Week 3 Notes: More Generalized Linear Models and Likelihood Theory

Antonio R. Linero
University of Texas at Austin

Goals

1. Learn basic theory underlying GLMs.
2. Learn how to use statistical theory to test simple hypotheses and perform inference.

Likelihood of a GLM

The likelihood is given by

$$L(\beta, \phi) = \prod_{i=1}^{N} \exp\left\{ \frac{Y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(Y_i; \phi/\omega_i) \right\},$$

- $\theta_i \equiv (b')^{-1}(\mu_i)$
- $\mu_i \equiv g^{-1}(X_i^\top \beta)$.

Score Function

The score function is given by

$$
\begin{aligned}
s(\beta, \phi) &= \frac{\partial}{\partial \beta} \log L(\beta, \phi) \\
&= \sum_{i=1}^{N} \frac{\partial}{\partial \beta} \frac{Y_i \theta_i - b(\theta_i)}{\phi / \omega_i} + c(Y_i; \phi/\omega_i). \\
&= \underbrace{\sum_{i=1}^{N} \frac{\omega_i (Y_i - \mu_i) X_i}{\phi V(\mu_i) g'(\mu_i)}}_{\text{weighted sum of residuals}} .
\end{aligned}
$$

The MLE corresponds to the solution to $\widehat{\beta}$ of $s(\beta, \phi) = 0$. It is an example of an *M-estimator!*

The Fisher Information

Exercise: Deriving the Fisher Information

We define the *expected* and *observed Fisher Information* to be

$$\mathcal{I}(\beta,\phi) = -\mathbb{E}\left\{\frac{\partial^2}{\partial\beta\partial\beta^\top}\log L(\beta,\phi) \mid \beta,\phi\right\}. \qquad \text{and} \qquad \mathcal{J}(\beta,\phi) = -\frac{\partial^2}{\partial\beta\partial\beta^\top}\log L(\beta,\phi).$$

Show that we have

$$\langle\mathcal{J}(\beta,\phi)\rangle_{jk} = \frac{1}{\phi}\sum_{i=1}^{N} X_{ij}X_{ik}\left\{\frac{\omega_i}{V(\mu_i)g'(\mu_i)^2} - \frac{\omega_i(Y_i-\mu_i)}{g'(\mu_i)}\left(\frac{\partial}{\partial\mu_i}\frac{1}{V(\mu_i)g'(\mu_i)}\right)\right\}$$

and

$$\langle\mathcal{I}(\beta,\phi)\rangle_{jk} = \frac{1}{\phi}\sum_{i=1}^{N} X_{ij}X_{ik}\frac{\omega_i}{V(\mu_i)g'(\mu_i)^2}$$

Show also that $\mathcal{I}(\beta,\phi) = \mathcal{J}(\beta,\phi)$ when the canonical link is used. Hence we can write

$$\mathcal{I}^{-1} = \phi(\boldsymbol{X}^\top D\boldsymbol{X})^{-1}$$

Aside: Likelihood-Based Inference

- Define $\mathcal{D} = \{Z_i : i = 1, \ldots, N\}$ iid from $f_{\theta_0}(z)$
- $\{f_\theta : \theta \in \Theta\}$ is a parametric family of densities.
- Likelihood theory quantities:

$$\ell(\theta) = \sum_{i=1}^{N} \log f(Z_i \mid \theta),$$

$$s(\theta) = \frac{\partial}{\partial \theta} \ell(\theta),$$

$$\mathcal{I}(\theta) = -\mathbb{E}\left\{ \frac{\partial^2}{\partial \theta \partial \theta^\top} \ell(\theta) \mid \theta \right\}.$$

Score Methods

Exercise: Score Methods

Using the multivariate central limit theorem, show that

$$s(\theta_0) \overset{\bullet}{\sim} \mathrm{Normal}\{0, \mathcal{I}(\theta_0)\},$$

but only if we plug in the true value θ_0 *Note:* this asymptotic notation means that $X \overset{\bullet}{\sim} \mathrm{Normal}(\mu, \Sigma)$ if-and-only-if $\Sigma^{-1/2}(X - \mu) \to \mathrm{Normal}(0, \mathrm{I})$ in distribution.

