# Generalized Linear Model & Logistic Regression
## Fundamental Techniques in Data Science

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

**Utrecht University**

# Outline

Generalized Linear Model

Logistic Regression

Classification

Evaluating Classification Performance

# General Linear Model

So far, we've been discussing models with this form:

$$Y = \beta_0 + \sum_{p=1}^{P} \beta_p X_p + \varepsilon$$

This type of model is known as the *general linear model*.

- All flavors of linear regression are general linear models.
  - ANOVA
  - ANCOVA
  - Multilevel linear regression models

# Components of the General Linear Model
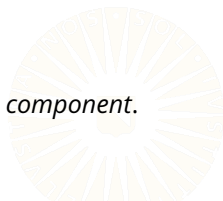
We can break our model into pieces:

$$\eta = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$$

$$Y = \eta + \varepsilon$$

Because $\varepsilon \sim \mathrm{N}(0, \sigma^2)$, we can also write:

$$Y \sim \mathrm{N}(\eta, \sigma^2)$$

In this representation:

- $\eta$ is the *systematic component* of the model
- The normal distribution, $\mathrm{N}(\cdot, \cdot)$, is the model's *random component*.

# Components of the General Linear Model

The purpose of general linear modeling (i.e., regression modeling) is to build a model of the outcome's mean, $\mu_Y$.

- In this case, $\mu_Y = \eta$.
- The systematic component defines the mean of $Y$.

The random component quantifies variability (i.e., error variance) around $\mu_Y$.

- In the general linear model, we assume that this error variance follows a normal distribution.
- Hence the normal random component.

# GENERALIZED LINEAR MODEL

# Extending the General Linear Model

We can generalize the models we've been using in two important ways:

1. Allow for random components other than the normal distribution.
2. Allow for more complicated relations between $\mu_Y$ and $\eta$.
   - Allow: $g(\mu_Y) = \eta$

These extensions lead to the class of *generalized linear models* (GLMs).

# Components of the Generalized Linear Model

The random component in a GLM can be any distribution from the so-called *exponential family*.

- The exponential family contains many popular distributions:
  - Normal
  - Binomial
  - Poisson
  - Many others...

The systematic component of a GLM is exactly the same as it is in general linear models:

$$\eta = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$$

# Link Functions

In GLMs, $\eta$ does not directly describe $\mu_Y$.

- We first transform $\mu_Y$ via a *link function*.
- $g(\mu_Y) = \eta$

The link function allows GLMs for outcomes with restricted ranges without requiring any restrictions on the range of the $\{X_p\}$.

- For strictly positive $Y$, we can use a *log link*:

$$\ln(\mu_y) = \eta.$$

- The general linear model employs the *identity link*:

$$\mu_y = \eta.$$

# Components of the Generalized Linear Model

Every GLM is built from three components:

1. The systematic component, $\eta$.
   - A linear function of the predictors, $\{X_p\}$.
   - Describes the association between $\mathbf{X}$ and $Y$.

2. The link function, $g(\mu_Y)$.
   - Transforms $\mu_Y$ so that it can take any value on the real line.

3. The random component, $P(Y|g^{-1}(\eta))$
   - The distribution of the observed $Y$.
   - Quantifies the error variance around $\eta$.

# General Linear Model $\subset$ Generalized Linear Model

The general linear model is a special case of GLM.

1. Systematic component:

$$\eta = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$$

2. Link function:

$$\mu_Y = \eta$$

3. Random component:

$$Y \sim \mathrm{N}(\eta, \sigma^2)$$

# LOGISTIC REGRESSION

# Logistic Regression

So why do we care about the GLM when linear regression models have worked thus far?
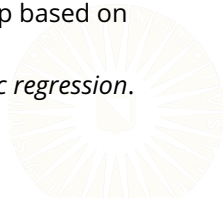
- In a word: Classification.

In the classification task, we have a discrete, qualitative outcome.

