

***Rotman***

# INTRO TO R PROGRAMMING

R Tutorial (RSM456) – Session 4

January 30, 2025 Prepared by Jay Cao / [TDMDAL](#)

Website: <https://tdmdal.github.io/r-intro-2025-winter/>



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.  $Y = \{0, 1\} = \{fail, pass\}$ ,  $Y = \{0, 1\} = \{down, up\}$ , etc.
- Let  $p = \text{prob}(Y = 1)$ ;  $\frac{p}{1-p}$  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  $p = \text{prob}(Y = 0)$
- Binary logistic regression models the logit-transformed probability as a linear function of the predictor variables
  - Coefficients  $(\beta_0 \dots \beta_k)$  are estimated using maximum likelihood method

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

# From Log Odds to Probability to Prediction

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

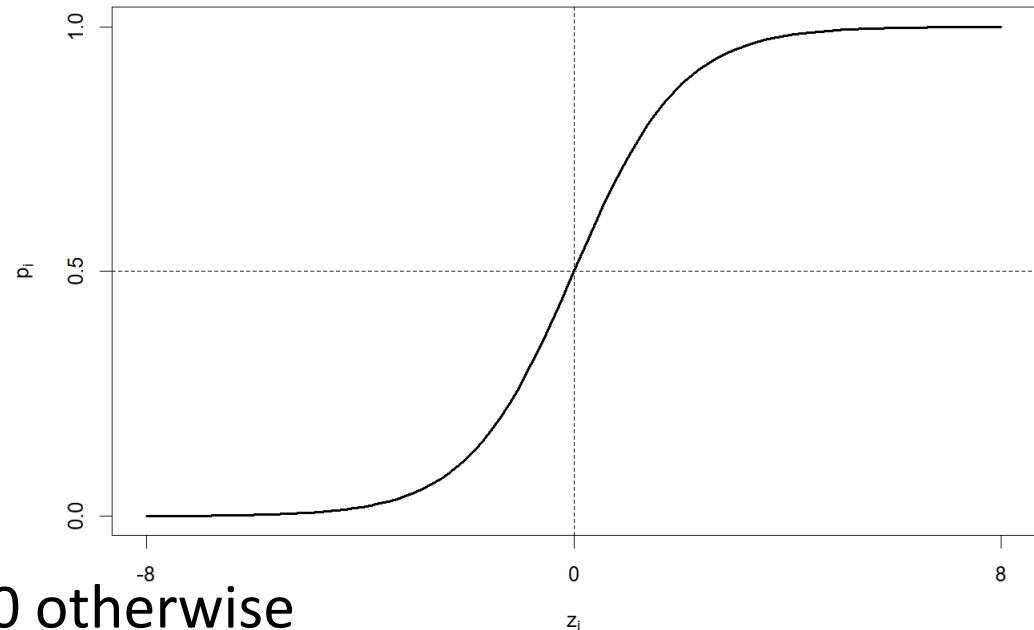
- Then,  $p_i = \frac{e^{z_i}}{1+e^{z_i}}$

- Note  $0 < p_i < 1$

- Threshold prob  $t$

- If the predicted  $\hat{p}_i \geq t$ , classify as 1, 0 otherwise

- $t$  is a hyper-parameter set by you (so you can tune it through validation)



# An Example - Setup & `glm()`

- 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

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

- R: `glm(formula, family = binomial, data)`

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

# Interpret the Coefficients Estimated

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

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

...

# 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, family = binomial, data)` and `predict()`
  - Construct confusion matrix and calculate accuracy rate