Rotman

INTRO TO R PROGRAMMING

R Tutorial (RSM358) – Session 3 (Optional Materials)

September 24, 2024 Prepared by Jay Cao / TDMDAL

Website: https://tdmdal.github.io/r-intro-2024-fall/

Rotman School of Management
UNIVERSITY OF TORONTO

Binomial Logistic Regression

- let Y be a binary outcome variable (i.e., a binary categorical variable)
  - e.g. $\mathrm{Y} = \{0, 1\} = \{fail,\ pass\}$, $\mathrm{Y} = \{0, 1\} = \{down,\ up\}$, etc.

- Let $p = prob(Y = 1)$; $\frac{p}{1-p}$ is then the odds of being 1
  - The category of $Y = 0$ is a reference category
  - Reference category is relative as you can instead set $p = prob(Y = 0)$

- Binary logistic regression models the logit-transformed probability as a linear function of the predictor variables
  - Coefficients $(\beta_0 \ldots \beta_k)$ are estimated using maximum likelihood method

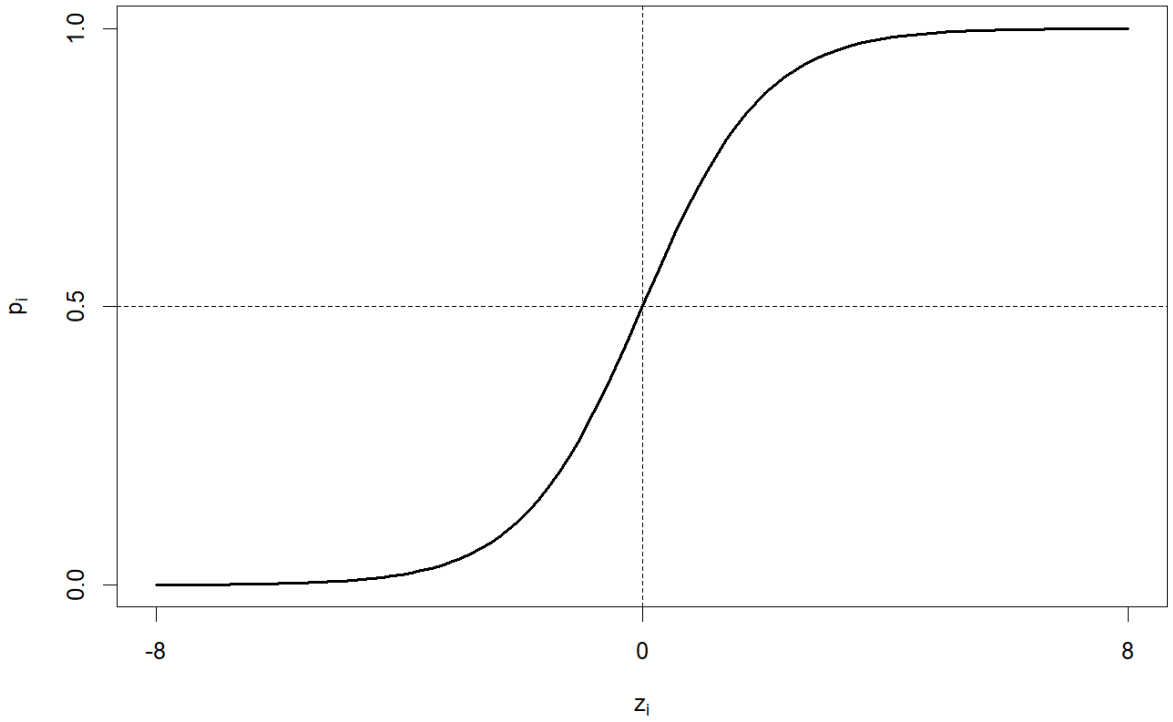
$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

From Log Odds to Probability to Prediction

- Let $z_i = logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}$.

- Then, $p_i = \frac{e^{z_i}}{1+e^{z_i}}$
  - Note $0 < p_i < 1$

- Threshold prob
  - It's a hyper-parameter



Interpret the Coefficients Estimated - 1

- An example: predict (or explain) if a student is in an honors class
  - Outcome variable: hon = {1-Yes, 0-No}. Set No to be the reference category.
  - Predictors are math score, female (1-yes, 0-no), and reading score

$$logit(p) = \beta_0 + \beta_1 math + \beta_2 female + \beta_3 read$$

Ref: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/

Interpret the Coefficients Estimated - 2

$$logit(p) = \beta_0 + \beta_1 math + \beta_2 female + \beta_3 read$$

```
Call:
glm(formula = hon ~ math + female + read, family = binomial, data = df)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.77025    1.71068  -6.880 5.97e-12 ***
math          0.12296    0.03128   3.931 8.44e-05 ***
female1       0.97995    0.42163   2.324   0.0201 *
read          0.05906    0.02655   2.224   0.0261 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

…
```

Ref: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/

Logistic Regression in R – Stock Market Ex.

- Import the `smarket.csv` data

- Prepare the data for logistic regression
  - Convert categorical variables to `factor` type (Y, and any predictors X)
  - Split data into training and test set

- Perform a logistic regression analysis
  - `glm(formula, data, family = binomial)` and `predict()`
  - Construct confusion matrix and calculate accuracy rate