Week 10: Logistic Regression

Francisco Richter Ernst Wit

Introduction to Data Science (MSc)

1 Binary Response Theory and Exponential Family Foundations

1.1 Binary Response Variables

Definition 1.1 (Binary Response Variable)**.** *A binary response variable Y takes values in $\{0, 1\}$ with:*

$$P(Y = 1) = p, \quad P(Y = 0) = 1 - p$$

where $p \in (0, 1)$ is the success probability.

Theorem 1.1 (Bernoulli Distribution as Exponential Family)**.** *The Bernoulli distribution belongs to the exponential family:*

$$f(y; p) = p^y (1 - p)^{1-y} = \exp\left\{ y \log\left(\frac{p}{1-p} \right) + \log(1 - p) \right\}$$

This gives us:

$$\theta = \log\left(\frac{p}{1-p} \right) \quad \textit{(canonical parameter)} \tag{1}$$

$$b(\theta) = \log(1 + e^\theta) \tag{2}$$

$$a(\phi) = 1 \quad \textit{(dispersion parameter)} \tag{3}$$

$$c(y, \phi) = 0 \tag{4}$$

1.2 Link Function Theory for Binary Data

Definition 1.2 (Link Functions for Binary Data)**.** *A link function g maps probabilities to the real line: $g : (0, 1) \to \mathbb{R}$, enabling the use of linear modeling techniques for binary responses.*

The choice of link function fundamentally determines how we model the relationship between predictors and response probabilities. The most commonly used link function is the **logit link**, defined as $g(p) = \log(p/(1 - p))$, which transforms the probability p into the log-odds. This transformation has profound theoretical and practical advantages that make it the canonical choice for binary regression.

Alternative link functions include the **probit link** $g(p) = \Phi^{-1}(p)$, where Φ represents the standard normal cumulative distribution function. The probit link arises naturally when we assume that there exists an underlying continuous latent variable following a normal distribution, and we observe a binary outcome based on whether this latent variable exceeds a threshold. The **complementary log-log link** $g(p) = \log(-\log(1 - p))$ is particularly useful when dealing with extreme probabilities or when the underlying process follows a Gumbel distribution.

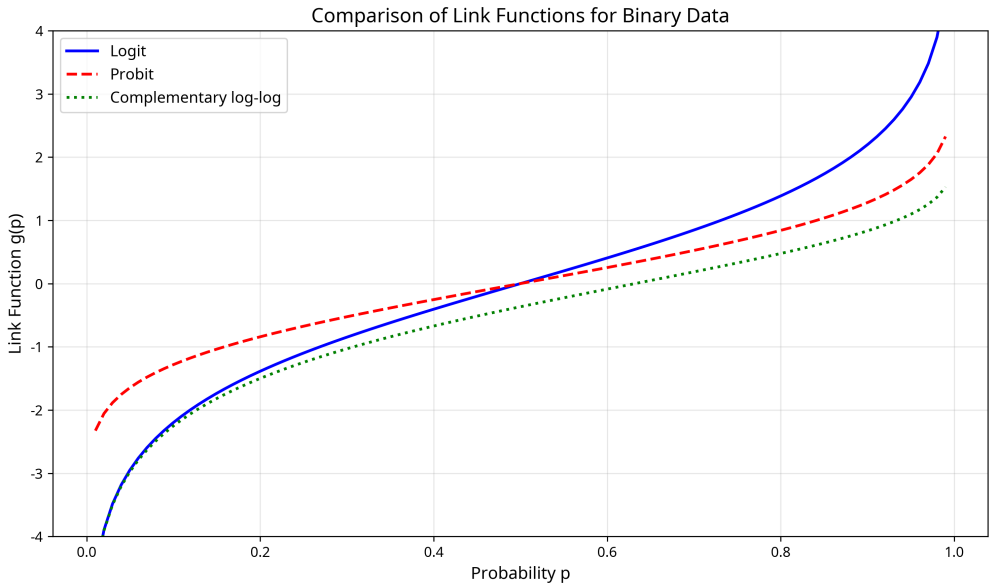

Figure 1: Comparison of Link Functions: The logit, probit, and complementary log-log functions map probabilities from $(0, 1)$ to $(-\infty, \infty)$ with different shapes. The logit function (solid line) provides symmetric transformation around $p = 0.5$, while the probit function (dashed line) has slightly heavier tails. The complementary log-log function (dotted line) is asymmetric, making it suitable for modeling rare events where probabilities are typically small.

