Day 3: GLMs for Binary Data

Outline

- GLM Overview
- GLMs for Binary Data
- Logistic Regression
- Model Checking

Regression Components

Recall that we described regression models as having the form:

$$E[Y] = \mu = \phi(X_1, X_2, \ldots, X_p)$$

Now, we can rearrange this model into three parts:

- Y is a random variable distributed with mean μ, so $E[Y] = \mu$.
- Linear function ϕ produces a value η, so now $\eta = \phi(X_1, X_2, \ldots, X_n)$.
- η, is equal to some function $g(\cdot)$ of the expected values of the response variable, μ, so $\eta = g(\mu)$.

Linear Regression

In the linear regression setting, we assume:

- $Y \sim \mathcal{N}(\mu, \sigma^2)$
- $\phi(X_1, X_2, \ldots, X_n) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$
- $g(\mu) = \mu$, so $\eta = \mu$

This gives us the familiar model form:

$$E[Y] = \mu = \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

Generalized Linear Models

GLMs relax the assumptions made for linear regression models:

- Y is described by some distribution in the exponential family
- $g(\mu)$ can take other functional forms.

The remainder of our sessions will cover commonly used cases within this framework.

Binary Response Data

In cases where we have a response variable that takes a binary outcome, we can no longer assume normality.

- Now, $Y_i \sim \text{Bin}(n, \pi_i)$.
- We still have $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$.
- What should $g(\cdot)$ be?

Exponential Family of Distributions

In the GLM context, it is helpful to express the likelihood in exponential family form

$$f(y_i|\theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\},$$

where

- θ_i is the natural parameter
- ϕ represents a dispersion parameter

Once the likelihood is in this form, one common link function choice is to set

$$g(\mu) = \theta = \eta.$$

This is known as the canonical link function.

Logit Link Function

With a binomial likelihood, we have

$$
\begin{aligned}
f(y_i|n,\pi_i) &= \binom{n}{y_i}\pi_i^{y_i}(1-\pi_i)^{n-y_i} \\
&\propto \pi_i^{y_i}(1-\pi_i)^{n-y_i} \\
&= \text{exp}\left\{y_i\text{log}(\pi_i) + (n-y_i)\text{log}(1-\pi_i)\right\} \\
&= \text{exp}\left\{y_i\left[\text{log}(\pi_i) - \text{log}(1-\pi_i)\right] + n\text{log}(1-\pi_i)\right\} \\
&= \text{exp}\left\{y_i\text{log}\left(\frac{\pi_i}{1-\pi_i}\right) + n\text{log}(1-\pi_i)\right\}
\end{aligned}
$$

We see that the natural parameter is equal to $\theta = \text{log}\left(\frac{\pi}{1-\pi}\right)$, so we use the link function

$$
g(\pi) = \theta = \eta = \text{log}\left(\frac{\pi}{1-\pi}\right) = \text{logit}(\pi).
$$

Logistic Regression

Using the information above, logistic regression takes the form

$$\eta = \mathsf{logit}(\pi) = \mathsf{log}\left(\frac{\pi}{1-\pi}\right) = \sum_{i=1}^{p} \beta_i X_i.$$

By inverting the link function, we have

$$\pi = \frac{\mathsf{exp}\left(\sum_{i=1}^{p} \beta_i X_i\right)}{1 + \mathsf{exp}(\sum_{i=1}^{p} \beta_i X_i)}.$$

Interpreting Coefficients

Unlike linear regression, the coefficients in logistic regression models are interpreted multiplicatively.

If we have a simple model

$$\text{logit}(\pi) = \alpha + \beta X,$$

then with $\pi_1 = P(Y = 1|X = 1)$ and $\pi_0 = P(Y = 1|X = 0)$,

$$\begin{aligned}\beta &= \log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\left(\frac{\pi_0}{1-\pi_0}\right)\\ &= \log\left(\frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}\right),\end{aligned}$$

β can be interpreted as a log odds ratio, and e^β is the odds ratio.

Other Link Functions

It is not necessary to use the canonical link function, only convenient. Other link functions commonly used for binary data are:

- Probit (inverse normal) function: $g(\pi) = \Phi^{-1}(\pi)$
- Complementary log-log function: $g(\pi) = \log\{-\log(1-\pi)\}$
- Log-log function: $g(\pi) = -\log\{-\log(\pi)\}$

Which link function you choose should depend on model fit, but the most common of these alternatives is the probit function.

Interpreting model coefficients is usually more difficult with these link functions.

Model Fitting

Unlike OLS models, it is usually impossible to find analytic solutions for GLM coefficients.

The most common approach is to use the Newton-Raphson algorithm to find the maximum-likelihood parameter estimates iteratively.

Model Checking

Checking the model fit is also much more difficult with GLMs. For binary data, there are a few common options:

- Model accuracy
- Sensitivity and Specificity / ROC Curve
- Alternative goodness of fit tests

Model Accuracy

Model accuracy is a measure of how frequently the model correctly predicts the response. This is simple to explain, but there are some problems with this approach:

- Unbalanced data can present issues
- Sometimes we may want to prioritize minimizing false positive or false negatives

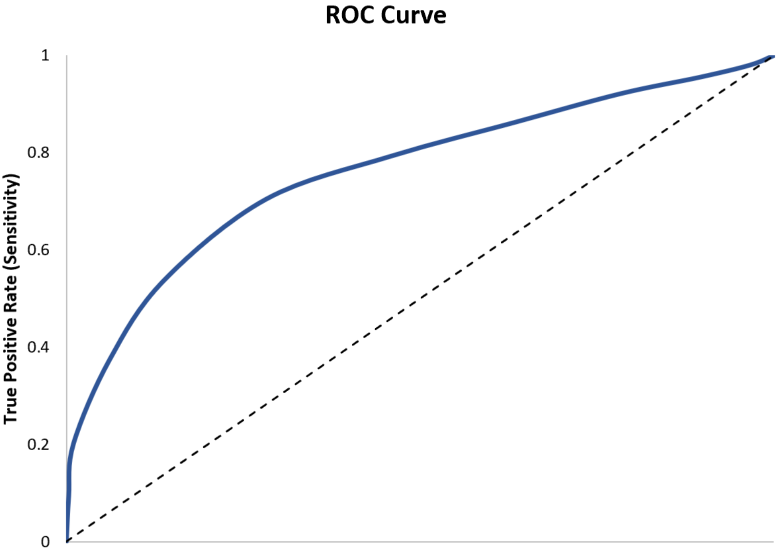
Sensitivity and Specificity

- Sensitivity: Probability of predicting a positive result given that the true outcome is positive
- Specificity: Probability of predicting a negative result given that the true outcome is negative

This approach allows us to avoid trouble from unbalanced datasets.

ROC Curve and AUC

The receiver operating characteristic (ROC) curve provides a way to check the benefit of prioritizing sensitivity versus specificity.



ROC Curve and AUC

The Area Under the ROC Curve (AUC) can be used to summarize the results shown and compare between models.

- AUC closer to 1 shows that the model needs to make less of a tradeoff between sensitivity and specificity.

ROC functions are included in the `pROC` and `verification` packages in R.

Goodness of Fit Tests

The Homer-Lemeshow Test is the most common statistical test for model fit:

- Break observations into groups based on deciles of fitted risk values
- Compare observed event rates with the predicted event rates
- If the two rates are close, then the model is a good fit