- We will begin with the situation of two-level outcomes.
  - Alive or Dead
  - Pass or Fail
  - Pay or Default

We want to build a model that predicts class membership based on some set of interesting features.
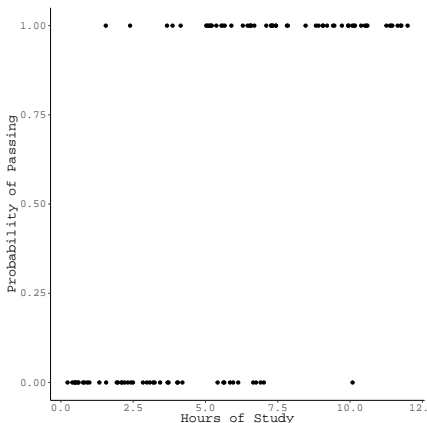
- To do so, we will use a very useful type of GLM: *logistic regression*.

# Classification Example

Suppose we want to know the effect of study time on the probability of passing an exam.
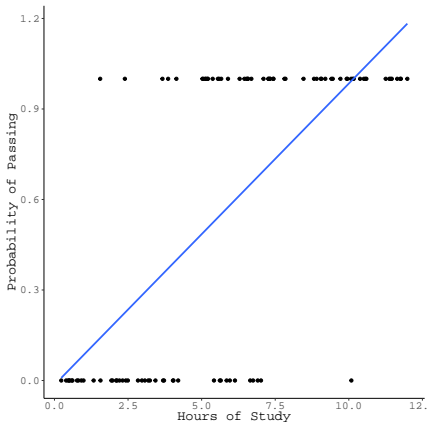
- The probability of passing must be between 0 and 1.

- We care about the probability of passing, but we only observe absolute success or failure.

  ○ $Y \in \{1, 0\}$

# Linear Regression for Binary Outcomes?

What happens if we try to model these data with linear regression?

• Hmm...notice any problems?

# Logistic Regression Visualized

We get a much better model using logistic regression.

- The link function ensures legal predicted values.

- The sigmoidal curve implies fluctuation in the effectiveness of extra study time.

  - More study time is most beneficial for students with around 5.5 hours of study.
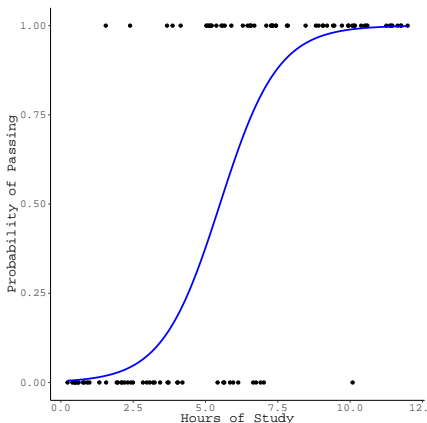
# Probabilities, Odds, & Odds-Ratios

In 2017, 2535 people participated in the *Ultra-Trail du Mont-Blanc*, but only 66.55% finished the race.

- Below, you can find a cross-tabulation of finishing status and sex.

|        | Finish |      |
| ------ | ------ | ---- |
| Sex    | No     | Yes  |
| Female | 95     | 147  |
| Male   | 753    | 1540 |

- What is the *probability* of finishing for each sex?

- What are the *odds* of finishing for each sex?

- What is the *odds ratio* of finishing for males vs. females?

# Defining the Logistic Regression Model

In logistic regression problems, we are modeling binary data:

- Usual coding: $Y \in \{1 = \text{"Success"}, 0 = \text{"Failure"}\}$.

The *Binomial* distribution is a good way to represent this kind of data.

- The systematic component in our logistic regression model will be the binomial distribution.

The mean of the binomial distribution (with $N = 1$) is the "success" probability, $\pi = P(Y = 1)$.

- We are interested in modeling $\mu_Y = \pi$:

$$g(\pi) = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$$

# Link Function for Logistic Regression

Because $\pi$ is bounded by 0 and 1, we cannot model it directly—we must apply an appropriate link function.

- Logistic regression uses the *logit link*.

- Given $\pi$, we can define the *odds* of success as:

$$O_s = \frac{\pi}{1 - \pi}$$

- Because $\pi \in [0, 1]$, we know that $O_s \geq 0$.

