Generalized Linear Models

Advanced Methods for Data Analysis (36-402/36-608)

Spring 2014

1 Generalized linear models

1.1 Introduction: two regressions

- So far we've seen two canonical settings for regression. Let $X \in \mathbb{R}^p$ be a vector of predictors. In linear regression, we observe $Y \in \mathbb{R}$, and assume a linear model:

$$\mathbb{E}(Y|X) = \beta^T X,$$

for some coefficients $\beta \in \mathbb{R}^p$. In logistic regression, we observe $Y \in \{0, 1\}$, and we assume a logistic model

$$\log\Big(\frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)}\Big) = \beta^T X.$$

- What's the similarity here? Note that in the logistic regression setting, $\mathbb{P}(Y=1|X) = \mathbb{E}(Y|X)$. Therefore, in both settings, we are assuming that a *transformation* of the conditional expectation $\mathbb{E}(Y|X)$ is a linear function of X, i.e.,

$$g\big(\mathbb{E}(Y|X)\big) = \beta^T X,$$

for some function g. In linear regression, this transformation was the identity transformation $g(u) = u$; in logistic regression, it was the logit transformation $g(u) = \log(u/(1-u))$

- Different transformations might be appropriate for different types of data. E.g., the identity transformation $g(u) = u$ is not really appropriate for logistic regression (why?), and the logit transformation $g(u) = \log(u/(1-u))$ not appropriate for linear regression (why?), but each is appropriate in their own intended domain

- For a third data type, it is entirely possible that transformation neither is really appropriate. What to do then? We think of another transformation g that is in fact appropriate, and this is the basic idea behind a generalized linear model

1.2 Generalized linear models

- Given predictors $X \in \mathbb{R}^p$ and an outcome Y, a *generalized linear model* is defined by three components: a *random component*, that specifies a distribution for $Y|X$; a *systematic component*, that relates a parameter η to the predictors X; and a *link function*, that connects the random and systematic components

- The random component specifies a distribution for the outcome variable (conditional on X). In the case of linear regression, we assume that $Y|X \sim N(\mu, \sigma^2)$, for some mean μ and variance σ^2. In the case of logistic regression, we assume that $Y|X \sim \text{Bern}(p)$ for some probability p.

In a generalized model, we are allowed to assume that $Y|X$ has a probability density function or probability mass function of the form

$$f(y;\theta,\phi) = \exp\Big(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\Big).$$

Here θ, ϕ are parameters, and a, b, c are functions. Any density of the above form is called an *exponential family density*. The parameter θ is called the *natural parameter*, and the parameter ϕ the *dispersion parameter*; it helps to think of the normal case, where $\theta = \mu$, and $\phi = \sigma$

- We will denote the expectation of this distribution as μ, i.e., $\mathbb{E}(Y|X) = \mu$. It will be our goal to estimate μ. This typically doesn't involve the dispersion parameter ϕ, so for simplicity, we'll assume this is known
- The systematic component relates a parameter η to the predictors X. In a generalized linear model, this is always done via

  $$\eta = \beta^T X = \beta_1 X_1 + \ldots + \beta_p X_p.$$

  Note that throughout we are conditioning on X, hence we think of it as systematic (nonrandom)
- Finally, the link component connects the random and systematic components, via a link function g. In particular, this link function provides a connection between μ, the mean of $Y|X$, and and η, as in

  $$g(u) = \eta.$$

  Again, it helps to think of the normal case, where $g(\mu) = \mu$, so that $\mu = \beta^T X$

2 Examples

- So many parameters: θ, ϕ, μ, η...! Let's do a little bit of sorting out.
  - First of all, remember that μ is the mean of $Y|X$, what we want to ultimately estimate
  - Now, θ is just a parameter we use to govern the shape of the density of $Y|X$, and so μ—the mean of this distribution—will obviously depend on θ. It can be the case that $\mu = \theta$ (e.g., normal), but this doesn't need to be true (e.g., Bernoulli, Poisson)
  - Recall that ϕ is a dispersion parameter of the density of $Y|X$, but we'll think of this as known, because to estimate μ we won't need to know its value
  - The parameter η may seem like a weird one. At this point, for generalized linear models, you can just think of it as short form for a linear combination of the predictors, $\eta = \beta^T X$. From a broader perspective, we're aiming to model a transformation of the mean μ by some function of X, written $g(\mu) = \eta(X)$. For generalized linear models, we are always modeling a transformation of the mean by a linear function of X, but this will change for generalized additive models

  Now it helps to go through several examples

2.1 Bernoulli

- Suppose that $Y \in \{0, 1\}$, and we model the distribution of $Y|X$ as Bernoulli with success probability p, i.e., 1 with probability p and 0 with probability $1 - p$. Then the probability mass function (not a density, since Y is discrete) is

  $$f(y) = p^y (1-p)^{1-y}$$

- We can rewrite to fit the exponential family form as

$$\begin{aligned} f(y) &= \exp\Big(y \log p + (1-y)\log(1-p)\Big) \\ &= \exp\Big(y \cdot \log\big(p/(1-p)\big) + \log(1-p)\Big) \end{aligned}$$

