7

The Generalized Linear Model

7.1 THE GENERALIZED LINEAR MODEL

In Part III, we focus on the most widespread modeling framework applying likelihood principles: the *Generalized Linear Model* (GLM). Any GLM has the following components:

- A specified probability model, $Y_i \sim f$ where f is a member of the *exponential family* of distributions,
- A *linear predictor* term, $\mathbf{x}_i^\intercal \boldsymbol{\beta}$,
- A *link function*, $g(\cdot)$, such that $\mathrm{E}[Y_i] = g^{-1}(\mathbf{x}_i^\intercal \boldsymbol{\beta})$.

In the language we have been using so far, the probability model describes the stochastic component of Y while the linear predictor represents the systematic component. The linear predictor and the link function emphasize that GLMs are models of the mean or expected response, conditional on covariates and parameters. The inverse link function serves to map the linear predictor values into intervals appropriate to the assumed probability model; it links the linear predictor with the conditional mean.

In case you were wondering … 7.1 Which "GLM"?

The Generalized Linear Model is not be confused with the "General Linear Model," "General Linear Methods" or "Generalized Least Squares."

General Linear Methods are a collection of tools for solving differential equations; they are outside the scope of this text.

Generalized Least Squares is a generalization of OLS to situations in which residuals are correlated. We introduced some of these ideas in Section 2.4.1 in the context of robust standard errors.

https://doi.org/10.1017/9781316888544.010 Published online by Cambridge University Press

The General Linear Model is the matrix generalization of standard OLS-based linear regression to the situation of q outcomes:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ is now a $n \times q$ matrix of response variables, $\mathbf{X}$ is $n \times k$ design matrix, $\boldsymbol{\beta}$ is a $k \times q$ matrix of to-be-estimated regression parameters, and $\boldsymbol{\varepsilon}$ is a $n \times q$ matrix of errors, assumed to follow multivariate normal distribution.

Nelder and Wedderburn (1972) coined the term "Generalized Linear Model" and, more importantly, built out the mathematics of the exponential family, developed the concept of a linear predictor and a link function, and showed that IWLS can be used to find the MLE for models in the exponential family.

7.2 THE EXPONENTIAL FAMILY

The *exponential family* is a class of probability distribution and mass functions that can be written in certain mathematically convenient forms. It is important to distinguish the exponential *family* from the exponential *distribution*. To further confuse matters, the exponential distribution is also a member of the exponential family.

In case you were wondering ... 7.2 Exponential distribution

Suppose $X \in [0, \infty)$. We say that X follows the exponential distribution with rate parameter λ, $X \sim f_e(x; \lambda)$, if

$$f_e(x) = \lambda \exp(-\lambda x),$$

with $\mathrm{E}[X] = \lambda^{-1}$ and $\mathrm{var}(X) = \lambda^{-2}$.

Definition 7.1 (Exponential Family)**.** A distribution or mass function $f(x; \boldsymbol{\theta})$ is of the *exponential family* if it can be written as

$$f(x; \boldsymbol{\theta}) = \exp\left[\eta(\boldsymbol{\theta})^{\intercal} h(x) - A(\boldsymbol{\theta}) + c(x)\right]$$

for some functions $\eta(\boldsymbol{\theta})$, $h(x)$, $A(\boldsymbol{\theta})$ and $c(x)$. The function $h(x)$ returns a vector of "sufficient statistics" of the same length as $\eta(\boldsymbol{\theta})$.

An extension of the exponential family commonly used in GLMs is the *exponential dispersion model.*

Definition 7.2 (Exponential dispersion model)**.** A distribution function $f(x; \theta, \phi)$ is said to be of the *natural exponential dispersion family* if it can be written of the form

https://doi.org/10.1017/9781316888544.010 Published online by Cambridge University Press

$$f(x;\theta,\phi) = \exp\left[\frac{x\theta - A(\theta)}{\phi} + c(x,\phi)\right]$$

for some functions $A(\theta)$ and $c(x,\phi)$.

Written in this form, we refer to θ as the *canonical parameter* and ϕ as the *dispersion parameter*. The key feature of exponential family distributions is that they can be factored into terms containing only data (x), only parameters (θ,ϕ), or constants that do not depend on either θ or the data. This implies that distributions in the exponential family have a particular relationship between their mean and variance. Specifically, for scalar θ, $\mathrm{E}[X] = \mu = \frac{dA(\theta)}{d\theta}$ while $\mathrm{var}(X) = \phi\frac{d^2A(\theta)}{d\theta^2}$. It is this relationship that makes modeling and estimation more convenient. Products of exponential family distributions are themselves in the exponential family. So it follows that the joint distribution of i.i.d. samples from an exponential family distribution is also in the exponential family.

