Financial Data Science

Lecture 4
Logistic Regression in Finance

Liu Peng
liupeng@smu.edu.sg

Logistic regression:
https://youtu.be/I51DDBeZ-VU?si=90DWcXLmGnCWJ0tQ

Two Types of Predictive Models

Regression

- Target variable is continuous y
- Examples: house-price prediction, temperature forecasting
- Common algorithms: linear/polynomial regression, ridge & lasso, SVR
- Evaluation metrics: MSE, RMSE, MAE, R2

Classification

- Target variable is categorical y
- Examples: spam detection, image recognition
- Common algorithms: logistic regression, k-NN, decision trees/random forests, SVM
- Evaluation metrics: accuracy, precision/recall/F1, confusion matrix, ROC-AUC

Q:

- Which type of task is more difficult?
- Which direction of conversion makes more sense?

Connecting Linear Regression to Classification

y ~ Bernoulli(p)

probability

if \hat{y} > 0.5 —> 1
if \hat{y} <= 0.5 —> 0

- Continuous to discrete:
  - Linear regression estimates
  - For classification, use a threshold: predict class 1 if prediction is above 0.5, else class 0.
- Decision boundary:
  - The threshold 0.5 defines a hyperplane, which separates the two classes.
- Not a good choice for classification:
  - Prediction is unbounded
  - Classification outcomes are Bernoulli, not Gaussian; linear regression's constant-variance assumption is violated

    y ~ N(. , .)
  - Linear regression doesn't constrain outputs to valid probability ranges, leading to negative or >1 "probabilities"

