Logistic Regression

STAT499 C DRP WI25

Catherine Zhang, Linyi Xia, Kayla Irish

OVERVIEW

1. **Intro&Data Structure for Linear vs. Logistic Regression**
2. **Why not linear regression** (how logistic regression is a potential solution)
3. **Define the Model Using Link Functions (logit)**
4. **Maximum Likelihood Estimation**
5. **Model setup**
6. **Example**
7. References

INTRO AND DATA STRUCTURE

$$p(X) = \beta_0 + \beta_1 X. \quad (4.1)$$

Linear regression model.

- **X** = independent variables/predictors in the model
- **Y** = dependent variable (continuous)
- **p(X)** = predicted value of with Y based on X

- **beta coeffs** β_0, β_1 = coeffs that measures the expected change in Y for a one-unit change in X, holding all other predictors constant (OLS)

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}. \quad (4.2)$$

Logistic function.

- **X** = independent variables/predictors in the model
- **Y** = dependent variable (binary)
- **p** = probability of Y being 1 given X
    - model p as p(X) which is a function of X
- use p(X) with X inputs to predict the probability of Y=1

- **beta coeffs** β_0, β_1 = parameters define the relationship between each predictor variable and the log odds of the dependent binary outcome (MLE)

INTRO AND DATA STRUCTURE

$$p(X) = \beta_0 + \beta_1 X. \tag{4.1}$$

Linear regression model.

- Best for continuous, quantitative outcomes
- Goal is to find a linear function that best fits the observed data

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}. \tag{4.2}$$

Logistic function.

- Useful for binary outcomes (e.g., yes/no, pass/fail)
- Goal is to find the **probability** that the observation belongs to one of the two classes

WHY NOT LINEAR REGRESSION?

Predicting **binary** medical conditions (e.g., stroke, drug overdose, epileptic seizure).

- Issue: Linear regression predictions may be outside the valid range (e.g., negative values or values greater than 1).

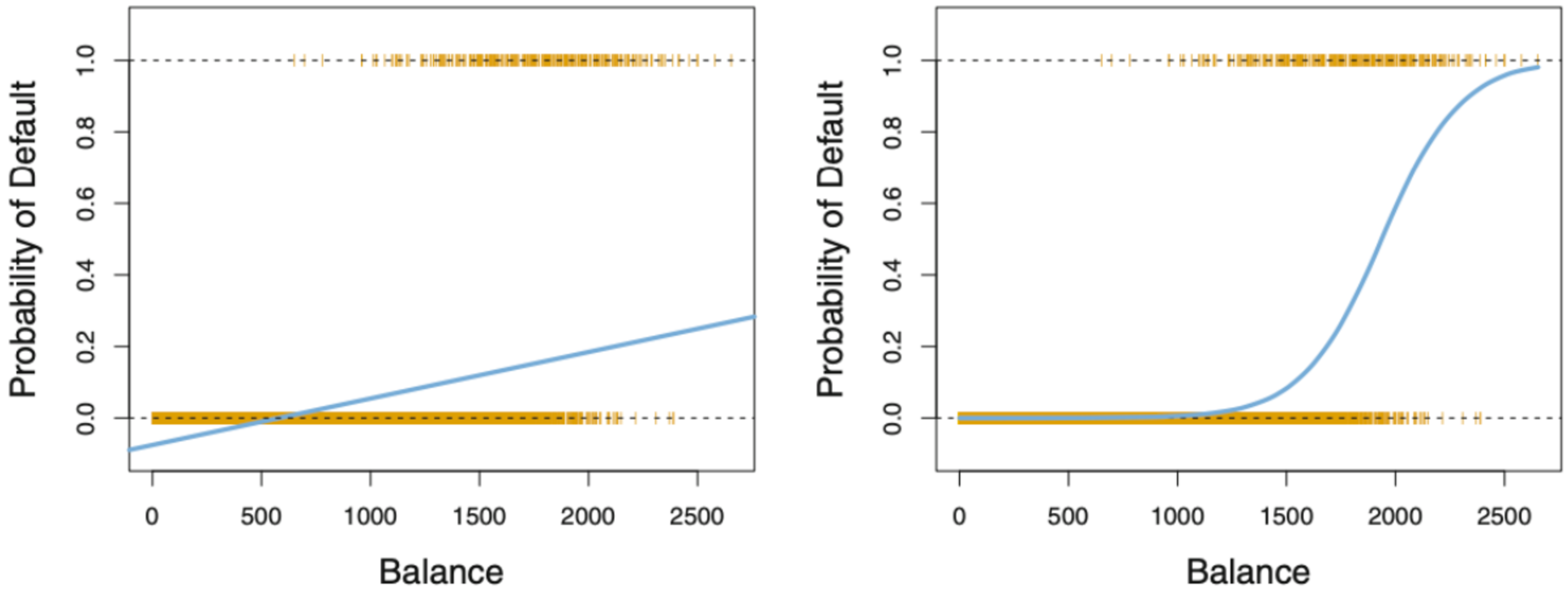
$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}. \qquad (4.2)$$

Logistic function.