- Here we would identify $\theta = \log(p/(1-p))$ as the natural parameter. Note that the mean here is $\mu = p$, and using the inverse of the above relationship, we can directly write the mean p as a function of θ, as in $p = e^\theta/(1+e^\theta)$. Hence $b(\theta) = \log(1-p) = -\log(1+e^\theta)$
- There is no dispersion parameter, so we can set $a(\phi) = 1$. Also, $c(y, \phi) = 0$

2.2 Poisson

- Now suppose that $Y \in \{0, 1, 2, 3, \ldots\}$, i.e., Y is a nonnegative count, and we model its distribution (conditional on X) as Poisson with mean μ. Its probability mass function is

$$f(y) = \frac{1}{y!} e^{-\mu} \mu^y$$

- Rewriting this,

$$f(y) = \exp\Big(y \log \mu - \mu - \log(y!)\Big)$$

- Hence $\theta = \log \mu$ is the natural parameter. Reciprocally, we can write the mean in terms of the natural parameter as $\mu = e^\theta$. Hence $b(\theta) = \mu = e^\theta$
- Again there is no dispersion parameter, so we can set $a(\phi) = 1$. Finally, $c(y, \phi) = -\log(y!)$

2.3 Gaussian

- A familiar setting is probably when $Y \in \mathbb{R}$, and we model $Y|X$ as $N(\mu, \sigma^2)$. The density is

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(y-\mu)^2}{2\sigma^2}\Big)$$

- This is pretty much already in exponential family form, but we can simply manipulate it a bit more to get

$$f(y) = \exp\Big(\frac{y\mu - \mu^2/2}{\sigma^2} - y^2/(2\sigma^2) - \log\sigma - \log\sqrt{2\pi}\Big)$$

- Now the natural parameter is simply $\theta = \mu$, and $b(\theta) = \theta^2/2$
- Here we have a dispersion parameter, $\phi = \sigma$, and $a(\phi) = \sigma^2$. Also $c(y, \phi) = y^2/(2\sigma^2) - \log\sigma - \log\sqrt{2\pi}$

2.4 Link functions

- What about the link function g? Well, we've already seen that for the normal case, the "right" choice of link function is the identity transform $g(\mu) = \mu$, so that we model $\mu = \beta^T X$; and for the Bernoulli case, the "right" choice of link function is the logit transform $g(\mu) = \log(\mu/(1-\mu))$, so that we model $\log(\mu/(1-\mu)) = \beta^T X$. We've already explained that each of these transformations is appropriate in their own context (recall that $\mu = p$, the success probability, in the Bernoulli case)

- But what about the Poisson case? And in general, given an exponential family, what is the "right" transform? Fortunately, there is a default choice of link function called the *canonical link*. We can define this implicitly by the link function that sets

$$\theta = \eta.$$

  In other words, the link function is defined via $g(\mu) = \theta$, by writing the natural parameter θ in terms of μ

- In many cases, we can read off the canonical link just from the term that multiplies y in the exponential family density or mass function. E.g., for normal, this is $g(\mu) = \mu$, for Bernoulli, this is $g(\mu) = \log(\mu/(1-\mu))$, and for Poisson, this is $g(\mu) = \log \mu$

- The canonical link is general and tends to work well. But it is important to note that the canonical link is not the only "right" choice of link function. E.g., in the Bernoulli setting, another common choice (aside from the logit link $g(\mu) = \log(\mu/(1-\mu))$) is the probit link, $g(\mu) = \Phi^{-1}(\mu)$, where Φ is the standard normal CDF

3 Estimation from samples

- Suppose that we are given independent samples (x_i, y_i), $i = 1, \ldots n$, with each $y_i|x_i$ having an exponential family density

$$f(y_i; \theta_i, \phi) = \exp\Big(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\Big), \quad i = 1, \ldots n.$$

  I.e., we fix an underlying exponential family distribution and common dispersion parameter ϕ, but allow each sample $y_i|x_i$ its own natural parameter θ_i. The analogy in the Gaussian case is $y_i|x_i \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots n$

- Our goal is to estimate the means $\mu_i = \mathbb{E}(y_i|x_i)$, for $i = 1, \ldots n$. Recall that we have a link function $g(\mu_i) = \eta_i$, which connects the mean μ_i to the parameter $\eta_i = x_i^T\beta$. Hence we can first estimate the coefficients β, as in $\hat{\beta}$, and then use these estimates to argue that

$$g(\hat{\mu}_i) = x_i^T\hat{\beta}, \quad i = 1, \ldots n$$

  i.e.,

$$\hat{\mu}_i = g^{-1}(x_i^T\hat{\beta}), \quad i = 1, \ldots n$$

- So how do we compute $\hat{\beta}$? We use maximum likelihood. The likelihood of the samples (x_i, y_i), $i = 1, \ldots n$, conditional on x_i, $i = 1, \ldots n$, is

$$L(\theta) = \prod_{i=1}^n f(y_i; \theta_i, \phi),$$

  written as a function of $\theta = (\theta_1, \ldots \theta_n)$ to denote the dependence the natural parameter.

