Logistic regression and Generalized Linear Models

Gioele La Manno

École Polytechnique Fédérale de Lausanne (EPFL)

School of Life Science (SV)

April 2025

EPFL - BMI - UPLAMANNO

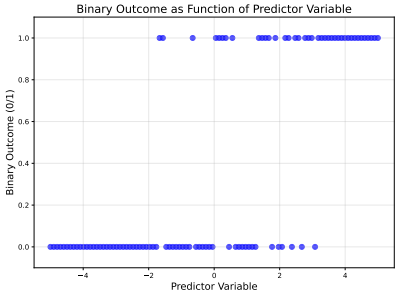
Contents

Introduction to Generalized Linear Models

Beyond Linear Regression

Linear regression is limited when modeling:

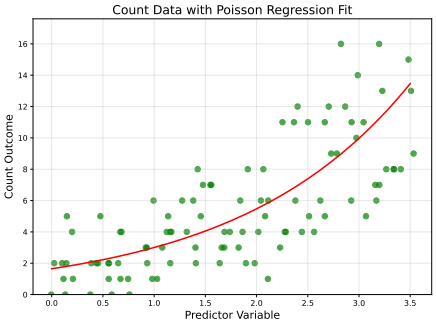
- Binary outcomes (differentiation/quiescence)
- Count data (sequencing reads)
- Strictly positive measurements



The Structure of GLMs

GLMs have three components:

1. **Random Component:** Distribution of response variable
2. **Systematic Component:** $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$
3. **Link Function:** $g(E[Y]) = \eta$



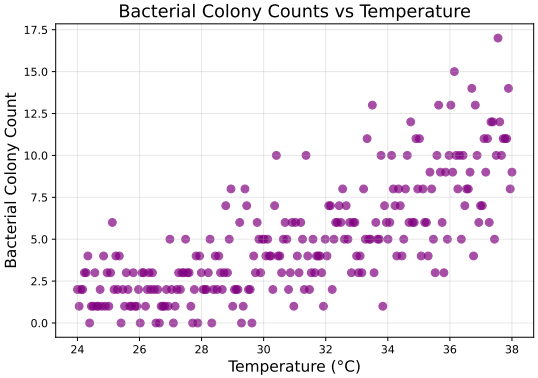
The Role of Link Functions

Link functions transform between bounded response space and unbounded linear space.

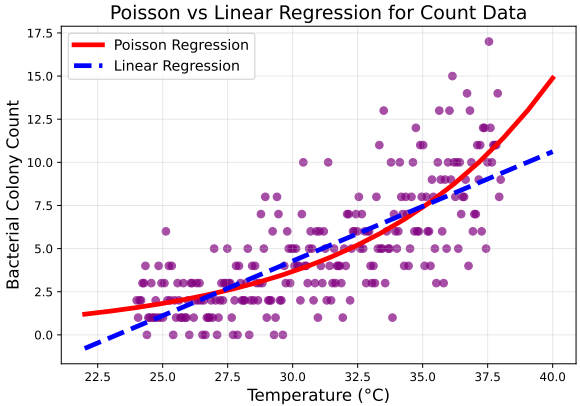
- **Normal:** Identity link $g(\mu) = \mu$
- **Binomial:** Logit link $g(\mu) = \log \frac{\mu}{1-\mu}$
- **Poisson:** Log link $g(\mu) = \log \mu$
- **Gamma:** Inverse link $g(\mu) = \frac{1}{\mu}$

Poisson Regression: Modeling Count Data

Ideal for count data like bacterial colonies or sequencing reads.

