Generalized Linear Models

Jacob M. Montgomery

2018

The generalized linear model

Overview

- Lastclass
  - How some covariate X relates to an outcome Y
  - Different ways to estimate and interpret such models
- Today we are going to try to generalize this a bit from a "traditional" GLM framework

The linear model redux

1. We set up a parametric model

$$y = \alpha + \beta x + \epsilon, \ \epsilon \sim N(0, \sigma^2)$$

The linear model redux

1. We set up a parametric model

$$y = \alpha + \beta x + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

2. Then we calcuate fitted values based on the (noisy) data we have

$$\hat{y} = \hat{\alpha} + \hat{\beta} x + \epsilon,\ \epsilon \sim N(0, \hat{\sigma}^2)$$

The linear model redux

1. We set up a parametric model

$$y = \alpha + \beta x + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

2. Then we calcuate fitted values based on the (noisy) data we have

$$\hat{y} = \hat{\alpha} + \hat{\beta} x + \epsilon,\ \epsilon \sim N(0, \hat{\sigma}^2)$$

3. We choose estimates that reduce some measure of "error" or discrepancy
   - L_2 norm is

$$S_2(y, \hat{y}) = \sum (y_i - \hat{y}_i)^2$$

   - L_1 norm is

$$S_1(y, \hat{y}) = \sum |y_i - \hat{y}_i|$$

Pulling past lectures together a bit

- For the classic model, this can re-framed as saying that
  - $\mu = \alpha + \beta x$
  - $$p(Y|\mu) \sim N(\mu, \sigma^2)$$

Pulling past lectures together a bit

- For the classic model, this can re-framed as saying that
  - $\mu = \alpha + \beta x$
  - $$p(Y|\mu) \sim N(\mu, \sigma^2) \propto \exp(-(y-\mu)^2/(2\sigma^2))$$
- If we regard x to be fixed, this will give us back the least squares criteria.
- We could also view this as giving us information on more or less likely functions of μ, which leads to an MLE or Bayesian approach.
- So we want a model that has good “fit” (reduces error)

Pulling past lectures together a bit

- For the classic model, this can re-framed as saying that
  - $\mu = \alpha + \beta x$
  - $$p(Y|\mu) \sim N(\mu, \sigma^2) \propto \exp(-(y-\mu)^2/(2\sigma^2))$$
- If we regard x to be fixed, this will give us back the least squares criteria.
- We could also view this as giving us information on more or less likely functions of μ, which leads to an MLE or Bayesian approach.
- So we want a model that has good "fit" (reduces error) BUT which also has good out of sample properties.

Formalizing a bit

- $\mathbf{y} = (y_1, \ldots, y_n)'$
- $\mathbf{X}$ is an $n \times p$ matrix of covariates (first column will be all ones)
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$

Formalizing a bit

- $\mathbf{y} = (y_1, \ldots, y_n)'$
- $\mathbf{X}$ is an $n \times p$ matrix of covariates (first column will be all ones)
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$
- $$e(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\beta$$

Formalizing a bit

- $\mathbf{y} = (y_1, \ldots, y_n)'$
- $\mathbf{X}$ is an $n \times p$ matrix of covariates (first column will be all ones)
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$
- $$e(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\beta$$
- And we minimize $\mathbf{e}'\mathbf{e} = \sum e_i^2$

Getting to the GLM for the classic regression model

We divide up the model into a "systematic" and "random" component

1. The random component: The components of $\mathbf{y}$ have indpendent Normal distributions with $E(\mathbf{y}) = \mu$ and constant variance σ^2

Getting to the GLM for the classic regression model

We divide up the model into a "systematic" and "random" component

1. The random component: The components of $\mathbf{y}$ have indpendent Normal distributions with $E(\mathbf{y}) = \mu$ and constant variance σ^2
2. The systematic component: covariates $\mathbf{x}_1, \ldots, \mathbf{x}_p$ produce a linear predictor η given by

$$\eta = \sum_{j=1}^{p} \mathbf{x}_j \beta_j$$

Getting to the GLM for the classic regression model

We divide up the model into a “systematic” and “random” component

1. The random component: The components of $\mathbf{y}$ have indpendent Normal distributions with $E(\mathbf{y}) = \mu$ and constant variance σ^2
2. The systematic component: covariates $\mathbf{x}_1, \ldots, \mathbf{x}_p$ produce a linear predictor η given by

$$\eta = \sum_{j=1}^{p} \mathbf{x}_j \beta_j$$

3. The **link** between the random and systematic component is: $\mu = \eta$

Getting to the GLM for the classic regression model

We divide up the model into a "systematic" and "random" component

1. The random component: The components of $\mathbf{y}$ have indpendent Normal distributions with $E(\mathbf{y}) = \mu$ and constant variance σ^2
2. The systematic component: covariates $\mathbf{x}_1, \ldots, \mathbf{x}_p$ produce a linear predictor η given by

$$\eta = \sum_{j=1}^{p} \mathbf{x}_j \beta_j$$

3. The **link** between the random and systematic component is: $\mu = \eta$

All of GLM modeling involves setting up these three components

Getting to the GLM for the classic regression model

We divide up the model into a "systematic" and "random" component

1. The random component: The components of $\mathbf{y}$ have indpendent Normal distributions with $E(\mathbf{y}) = \mu$ and constant variance σ^2
2. The systematic component: covariates $\mathbf{x}_1, \ldots, \mathbf{x}_p$ produce a linear predictor η given by

$$\eta = \sum_{j=1}^{p} \mathbf{x}_j \beta_j$$

3. The **link** between the random and systematic component is: $\mu = \eta$

All of GLM modeling involves setting up these three components and then estimating the parameters.