- We take the natural log of the odds as the last step to fully map $\pi$ to the real line.

$$\text{logit}(\pi) = \ln\left(\frac{\pi}{1 - \pi}\right)$$

# Fully Specified Logistic Regression Model

Our final logistic regression model is:

$$Y \sim \text{Bin}(\pi, 1)$$

$$\text{logit}(\pi) = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$$

The fitted model can be represented as:

$$\text{logit}(\hat{\pi}) = \hat{\beta}_0 + \sum_{p=1}^{P} \hat{\beta}_p X_p$$

The fitted coefficients, $\{\hat{\beta}_0, \hat{\beta}_p\}$, are interpreted in units of *log odds*.

# Logistic Regression Example

If we fit a logistic regression model to the test-passing data plotted above, we get:

$$\text{logit}(\hat{\pi}_{pass}) = -3.414 + 0.683 X_{study}$$

- A student who does not study at all has -3.414 log odds of passing the exam.
- For each additional hour of study, a student's log odds of passing increase by 0.683 units.

Log odds do not lend themselves to interpretation.

- We can convert the effects back to an odds scale by exponentiation.
- $\hat{\beta}$ has log odds units, but $e^{\hat{\beta}}$ has odds units.
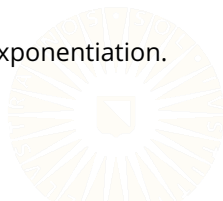
## Interpretations

Exponentiating the coefficients also converts the additive effects to multiplicative effects.

- $\ln(AB) = \ln(A) + \ln(B)$
- We can interpret $\hat{\beta}$ as we would in linear regression:
  - A unit change in $X_p$ produces an expected change of $\hat{\beta}_p$ units in $\text{logit}(\pi)$.
- After exponentiation, however, unit changes in $X_p$ imply multiplicative changes in $O_s = \pi/(1 - \pi)$.
  - A unit change in $X_p$ results in multiplying $O_s$ by $e^{\hat{\beta}_p}$.

## Interpretations

Exponentiating the coefficients in our toy test-passing example produces the following interpretations:

- A student who does not study is expected to pass the exam with odds of 0.033.
- For each additional hour a student studies, their odds of passing increase by 1.98 *times*.
  - Odds of passing are *multiplied* by 1.98 for each extra hour of study.

# Interpretations

Exponentiating the coefficients in our toy test-passing example produces the following interpretations:

- A student who does not study is expected to pass the exam with odds of 0.033.
- For each additional hour a student studies, their odds of passing increase by 1.98 *times*.
  - Odds of passing are *multiplied* by 1.98 for each extra hour of study.

Due to the confusing interpretations of the coefficients, we often focus on the valance of the effects:

- Additional study time is associated with increased odds of passing.
- $\hat{\beta}_p > 0$ = "Increased Success", $e^{\hat{\beta}_p} > 1$ = "Increased Success"
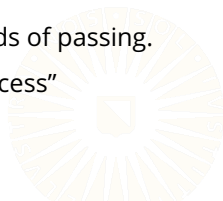
# Multiple Logistic Regression

The preceding example was a *simple logistic regression*.

- Including multiple predictor variables in the systematic component leads to *multiple logistic regression*.
- The relative differences between simple logistic regression and multiple logistic regression are the same as those between simple linear regression and multiple linear regression.
  - The only important complication is that the regression coefficients become partial effects.

# Multiple Logistic Regression Example

Suppose we want to predict the probability of a patient having "high" blood glucose from their age, BMI, and average blood pressure.
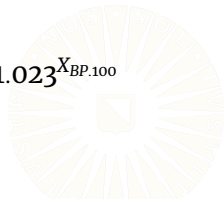
- We could do so with the following model:

$$\text{logit}(\pi_{hi.gluc}) = \beta_0 + \beta_1 X_{age.40} + \beta_2 X_{BMI.25} + \beta_3 X_{BP.100}$$

- By fitting this model to our usual "diabetes" data we get:

$$\text{logit}(\hat{\pi}_{hi.gluc}) = -0.155 + 0.035 X_{age.40} + 0.107 X_{BMI.25} + 0.023 X_{BP.100}$$