What can we do with this?

Wald Methods

Exercise: Wald Methods

Using Taylor's theorem, we have

$$s(\theta_0) = s(\widehat{\theta}) - \mathcal{J}(\theta^\star)(\theta_0 - \widehat{\theta}) = -\mathcal{J}(\theta^\star)(\theta_0 - \widehat{\theta}).$$

where $\theta^\star$ lies on the line segment connecting θ_0 and $\widehat{\theta}$. Now, assume that we know somehow that $\widehat{\theta}$ is a *consistent* estimator of θ_0. Show that

$$\widehat{\theta} \stackrel{\bullet}{\sim} \mathrm{Normal}(\theta_0, \mathcal{I}(\theta_0)^{-1}).$$

What can we do with this?

LRT Methods

Exercise: Likelihood Ratio Methods

Consider the Taylor expansion

$$\ell(\theta_0) = \ell(\widehat{\theta}) + s(\widehat{\theta})^\top(\theta_0 - \widehat{\theta}) - \frac{1}{2}(\theta_0 - \widehat{\theta})^\top \mathcal{J}(\theta^\star)(\theta_0 - \widehat{\theta})$$

where $\theta^\star$ lies on the line segment connecting $\widehat{\theta}$ and θ_0. Show that

$$-2\{\ell(\theta_0) - \ell(\widehat{\theta})\} \to \chi^2_P.$$

in distribution, where $P = \dim(\theta)$. Recall here that the χ^2_P distribution is the distribution of $\sum_{i=1}^P U_i^2$ where $U_1, \ldots, U_P \stackrel{\text{iid}}{\sim} \text{Normal}(0, 1)$.

Wilk

Theorem: Wilk's Theorem

Suppose that $\{f_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$ is a parametric family satisfying certain regularity conditions. Consider the null hypothesis $H_0 : \eta = \eta_0$, let $\widehat{\theta}_0$ denote the MLE obtained under the null model, and let $(\widehat{\theta}, \widehat{\eta})$ denote the MLE under the unrestricted model. Then, if (θ_0, η_0) denote the values of the parameters that generated the data (so that H_0 is true) then

$$-2\{\ell(\widehat{\theta}_0, \eta_0) - \ell(\widehat{\theta}, \widehat{\eta})\} \overset{\bullet}{\sim} \chi^2_D$$

where $D = \dim(\eta)$, as the amount of data tends to ∞.

- Note vagueness!
- Great for hypothesis testing!

Likelihood-Based Inference for GLMs

- Life is easy for Bayesians: all inference flows from posterior.

Likelihood-Based Inference for GLMs

- Life is easy for Bayesians: all inference flows from posterior.
- Frequentist inference usually depends on the asymptotics in practice.

Likelihood-Based Inference for GLMs

- Life is easy for Bayesians: all inference flows from posterior.
- Frequentist inference usually depends on the asymptotics in practice.

Definition: Deviance of a GLM

The *saturated model* has a separate parameter for all unique values of x in $\mathcal{D}$:

$$f(y \mid x, \phi/\omega) = \exp\left\{ \frac{y\theta_x - b(\theta_x)}{\phi/\omega} + c(y; \phi/\omega). \right\}.$$

The *residual deviance* of a model is defined by

$$D = -2\phi \left\{ \ell(\widehat{\theta}) - \ell(\widehat{\theta}_x) \right\}$$

where $\ell(\theta) = \sum_{i=1}^{N} \dfrac{\omega_i(Y_i\theta_i - b(\theta_i))}{\phi}$ is the log-likelihood of θ and $\widehat{\theta}_{xi} = (b')^{-1}(Y_i)$.

