Generalized Linear Models

Goal: To find the best fitting and most parsimonious, clinically interpretable model to describe the relationship between an outcome (dependent or response) variable and a set of independent predictor variables.

In the context of regression models, a parsimonious model is one that accomplishes a desired level of explanation or prediction with as few predictor variables as possible.

The independent variables are often called predictors, explanatory variables, or covariates.

Why Models?

(1) Structural form of the model describes the patterns of association and interaction

(2) Sizes of model parameters determine the strength and importance of effects

(3) Inferences about parameters evaluate which explanatory variables are truly associated with the response variable after controlling for potential confounders

(4) The model's predicted values smooth the data and provide improved estimates.

Components of a Generalized Linear Model (GLM)

Random component, linear predictor, link function

Random Component

The random component of a GLM identifies the response variable Y and its associated probability distribution. Suppose that we observe $Y_1, Y_2, \ldots, Y_n$ (independent).

Type of Observation	**Probability Distribution**
Binary (success/failure) or # successes in n trials	Binomial
Counts	Poisson or negative binomial
Continuous	Normal

Linear Predictor

The linear predictor of a GLM specifies the set of explanatory variables

$$\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

Note that some of the terms in the model can be functions of others. For example, we could include an interaction term like $x_3 = x_1 x_2$.

Link Function

The mean / expected value of Y has a value that varies according to the values of the explanatory variables. The link function connects the random component of the model (Y) with the linear predictor function of explanatory variables.

For example, in linear regression, our model is:

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

This is called the identity link function (i.e. the function $Y = Y$). But it's not the only option! Remember how we talked about variable transformation on Y, usually in the context of variance stabilizing transformation.

Set $Y = \log(Y)$ or $Y = \sqrt{Y}$. In this situation, our function is of the form:

$$\log(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

or

$$\sqrt{Y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

These represent different link functions, both of which are nonlinear.

Definition: A *loglinear* model uses a log link function ($Y = \log(Y)$) Since logarithms only apply to positive numbers, a loglinear model is most appropriate when Y cannot be negative (e.g. count data).

Definition: A logistic regression model uses a *logit link* function

A logit link function models the log of an odds, and is most appropriate when Y is between 0 and 1 (e.g probabilities)
• For the binomial distribution, the "Y" that we want to model is actually the probability of success, so the model is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

Note that linear and logistic regression are both special cases of generalized linear models!

• Linear regression uses a normal random component and the identity link function, but it doesn't have to be this way!

• Historically, the idea was to transform the data to be normal and then apply linear regression, but in the GLM setting, we can use maximum likelihood methods on your choice of a random component
- The link function is a separate choice that is meant to make the functional form of the relationship linear, not to produce normality or stabilize the variance.

Generalized Linear Models for Binary Data

Setup: Many categorical response variables have only two categories of outcomes (success/failure).

Denote: $Y = 1$ if success and $Y = 0$ if failure.

and define probabilities

$P(y) = \pi$ if $Y = 1$ and $P(y) = 1 - \pi$ if $Y = 0$.

The mean of the distribution is $E(Y) = \pi$, so this is the quantity that we want to estimate with our linear model. Note that:
• For n independent observations, the number of successes is Binomial(n, π).

• Each binary observation is binomial with $n = 1$.

Linear Probability Model

$$\pi = p(Y = 1) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

• This is a GLM with a binomial random component and an identity link function

Pros: Simple and easy to interpret! Here, the probability of success changes linearly in each explanatory variable, e.g.

• β_1 represents the change in $P(Y = 1)$ for a 1 unit change in X_1, adjusting for all other variables in the model.

Cons: Suppose that our model is $\pi = 0.25 + 0.05X$.

What is the value of π when $X = 4$? How about if $X = 15$? Or $X = -6$? Any problems?