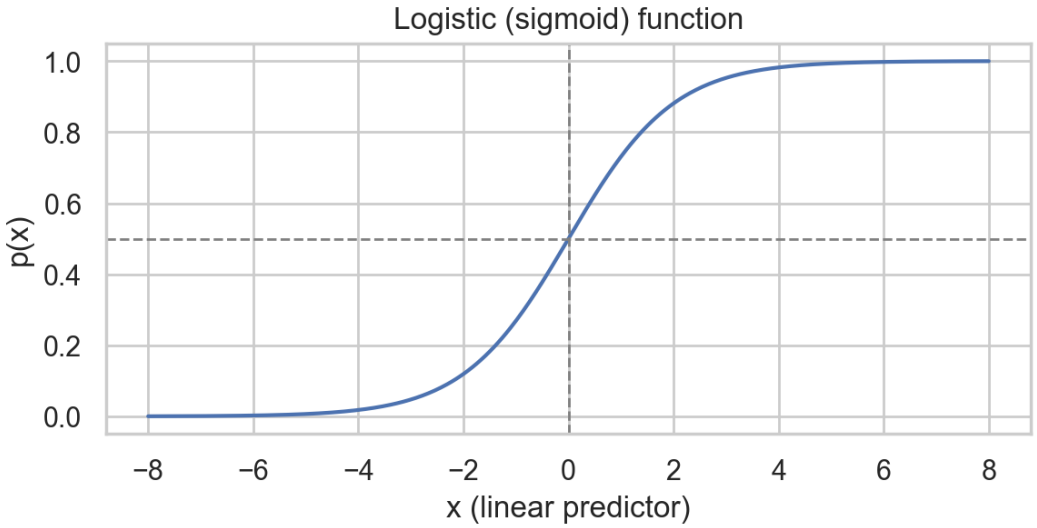

Figure 2: Logistic (sigmoid) function $p(x) = 1/(1+e^{-x})$ mapping linear predictors to probabilities; most sensitive near $p = 0.5$.

Theorem 1.2 (Properties of the Logit Link). *The logit link $g(p) = \log(p/(1-p))$ possesses several mathematically elegant and practically important properties that establish it as the canonical choice for binary regression.*

The logit link's canonical property stems from its direct correspondence to the natural parameter θ of the Bernoulli distribution in exponential family form. When we write the Bernoulli distribution as $f(y;p) = \exp\{y\log(p/(1-p)) + \log(1-p)\}$, we immediately see that the canonical parameter is precisely $\theta = \log(p/(1-p))$, which is the logit transformation. This canonical property ensures that the log-likelihood function is concave in the regression parameters, guaranteeing the existence and uniqueness of the maximum likelihood estimator.

The symmetry property $g(p) = -g(1-p)$ reflects the balanced treatment of success and failure outcomes. This symmetry means that the logit of the probability of success is the negative of the logit of the probability of failure, providing an intuitive interpretation where positive coefficients increase the log-odds of success while negative coefficients decrease them.

The interpretability of logistic regression coefficients as log-odds ratios provides direct practical meaning. When we increase a predictor x_j by one unit, the log-odds change by β_j, which corresponds to multiplying the odds by $\exp(\beta_j)$. This multiplicative interpretation on the odds scale is both mathematically tractable and practically meaningful for decision-making.

1.3 Odds and Odds Ratios

Definition 1.3 (Odds and Odds Ratio). *For probability p:*

$$\textit{Odds} = \frac{p}{1-p} \tag{5}$$

$$\textit{Odds Ratio} = \frac{\textit{Odds}_1}{\textit{Odds}_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \tag{6}$$

Theorem 1.3 (Odds Ratio in Logistic Regression). *For logistic regression $\textit{logit}(p) = \mathbf{x}^T\boldsymbol{\beta}$:*

$$OR_{x_j} = \exp(\beta_j)$$

represents the multiplicative change in odds for a one-unit increase in x_j.

