21:06 Thursday 17th March, 2016

Chapter 12

Generalized Linear Models and Generalized Additive Models

[[TODO: Merge with logistic chapter]]

12.1 Generalized Linear Models and Iterative Least Squares

Logistic regression is a particular instance of a broader kind of model, called a **generalized linear model** (GLM). You are familiar, of course, from your regression class with the idea of transforming the response variable, what we've been calling Y, and then predicting the transformed variable from X. This was *not* what we did in logistic regression. Rather, we transformed the *conditional expected value*, and made *that* a linear function of X. This seems odd, because it *is* odd, but it turns out to be useful.

Let's be specific. Our usual focus in regression modeling has been the conditional expectation function, $\mu(x) = \mathbb{E}[Y|X=x]$. In plain linear regression, we try to approximate $\mu(x)$ by $\beta_0 + x \cdot \beta$. In logistic regression, $\mu(x) = \mathbb{E}[Y|X=x] = \Pr(Y=1|X=x)$, and it is a transformation of $\mu(x)$ which is linear. The usual notation says

$$\eta(x) = \beta_0 + x \cdot \beta \tag{12.1}$$

$$\eta(x) = \log \frac{\mu(x)}{1-\mu(x)} \tag{12.2}$$

$$= g(\mu(x)) \tag{12.3}$$

defining the logistic **link function** by $g(m) = \log m/(1-m)$. The function $\eta(x)$ is called the **linear predictor**.

Now, the first impulse for estimating this model would be to apply the transformation g to the response. But Y is always zero or one, so $g(Y) = \pm\infty$, and regression will not be helpful here. The standard strategy is instead to use (what else?) Taylor expansion. Specifically, we try expanding $g(Y)$ around $\mu(x)$, and stop at first order:

$$g(Y) \approx g(\mu(x)) + (Y - \mu(x))g'(\mu(x)) \tag{12.4}$$

$$= \eta(x) + (Y - \mu(x))g'(\mu(x)) \equiv z \tag{12.5}$$

We *define* this to be our effective response after transformation. Notice that if there were no noise, so that y was always equal to its conditional mean $\mu(x)$, then regressing z on x would give us back exactly the coefficients β_0, β. What this suggests is that we can *estimate* those parameters by regressing z on x.

The term $Y - \mu(x)$ has expectation zero, so it acts like the noise, with the factor of g' telling us about how the noise is scaled by the transformation. This lets us work out the variance of z:

$$\begin{aligned} \mathbb{V}[Z|X=x] &= \mathbb{V}[\eta(x)|X=x] + \mathbb{V}\left[(Y-\mu(x))g'(\mu(x))|X=x\right] \quad (12.6) \\ &= 0 + (g'(\mu(x)))^2 \mathbb{V}[Y|X=x] \quad (12.7) \end{aligned}$$

For logistic regression, with Y binary, $\mathbb{V}[Y|X=x] = \mu(x)(1-\mu(x))$. On the other hand, with the logistic link function, $g'(\mu(x)) = \frac{1}{\mu(x)(1-\mu(x))}$. Thus, for logistic regression, $\mathbb{V}[Z|X=x] = [\mu(x)(1-\mu(x))]^{-1}$.

Because the variance of Z changes with X, this is a heteroskedastic regression problem. As we saw in chapter 7, the appropriate way of dealing with such a problem is to use weighted least squares, with weights inversely proportional to the variances. This means that, in logistic regression, the weight at x should be proportional to $\mu(x)(1-\mu(x))$. Notice two things about this. First, the weights depend on the current guess about the parameters. Second, we give lots of weight to cases where $\mu(x) \approx 0$ or where $\mu(x) \approx 1$, and little weight to those where $\mu(x) = 0.5$. This focuses our attention on places where we have a lot of potential information — the distinction between a probability of 0.499 and 0.501 is just a lot easier to discern than that between 0.001 and 0.003!