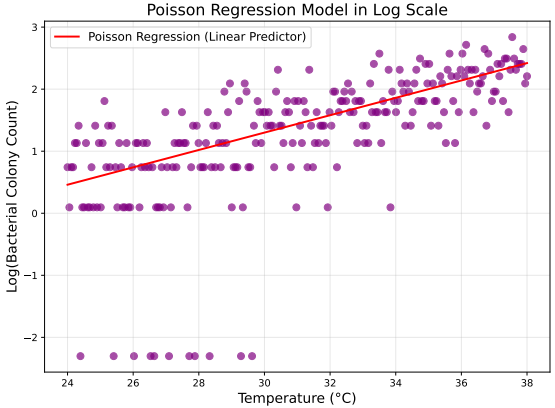


Poisson vs. Linear Regression for Count Data



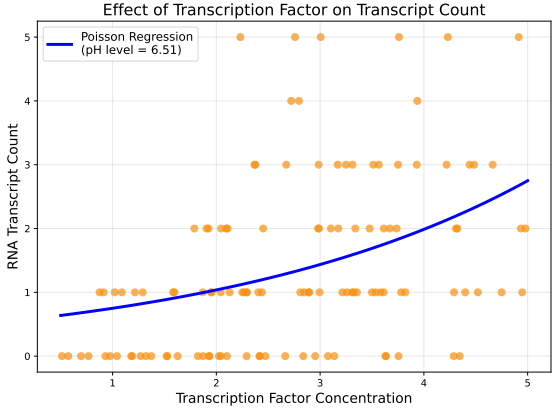
Poisson regression ensures positive predictions and models increasing variance.

Log Link in Poisson Regression



The log link creates a linear relationship on the log scale.

Marginal Effects in Poisson Regression

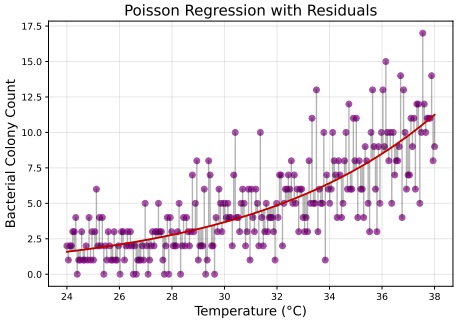


The curve shows multiplicative effects on the response.

Heteroscedasticity in Count Data

Poisson regression handles increasing variance with the mean:

$$\text{Var}(Y) = E[Y] = \lambda$$



Logistic Regression: Modeling Binary Outcomes

From Regression to Classification

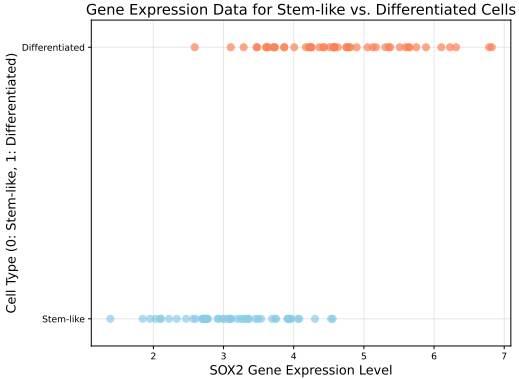
Logistic regression addresses a classification problem:

- Predicting stem cell differentiation
- Diagnosing disease from biomarkers
- Classifying protein sequences

Instead of predicting classes directly, logistic regression models the probability of belonging to the positive class.

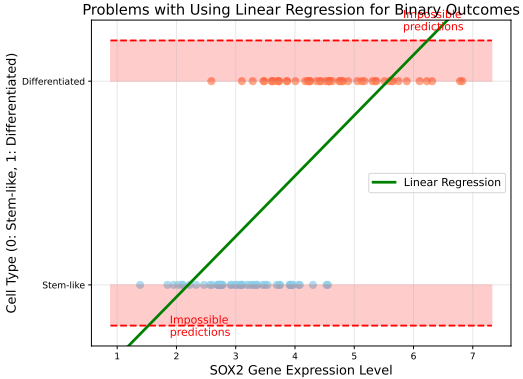
Binary Outcomes in Biology

Binary nature of the data: each observation belongs to exactly one class.



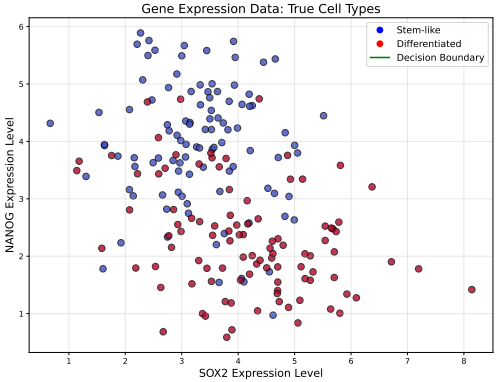
Problems with Linear Regression for Binary Data

Linear regression is problematic for binary outcomes:



Classification in Higher Dimensions

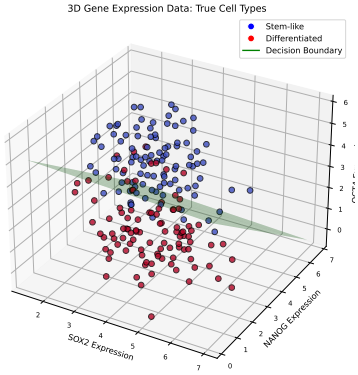
In 2D, we look for a line that separates classes.