1.4 Logistic Function Properties

Theorem 1.4 (Logistic Function Derivatives). *For $p(x) = \frac{e^x}{1+e^x}$:*

$$p'(x) = p(x)(1-p(x)) \tag{7}$$

$$p''(x) = p(x)(1-p(x))(1-2p(x)) \tag{8}$$

$$\max_x p'(x) = \frac{1}{4} \text{ at } x = 0 \tag{9}$$

Corollary 1.1 (Maximum Slope Property). *The logistic function has maximum slope at $p = 0.5$, making it most sensitive to changes in the linear predictor when probabilities are near 0.5.*

2 Advanced Maximum Likelihood Theory

2.1 Detailed Likelihood Derivation

Theorem 2.1 (Bernoulli Likelihood for Logistic Regression). *For independent observations $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, where $Y_i \sim \textit{Bernoulli}(p_i)$:*

The likelihood function is:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

Substituting $p_i = \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}$*:*

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\right)^{y_i} \left(\frac{1}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\right)^{1-y_i}$$

Theorem 2.2 (Log-Likelihood Simplification)**.** *The log-likelihood simplifies to:*

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \mathbf{x}_i^T\boldsymbol{\beta} - \log(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}})\right]$$

This can be written in exponential family form:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i\theta_i - b(\theta_i)\right]$$

where $\theta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ *and* $b(\theta) = \log(1+e^{\theta})$*.*

2.2 Score Function and Hessian

Theorem 2.3 (Score Function Derivation)**.** *The score function (first derivative) is:*

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}(y_i - p_i)\mathbf{x}_i = \mathbf{X}^T(\mathbf{y} - \boldsymbol{\pi})$$

where $\boldsymbol{\pi} = (p_1, \ldots, p_n)^T$ *and* $p_i = \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}$*.*

Proof.

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n}\left[y_i x_{ij} - \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}} x_{ij}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\right] \tag{10}$$

$$= \sum_{i=1}^{n} x_{ij}(y_i - p_i) \tag{11}$$

□

Theorem 2.4 (Hessian Matrix)**.** *The Hessian (second derivative matrix) is:*

$$\mathbf{H} = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\mathbf{X}^T\mathbf{W}\mathbf{X}$$

where $\mathbf{W} = diag(w_1, \ldots, w_n)$ *with* $w_i = p_i(1-p_i)$*.*

Corollary 2.1 (Concavity of Log-Likelihood)**.** *Since* $w_i = p_i(1-p_i) > 0$ *for all* $p_i \in (0,1)$*, the matrix* $\mathbf{W}$ *is positive definite, making* $\mathbf{H}$ *negative definite. Therefore, the log-likelihood is strictly concave, ensuring a unique global maximum.*

2.3 Fisher Information and Asymptotic Properties

Theorem 2.5 (Fisher Information Matrix)**.** *The Fisher information matrix is:*

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbb{E}[-\mathbf{H}] = \mathbf{X}^T\mathbf{W}\mathbf{X}$$

where the expectation is taken over the distribution of $\mathbf{Y}$*.*

Theorem 2.6 (Asymptotic Distribution of MLE)**.** *Under regularity conditions, as* $n \to \infty$*:*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\beta}))$$

For finite samples, approximately:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}))$$

3 Introduction to Logistic Regression

Logistic regression is a fundamental statistical method for modeling binary and categorical outcomes. Unlike linear regression, which models the mean of a continuous response, logistic regression models the probability of an event occurring, making it essential for classification problems in data science and machine learning.

The key insight underlying logistic regression is the recognition that while linear regression predicts values that can range from negative to positive infinity, probabilities must be constrained between 0 and 1. Logistic regression elegantly solves this constraint problem by using the logistic function to transform linear combinations of predictors into valid probabilities. This transformation ensures that no matter what values the predictors take, the predicted probabilities will always fall within the required range.