We can now put all this together into an estimation strategy for logistic regression.

1. Get the data $(x_1, y_1), \ldots (x_n, y_n)$, and some initial guesses β_0, β.
2. until β_0, β converge
   (a) Calculate $\eta(x_i) = \beta_0 + x_i \cdot \beta$ and the corresponding $\widehat{\mu}(x_i)$
   (b) Find the effective transformed responses $z_i = \eta(x_i) + \frac{y_i - \widehat{\mu}(x_i)}{\widehat{\mu}(x_i)(1-\widehat{\mu}(x_i))}$
   (c) Calculate the weights $w_i = \widehat{\mu}(x_i)(1-\widehat{\mu}(x_i))$
   (d) Do a weighted linear regression of z_i on x_i with weights w_i, and set β_0, β to the intercept and slopes of this regression

Our initial guess about the parameters tells us about the heteroskedasticity, which we use to improve our guess about the parameters, which we use to improve our guess about the variance, and so on, until the parameters stabilize. This is called **iterative reweighted least squares** (or "iterative weighted least squares", "iteratively weighted least squares", "iterated reweighted least squares", etc.), abbreviated IRLS, IRWLS, IWLS, etc. As mentioned in the last chapter, this turns out to be *almost* equivalent to Newton's method, at least for this problem.

12.1.1 GLMs in General

The set-up for an arbitrary GLM is a generalization of that for logistic regression. We need

- A **linear predictor**, $\eta(x) = \beta_0 + x \cdot \beta$
- A **link function** g, so that $\eta(x) = g(\mu(x))$. For logistic regression, we had $g(\mu) = \log \mu/(1-\mu)$.
- A **dispersion scale function** V, so that $\mathbb{V}[Y|X=x] = \sigma^2 V(\mu(x))$. For logistic regression, we had $V(\mu) = \mu(1-\mu)$, and $\sigma^2 = 1$.

With these, we know the conditional mean and conditional variance of the response for each value of the input variables x.

As for estimation, basically everything in the IRWLS set up carries over unchanged. In fact, we can go through this algorithm:

1. Get the data $(x_1, y_1), \ldots (x_n, y_n)$, fix link function $g(\mu)$ and dispersion scale function $V(\mu)$, and make some initial guesses β_0, β.
2. Until β_0, β converge:
   (a) Calculate $\eta(x_i) = \beta_0 + x_i \cdot \beta$ and the corresponding $\widehat{\mu}(x_i)$
   (b) Find the effective transformed responses $z_i = \eta(x_i) + (y_i - \widehat{\mu}(x_i)) g'(\widehat{\mu}(x_i))$
   (c) Calculate the weights $w_i = [(g'(\widehat{\mu}(x_i))^2 V(\widehat{\mu}(x_i))]^{-1}$
   (d) Do a weighted linear regression of z_i on x_i with weights w_i, and set β_0, β to the intercept and slopes of this regression

Notice that even if we don't know the over-all variance scale σ^2, that's OK, because the weights just have to be *proportional* to the inverse variance.

12.1.2 Examples of GLMs

12.1.2.1 Vanilla Linear Models

To re-assure ourselves that we are not doing anything crazy, let's see what happens when $g(\mu) = \mu$ (the "identity link"), and $\mathbb{V}[Y|X=x] = \sigma^2$, so that $V(\mu) = 1$. Then $g' = 1$, all weights $w_i = 1$, and the effective transformed response $z_i = y_i$. So we just end up regressing y_i on x_i with no weighting at all — we do ordinary least squares. Since neither the weights nor the transformed response will change, IRWLS will converge exactly after one step. So if we get rid of all this nonlinearity and heteroskedasticity and go all the way back to our very first days of doing regression, we get the OLS answers we know and love.