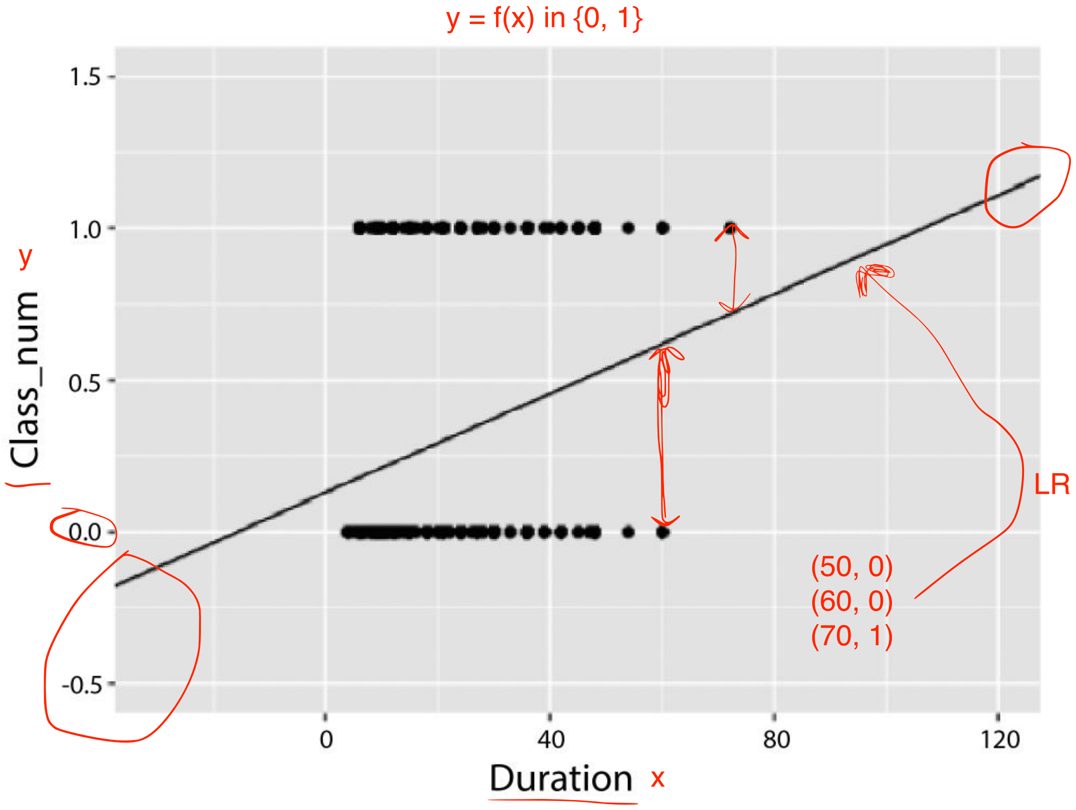


Figure 13.4 – Visualizing the linear regression model with extended range

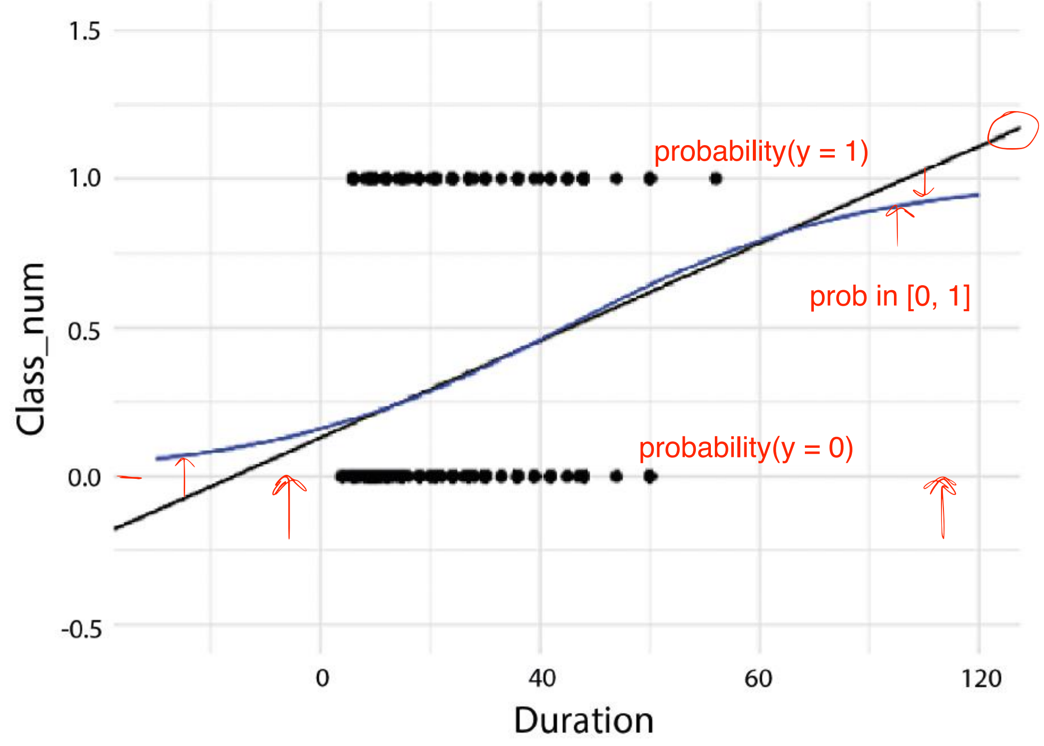


Figure 13.6 – Visualizing the logistic regression model with extended range

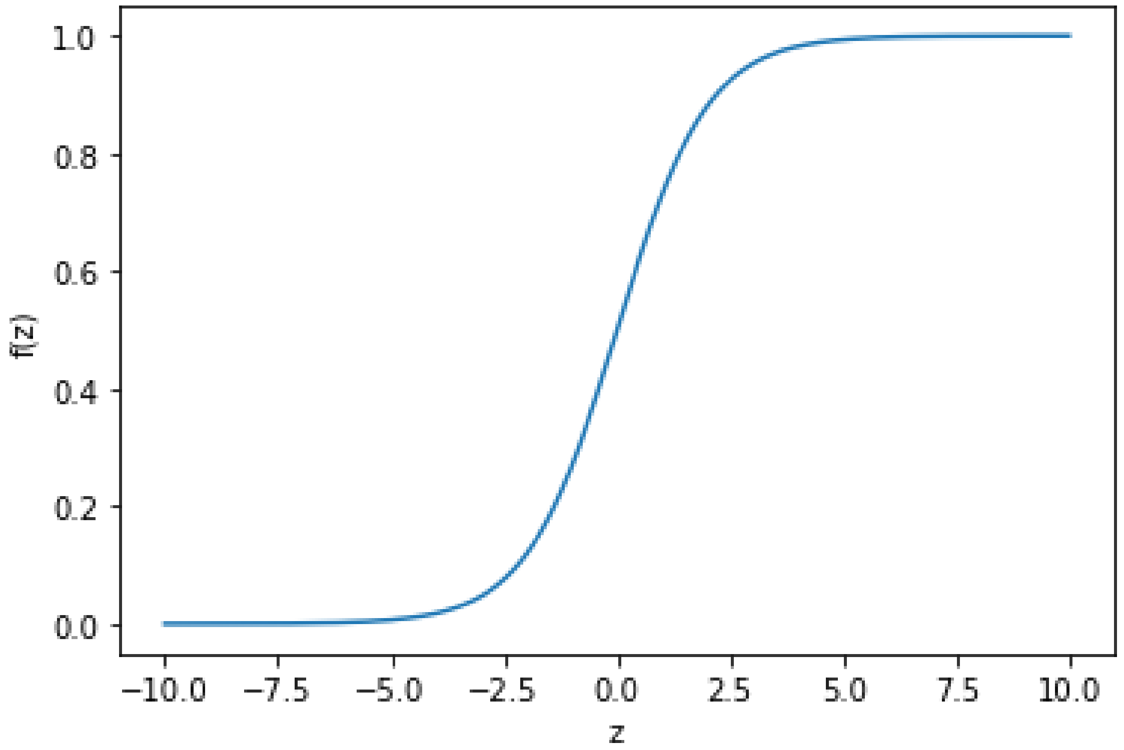
sigmoid fn

(-inf, +inf) —> (0, 1)

How to turn a number to a probability?

Introducing the Sigmoid Function

- $p = \sigma(z) = \frac{1}{1+e^{-z}}$
- $z \rightarrow \infty, \sigma(z) \rightarrow 1$; $z \rightarrow -\infty$, $\sigma(z) \rightarrow 0$
- Hence $\sigma(z) \in [0,1]$
- $\sigma(z)$ represents the probability of an event



Q: what is $\sigma'(z)$? = sigma(z) (1 - sigma(z))

In-class Quiz

- Q1-4

Introducing Logistic Regression

- Binary outcome that assumes a Bernoulli distribution
- Sigmoid (Logistic) function
- Joint Likelihood
- Log-likelihood and Cross-entropy loss

Binary outcome (Bernoulli distribution)

The binary variable y follows a Bernoulli distribution with probability p:

$$y \sim \mathrm{Bernoulli}(p), \quad y \in \{0, 1\}$$

The probability mass function (PMF) is:

P(y = 1) = p
P(y = 0) = 1 - p

$$P(y \mid p) = p^y (1 - p)^{1-y}$$

Logistic regression model with Sigmoid (Logistic) function

odds

P(y = 1) = 80% ⟷ lg 8 : 2 = 4

In logistic regression, we link the linear combination $z = x^\top \beta$ to probability $p(x)$ using the sigmoid function:

2 1

$$p(x) = \sigma(z) = \frac{1}{1 + e^{-z}}, \quad \text{where} \quad z = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

This ensures $0 < p(x) < 1$, making it a valid Bernoulli parameter.

prob log odds

In other words: $P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}}$ Or equivalently, $\log \frac{P(y = 1)}{P(y = 0)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

