Logistic Regression

@S S Roy,
3rd Aug,2025

📘 Title: Logistic Regression Explained — Based on Cox's 1958 Foundational Paper

---

🔰 Objective

To understand how to analyze binary outcomes (like success/failure) in relation to one or more predictors using a model that ensures predicted probabilities remain between 0 and 1.

🧩 Prerequisite Concepts

- **Binary variable**: Variable that takes on only two values (e.g., 0 = failure, 1 = success).
- **Regression**: Explains a dependent variable using one or more independent variables.
- **Odds**: Ratio of probability of success to failure, $\frac{p}{1-p}$.
- **Logit function**: $\log\left(\frac{p}{1-p}\right)$, maps probabilities (0,1) to real numbers.

📚 The Core Idea

Suppose we observe sequences of binary outcomes (e.g., success/failure). Cox introduced a model to **estimate how the probability of success varies with predictors**, like time, age, treatment, etc.

🔑 Key Model

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \alpha + \beta x_i$$

- p_i: probability of success for observation i

- x_i: predictor variable
- α: intercept (nuisance parameter)
- β: slope (our main interest)

This is the **logistic regression model**, ensuring:

- Output is always a valid probability
- Relationship between predictors and probability is nonlinear but interpretable

---

📌 Why Linear Models Fail Here

A linear model like $p_i = \alpha + \beta x_i$ can produce $p_i < 0$ or $p_i > 1$, which makes no sense for probabilities. The logistic model avoids this by using the logit transformation.

🛠 How Cox Approaches Estimation

Step 1: Binary Sequences

We observe a set of binary outcomes:

$$Y = \{y_1, y_2, \ldots, y_n\}, \quad y_i \in \{0, 1\}$$

with corresponding covariates:

$$X = \{x_1, x_2, \ldots, x_n\}$$

Step 2: Likelihood Function

The joint likelihood under independence:

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

This simplifies to:

$$L(\alpha, \beta) = \exp\left[\sum y_i(\alpha + \beta x_i)\right] \prod (1 + e^{\alpha+\beta x_i})^{-1}$$

🧪 Hypothesis Testing and Inference

Goal:

Test whether $\beta = 0$ (no relationship between outcome and predictor).

Approach:

Use **conditional inference**:

- Treat the total number of successes $y = \sum y_i$ as fixed.
- Focus on the distribution of the **test statistic** $X = \sum x_i y_i$ given y.

This leads to a **hypergeometric**-like model under the null hypothesis and helps eliminate nuisance parameter α.

🧮 Approximate Solutions

Cox derived approximations for small or large samples:

- **Normal approximation** for test statistic X
- **Cumulant expansions** to estimate mean/variance under logistic alternatives

---

📊 Practical Example (2×2 Table)

Suppose we have:

	Success (1)	Failure (0)
Group A	3	11
Group B	60	32

This becomes a **logistic regression** problem with group indicator as predictor. The odds ratio:

$$\text{OR} = \frac{60 \times 11}{32 \times 3}$$

Cox shows how to compute confidence intervals for β, which leads to CI for the **odds ratio**, and compares them with exact and approximate methods.

🧠 Extensions Covered by Cox

1. Multiple Predictors

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots$$

2. Markov Dependence

- Probability of success depends on outcome of previous trial:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta y_{i-1}$$

3. Learning Effect / Cumulative Scores

- Let success depend on **number of past successes**:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta \cdot (\text{number of past 1's})$$

🧾 Significance of Cox's Work

- First general **formulation of logistic regression**.
- Established the **likelihood approach** for binary outcomes.
- Demonstrated that **non-parametric tests** (e.g., Wilcoxon) align with logistic assumptions.
- Introduced concepts that predate modern GLMs (Generalized Linear Models).

🧰 Tools & Techniques Cox Introduced

Technique	Description
Logistic law	Ensures probability lies in (0,1)
Conditional inference	Eliminates nuisance parameters
Cumulant expansions	Approximates distribution of test statistics
Sampling without replacement	Basis for exact tests
Multiple regression on logits	Extension to multiple variables

---

✅ Summary Table

Concept	Summary
Model	$\log\left(\frac{p}{1-p}\right) = \alpha + \beta x$
Target	Estimate/test effect of x on binary y
Estimator	Maximum likelihood / conditional method
Testing	Conditional on total y; uses distribution of $\sum x_i y_i$
Assumptions	Independence, binary outcome
Extensions	Markov dependence, cumulative response, multiple predictors

---

🧠 Final Thoughts

Cox's 1958 paper didn't just invent logistic regression—it provided a **complete statistical framework** for analyzing binary outcomes with covariates. It's **robust, interpretable, and foundational** to modern machine learning and statistics.

@S S Roy
3rd Aug,2025