12.1.2.2 Binomial Regression

In many situations, our response variable y_i will be an integer count running between 0 and some pre-determined upper limit n_i. (Think: number of patients in a hospital ward with some condition, number of children in a classroom passing a test, number of widgets produced by a factory which are defective, number of people in a village with some genetic mutation.) One way to model this would be as a binomial random variable, with n_i trials, and a success probability p_i which is a logistic function of predictors x. The logistic regression we have done so far is the special case where $n_i = 1$ always. I will leave it as an EXERCISE (1) for you to work out the link function and the weights for general binomial regression, where the n_i are treated as known.

One implication of this model is that each of the n_i "trials" aggregated together in y_i is independent of all the others, at least once we condition on the predictors x. (So, e.g., whether any student passes the test is independent of whether any of their classmates pass, once we have conditioned on, say, teacher quality and average previous knowledge.) This may or may not be a reasonable assumption. When the successes or failures are dependent, even after conditioning on the predictors, the binomial model will be mis-specified. We can either try to get more information, and hope that conditioning on a richer set of predictors makes the dependence go away, or we can just try to account for the dependence by modifying the variance ("overdispersion" or "underdispersion"); we'll return to both topics in §12.1.4.

12.1.2.3 Poisson Regression

Recall that the Poisson distribution has probability mass function

$$p(y) = \frac{e^{-\mu}\mu^y}{y!} \tag{12.8}$$

with $\mathbb{E}[Y] = \mathbb{V}[Y] = \mu$. As you remember from basic probability, a Poisson distribution is what we get from a binomial if the probability of success per trial shrinks towards zero but the number of trials grows to infinity, so that we keep the mean number of successes the same:

$$\text{Binom}(n, \mu/n) \rightsquigarrow \text{Pois}(\mu) \tag{12.9}$$

This makes the Poisson distribution suitable for modeling counts with no fixed upper limit, but where the probability that any one of the many individual trials is a success is fairly low. If μ is allowed to change with the predictor variables, we get Poisson regression. Since the variance is equal to the mean, Poisson regression is always going to be heteroskedastic.

Since μ has to be non-negative, a natural link function is $g(\mu) = \log\mu$. This produces $g'(\mu) = 1/\mu$, and so weights $w = \mu$. When the expected count is large, so is the variance, which normally would reduce the weight put on an observation in regression, but in this case large expected counts also provide more information about the coefficients, so they end up getting increasing weight.

12.1.3 Uncertainty

Standard errors for coefficients can be worked out as in the case of weighted least squares for linear regression. Confidence intervals for the coefficients will be approximately Gaussian in large samples, for the usual likelihood-theory reasons, when the model is properly specified. One can, of course, also use either a parametric bootstrap, or resampling of cases/data-points to assess uncertainty.

Resampling of residuals can be trickier, because it is not so clear what counts as a residual. When the response variable is continuous, we can get "standardized" or "Pearson" residuals, $\hat{\epsilon}_i = \frac{y_i - \widehat{\mu}(x_i)}{\sqrt{\widehat{V(\mu(x_i))}}}$, resample them to get $\tilde{\epsilon}_i$, and then add $\tilde{\epsilon}_i \sqrt{\widehat{V(\mu(x_i))}}$ to the fitted values. This does not really work when the response is discrete-valued, however.

[[ATTN: Look up if anyone has a good trick for this]]

12.1.4 Modeling Dispersion

When we pick a family for the conditional distribution of Y, we get a predicted conditional variance function, $V(\mu(x))$. The actual conditional variance $\mathbb{V}[Y|X=x]$ may however not track this. When the variances are larger, the process is **over-dispersed**; when they are smaller, **under-dispersed**. Over-dispersion is more common and more worrisome. In many cases, it arises from some un-modeled aspect of the process — some unobserved heterogeneity, or some missed dependence. For instance, if we observe count data with an upper limit and use a binomial model, we're assuming that each "trial" within a data point is independent; positive correlation between the trials will give larger variance around the mean that the $mp(1-p)$ we'd expect[1].