- constraint output between 0 and 1

WHY LOGISTIC REGRESSION

- Example: Default prediction in finance.
- Formula: $\Pr(\text{default} = \text{Yes} \mid \text{balance})$ is modeled as a function of balance.



LINK FUNCTION

A wide choice of link functions $g(\pi)$ is available. Three functions commonly used in practice are

1. the logit or logistic function

$$g_1(\pi) = \log\{\pi/(1-\pi)\};$$

2. the probit or inverse Normal function

$$g_2(\pi) = \Phi^{-1}(\pi);$$

3. the complementary log-log function

$$g_3(\pi) = \log\{-\log(1-\pi)\}.$$

A fourth possibility, the log-log function

$$g_4(\pi) = -\log\{-\log(\pi)\},$$

which is the natural counterpart of the complementary log-log function, is seldom used because its behaviour is inappropriate for $\pi < \frac{1}{2}$, the region that is usually of interest. All four functions can

- π: probability of the occurrence of an event; denotes the probability that Y=1 for a given set of predictor variables (i.e., in medical context, the probability that a patient has a disease, given their symptoms and test results.)
- **mainly use the logit**
- The other 3 also constrain the predicted outcome from the model between 0 and 1.

A wide choice of link functions $g(\pi)$ is available. Three functions commonly used in practice are

1. the logit or logistic function

$$g_1(\pi) = \log\{\boxed{\pi/(1-\pi)}\};$$

2. the probit or inverse Normal function

$$g_2(\pi) = \Phi^{-1}(\pi);$$

3. the complementary log-log function

$$g_3(\pi) = \log\{-\log(1-\pi)\}.$$

A fourth possibility, the log-log function

$$g_4(\pi) = -\log\{-\log(\pi)\},$$

which is the natural counterpart of the complementary log-log function, is seldom used because its behaviour is inappropriate for $\pi < \frac{1}{2}$, the region that is usually of interest. All four functions can

- maps this probability π to the log odds of Y=1
- maps probabilities from the interval (0, 1) to the entire real line which is useful for modeling binary outcomes (i.e., success/failure, yes/no) where π is the probability of success
- output the log-odds = logarithm of the **odds (ratio)** of the event occurring versus not occurring
  - positive $\longrightarrow$ the odds of the event occurring are greater than the odds of it not occurring

once we have our predictor/ covariates Xs and outcome Y, we can set up the logistic model using the logit function

$$\boxed{\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2}$$

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2).$$

$$\boxed{\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}.}$$

$$\boxed{p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.}$$

we solve for \pi and and thus get the probability of a positive response (Y=1) using this model

MLE METHOD

for all individuals who did not. This intuition can be formalized using a mathematical equation called a *likelihood function*:

$$\ell(\beta_0, \beta_1) = \boxed{\prod_{i:y_i=1} p(x_i)} \boxed{\prod_{i':y_{i'}=0} (1 - p(x_{i'}))}. \qquad (4.5)$$

likelihood function

- **Goal**: find the optimal parameters (betas) of the model that **maximize** this function
- captures the probability of observing the specific set of outcomes given the predictor values and model parameters (betas)
- based on the product of probabilities for each individual observation in the dataset

- **runs over all cases where the observed outcome is 1 - model's estimated probability that Y=1**
- **runs over all cases where the observed outcome is 0 - model's estimated probability that Y=0**

$$\boxed{p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.}$$

MODEL SETUP

• **Dataset:** Heart Failure Clinical Records

• **Objective:** Identify key predictors of death in heart failure patients

• Key Variables:
  - **Outcome (Y): Death event** (1 = Yes, 0 = No)
  - Predictors (X): We had a lot of variables to choose from as predictors but hypothesized that: **serum creatinine, ejection fraction, serum sodium** best predicted the outcome