4 The Logistic Function

4.1 Mathematical Foundation

Definition 4.1 (Logistic Function)**.** *The logistic function is defined as:*

$$p(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}$$

This function maps any real number to the interval $(0, 1)$*.*

4.2 Properties

Key properties of the logistic function:

- $\lim_{x \to -\infty} p(x) = 0$
- $\lim_{x \to \infty} p(x) = 1$
- $p(0) = 0.5$
- The function is monotonically increasing
- It has an S-shaped (sigmoid) curve

4.3 The Logit Transformation

Definition 4.2 (Logit Function)**.** *The logit function is the inverse of the logistic function:*

$$\textit{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

where $\frac{p}{1-p}$ *is called the odds ratio.*

The logit transformation maps probabilities from $(0, 1)$ to $(-\infty, \infty)$, allowing us to use linear modeling techniques.

5 Binary Logistic Regression

5.1 Model Specification

Definition 5.1 (Binary Logistic Regression Model)**.** *For a binary response* $Y \in \{0, 1\}$ *and predictors* $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$*:*

$$\textit{logit}(P(Y = 1|\mathbf{x})) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Equivalently:

$$P(Y = 1|\mathbf{x}) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

5.2 Interpretation of Coefficients

Theorem 5.1 (Odds Ratio Interpretation)**.** *For a one-unit increase in* x_j *(holding other variables constant):*

$$\textit{Odds Ratio} = e^{\beta_j}$$

This represents the multiplicative change in odds.

Example 5.1 (Coefficient Interpretation)**.** *If* $\beta_1 = 0.693$*, then* $e^{0.693} = 2$*. This means a one-unit increase in* x_1 *doubles the odds of the event occurring.*

6 Maximum Likelihood Estimation

6.1 Likelihood Function

For independent observations $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$:

Definition 6.1 (Likelihood for Logistic Regression)**.**

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$$

where $p_i = P(Y_i = 1|\mathbf{x}_i) = \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}$.

6.2 Log-Likelihood

Definition 6.2 (Log-Likelihood)**.**

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

Substituting the logistic form:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \mathbf{x}_i^T \boldsymbol{\beta} - \log(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}) \right]$$

6.3 Score Function and Information Matrix

Theorem 6.1 (Score Function)**.** *The score function (gradient of log-likelihood) is:*

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - p_i)\mathbf{x}_i = \mathbf{X}^T(\mathbf{y} - \boldsymbol{\pi})$$

where $\boldsymbol{\pi} = (p_1, p_2, \ldots, p_n)^T$.

Theorem 6.2 (Fisher Information Matrix)**.**

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{W} \mathbf{X}$$

where $\mathbf{W} = diag(p_1(1 - p_1), p_2(1 - p_2), \ldots, p_n(1 - p_n))$.

7 Newton-Raphson Algorithm

Since there's no closed-form solution, we use iterative methods:

Definition 7.1 (Newton-Raphson for Logistic Regression)**.**

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + [\mathbf{I}(\boldsymbol{\beta}^{(k)})]^{-1} \mathbf{U}(\boldsymbol{\beta}^{(k)})$$

This is equivalent to Iteratively Reweighted Least Squares (IRLS):

$$\boldsymbol{\beta}^{(k+1)} = (\mathbf{X}^T \mathbf{W}^{(k)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(k)} \mathbf{z}^{(k)}$$

where $\mathbf{z}^{(k)} = \mathbf{X}\boldsymbol{\beta}^{(k)} + [\mathbf{W}^{(k)}]^{-1}(\mathbf{y} - \boldsymbol{\pi}^{(k)})$ is the working response.