The *scaled deviance* is $D^\star = D/\phi$: it is the LRT statistic for comparing the model with the saturated model which has the maximal number of model parameters in the GLM.

Estimating the Dispersion

Exercise: Estimating the Dispersion

Show that the quantity

$$\widetilde{\phi} = \frac{1}{N} \sum_i \frac{\omega_i (Y_i - \mu_i)^2}{V(\mu_i)}$$

is unbiased for ϕ. We don't use $\widetilde{\phi}$ because we don't know the μ_i's, so the modified denominator in $\widehat{\phi}$ compensates for the "degrees of freedom" used to estimate β.

In practice: $\widehat{\phi} = \frac{1}{N-P} \sum_i \frac{(Y_i - \widehat{\mu}_i)^2}{V(\mu_i)}$.

Analysis of Deviance

Analysis of Deviance

1. Goodness-of-fit test with nonparametric alternative:
   sometimes, $D^\star \overset{\bullet}{\sim} \chi^2_{N-P}$ under null that model is correct.

Analysis of Deviance

1. Goodness-of-fit test with nonparametric alternative: sometimes, $D^\star \stackrel{\bullet}{\sim} \chi^2_{N-P}$ under null that model is correct.
2. If model $\mathcal{M}_0$ is a submodel of $\mathcal{M}_1$ then the LRT statistic for comparing these models is $D_0^\star - D_1^\star$. Under very weak conditions, we have $D_0^\star - D_1^\star \stackrel{\bullet}{\sim} \chi^2_K$ where K is the difference in the number of parameters between the two models.

More Ships

```r
## Load
ships <- MASS::ships

## Fit GLM (see previous notes)
ships_glm <- glm(
  incidents ~ type + factor(period) + factor(year),
  family = poisson,
  offset = log(service),
  data = dplyr::filter(ships, service > 0)
)

anova(ships_glm, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: incidents
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                              33    146.328
## type            4   55.439        29     90.889 2.629e-11 ***
## factor(period)  1   20.786        28     70.103 5.135e-06 ***
## factor(year)    3   31.408        25     38.695 6.975e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Go over table, and goodness of fit.

Goodness of Fit Conditions

- Number of observations is small relative to number of parameters...
- Can be shown that things would be OK if the counts are at least large.

```
print(ships$incidents)
```

```
##  [1]  0  0  3  4  6 18  0 11 39 29 58 53 12 44  0 18  1  1  0  1  6
## [26]  0  0  0  2 11  0  4  0  0  7  7  5 12  0  1
```

Likelihood-Based Confidence Intervals

Confidence Set:

$$\{\beta_{01} : \text{The LRT fails to reject } H_0 : \beta_0 = \beta_{01}\}.$$

If the LRT has Type I error rate α for all β_{01} then the above set is guaranteed to be a $100(1-\alpha)\%$ confidence set.

```
confint(ships_glm)
```

```
## Waiting for profiling to be done...

##                         2.5 %      97.5 %
## (Intercept)       -6.84305161 -5.98968373
## typeB             -0.88135891 -0.18353080
## typeC             -1.37649167 -0.07452031
## typeD             -0.67151807  0.47524605
## typeE             -0.14346972  0.78520455
## factor(period)75   0.15339419  0.61740478
## factor(year)65     0.40752296  0.99512708
## factor(year)70     0.48728088  1.15369754
## factor(year)75    -0.01234169  0.90386446
```

Drop-1 Tests

- `anova` does sequential tests.
- `drop1` does "leave one out" tests

```
drop1(ships_glm, test = "LRT")
```

```
## Single term deletions
##
## Model:
## incidents ~ type + factor(period) + factor(year)
##                Df Deviance    AIC    LRT  Pr(>Chi)
## <none>             38.695 154.56
## type            4   62.365 170.23 23.670 9.300e-05 ***
## factor(period)  1   49.355 163.22 10.660  0.001095 **
## factor(year)    3   70.103 179.97 31.408 6.975e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```