The most satisfying solution to over-dispersion is to actually figure out where it comes from, and model its origin. Failing that, however, we can fall back on more "phenomenological" modeling. One strategy is to say that

$$\mathbb{V}[Y|X=x] = \phi(x) V(\mu(x)) \tag{12.10}$$

and try to estimate the function ϕ — a modification of the variance-estimation idea we saw in §7.3. In doing so, we need a separate estimate of $\mathbb{V}\left[Y|X=x_i\right]$. This can come from repeated measurements at the same value of x, or from the squared residuals at each data point. Once we have some noisy but independent estimate of $\mathbb{V}\left[Y|X=x_i\right]$, the ratio $\mathbb{V}\left[Y|X=x_i\right]/V(\mu(x_i))$ can be regressed on x_i to estimate ϕ. Some people recommend doing this step, itself, through a generalized linear or generalized additive model, with a gamma distribution for the response, so that the response is guaranteed to be positive.

12.1.5 Likelihood and Deviance

When dealing with GLMs, it is conventional to report not the log-likelihood, but the **deviance**. The deviance of a model with parameters (β_0, β) is defined as

$$D(\beta_0, \beta) = 2[\ell(\text{saturated}) - \ell(\beta_0, \beta)] \tag{12.11}$$

[1] If (for simplicity) all the trials have the same covariance ρ, then the variance of their sum is $mp(1-p) + m(m-1)\rho$ (why?).

Here, $\ell(\beta_0, \beta)$ is the log-likelihood of our model, and ℓ(saturated) is the log-likelihood of a **saturated** model which has one parameter per data point. Thus, models with high likelihoods will have low deviances, and vice versa. If our model is correct and has $p+1$ parameters in all (including the intercept), then the deviance will generally approach a χ^2 distribution asymptotically, with $n-(p+1)$ degrees of freedom (Appendix I); the factor of 2 in the definition is to ensure this.

For discrete response variables, the saturated model can usually ensure that $\Pr(Y = y_i | X = x_i) = 1$, so $\ell(\text{saturated}) = 0$, and deviance is just twice the negative log-likelihood. If there are multiple data points with the same value of x but different values of y, then $\ell(\text{saturated}) < 0$. In any case, even for repeated values of x or even continuous response variables, differences in deviance are just twice differences in log-likelihood: $D(\text{model}_1) - D(\text{model}_2) = 2[\ell(\text{model}_2) - \ell(\text{model}_1)]$.

12.1.5.1 Maximum Likelihood and the Choice of Link Function

Having chosen a family of conditional distributions, it may happen that when we write out the log-likelihood, the latter depends on the *both* the response variables y_i and the coefficients only through the product of y_i with some transformation of the conditional mean $\widehat{\mu}$:

$$\ell = \sum_{i=1}^{n} f(y_i, x_i) + y_i g(\widehat{\mu}_i) + h(\theta) \tag{12.12}$$

In the case of logistic regression, examining Eq. 11.8 (§11.2.1, p. 255) shows that the log-likelihood can be put in this form with $g(\widehat{\mu}_i) = \log \widehat{\mu}_i/(1-\widehat{\mu}_i)$. In the case of a Gaussian conditional distribution for Y, we would have $f = -y_i^2/2$, $g(\widehat{\mu}_i) = \widehat{\mu}_i$, and $h(\theta) = -\widehat{\mu}_i^2$. When the log-likelihood can be written in this form, $g(\cdot)$ is the "natural" transformation to apply to the conditional mean, i.e., the natural link function, and assures us that the solution to iterative least squares will converge on the maximum likelihood estimate.[2] Of course we are free to nonetheless use other transformations of the conditional expectation.