- I.e., the log likelihood is

$$\ell(\theta) = \sum_{i=1}^n \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^n c(y_i, \phi)$$

- We want to maximize this log likelihood over all choices of coefficients $\beta \in \mathbb{R}^p$; this is truly a function of β, because each natural parameter θ_i can be written in terms of the mean μ_i of the exponential family distribution, and $\mu_i = x_i^T\beta$, $i = 1, \ldots n$. Therefore we can write

$$\ell(\beta) = \sum_{i=1}^n y_i\theta_i - b(\theta_i)$$

  where we have discarded terms that don't depend on θ_i, $i = 1, \ldots n$

- To be more concrete, suppose that we are considering the canonical link function g; recall that this is the link function that sets $\theta_i = \eta_i = x_i^T\beta$, $i = 1, \ldots n$. Then the log likelihood, to maximize over β, is

$$\ell(\beta) = \sum_{i=1}^n y_i x_i^T\beta - b(x_i^T\beta) \tag{1}$$

- Some quick checks: in the Bernoulli model, this is

$$\sum_{i=1}^n y_i x_i^T\beta - \log\big(1 + \exp(x_i^T\beta)\big),$$

  which is precisely what we saw in logistic regression. In the Gaussian model, this is

$$\sum_{i=1}^n y_i x_i^T\beta - (x_i^T\beta)^2/2,$$

  and maximizing the above is just the same as minimizing the least squares criterion

$$\sum_{i=1}^n (y_i - x_i^T\beta)^2,$$

  as in linear regression. Finally, in the Poisson model, we get

$$\sum_{i=1}^n y_i x_i^T\beta - \exp(x_i^T\beta),$$

  which is new (for us, so far!), and called *Poisson regression*

- How do we maximize the likelihood (1) to form our estimate $\hat{\beta}$? In general, there is no closed-form for its maximizer (as in there is in the Gaussian least squares case). Therefore we must run an optimization algorithm to compute its maximizer. Fortunately, though, it turns out we can maximize $\ell(\beta)$ by repeatedly performing weighted least squares regressions! This is the same as what we described in logistic regression, and is an instance of Newton's method for maximizing (1) that we call iteratively reweighted least squares, or IRLS

- Aside from being computationally convenient (and efficient), the IRLS algorithm is very helpful from the perspective of statistical inference: we simply treat the coefficients $\hat{\beta}$ as a result of a single weighted least squares regression—the last one in the IRLS sequence—and now apply all of our inferential tools from linear regression (e.g., form confidence intervals) accordingly. This is the standard in software, and we will not go into detail, but it helps to know where such constructs come from

4 Generalized additive models

- Recall that we saw we could augmented the standard linear model

$$\mathbb{E}(y_i|x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}, \quad i = 1, \ldots n,$$

  with the additive model

$$\mathbb{E}(y_i|x_i) = \beta_0 + r_1(x_{i1}) + r_2(x_{i2}) + \ldots + r_p(x_{ip}),$$

  where each r_j is an arbitrary (univariate) regression function, $j = 1, \ldots p$

- The same extension can be applied to the generalized linear model, yielding what we call a *generalized additive model*. The only change is in the parameter η, which we now define as

$$\eta_i = \beta_0 + r_1(x_{i1}) + r_2(x_{i2}) + \ldots + r_p(x_{ip}), \quad i = 1, \ldots n$$

- That is, the link function g now connects the random component, the mean of the exponential family distribution $\mu_i = \mathbb{E}(y_i|x_i)$, to the systematic component η_i, via

$$g(\mu_i) = \beta_0 + r_1(x_{i1}) + r_2(x_{i2}) + \ldots + r_p(x_{ip}), \quad i = 1, \ldots n$$

- In the Gaussian case, the above reduces to the additive model that we've already studied. In the Bernoulli case, with canonical link, this model becomes

$$\log \frac{p_i}{1 - p_i} = \beta_0 + r_1(x_{i1}) + r_2(x_{i2}) + \ldots + r_p(x_{ip}), \quad i = 1, \ldots n,$$

  were $p_i = \mathbb{P}(y_i = 1|x_i)$, $i = 1, \ldots n$, which gives us *additive logistic regression*. In the Poisson case, with canonical link, this becomes

$$\log \mu_i = \beta_0 + r_1(x_{i1}) + r_2(x_{i2}) + \ldots + r_p(x_{ip}), \quad i = 1, \ldots n,$$

  which is called *additive Poisson regression*

- Generalized additive models are a marriage of two worlds: generalized linear models and additive models, and is altogether a very flexible, powerful platform for modeling. The generalized linear model element allows us to account for different types of outcome data y_i, $i = 1, \ldots n$, and the additive model element allows us to consider a transformation of the mean $\mathbb{E}(y_i|x_i)$ as a nonlinear (but additive) function of $x_i \in \mathbb{R}^p$, $i = 1, \ldots n$