Lecture: Generalized Linear Models

Denis Cohen[*]

Generalized linear models

Goals for this session

- To understand generalized linear models (GLM) as a unified methodology for producing parameter estimates (Gill 2001, Gill and Torres 2019).
- To understand that all GLM produce two important, intuitive, and substantively meaningful quantities of interest that can be derived from the parameter estimates and the data (King et al. 2000):

  1. Expected values (conditional expectations)
  2. Average marginal effects or first differences

- To learn how to estimate these quantities and how to specify the uncertainty about our inferences using simulation techniques (King et al. 2000).

The three parts of every GLM

All generalized linear models have three characteristic parts:

Family

- The family stipulates a stochastic process that can plausibly generate an outcome y
- This means we choose a pdf or pmf for y given some parameters: $y_i \sim \text{f}(\theta_i, \psi)$

[*] Mannheim Centre for European Social Research, University of Mannheim, 68131 Mannheim, Germany. denis.cohen@uni-mannheim.de.

- The choice usually depends on the distributional properties of y
- Aliases: data-generating process, generative model, likelihood function

Linear component

- A linear model $y_i^* = \mathbf{x}_i'\beta + \epsilon_i$
- The goal of inference is the estimation of β
- From, this, we can derive our *systematic component* or *linear predictor*, $\eta_i = \mathbf{x}_i'\beta$

Inverse link function

- A function that transforms the systematic component η_i such that it represents a characteristic *parameter* θ_i of the family
- $\theta_i = g^{-1}(\eta_i)$

In a nutshell

Putting it all together, a GLM is given by

$$y_i \sim \mathrm{f}(\theta_i = g^{-1}(\mathbf{x}_i'\beta), \psi)$$

where ψ is an auxiliary parameter that will sometimes be estimated (e.g., σ^2 in the linear model) and sometimes be fixed.

Every generalized linear model is a *special case* of this general framework.

Example 1: The linear model

- While every GLM is a special case, the linear model is arguably a *very* special case.
- Why? Because its link function is the *identity function.*
- This not only makes the notation easier, but also means that the β's are *directly interpretable* on the scale of the outcome.

The three parts:

- Family:

$$y_i \sim \mathrm{N}(\eta_i, \sigma^2)$$

- Linear component:

$$y_i = \underbrace{\mathbf{x}_i'\beta}_{\eta_i} + \underbrace{\epsilon_i}_{\sim \mathrm{N}(0,\sigma^2)}$$

- Inverse link function:

$$\eta_i = \mathrm{id}(\eta_i)$$

Thus, the linear model is given by

$$y_i \sim \mathrm{N}(\mathbf{x}_i'\beta, \sigma^2)$$

where both β and σ^2 are being estimated.

Example 2: The probit model

The probit model is a popular choice for modeling idiosyncratic binary choices.

The three parts:

- Family:

$$y_i \sim \mathrm{Bernoulli}(\pi_i)$$

- Linear component:

$$y_i^* = \underbrace{\mathbf{x}_i'\beta}_{\eta_i} + \underbrace{\epsilon_i}_{\sim \mathrm{N}(0,1)}$$

- Inverse link function:

$$\pi_i = \Phi(\eta_i)$$

Thus, the probit model is given by

$$y_i \sim \mathrm{Bernoulli}(\Phi(\mathbf{x}_i'\beta))$$

where the β vector is being estimated. Φ is the standard normal CDF. Note that the standard normal CDF follows from fixing the variance of the error term, $\epsilon \sim \text{N}(0, 1)$.

Example 3: The logit model

With a slight change in the error distribution in the linear component and a corresponding change in the inverse link function, we can derive the logit model for binary choices.

The three parts:

- Family:

$$y_i \sim \text{Bernoulli}(\pi_i)$$

- Linear component:

$$y_i^* = \underbrace{\mathbf{x}_i'\beta}_{\eta_i} + \underbrace{\epsilon_i}_{\sim\text{Logistic}(0,1)}$$

- Inverse link function:

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

Thus, the logit model is given by

$$y_i \sim \text{Bernoulli}\left(\frac{\exp(\eta_i)}{1 + \exp(\eta_i)}\right)$$

where the β vector is being estimated. $\frac{\exp(\cdot)}{1+\exp(\cdot)}$ is the standard logistic CDF. Shorthand: $\text{logit}^{-1}(\cdot)$.

As you can see, the *assumed* distribution of the error term on the latent variable y_i^* dictates our choice of the inverse link function.

As the error distribution is fixed and its parameters are not being estimated, it is not in itself of substantive interest.

GLM Typology

Single-family models

We will first focus on models whose likelihood function follows a single pdf or pmf.