Many of the most commonly used distributions are members of the exponential family, including the normal, log-normal, Bernoulli, beta, χ^2, Dirichlet, gamma, and Poisson. Distributions where the support for the distribution involves θ (e.g., the uniform distribution) are not in the exponential family. Neither are the Cauchy nor t-distributions.

7.3 PAST EXAMPLES AS GLMS

In Chapter 3 we derived the Bernoulli GLM. The Bernoulli distribution can be written as

$$\begin{aligned}
f_B(x;\theta) &= \theta^x(1-\theta)^{1-x} \\
&= \exp\left[\log(\theta^x(1-\theta)^{1-x})\right] \\
&= \exp\left[x\log\theta + (1-x)\log(1-\theta)\right] \\
&= \exp\left[x\log\frac{\theta}{1-\theta} + \log(1-\theta)\right] \\
&= \exp\left[x\mathrm{logit}(\theta) + \log(1-\theta)\right].
\end{aligned}$$

From here we see that the Bernoulli distribution is in the exponential family, with dispersion fixed at 1. This expression also makes it easy to see how we arrive at the logit as the link function for the Bernoulli GLM.

In deriving the normal-linear model in Chapter 1, we actually specified a GLM in which $Y_i \sim f_{\mathcal{N}}(y_i;\boldsymbol{\theta}_i)$ and where $\boldsymbol{\theta}_i = (\mu_i,\sigma^2)$. To see that the normal distribution (with unknown mean and variance) is of the exponential family, note that we can factor the distribution to be

$$f_{\mathcal{N}}(y;\mu,\sigma^2) = \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{\log 2\pi\sigma^2}{2} - \frac{y^2}{2\sigma^2}\right],$$

https://doi.org/10.1017/9781316888544.010 Published online by Cambridge University Press

so for the normal distribution, we have canonical parameter μ and dispersion parameter σ^2.

For the normal GLM, $\mathrm{E}[Y_i] = \mu_i = \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}$. Since the normal distribution has support over the entire real line it can take any real number as a mean value. The link function for the normal GLM is correspondingly the identity $\mathbf{I}_n\mathbf{X}\boldsymbol{\beta} = \mathbf{I}_n^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

7.3.1 GLMs in $\mathcal{R}$

The syntax of the $\mathcal{R}$ function `glm` is built on the theory of the GLM. In particular, `glm` requires that the user:

- Supply a linear predictor term in the form of an $\mathcal{R}$ formula. In the Mroz example in Chapter 3 we specified `lfpbin~young.kids + school.kids + age + college.woman + wage`.
- Choose a probability model using the `family` argument. In the Mroz example we specified `family=binomial`.
- Choose a link function (or write your own). In the Mroz example we chose the logit link, `family=binomial(link='logit')`, but others are available, including the probit and complementary log-log.

In summarizing a model fit with `glm`, $\mathcal{R}$ returns information about the dispersion parameter for that family. For example, using `glm` to estimate the LPM described in Table 3.4, $\mathcal{R}$ returns

```
(Dispersion parameter for Gaussian family taken to be
   0.2155109)
```

As we have just seen, in the normal model the dispersion parameter corresponds to σ^2. Thus $0.22 = \hat{\sigma}^2$ in the Mroz LPM example. The square root of this quantity for a linear-normal GLM corresponds to the standard error of the regression. Table 3.4 reports a regression standard error of 0.46, which is approximately $\sqrt{0.2155109}$. In fitting the logit model for the Mroz example the code chunk in Section 3.4.2 shows that $\mathcal{R}$ reports the dispersion parameter fixed at 1, as the binomial GLM requires.