- Exponentiating the coefficients produces:

$$\frac{\hat{\pi}_{hi.gluc}}{1 - \hat{\pi}_{hi.gluc}} = 0.857 \times 1.035^{X_{age.40}} \times 1.113^{X_{BMI.25}} \times 1.023^{X_{BP.100}}$$

# Exponentiating the Systematic Component

$$\text{logit}(\hat{\pi}_{hi.gluc}) = -0.155 + 0.035X_{age.40} + 0.107X_{BMI.25} + 0.023X_{BP.100}$$

$$e^{\text{logit}(\hat{\pi}_{hi.gluc})} = e^{(-0.155 + 0.035X_{age.40} + 0.107X_{BMI.25} + 0.023X_{BP.100})}$$

$$\frac{\hat{\pi}_{hi.gluc}}{1 - \hat{\pi}_{hi.gluc}} = e^{-0.155} \times e^{0.035X_{age.40}} \times e^{0.107X_{BMI.25}} \times e^{0.023X_{BP.100}}$$

$$= (e^{-0.155}) \times (e^{0.035})^{X_{age.40}} \times (e^{0.107})^{X_{BMI.25}} \times (e^{0.023})^{X_{BP.100}}$$

$$= 0.857 \times 1.035^{X_{age.40}} \times 1.113^{X_{BMI.25}} \times 1.023^{X_{BP.100}}$$

# CLASSIFICATION

# Predictions from Logistic Regression

Given a fitted logistic regression model, we can get predictions for new observations of $\{X_p\}$, $\{X'_p\}$.

- Directly applying $\{\hat{\beta}_0, \hat{\beta}_p\}$ to $\{X'_p\}$ will produce predictions on the scale of $\eta$:

$$\hat{\eta}' = \hat{\beta}_0 + \sum_{p=1}^{P} \hat{\beta}_p X'_p$$

- By applying the inverse link function, $g^{-1}(\cdot)$, to $\hat{\eta}'$, we get predicted success probabilities:

$$\hat{\pi}' = g^{-1}(\hat{\eta}')$$

# Predictions from Logistic Regression

In logistic regression, the inverse link function, $g^{-1}(\cdot)$, is the *logistic function*:

$$\text{logistic}(X) = \frac{e^X}{1 + e^X}$$

So, we convert $\hat{\eta}'$ to $\hat{\pi}'$ by:

$$\hat{\pi}' = \frac{e^{\hat{\eta}'}}{1 + e^{\hat{\eta}'}} = \frac{\exp\left(\hat{\beta}_0 + \sum_{p=1}^{P} \hat{\beta}_p X_p'\right)}{1 + \exp\left(\hat{\beta}_0 + \sum_{p=1}^{P} \hat{\beta}_p X_p'\right)}$$

# Classification with Logistic Regression

Once we have computed the predicted success probabilities, $\hat{\pi}'$, we can use them to classify new observations.

- By choosing a threshold on $\hat{\pi}'$, say $\hat{\pi}' = t$, we can classify the new observations as "Successes" or "Failures":

$$\hat{Y}' = \begin{cases} 1 & if \quad \hat{\pi}' \geq t \\ 0 & if \quad \hat{\pi}' < t \end{cases}$$
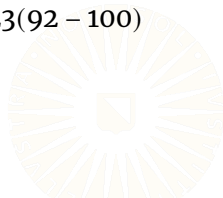
# Classification Example

Say we want to classify a new patient into either the "high glucose" group or the "not high glucose" group using the model fit above.

- Assume this patient has the following characteristics:
    - They are 57 years old
    - Their BMI is 28
    - Their average blood pressure is 92

First we plug their predictor data into the fitted model to get their model-implied $\eta$:

$$\hat{\eta} = -0.155 + 0.035(57 - 40) + 0.107(28 - 25) + 0.023(92 - 100)$$
$$= 0.572$$

# Classification Example

Next we convert the predicted $\eta$ value into a model-implied success probability by applying the logistic function:

$$\frac{e^{0.572}}{1 + e^{0.572}} = 0.639$$

Finally, to make the classification, assume a threshold of $\hat{\pi}' = 0.5$ as the decision boundary.

- Because $0.639 > 0.5$ we would classify this patient into the "high glucose" group.