Univariate η_i, univariate θ_i

Among the simplest GLM are those models that require

- a single family
- a univariate systematic component η_i
- a univariate parameter θ_i

Examples include the three models discussed above:

- The linear model ($\eta_i = \theta_i = \mu_i$)
- The probit model ($\eta_i = \mathbf{x}_i'\beta$, $\theta_i = \Phi(\eta_i)$)
- The logit model ($\eta_i = \mathbf{x}_i'\beta$, $\theta_i = \text{logit}^{-1}(\eta_i)$)

An additional example would be the *Poisson model* for counts:

- $y_i \sim \text{Poisson}(\exp(\theta_i = \mathbf{x}_i'\beta))$

Univariate η_i, multivariate θ_i

Things get a bit more intricate when we model multivariate outcomes, e.g., discrete choice across multiple categories.

Multi-categorical discrete choice outcomes typically require that we stipulate a *categorical distribution* which requires choice-specific probability parameters $\theta_{ij} = \Pr(Y_i = j)$.

Thus, θ_i is a length-J vector for each $i = 1, .., N$: $\theta_i = \begin{bmatrix}\theta_{i1} & \dots & \theta_{iJ}\end{bmatrix}'$, where $\sum_{j=1}^{J} \theta_{ij} = 1$ for each $i = 1, .., N$.

A model that accommodates this while using a single univariate linear predictor η_i is the *ordered logit model*, which can be used for modeling ordered outcomes with J categories.

Family

$$y_{ij} \sim \text{Categorical}(\theta_{ij})$$

Systematic component The linear predictor is $\eta_i = \mathbf{x}_i' \beta$, where β does *not* include an intercept and $\mathbf{x}_i'$ does not include a leading one.

In place of an intercept, the model produces $J-1$ ordered threshold parameters κ.

Link function We then use the inverse logit link function to retrieve J probabilities θ_{ij} that η_i exceeds a given threshold κ:

$$\begin{aligned}\Pr(y_i = j | \mathbf{x}_i) &= \Pr(\kappa_{j-1} < \eta_i \leq \kappa_j) \\ &= \text{logit}^{-1}(\kappa_j - \eta_i) - \text{logit}^{-1}(\kappa_{j-1} - \eta_i)\end{aligned}$$

Multivariate η_i, multivariate θ_i

When moving from ordered to unordered discrete choices, we not only need to model choice-specific probability parameters θ_{ij} but also choice-specific linear predictors η_{ij}.

An example is the *multinomial logistic regression model.*

Family

$$y_{ij} \sim \text{Categorical}(\theta_{ij})$$

Systematic component The linear predictor is $\eta_{ij} = \mathbf{x}_i' \beta_j + \mathbf{z}_{ij}' \gamma$. For statistical identification, we must set the β vector for one category to zero, e.g., $\beta_J = \mathbf{0}$.

Link function The link function is the *softmax* function, a multivariate generalization of the inverse logit function:

$$\Pr(y_i = j) = \text{softmax}(\eta_{ij}) = \frac{\exp(\eta_{ij})}{\sum_{j=1}^{J} \exp(\eta_{ij})}$$

Multiple families

Some models stipulate complex data generating processes. They combine multiple families f in the likelihood (which means that the likelihood will be a mixture of the constitutive likelihoods).

Univariate η_i, multivariate θ_i

These models estimate only one set of parameters β, but use different link functions for translating the resulting linear predictor η_i into different parameters θ_i^f that match the stipulated data-generating processes f.

A well-known case is the *tobit model* for censored data. For instance, a left-censored tobit model with a lower bound at $y_L = 0$ jointly accommodates a Bernoulli data-generating process for $\Pr(y > y_L)$ and a normal data-generating process for the variation in y given $y > y_L$. By assumption, both data-generating processes are governed by the same parameters β.

Similar logics apply to other models that involve censored data, e.g., in survival analysis.

Multivariate η_i, multivariate θ_i

Two-part models A generalization of this are *two-part models.* Instead of estimating one set of parameters β and using different link functions for translating η_i into θ_i^f, these models estimate distinct sets of parameters β^f for each of the stipulated data-generating processes.

An example is a *hurdle model.* Unlike a left-censored tobit model, this model allows for the possibility that different sets of parameters govern the data-generating process for $\Pr(y > y_L)$ and the normal data-generating process for the variation in y given $y > y_L$.

Finite mixtures of identical families

A different intuition underlies finite mixture models. Rather than stipulating different *families* depending on the observed values of each unit, finite mixture models stipulate that substantively different data generating processes of the *same family* may generate the observed outcomes of all observations.

Quantities of interest

Expected values

- The *expected value* tells you where to expect the *conditional mean* of y given some covariate values $\mathbf{x}$ on the scale of y.