8 Statistical Inference

8.1 Asymptotic Properties

Theorem 8.1 (Asymptotic Distribution of MLE)**.** *Under regularity conditions:*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\beta}))$$

8.2 Hypothesis Testing

8.2.1 Wald Test

For testing $H_0 : \beta_j = 0$:

$$W = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)} \sim N(0,1) \text{ asymptotically}$$

8.2.2 Likelihood Ratio Test

For testing $H_0 : \boldsymbol{\beta}_{\text{sub}} = \mathbf{0}$:

$$\text{LRT} = 2[\ell(\hat{\boldsymbol{\beta}}_{\text{full}}) - \ell(\hat{\boldsymbol{\beta}}_{\text{reduced}})] \sim \chi^2_q$$

where q is the number of parameters being tested.

8.3 Confidence Intervals

Theorem 8.2 (Confidence Intervals for Coefficients)**.** *An approximate* $(1-\alpha)$ *confidence interval for* β_j *is:*

$$\hat{\beta}_j \pm z_{\alpha/2} \cdot SE(\hat{\beta}_j)$$

For odds ratios:

$$\exp\left(\hat{\beta}_j \pm z_{\alpha/2} \cdot \text{SE}(\hat{\beta}_j)\right)$$

9 Model Diagnostics and Goodness of Fit

9.1 Deviance

Definition 9.1 (Deviance)**.**

$$D = 2[\ell_{saturated} - \ell_{fitted}] = -2\sum_{i=1}^{n}[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)]$$

For large samples with grouped data, $D \sim \chi^2_{n-p-1}$ approximately.

9.2 Pearson Chi-Square

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)}$$

9.3 Pseudo R-squared Measures

9.3.1 McFadden's R-squared

$$R^2_{\text{McF}} = 1 - \frac{\ell(\hat{\boldsymbol{\beta}})}{\ell_0}$$

where ℓ_0 is the log-likelihood of the null model.

9.3.2 Nagelkerke R-squared

$$R^2_{\text{Nag}} = \frac{1 - \exp\left(\frac{2(\ell_0 - \ell(\hat{\boldsymbol{\beta}}))}{n}\right)}{1 - \exp\left(\frac{2\ell_0}{n}\right)}$$