# Evaluating Classification Performance

# Confusion Matrix

One of the most direct ways to evaluate classification performance is to tabulate the true and predicted classes.

- Such a cross-tabulation is called a *confusion matrix*.

|       | Predicted | |
|-------|-----|------|
| True  | Low | High |
| Low   | 123 | 82   |
| High  | 62  | 175  |

Confusion Matrix of Blood Glucose Level

# Confusion Matrix

One of the most direct ways to evaluate classification performance is to tabulate the true and predicted classes.

- Such a cross-tabulation is called a *confusion matrix*.

|        | Predicted |      |
|--------|-----------|------|
| True   | Low       | High |
| Low    | 123       | 82   |
| High   | 62        | 175  |

Confusion Matrix of Blood Glucose Level

We can summarize the confusion matrix in many ways.

- Different summaries highlight different aspects of the classifier's performance.

# Summarizing the Confusion Matrix

Sensitivity (Recall, Hit Rate, True-Positive Rate):

$$Sensitivity = \frac{True\ Postives}{True\ Positives + False\ Negatives} = \frac{True\ Postives}{Total\ Positives}$$

Specificity (Selectivity, True-Negative Rate):

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} = \frac{True\ Negatives}{Total\ Negatives}$$

# Summarizing the Confusion Matrix

Accuracy:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{TP + TN + FP + FN} = \frac{Correct\ Classifications}{Total\ Cases}$$
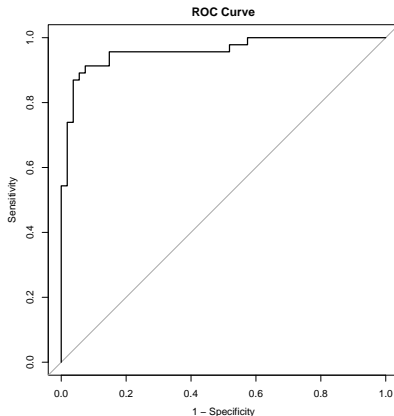
Error Rate:

$$Error\ Rate = \frac{False\ Positives + False\ Negatives}{TP + TN + FP + FN} = \frac{Incorrect\ Classifications}{Total\ Cases}$$

# ROC Curve

We can visualize a classifier's potential performance via a *Receiver Operating Characteristic* Curve.

- Y-Axis: True-Positive Rate
  - Sensitivity

- X-Axis: False-Positive Rate
  - 1 - Specificity

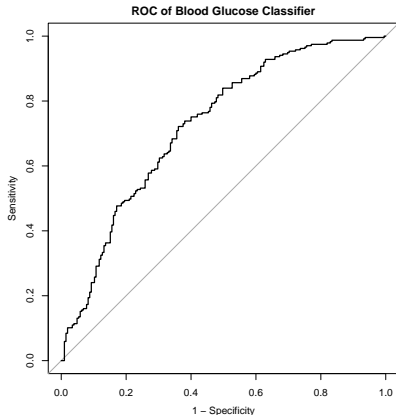- Area Under the ROC Curve (AUC) summarizes the classifier's discrimination

# Example

$$Sensitivity = \frac{175}{175 + 62} = 0.738$$

$$Specificity = \frac{123}{123 + 82} = 0.6$$

$$AUC = 0.725$$



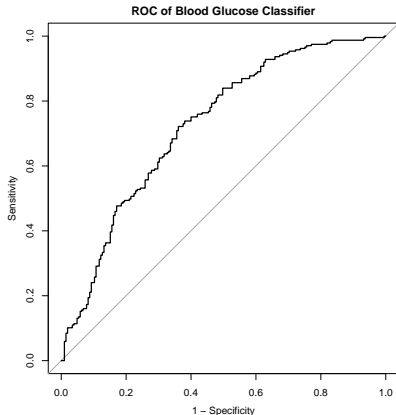ROC of Blood Glucose Classifier

# Example

$$Sensitivity = \frac{175}{175 + 62} = 0.738$$

$$Specificity = \frac{123}{123 + 82} = 0.6$$

$$AUC = 0.725$$



ROC of Blood Glucose Classifier

What is the accuracy? What is the error rate?