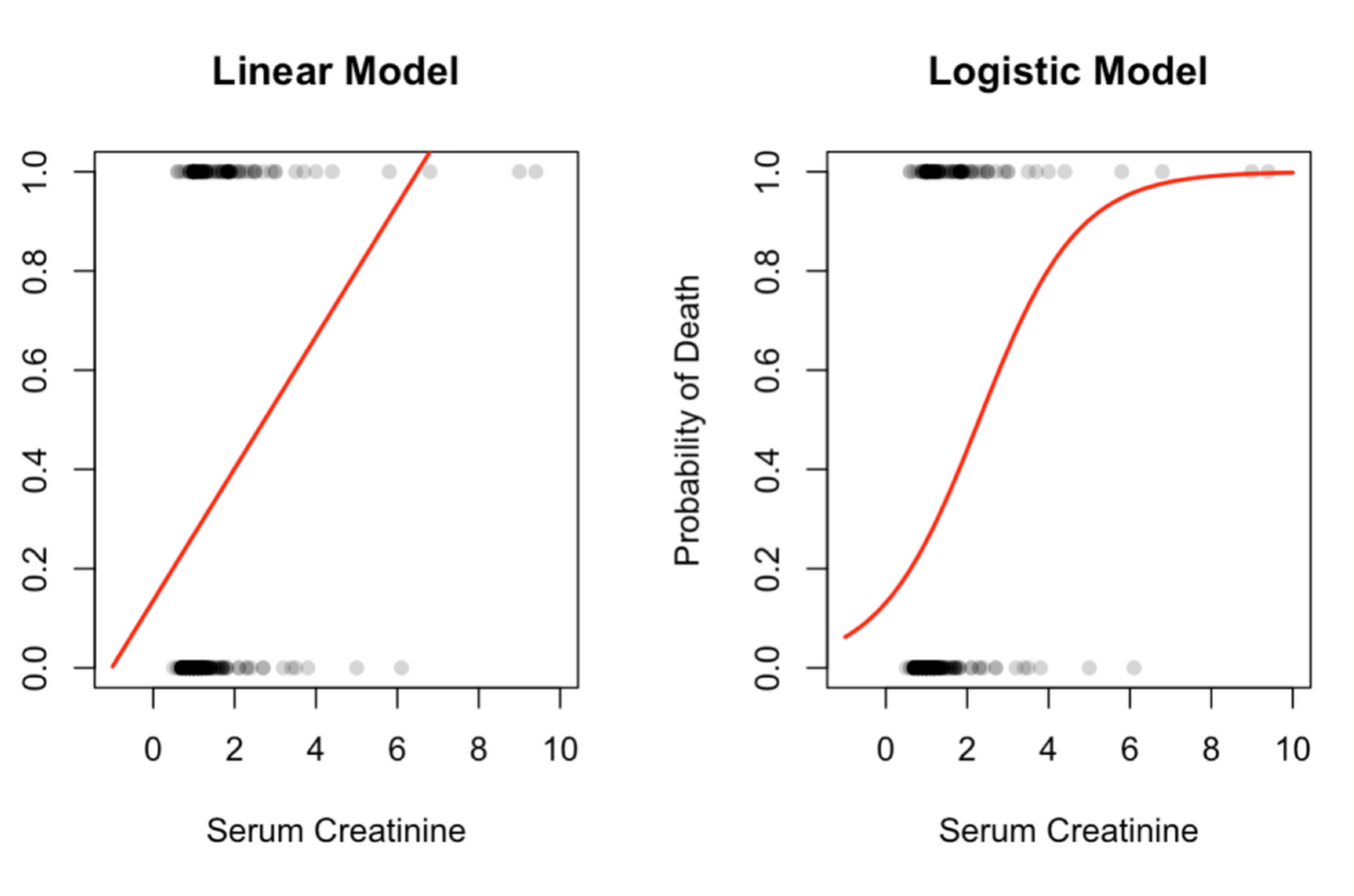
Used forward/backward selection to identify significant predictors

• Final Predictors:
  - **Serum Creatinine & Ejection Fraction**
 (Serum sodium was removed due to insignificance)

• Run a univariate logistic regression for each predictor

• **Result:** Higher serum creatinine & lower ejection fraction increase mortality risk