Some things I have to show you

- For a distribution that is in the exponential family, we can re-write

$$f(y|\theta,\phi) = \exp\left((y\theta - b(\theta))/a(\phi) + c(y,\phi)\right)$$

  - θ is the "canonical" parameter (or other terms)
- We know that $E(\frac{\partial \mathcal{L}}{\partial \theta}) = 0$. Find this, and solve for μ.

Some things I have to show you

- For a distribution that is in the exponential family, we can re-write

$$f(y|\theta, \phi) = \exp\left((y\theta - b(\theta))/a(\phi) + c(y, \phi)\right)$$

  - θ is the "canonical" parameter (or other terms)
- We know that $E(\frac{\partial \mathcal{L}}{\partial \theta}) = 0$. Find this, and solve for μ.
- $\mu = b'(\theta)$ is called the "canonical link"
- We can follow a similar proof to get that:

$$Var(Y) = b''(\theta)a(\phi)$$

- Recall that $\eta = \sum_{j=1}^{p} \mathbf{x}_j \beta_j$
- Here are some canonical links:

Distribution	link	name
normal	$\eta = \mu$	identity
Poisson	$\eta = log \mu$	log
binomial	$\eta = log(\pi/(1-\pi))$	logit
gamma	$\eta = \mu^{-1}$	reciprical
inverse Gaussian	$\eta = \mu^{-2}$	whatever

- Some links are used, but don't fall out quite so easily

1. probit
$$\eta = \Phi^{-1}(\mu)$$
2. complementary log-log
$$\eta = log(-log(1-\mu))$$

Things to remember

- All of this is for *fitting* the model
- For *understanding* the model, we are often going to want to understand how μ changes as a function of some x.
- For that, we are going to need the **inverse** link function.

Fitting the model

- It turns out that most of the models you will present don't actually use any of the methods we like to teach.
- But it is still good to go through this a bit to get a feeling for it.
- Focus on the concepts of "fit" and the concepts/vocabulary rather than the formulas and their origins.

Deviance

- We are going to compare how well our model does versus a “full” model (where each observation is estimated by itself)
  - $\mathcal{L}(\hat{\mu}, \phi : \mathbf{y})$
  - $\mathcal{L}(\mathbf{y}, \phi : \mathbf{y})$
- Let $\hat{\theta} = \theta(\hat{\mu})$, $\tilde{\theta} = \theta(y)$, and $a_i(\phi) = \phi/w_i$, then the discrepency between the two models can be written as

$$\sum 2w_i\big[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\big]/\phi = D(\mathbf{y}|\hat{\boldsymbol{\mu}})/\phi$$

Deviance 2

- For the poisson, recall that $b(\theta) = \exp(\theta)$
- Recall that $\theta = \log(\mu)$
- Plugging into the formula, we get:

$$\sum y_i(\log(y_i) - \log(\hat{\mu}_i) - \exp(\log(y_i)) + \exp(\log(\hat{\mu}_i))$$

$$\sum y_i(\log(y_i/\hat{\mu}_i)) - (y_i - \hat{\mu}_i)$$

Actually fitting the thing (MLE style)

The "classic" way that GLM models are fit are with iterated weighted least squares (IWLS). This works as follows.

1. Let $\hat{\eta}_0$ be the current estimate of $\eta = g(\mu)$. We calculate the "adjusted" dependent variable

$$z_0 = \hat{\eta}_0 + (y - \mu_0)\left(\frac{\partial \eta}{\partial \mu}\right)_0$$

2. We are going to calculate "weights" for our observations such that

$$W_0^{-1} = (\partial \eta / \partial \mu)_0^2 \, V_0$$

, where V_0 is the variance function evaluated at $\hat{\mu}_0$

3. Now we do a weighted regression where z_0 is the dependent variables, with explanatory variables $\mathbf{X}$ and weights W_0.
4. Repeat until the changes are sufficiently small.

Let's go over that a bit just in terms of a logit model

1. Choose a starting value for β
2. Compute π_{i0} based on this for each observation
3. Use taylor series expansion to build a new variable $z = plogis(mu) + (y - \pi_{i0})/(\pi_0(1 - \pi_{i0}))$
4. Uncertainty in Z varies, so we need weights which are $\pi_i(1 - \pi_i)$
5. We run a **weighted** regression of X on Z and update Beta
6. Repeat 2-5 until convergence

- Recall that

$$f(x) \approx f(a) + f'(a)(x-a) + f''(a)(x-a)^2/2 \ldots$$

- Let $\eta = \beta_0 + x\beta$ and $g(\pi) = \beta_0 + x\beta$
- $g(y) \approx g(\mu) + (y-\mu)g'(\mu) \equiv z$
- But we have different uncertainty for different values of μ, so we weight each observation by:
- Note that this is fairly close to the formula:

$$\beta^{(t+1)} = \beta^{(t)} - f'(\beta^{(t)})/f''(\beta^{(t)})$$

- In fact, this is is just a slightly adjusted version of the Newton method.

Let (t) index iteration

1. $$\eta_i^{(t)} = \sum_j \beta_j^{(t)} x_{ij}$$

2. $$\pi_i^{(t)} = [1 + \exp(-\eta_i)]^{-1}$$

3. $$\nu_i^{(t)} = \hat{\pi}^{(t)}(1 - \hat{\pi}^{(t)})$$

4. $$z_i^{(t)} = \eta_i + (y_i - \hat{\pi}^{(t)})/\nu_i^{(t)}$$

5. $$\underset{\beta \in R^p}{\operatorname{argmin}}(\nu_i^{(t)}(z_i^{(t)} - \mathbf{x}_i'\beta^{(t)})^2)$$

6. Repeat until convergence