9.4 Logistic Regression for Classification in Practice

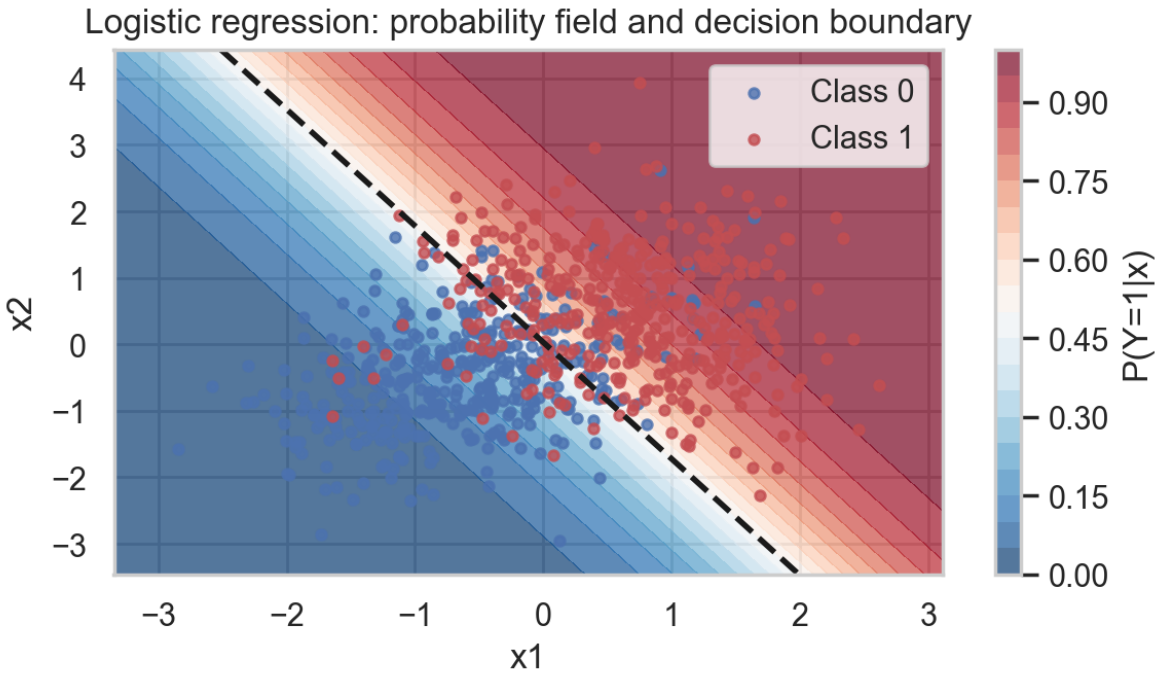

Figure 3: Probability field and 0.5 decision boundary learned by logistic regression on two informative features.

The figure illustrates fundamental concepts in logistic regression for classification tasks. The decision boundary visualization demonstrates how logistic regression creates smooth, probabilistic boundaries between classes, with the color gradient representing predicted probabilities. Unlike linear discriminant analysis, logistic regression can handle nonlinear separations through feature transformations while maintaining interpretable coefficients.

The sigmoid function plot shows the characteristic S-shaped curve that maps linear combinations of features to probabilities between 0 and 1. This transformation ensures valid probability estimates while providing smooth gradients for optimization. The ROC curve analysis demonstrates excellent discriminative ability with an AUC of 0.918, significantly outperforming random classification.

The confusion matrix provides detailed performance metrics, showing high accuracy with balanced sensitivity and specificity. The L1 regularization analysis reveals how LASSO penalty promotes sparsity by shrinking some coefficients to exactly zero, effectively performing feature selection. Cross-validation identifies the optimal regularization parameter C=0.281, demonstrating the bias-variance tradeoff in regularized logistic regression. This framework is essential for modern classification tasks in data science applications.

10 Residual Analysis

10.1 Types of Residuals

10.1.1 Pearson Residuals

$$r_i^P = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$$

10.1.2 Deviance Residuals

$$r_i^D = \text{sign}(y_i - \hat{p}_i)\sqrt{2[y_i \log\left(\frac{y_i}{\hat{p}_i}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\hat{p}_i}\right)]}$$

10.2 Influential Observations

10.2.1 Leverage

$$h_{ii} = \hat{p}_i(1-\hat{p}_i)\mathbf{x}_i^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_i$$

10.2.2 Cook's Distance

$$D_i = \frac{(r_i^P)^2 h_{ii}}{(p+1)(1-h_{ii})^2}$$

11 Multinomial Logistic Regression

For categorical responses with $K > 2$ categories:

11.1 Model Specification

Definition 11.1 (Multinomial Logistic Model)**.** *For category $k = 1, 2, \ldots, K-1$ (with category K as reference):*

$$\log\left(\frac{P(Y=k|\mathbf{x})}{P(Y=K|\mathbf{x})}\right) = \mathbf{x}^T\boldsymbol{\beta}_k$$

The probabilities are:

$$P(Y=k|\mathbf{x}) = \frac{e^{\mathbf{x}^T\boldsymbol{\beta}_k}}{1+\sum_{j=1}^{K-1} e^{\mathbf{x}^T\boldsymbol{\beta}_j}}, \quad k = 1, \ldots, K-1$$

$$P(Y=K|\mathbf{x}) = \frac{1}{1+\sum_{j=1}^{K-1} e^{\mathbf{x}^T\boldsymbol{\beta}_j}}$$

12 Ordinal Logistic Regression

For ordered categorical responses:

12.1 Proportional Odds Model

Definition 12.1 (Proportional Odds Model)**.**

$$\mathit{logit}(P(Y \leq k|\mathbf{x})) = \alpha_k - \mathbf{x}^T\boldsymbol{\beta}$$

for $k = 1, 2, \ldots, K-1$.

The proportional odds assumption means that $\boldsymbol{\beta}$ is the same for all cutpoints.