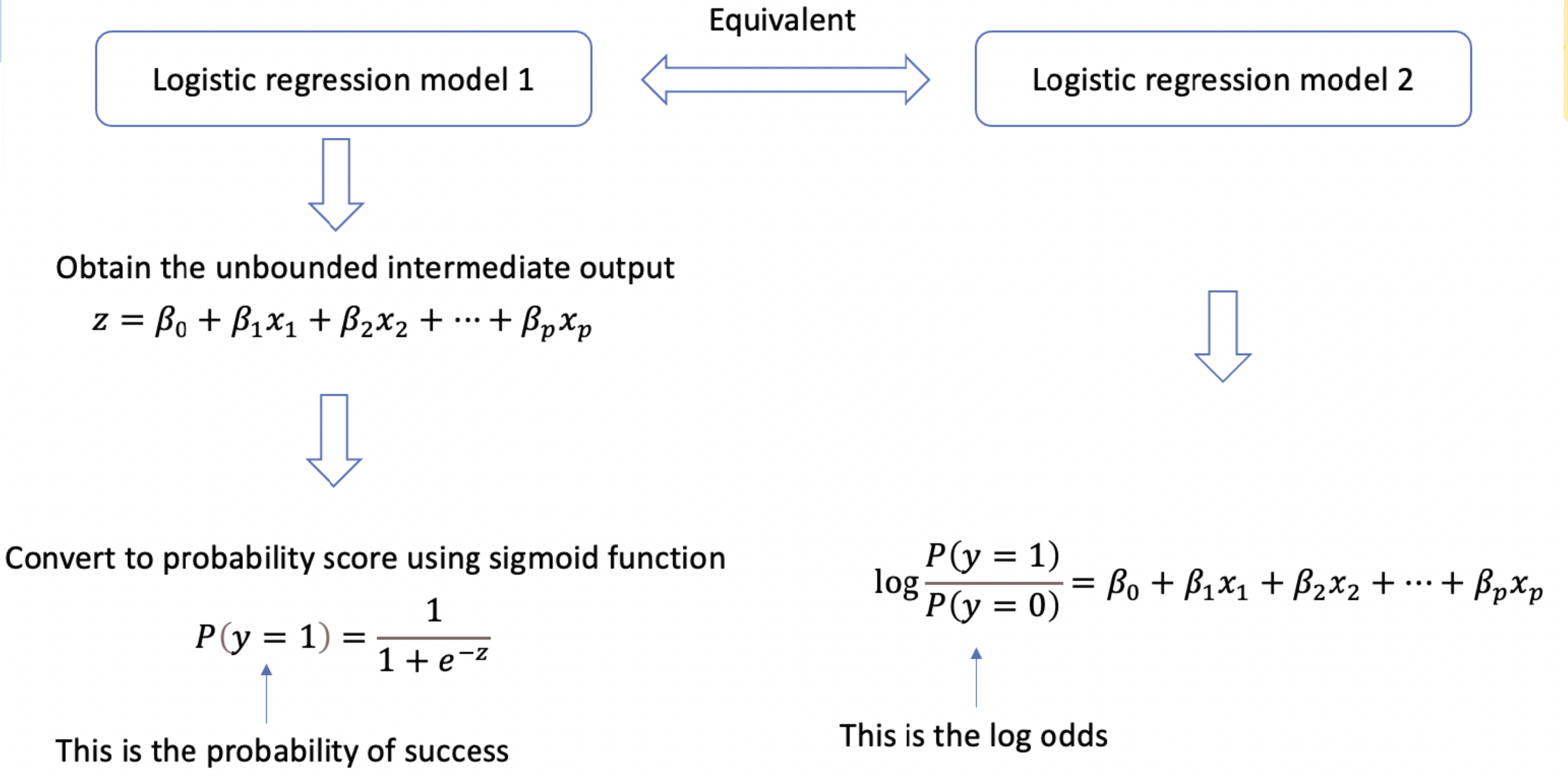


Figure 13.2 – Summarizing the logistic regression model

Bernoulli Likelihood for logistic regression

For n independent observations $\{(x_i, y_i)\}_{i=1}^{n}$, the joint likelihood is:

$$L(\beta) = \prod_{i=1}^{n} p(x_i)^{y_i} [1 - p(x_i)]^{1-y_i}$$

Plugging the sigmoid function:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{1}{1 + e^{-x_i^\top \beta}}\right]^{y_i} \left[1 - \frac{1}{1 + e^{-x_i^\top \beta}}\right]^{1-y_i}$$