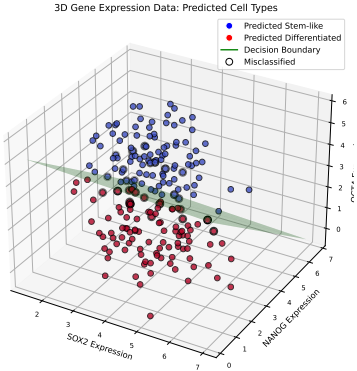
Predicted Classification Boundaries

3D Classification Problems

In 3D, we look for a separating plane.



3D Classification with Decision Boundary



Logistic Regression Structure

Logistic regression follows the GLM structure:

1. **Random Component:** Bernoulli distribution

$$Y \sim \text{Bernoulli}(p)$$

2. **Systematic Component:**

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

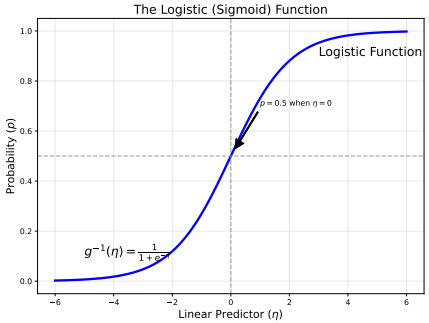
3. **Link Function:** Logit link

$$g(p) = \log\left(\frac{p}{1-p}\right) = \eta$$

The Logistic Function

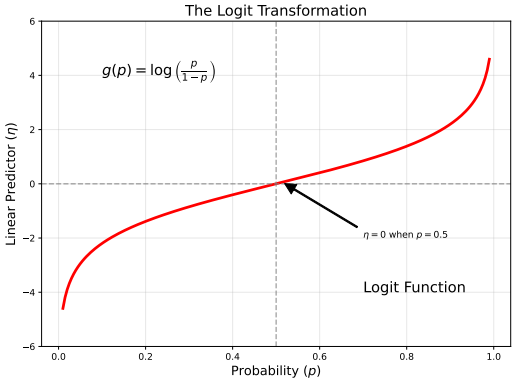
The inverse link transforms from the linear scale to probabilities:

$$p = \frac{1}{1 + e^{-\eta}}$$



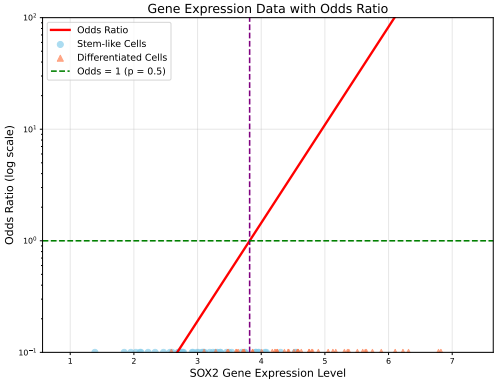
The Logit Transformation

Converting from probability to linear scale:



Logistic Regression Fit

Odds Ratio Interpretation



Interpreting Regression Coefficients

In logistic regression, coefficients are interpreted in terms of odds ratios:

- **Odds:** odds $= \frac{p}{1-p}$
- **Logistic model:** $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots$
- **For coefficient interpretation:** e^{β_j} is the multiplicative effect on odds when X_j increases by one unit

Example: If $\beta_1 = -0.8$ for Oct4, then $e^{-0.8} \approx 0.45$, meaning a one-unit increase in Oct4 reduces differentiation odds by 55

Evaluating Classification Models

From Probabilities to Decisions

To convert probabilities to binary predictions, we apply a threshold:

$$\hat{y}_i = \begin{cases} 1 & \text{if } \hat{p}_i \geq c \\ 0 & \text{if } \hat{p}_i < c \end{cases}$$

Typically $c = 0.5$, but the optimal threshold depends on the relative costs of false positives and false negatives.