[2] To be more technical, we say that a distribution with parameters θ is an **exponential family** if its probability density function at x is $\exp f(x) + T(x) \cdot g(\theta)/z(\theta)$, for some vector of statistics T and some transformation g of the parameters. (To ensure normalization, $z(\theta) = \int \exp(f(x) + T(x) \cdot g(\theta)) dx$. Of course, if the sample space x is discrete, replace this integral with a sum.) We then say that $T(\cdot)$ are the "natural" or "canonical" **sufficient statistics**, and $g(\theta)$ are the "natural" parameters. Eq. 12.12 is picking out the natural parameters, presuming the response variable is itself the natural sufficient statistic. Many of the familiar families of distributions, like Gaussians, exponentials, gammas, Paretos, binomials and Poissons are exponential families. Exponential families are very important in classical statistical theory, and have deep connections to thermodynamics and statistical mechanics (where they're called "canonical ensembles", "Boltzmann distributions" or "Gibbs distributions" (Mandelbrot, 1962)), and to information theory (where they're "maximum entropy distributions", or "minimax codes" (Grünwald, 2007)). Despite their coolness, they are a rather peripheral topic for our sort of data analysis — though see Guttorp (1995) for examples of using them in modeling discrete processes. Any good book on statistical theory (e.g., Casella and Berger 2002) will have a fairly extensive discussion; Barndorff-Nielsen (1978) and Brown (1986) are comprehensive treatments.

12.1.6 R: `glm`

As with logistic regression, the workhorse R function for all manner of GLMs is, simply, `glm`. The syntax is strongly parallel to that of `lm`, with the addition of a `family` argument that specifies the intended distribution of the response variable (`binomial`, `gaussian`, `poisson`, etc.), and, optionally, a link function appropriate to the family. (See `help(family)` for the details.) With `family="gaussian"` and an identity link function, its intended behavior is the same as `lm`.

12.2 Generalized Additive Models

In the development of generalized linear models, we use the link function g to relate the conditional mean $\widehat{\mu}(x)$ to the linear predictor $\eta(x)$. But really nothing in what we were doing required η to be *linear* in x. In particular, it all works perfectly well if η is an additive function of x. We form the effective responses z_i as before, and the weights w_i, but now instead of doing a linear regression on x_i we do an additive regression, using backfitting (or whatever). This gives us a generalized additive model (GAM).

Essentially everything we know about the relationship between linear models and additive models carries over. GAMs converge somewhat more slowly as n grows than do GLMs, but the former have less bias, and strictly include GLMs as special cases. The transformed (mean) response is related to the predictor variables not just through coefficients, but through whole partial response functions. If we want to test whether a GLM is well-specified, we can do so by comparing it to a GAM, and so forth.

In fact, one could even make $\eta(x)$ an arbitrary smooth function of x, to be estimated through (say) kernel smoothing of z_i on x_i. This is rarely done, however, partly because of curse-of-dimensionality issues, but also because, if one is going to go that far, one might as well just use kernels to estimate conditional distributions, as we will see in Chapter 14.

12.3 Further Reading

At our level of theory, good references on generalized linear and generalized additive models include Faraway (2006) and Wood (2006), both of which include extensive examples in R. Tutz (2012) offers an extensive treatment of GLMs with categorical response distributions, along with comparisons to other models for that task.

Overdispersion is the subject of a large literature of its own. All of the references just named discuss methods for it. Lambert and Roeder (1995) is worth mentioning for introducing some simple-to-calculate ways of detecting and describing overdispersion which give some information about *why* the response is over-dispersed. One of these (the "relative variance curve") is closely related to the idea sketched above about estimating the dispersion factor.

12.4 Exercises

1. In binomial regression, we have $Y|X = x \sim \text{Binom}(n, p(x))$, where $p(x)$ follows a logistic model. Work out the link function $g(\mu)$, the variance function $V(\mu)$, and the weights w, assuming that n is known and not random.

2. Problem set A.14, on predicting the death rate in Chicago, is a good candidate for using Poisson regression. Repeat the exercises in that problem set with Poisson-response GAMs. How do the estimated functions change? Why is this any different from just taking the log of the death counts, as we did in the homework?