- For most (but not all) GLM, the expected value is directly given by our estimate of the parameter $\theta_i = g^{-1}(\mathbf{x}_i'\beta)$:
  - Linear model: $\mathbb{E}[y|\mathbf{x}] = \mathbf{x}_i'\beta$ ("predicted value")
  - Probit model: $\mathbb{E}[y|\mathbf{x}] = \Phi(\mathbf{x}_i'\beta)$ ("predicted probability")
  - Logit model: $\mathbb{E}[y|\mathbf{x}] = \text{logit}^{-1}(\mathbf{x}_i'\beta)$ ("predicted probability")

First differences

- A first difference is the difference between two expected values.
- It usually gives an estimate of how changing one covariate d affects our conditional expectation of y while holding all else ($\mathbf{x}$) constant.
  - Linear model: $\mathbb{E}[y|d_1, \mathbf{x}] - \mathbb{E}[y|d_0, \mathbf{x}] = (\alpha + \tau d_1 + \mathbf{x}_i'\beta) - (\alpha + \tau d_0 + \mathbf{x}_i'\beta) = \tau(d_1 - d_0)$
  - Probit model: $\mathbb{E}[y|d_1, \mathbf{x}] - \mathbb{E}[y|d_0, \mathbf{x}] = \Phi(\alpha + \tau d_1 + \mathbf{x}_i'\beta) - \Phi(\alpha + \tau d_0 + \mathbf{x}_i'\beta)$
  - Logit model: $\mathbb{E}[y|d_1, \mathbf{x}] - \mathbb{E}[y|d_0, \mathbf{x}] = \text{logit}^{-1}(\alpha + \tau d_1 + \mathbf{x}_i'\beta) - \text{logit}^{-1}(\alpha + \tau d_0 + \mathbf{x}_i'\beta)$

An important insight is that the first difference in the linear model does *not* depend on the values of other covariates $\mathbf{x}$.

In all other GLM, the presence of an inverse link function makes first differences sensitive to the choice of covariates $\mathbf{x}$!

Marginal effects

The marginal effect of a variable d on the expected value $\mathbb{E}[y|d, \mathbf{x}]$ is given by the marginal rate of change in $\mathbb{E}[y|d, \mathbf{x}]$ for an infinitesimal change in d:

$$\frac{\mathbb{E}[y|d + \Delta_d, \mathbf{x}] - \mathbb{E}[y|d, \mathbf{x}]}{\Delta_d}$$

As $\Delta_d \rightarrow 0$, this becomes $\frac{\partial \mathbb{E}[y|d, \mathbf{x}]}{\partial d}$.

For the linear model, this is mathematically straightforward:

$$\frac{\partial(\alpha + \tau d + \mathbf{x}_i'\beta)}{\partial d} = \tau$$

For all other GLM, we rely on *normalized first differences* for a freely chosen shift Δ_d.

For instance, the marginal effect in a probit model for a *standard deviation increase* in d is given by

$$\frac{\mathbb{E}[y|d+\mathrm{sd}(d),\mathbf{x}]-\mathbb{E}[y|d,\mathbf{x}]}{\mathrm{sd}(d)}=\frac{\Phi(\alpha+\tau(d+\mathrm{sd}(d))+\mathbf{x}_i'\beta)-\Phi(\alpha+\tau d+\mathbf{x}_i'\beta)}{\mathrm{sd}(d)}$$

As with first differences, the marginal effect is thus sensitive to the choice of covariates $\mathbf{x}$!

Average quantities of interest

The sensitivity of first differences and marginal effects to the choice of $\mathbf{x}$ is problematic because makes these quantities of interest dependent on subjective judgment.

For instance, the estimates of these quantities may differ dramatically depending on whether we set all variables in $\mathbf{x}$ to their sample means, sample minimums, or sample maximums.

A remedy is the *observed values approach* (Hanmer and Kalkan 2013), which involves two steps:

1. Calculate unit-specific quantities of interest at the observed values $\mathbf{x}_i$ for each observation $i=1,...,N$
2. Average across the N unit-specific quantities to obtain the *average* quantities of interest

So for instance, the average first difference for a binary variable d in a probit model is given by

$$\frac{1}{N}\sum_{i=1}^{N}\Phi(\alpha+\tau\times 1+\mathbf{x}_i'\beta)-\Phi(\alpha+\tau\times 0+\mathbf{x}_i'\beta)$$

Analogously, the average marginal effect for a one unit increase in a continuous variable d is given by

$$\frac{1}{N}\sum_{i=1}^{N}\Phi(\alpha+\tau(d_i+1)+\mathbf{x}_i'\beta)-\Phi(\alpha+\tau d_i+\mathbf{x}_i'\beta)$$

Quantities of interest: why?