7.4 "QUASI-" AND "PSEUDO-"LIKELIHOOD

The functional relationship between the mean and variance among exponential family distributions has direct application in the "quasilikelihood" framework, sometimes called "pseudolikelihood." Rather than specify a full probability model (and thus a likelihood function), a quasilikelihood only requires a model for the mean and a function that links the mean to the variance:

$$\mathrm{E}[Y_i] = \mu_i(\beta) = g^{-1}(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}),$$
$$\mathrm{var}(Y_i) = \phi V(\mu_i).$$

Definition 7.3 (Quasilikelihood)**.** The *quasilikelihood function* $Q(y_i;\mu_i)$ is defined by the relation

$$\frac{\partial Q_i}{\partial \mu_i} = \frac{y_i - \mu_i}{\phi V(\mu_i)},$$

where $V(\mu_i)$ is assumed known. If $\mu_i \equiv g^{-1}(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta})$, then the quasilikelihood can be expressed with the relation

$$\frac{\partial Q_i}{\partial \beta_j} = \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j}. \tag{7.1}$$

Equation 7.1 defines the *quasiscore* function, $U(\boldsymbol{\beta})$, which is a vector of length k. Setting $U(\boldsymbol{\beta}) = \mathbf{0}$ gives a set of *quasiscore* or estimating equations. The $\hat{\boldsymbol{\beta}}$ that satisfies the quasiscore equations gives the quasilikelihood estimates of $\boldsymbol{\beta}$.

Based on the quasiscore we can derive the quasi-information as

$$-\frac{\partial U}{\partial \boldsymbol{\beta}} = \frac{\mathbf{D}\mathbf{V}^{-1}\mathbf{D}}{\phi}, \tag{7.2}$$

where $\mathbf{V}$ is a diagonal matrix with ith entry $V(\mu_i)$ and $\mathbf{D}$ is the $n \times k$ matrix with (i,j) element $\partial\mu_i(\beta)/\partial\beta_j$. The dispersion parameter is typically estimated as

$$\hat{\phi} = \frac{1}{n-k}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}. \tag{7.3}$$

Using methods analogous to those in Chapter 2, it can be shown that if the model for the mean is correct, then the maximum quasilikelihood estimator is consistent for β and asymptotically normal, with covariance matrix given by the inverse quasi-information. Note the close correspondence between this variance formula and the "sandwich" covariance estimator from Section 2.4.1. The chief difference is the form of the "meat" in the sandwich and the inclusion of the dispersion parameter.

In the quasilikelihood world the specification of the variance function determines the "family." For example, if $V(\mu_i) = \mu_i(1 - \mu_i)$, then we have a quasibinomial model (see Chapter 3). If $V(\mu_i) = \mu_i$, then we have a quasipoisson (see Chapter 10).

Because of the functional relation between the mean and variance in the exponential family, the quasiscore equation turns out to be exactly the score equation for the log-likelihood of a particular distribution. As a result, the $\hat{\boldsymbol{\beta}}$ from maximum likelihood will equal the $\hat{\boldsymbol{\beta}}$ from a quasilikelihood estimation among the exponential family. In fact, the consistency of the quasilikelihood

https://doi.org/10.1017/9781316888544.010 Published online by Cambridge University Press

estimates relies on the functional connection between the mean and variance that the exponential family provides. What differs between MLE and quasilikelihood is the covariance matrix and therefore the standard errors. The chief benefit of a quasilikelihood approach is the ability to partially relax some of the distributional assumptions inherent in a fully-specified likelihood. For example, quasilikelihood allows us to account for overdispersion in binary or count data. But the quasilikelihood does not calculate an actual likelihood, so likelihood-based quantities like the AIC and BIC as well as likelihood-based tests are not available.

7.5 CONCLUSION

This brief chapter introduced the concept and notation of the Generalized Linear Model, something we have already seen in the context of linear and logistic regression. The GLM provides a unified way of specifying and estimating a variety of models with different distributional assumptions. So long as we are hypothesizing probability models that rely on distributions in the exponential family, we can decompose our model building into the specification of a linear predictor term – covariates and regression parameters – and the link function that maps the linear predictor into the expected value.

Subsequent chapters in this part of the book introduce commonly used GLMs for outcomes that are integer counts as well as ordered and unordered categorical variables. Other GLMs exist for outcomes that are strictly positive, negative, or bounded, as well as many others, a testament to the flexibility of the exponential family of distributions. But it is also important to recognize that GLMs are not the only ways we can apply the method of maximum likelihood. We take up more complicated modeling tools in the last part of the book.

7.6 FURTHER READING

Advanced Study

McCullagh and Nelder (1989) is the canonical citation for the GLM. A variety of subsequent texts have expanded on these ideas, including Agresti (2002).

Wedderburn (1974) introduced quasilikelihood. A widely used extension of these ideas appears in the theory and implementation of Generalized Estimating Equations (GEE) (Hardin and Hilbe, 2012; Ziegler, 2011). Zorn (2000) summarizes the GEE approach with applications to political science.

https://doi.org/10.1017/9781316888544.010 Published online by Cambridge University Press