Log-likelihood and Cross-entropy loss

Taking logs to simplify optimization gives the log-likelihood:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i))]$$

Maximizing the log-likelihood $\ell(\beta)$ is equivalent to minimizing the **binary cross-entropy loss**:

J(beta) = $-\ell(\beta) = -\sum_{i=1}^{n} [y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i))]$

Why does it make sense?

CEL

Total CEL	Sum over all observations	Sum over all classes	Binary indicator on class k	Predicted probability for class k

$$Q(\mathbf{w}) = -\sum_{i=1}^{N} \sum_{k=1}^{C} y_{i,k} \log(p_{i,k})$$

The resulting cost will be...	if...	and...
Low	1	≈1
High	1	≈0
High	0	≈1
Low	0	≈0

Figure 13.9 – Illustrating the CEL

if y = 1, P(y = 1) —> 1: close to 0
if P(y = 1) —> 0, then very high

In-class Quiz

- Q5-8

Model Evaluation for Logistic Regression

- Confusion Matrix
  - Tabulates True Positives, True Negatives, False Positives, False Negatives
  - Basis for all threshold-dependent metrics
- Threshold-Dependent Metrics
  - Accuracy = (TP+TN)/(TP+TN+FP+FN)
  - Precision (PPV) = TP/(TP+FP)
  - Recall (TPR / Sensitivity) = TP/(TP+FN)
  - F_1 Score = 2·(Precision·Recall)/(Precision+Recall)
- Threshold-Independent Metrics
  - ROC Curve: plots TPR vs. FPR as threshold varies
  - AUC (Area Under ROC): single-number measure of discrimination

Confusion matrix

	Predicted $\hat{y} = 0$	Predicted $\hat{y} = 1$	Total
Non-event $y = 0$	a	c	$a + c$
Event $y = 1$	b	d	$b + d$
Total	$a + b$	$c + d$	n = a + b + c + d

$$\text{Accuracy} = \frac{a+d}{n}$$

$$\text{Error rate} = \frac{b+c}{n}$$

$$\text{Recall} = \frac{d}{b+d}$$ y = 1

$$\text{Precision} = \frac{d}{c+d}$$

$$\text{Specificity} = \frac{a}{a+c}$$

reward = 100 x recall + 1 x precision

Figure 13.10 – Illustrating the confusion matrix and common evaluation metrics for binary classification tasks

Group Discussion - How to Choose the Evaluation Metrics?

In-class Quiz

- Q9-12

Reading materials

- Chapter 13, The Statistics and Machine Learning with R Workshop

Group Homework

- Predict whether a customer will default on a loan (yes/no) using logistic regression
- Dataset
  - Choose a public credit-default dataset (e.g. UCI Credit Card Default, Kaggle “Give Me Some Credit,” or a comparable financial dataset).
- Data processing
  - Clean and impute missing values.
  - Encode categorical variables (one-hot, ordinal, target encoding).
  - Standardize or normalize numeric predictors as needed.
- Modeling
  - Fit a baseline logistic regression (no regularization).
  - Fit L1 (lasso) and L2 (ridge)–penalized logistic models; use cross-validation to select the penalty strength.
- Evalution
  - Accuracy, Precision, Recall, F1–score at a 0.5 threshold
  - ROC curve & AUC

Homework

- Second group homework to submit by one day before class starts next week
- Post learning reflections and questions in the group chat if any
- Review course contents and recording