The answers are simple:

- Your readers have a right to know!
- It makes your research accessible to non-technical audiences.
- Chances are: You will not truly understand your own findings without it.

Do not: "The logit-coefficient of our information treatment on turnout is $b = 1.5$ ($p < .05$). We thus conclude that there is a considerable treatment effect."

Do: "Our logit model yields a predicted turnout probability of 0.88 $[0.85, 0.91]$ in the treatment group, compared to 0.63 $[0.59, 0.67]$ in the control group. The corresponding difference of 0.25 $[0.22, 0.28]$ is of considerable substantive magnitude and statistically significant at the 95% level."

The simulation approach

Inferential uncertainty vs fundamental uncertainty

King et al. (2000) distinguish *inferential uncertainty* and *fundamental uncertainty.*

- *Inferential uncertainty* describes the problem that we never know our parameters exactly.
  - Instead, we estimate them with some uncertainty that is represented in their respective *sampling distribution.*
  - Example: In large samples, we assume that the mean of age in Germany has a normal sampling distribution such that $\hat{\mu} \sim \mathrm{N}(\bar{x}, \sigma_{\bar{x}}^2)$, where $\sigma_{\bar{x}}$ is the standard error of the mean.
  - Inferential uncertainty thus describes our uncertainty about *estimates* via *sampling distributions.*
- *Fundamental uncertainty* describes the randomness that comes from the fact that we stipulate stochastic data-generating processes.
  - Example: Over and beyond our uncertainty regarding the true mean of age in Germany, we have a stochastic model in mind that generates age, e.g.: $y \sim \mathrm{N}(\hat{\mu}, \hat{\sigma}^2)$. So we could generate a predictive distribution of age from a normal distribution whose mean is the sample mean and whose variance is the sample variance.

- Fundamental uncertainty thus describes our uncertainty about *predictions* via *predictive distributions.*

The GLM context

For a generic GLM, this means:

$$\underbrace{\underbrace{y_i \sim}_{\text{fund.}} \mathrm{f}(\theta_i = g^{-1}(\mathbf{x}_i' \underbrace{\beta), \psi}_{\text{inf.}})}_{\text{Total uncertainty}}$$

1. Inferential uncertainty about the model parameters can be simulated by taking draws from their joint sampling distribution.
2. Fundamental uncertainty can be simulated by taking draws of $\mathbf{y}$ from $\mathbf{y} \sim f(\theta, \psi)$.
3. Total uncertainty can be simulated by simulating (1) within (2)

Inferential uncertainty in GLM

Quantities of interest such as expected values, (average) marginal effects, or (average) first differences are estimates of population parameters.

Unless we engage in predictive modeling, we thus care primarily about inferential uncertainty.

However:

- In GLM, we rarely estimate our quantities of interest directly.
- We usually estimate our *model coefficients* β, from which we derive our quantities of interest.
- We thus only have information on the sampling distribution of β, $\beta \sim \text{MVN}(\hat{\beta}, \hat{\Sigma})$, not the sampling distribution of our quantity of interest.

So how can we get there?

- Beyond OLS, deriving the sampling distribution of quantities of interest analytically is rather painful.

- Analytical normal-approximation confidence intervals are also imprecise when the asymptotic properties of estimators do not hold in finite samples (e.g., 95% confidence intervals that include predicted probabilities outside of $[0, 1]$).
- So we will use a flexible approach that allows us to get sampling distributions for *any* quantity of interest: **Parameter simulation**.

The algorithm

The algorithm presented by King et al. (2000) contains five steps:

1. Simulate the sampling distribution of the model parameters by taking S draws from $\beta \sim \text{MVN}(\hat{\beta}, \hat{\Sigma})$
2. Choose a *covariate scenario*, i.e., specify a vector $\mathbf{x}^*$ or a matrix $\mathbf{X}^*$.
3. For each simulation β^s for $s = 1, ..., S$, calculate $\theta^s = g^{-1}(\mathbf{x}^{*\prime}\beta^s)$.

Whenever θ gives a direct estimate of $\mathbb{E}[y|\mathbf{x}^*]$, these three steps will be sufficient!

In some instances involving non-symmetrical transformations, we need to go the extra mile:

4. Simulate the predictive distribution M times for each simulation θ^s.
5. Average over the M predictive draws within each simulation s.

We will *not* cover any such examples. So in our upcoming applied logistic regression example, we can produce inferential uncertainty with the simpler three-step algorithm.

You can easily see why this the case: The expected value of a Bernoulli distribution with parameter π is π itself: $\mathbb{E}[\text{Bernoulli}(\pi)] = \pi$. So drawing many 0's and 1's in step 4 just to average them back to a proportion that will be (approximately) equal to π is redundant.