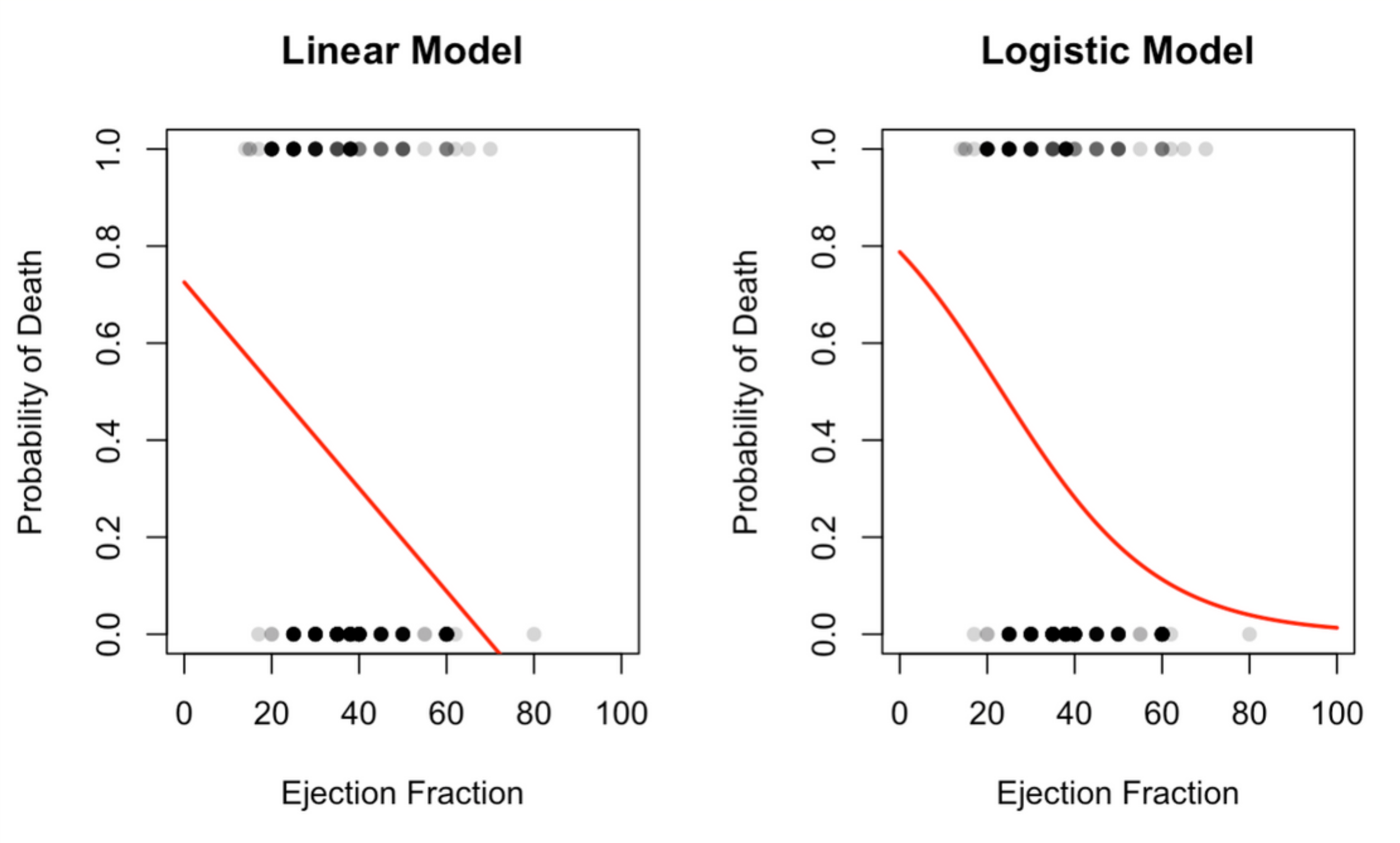
EXAMPLE



- Linear Model: p(X) = 0.136 + 0.133X
  - Does not fit the data well; predictions exceed probability limits
  - Not suitable for binary outcomes

- **Logistic Model**: $\log\left(\frac{p(X)}{1-p(X)}\right) = -1.9 + 0.8X$
  - Provides a better fit for binary classification
  - Coefficients are interpretable as log-odds

- Coefficient Interpretation:
  - Serum Creatinine Coefficient (0.8): Each unit increase in serum creatinine increases log-odds by 0.8
  - Odds of death increase by exp(0.8) ≈ 2.23 per unit increase in serum creatinine



- Linear Model: p(X) = 0.73 - 0.01X
  - Does not fit the data well; predictions exceed probability limits
  - Not suitable for binary outcomes

- **Logistic Model**: $\log\left(\frac{p(X)}{1-p(X)}\right) = 1.31 - 0.056X$
  - Provides a better fit for binary classification
  - Coefficients are interpretable as log-odds

- Coefficient Interpretation:
  - Ejection Fraction Coefficient (-0.056): Each unit increase in ejection fraction decreases log-odds by -0.056
  - Odds of death decrease by approximately 5.45% for each 1% increase in ejection fraction

REFERENCES

James, Gareth, et al. An Introduction to Statistical Learning: With Applications in R. Springer, 2013.

McCullagh, P. Generalized Linear Models. CRC Press LLC, 1989. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/washington/detail.action?docID=5631551. Accessed 22 Jan. 2025.

Thank you!

Q&A?