Probabilities have values between 0 and 1, but this model can predict values $\pi = P(Y = 1) < 0$ or $\pi > 1$. So the model only works over a restricted range of the explanatory variables

The linear probability model has the expression

$$\hat{\pi} = \alpha + \beta x.$$

Logistic Regression Model

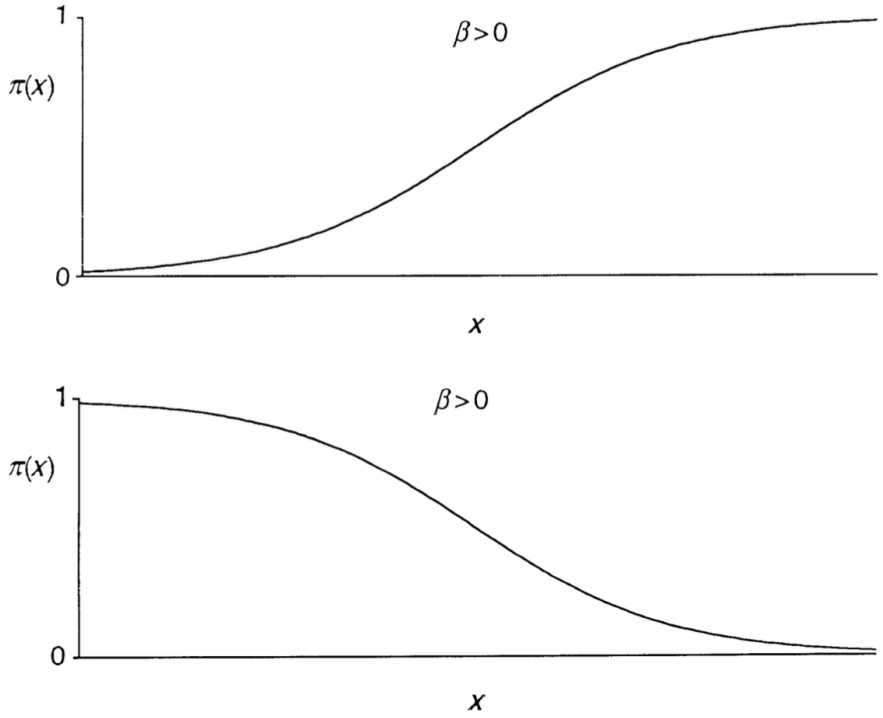
In practice, the effects of an explanatory variable are generally nonlinear.

• A fixed change in X will have less of an impact on an event that is highly likely ($\pi \approx 1$) or highly unlikely ($\pi \approx 0$) to occur than when $\pi \approx 0.5$.
Consider a sporting event, where one team is way up at the end of the game, so there is a high probability of a particular team winning a game. If you hustle for the rebound, will that have a dramatic impact on the outcome of the game? How about if the game was tied?

Let's stick with a basketball example and think about this idea mathematically.

Suppose that we define $X =$ the difference in the score at a certain point in the game. The probability of winning, $\pi = P(Y = 1)$ increases continuously (or decreases continuously) as x increases.

In practice, $\pi(x)$ often either increases continuously or decreases continuously as x increases. The S-shaped curves displayed in the following figure are often realistic shapes for the relationship.



An important mathematical function with an S-shaped curve has the formula using the exponential function:

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}.$$

This is called the *logistic regression* function. The corresponding logistic regression model form is

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x.$$

We write the expression as

$$\text{logit}[\hat{\pi}(x)] = \alpha + \beta x..$$

Here, β represents the rate of increase or decrease of a curve. The magnitude of β determines how fast the curve increases or decreases. As $|\beta|$ increases, the curve has a steeper rate of change. When $\beta = 0$, the curve flattens to a horizontal straight line.

Again, this is a special case of the GLM with a binomial distribution for the random component (success/failure) and a logit link function. Logistic regression models are often called logit models.

Pros: Unlike the probability model, $0 \leq \pi \leq 1$, the potential range for the linear predictors is all real number. Therefore, we don't have the same structural limitation as the probability model.

Cons:

- Less straightforward interpretation
- Math can be more challenging

Probit Regression Model

Another model that has the S-shaped curves of the figures above is called the probit model. The link function for the model, called the probit link, transforms probabilities to z-scores from the standard normal distribution. The probit model has expression

$$\text{probit}[\pi(x)] = \alpha + \beta x$$

. The probit link function applied to $\pi(x)$ gives the standard normal z−score at which the left-tail probability equals $\pi(x)$.