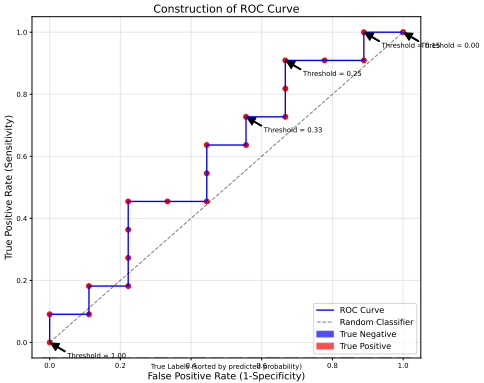
The Confusion Matrix

	Actual Positive	**Actual Negative**
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

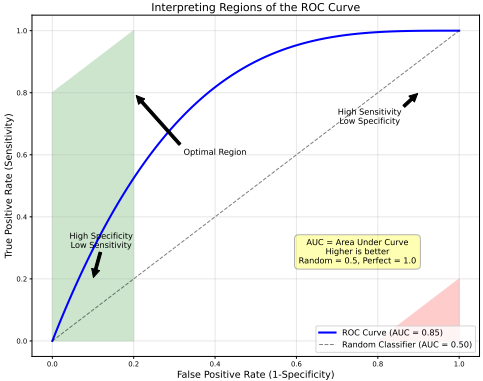
Key metrics:

- **Accuracy** $= \frac{TP+TN}{TP+FP+FN+TN}$
- **Sensitivity** $= \frac{TP}{TP+FN}$
- **Specificity** $= \frac{TN}{TN+FP}$
- **Precision** $= \frac{TP}{TP+FP}$

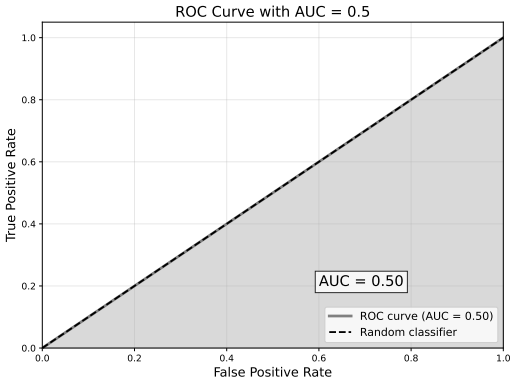
ROC Curve Construction



ROC Curve Interpretation

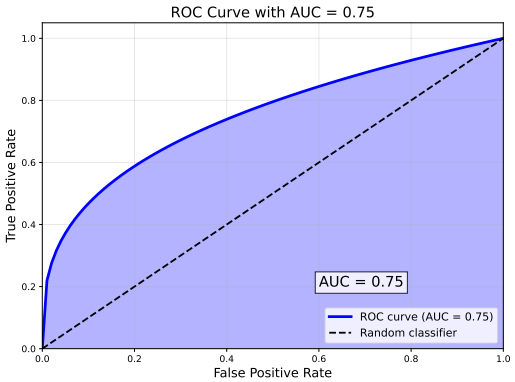


ROC Curves and AUC: Random Classifier



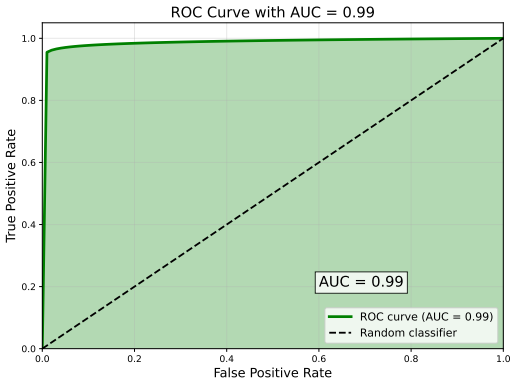
AUC = 0.5: No discriminative ability (equivalent to random guessing)

ROC Curves and AUC: Decent Classifier



AUC = 0.75: Good discriminative ability

ROC Curves and AUC: Excellent Classifier



AUC = 0.99: Nearly perfect discrimination