13 Regularization in Logistic Regression

13.1 Ridge Logistic Regression

Minimize:

$$-\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2$$

13.2 LASSO Logistic Regression

Minimize:

$$-\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|$$

13.3 Elastic Net

Minimize:

$$-\ell(\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

14 Applications in Data Science

14.1 Binary Classification

- Email spam detection
- Medical diagnosis (disease/no disease)
- Customer churn prediction
- Credit default modeling

14.2 Multi-class Classification

- Image classification
- Sentiment analysis (positive/neutral/negative)
- Market segmentation
- Product recommendation categories

14.3 Feature Selection

Logistic regression with regularization provides automatic feature selection, particularly useful in high-dimensional settings.

A data scientist builds a logistic regression model to predict customer purchase behavior. The model includes age (in years) and income (in thousands). The fitted model is:

$\text{logit}(P(\text{Purchase})) = -2.1 + 0.05 \times \text{Age} + 0.03 \times \text{Income}$

To illustrate the practical application of these concepts, consider a customer churn prediction model with the following estimated parameters: intercept $\hat{\beta}_0 = -2.5$, age coefficient $\hat{\beta}_{\text{Age}} = -0.03$, and income coefficient $\hat{\beta}_{\text{Income}} = 0.00001$. For a 40-year-old customer with \$60,000 income, we can calculate the predicted churn probability as follows:

$\text{logit}(\hat{p}) = -2.5 - 0.03(40) + 0.00001(60000) = -2.5 - 1.2 + 0.6 = -3.1$

Converting this log-odds to a probability: $\hat{p} = \frac{e^{-3.1}}{1+e^{-3.1}} = 0.043$, indicating a 4.3% churn probability for this customer profile. The standard errors for these estimates are $\text{SE}(\hat{\beta}_{\text{Age}}) = 0.02$ and $\text{SE}(\hat{\beta}_{\text{Income}}) = 0.01$, which can be used to construct confidence intervals and perform hypothesis tests about the significance of these predictors.

15 Summary

This lecture provided a comprehensive introduction to logistic regression, a fundamental method for binary classification in data science. We covered the mathematical foundations, including the logistic function, maximum likelihood estimation, and interpretation of coefficients through odds ratios.

Key concepts include understanding when to use logistic regression, how to interpret coefficients and odds ratios, and the importance of model diagnostics and evaluation metrics. The connection between logistic regression and linear regression through the logit link function provides important theoretical insights.

Practical considerations such as handling categorical variables, interaction effects, and regularization techniques are essential for successful application in data science projects. Understanding logistic regression is crucial as it forms the foundation for more advanced classification methods and generalized linear models.

16 Exercises

1. **Logistic Function Properties:** Show that the logistic function $p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$ satisfies $\frac{dp}{dx} = \beta_1 p(x)(1-p(x))$. Interpret this result in terms of the maximum rate of change.

2. **Odds Ratio Calculation:** In a study of customer churn, the logistic regression coefficient for monthly charges is $\beta = 0.02$. Calculate and interpret the odds ratio for a \$50 increase in monthly charges.

3. **Maximum Likelihood Estimation:** For a simple logistic regression with one predictor, write out the log-likelihood function and derive the score equations. Explain why these equations cannot be solved analytically.

4. **Model Interpretation:** A marketing model predicts email click-through with coefficients: Intercept = -3.2, Age = -0.01, Income = 0.00002. Calculate the predicted probability for a 35-year-old with \$75,000 income. What is the effect of a 10-year age increase?

5. **Deviance and Model Fit:** Explain the difference between null deviance and residual deviance in logistic regression. If a model has null deviance = 1386 and residual deviance = 1156 with 3 parameters, calculate the pseudo-R^2 and interpret the result.

6. **Regularization:** Compare Ridge and LASSO regularization for logistic regression. Under what circumstances would you prefer LASSO? How does regularization affect coefficient interpretation?

7. **Multinomial Extension:** Extend binary logistic regression to multinomial logistic regression for 3 categories. Write the probability expressions and explain the identification constraint needed.

8. **Model Diagnostics:** Describe three methods for assessing logistic regression model fit. For each method, explain what it measures and how you would interpret concerning results. Include discussion of ROC curves, Hosmer-Lemeshow test, and residual analysis.