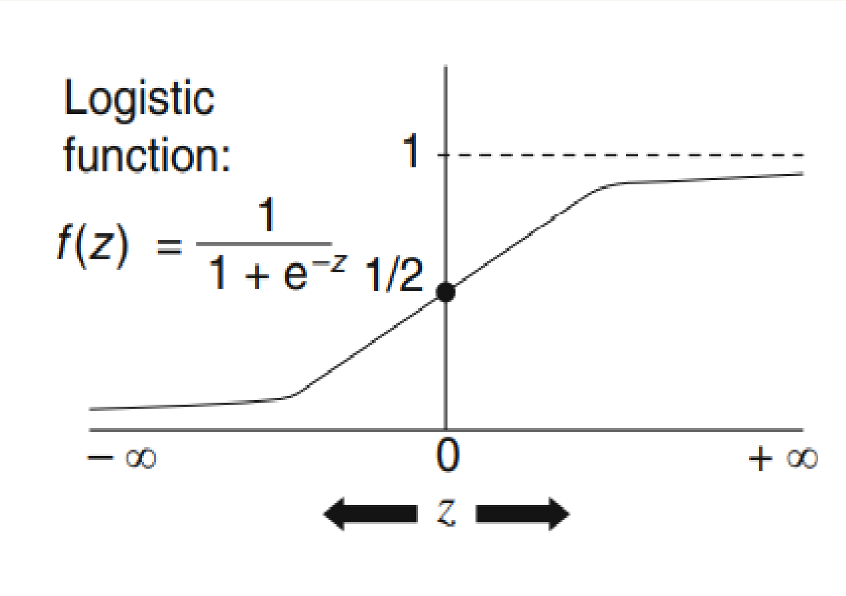
- e is not normally distributed because Y takes on only two values
  - violating another "classical regression assumption"
- The predicted probabilities can be greater than 1 or less than 0
  - which *can be a problem* if the predicted values are used in a subsequent analysis

The Logistic Regression Model

- The logit (logistic function) model solves these problems:

  $\ln[p/(1-p)] = \alpha + \beta X + e$

- p is the probability that the event Y occurs, p(Y=1)
- p/(1-p) is the "odds ratio"
- ln[p/(1-p)] is the log odds ratio, or "logit"



The Logistic Regression Model

More:

- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
- The estimated probability is:

$$p = 1/[1 + \exp(-\alpha - \beta X)]$$

- if you let $\alpha + \beta X = 0$, then $p = 0.50$
- as $\alpha + \beta X$ gets really big, p approaches 1
- as $\alpha + \beta X$ gets really small, p approaches 0

DEFINITION

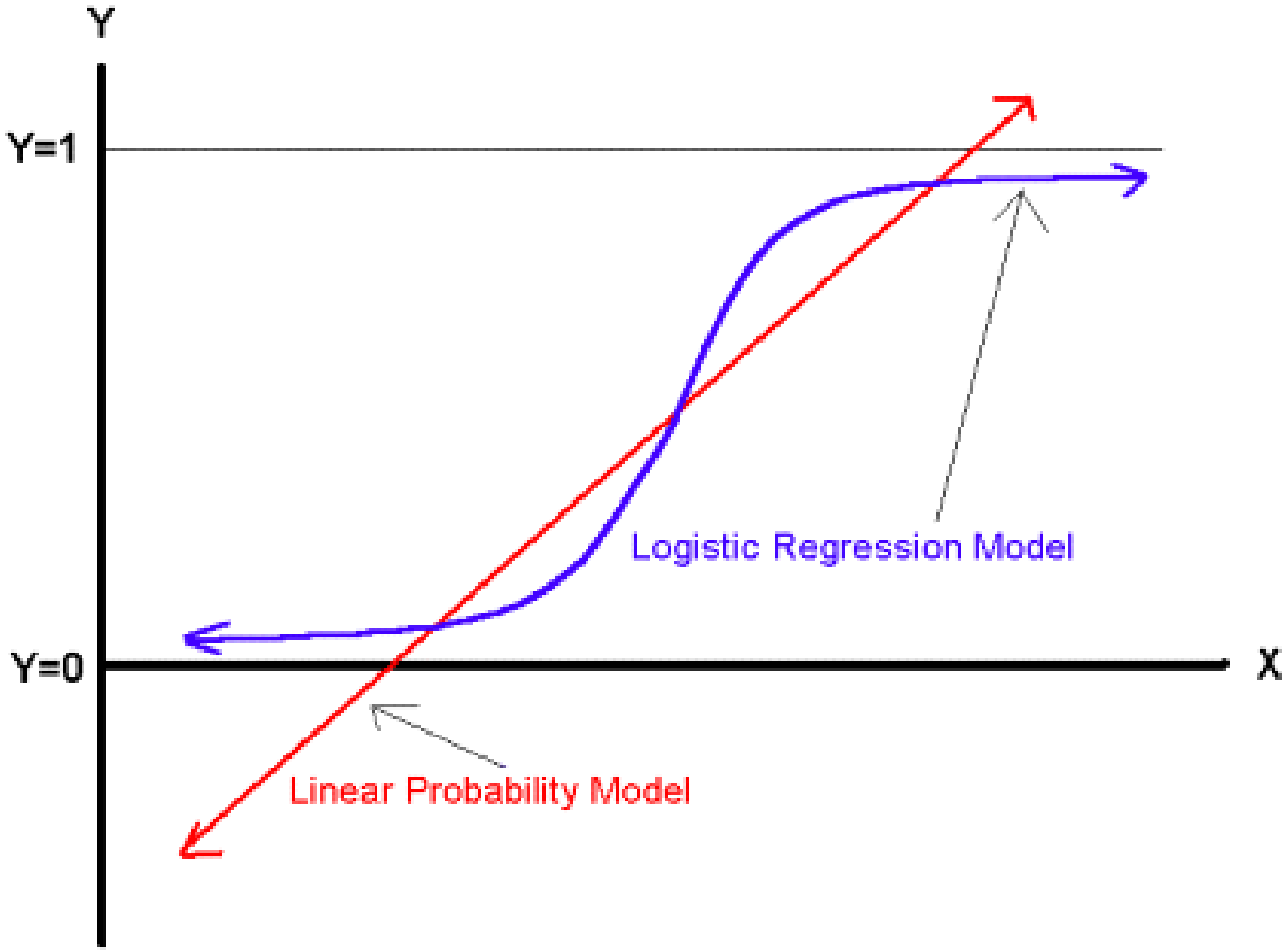
Logistic model:

$$P(D = 1 | X_1, X_2, \ldots, X_k) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}$$

$\uparrow \quad \uparrow$

unknown parameters

Comparing the LP and Logit Models



Maximum Likelihood Estimation (MLE)

- MLE is a statistical method for estimating the coefficients of a model.
- The likelihood function (L) measures
  - *the probability of observing the particular set of dependent variable values (p_1, p_2, ..., p_n) that occur in the sample:*

    $$L = Prob\ (p_1 * p_2 * * * p_n)$$
- The higher the L, the higher the probability of observing the particular set in the sample.

Maximum Likelihood Estimation (MLE)

More:

- MLE involves finding the coefficients (α, β)
  - *that makes the log of the likelihood function (LL < 0) as large as possible*
- Or, finds the coefficients (α, β)
  - *that make -2 times the log of the likelihood function (-2LL) as small as possible*
- The maximum likelihood estimates solve the following condition:

$$\{Y - p(Y=1)\}X_i = 0$$

summed over all observations, i = 1,…,n

Interpreting Coefficients

- Since:

  $\ln[p/(1-p)] = \alpha + \beta X + e$

  The slope coefficient (β) is interpreted as the rate of change in the "log odds" as X changes … not very useful.

- Since:

  $p = 1/[1 + \exp(-\alpha - \beta X)]$

  The marginal effect of a change in X on the probability is: $dp/dX = f(\beta X)\ \beta$

Interpreting Coefficients

More:

- An interpretation of the logit coefficient which is usually more intuitive is the "odds ratio"
- Since:

$$[p/(1-p)] = \exp(\alpha + \beta X)$$

$\exp(\beta)$ is the effect of the independent variable on the "odds ratio"

Code

```python
# numpy is used for working with arrays
# sklearn is used for machine learning and statistical modeling
import numpy
from sklearn import linear_model

# Reshaped for Logistic function.

# X represents the size of a tumor in centimeters.
X = numpy.array([3.78, 2.44, 2.09, 0.14, 1.72, 1.65, 4.92, 4.37, 4.96, 4.52, 3.69, 5.88]).reshape(-1,1)

# Note: X has to be reshaped into a column from a row for the LogisticRegression() function to work.
# y represents whether or not the tumor is cancerous (0 for "No", 1 for "Yes").
# each item corresponds to one observation
y = numpy.array([0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1])

# LogisticRegression() to create a logistic regression object.
# fit() that takes the independent and dependent values
# as parameters and fills the regression object with data that describes the relationship
logr = linear_model.LogisticRegression()
logr.fit(X,y)

# the coefficient is the expected change in log-odds of having the outcome per unit change in X.
# coefficient and intercept values can be used to find the probability that each tumor is cancerous

# Create a function that uses the model's coefficient and intercept values
# to return probability that the given observation is a tumor
def logit2prob(logr, X):
  log_odds = logr.coef_ * X + logr.intercept_
  odds = numpy.exp(log_odds)
  probability = odds / (1 + odds)
  return(probability)

print(logit2prob(logr, X))

#Output/Result Explained
#3.78 0.61 The probability that a tumor with the size 3.78cm is cancerous is 61%.
#2.44 0.19 The probability that a tumor with the size 2.44cm is cancerous is 19%.
```

Applications of Logistic Regression

- Predicting a probability of a person having a heart attack
- Predicting a customer’s propensity to purchase a product or halt a subscription.
- Predicting the probability of failure of a given process or product

References

- Artificial Intelligence - A Modern Approach, Third Edition, Stuart J. Russell and Peter Norvig
- Logistic Regression: A Self-Learning Text 3rd ed. 2010 Edition by David G. Kleinbaum and Mitchel Klein
- https://realpython.com/logistic-regression-python/#logistic-regression-in-python
- https://www.appstate.edu/~whiteheadjc/service/logit/intro.htm
- https://www.w3schools.com/python/python_ml_logistic_regression.asp