Generalized Linear Models

Inference and Residuals

Demetris Athienitis

UNIVERSITY *of* FLORIDA

Section 1

1. Framework
2. Wald
3. LRT
   - Deviance
   - Parameter Testing
4. Residuals

Inference with GLMs

Three main strategies for inference:

(i) Wald

(ii) Score

(iii) Likelihood Ratio Test (LRT)

- Some are easier than others, but software can do all three
- We will most often perform Wald or LRT for this class
- Each method comes with it's own assumptions
- In big sample sizes they should perform similarly
- All of them rely on asymptotic approximations

Framework

Since parameter estimation is done via ML, and MLEs are asymptotically normal, inference is done in the traditional way.

Inference on parameters

Let $\boldsymbol{\theta} = (\alpha, \beta)$ denote the parameter vector

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(0, \frac{1}{I(\boldsymbol{\theta}_0)}\right)$$

where $I(\boldsymbol{\theta}_0)$ is the *Fisher information* evaluated at $\boldsymbol{\theta}_0$ (not covered in this class).

Section 2

1. Framework
2. Wald
3. LRT
   - Deviance
   - Parameter Testing
4. Residuals

Wald

- To test $\mathrm{H}_0 : \beta = \beta_0$ you can create the test statistic

$$TS = \frac{\hat{\beta} - \beta_0}{s_{\hat{\beta}}} \overset{\mathrm{H}_0}{\sim} N(0, 1)$$

  and obtain p-value in traditional way.

- A $100(1 - \alpha)\%$ CI on β can also be created

$$\hat{\beta} \mp z_{1-\alpha/2} \left(s_{\hat{\beta}} \right)$$

These methods can be extended to one-sided tests.

Example (Infant malformatrion continued)

```
> summary(malform.logit)


Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9605     0.1154 -51.637   <2e-16 ***
Alcohol       0.3166     0.1254   2.523   0.0116 *
```

we can create a 95% CI on β

$$0.3166 \mp 1.96(0.1254) \longrightarrow (0.070816, 0.562384)$$

We have seen and will see `R` functions that create CIs but their default is not the Wald method.

Section 3

1. Framework
2. Wald
3. LRT
   - Deviance
   - Parameter Testing
4. Residuals

Deviance

Deviance is actually the LRT for *goodness of model fit*, seen in Chapter 2 as $G^2 = -2 \log \Lambda \underset{\text{approx.}}{\overset{\text{H}_0}{\sim}} \chi^2_{df}$ for hypothesis

H$_0$: model adequately fits

$$D(y; \hat{\mu}) := G^2 = -2[L(\hat{\mu}; y) - L(y; y)] \xrightarrow[\text{H}_0]{d} \chi^2_{df}$$

with p-value being $P\left(\chi^2_{df} \geq G^2\right)$

- $L(\hat{\mu}; y)$ is the log-likelihood of the fitted model
- $L(y; y)$ is the log-likelihood of the *saturated* model, that is the model that has a separate parameter for each observation giving a perfect fit but with 0 degrees of freedom (so no inference)
- *df* as discussed in notes, no. of observations - no. of parameters

Goodness of Fit

Remark

A *goodness of fit* can be used only if the number of predictor levels is fixed and relatively small to the overall sample size. Either, X^2 or G^2 can be used to compare the observed counts to the values predicted by the fitted model.

Remark

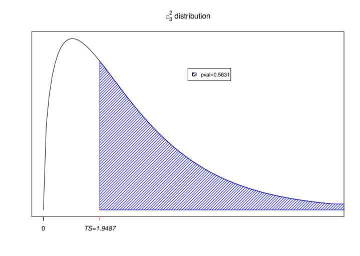
If the data is not grouped you may still perform a goodness of fit by

- grouping your predictor(s). For example, for temperature you could create groups 31-40, 41-50, . . . and then create scores such as 35, 45, . . . ensuring that the number of predictor levels remains fixed.
- comparing current model to a “fuller” model rather than to a saturated model (fullest). A fuller model could be one with polynomial terms, interactions, etc.

Example (revisited)

- (Infant Malformation) A GoF can be used (with either X^2 or G^2) as there are only 5 binomials, with predictor levels 0, 0.5, 1.5, 4.0, 7.0, and as more women are surveyed/sampled the number of binomials (rows of data) remains fixed.

```
Residual deviance: 1.9487  on 3  degrees of freedom
>  sum(resid(malform.logit,type=‘‘pearson’’)^2)
[1] 2.20523
```



$$G^2 = 1.9487 \; (X^2 = 2.20523) \text{ with } df = 3 \;\longrightarrow\; \text{p-value } = 0.5831$$

Example (revisited)

- (Challenger disaster) A GoF is not adequate as each row corresponds to a Bernoulli trial, that is a 0 or 1, and the temperature (predictor) was on a continuous, non-grouped, scale. As sample size increases so will the number of rows in the data.

Remark

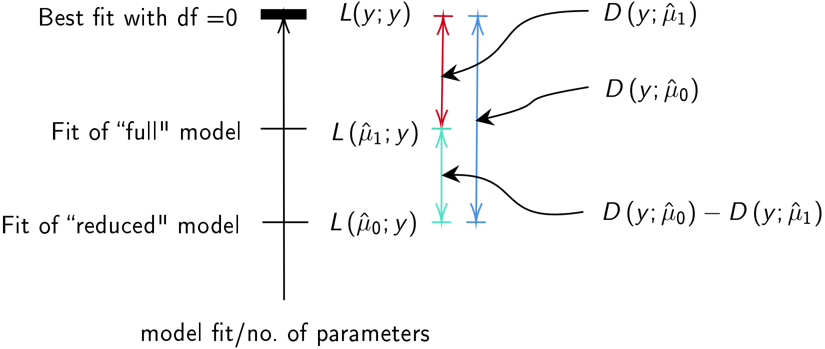
Goodness of fit can also be performed - preferred even - by using X^2 instead of G^2, as X^2 converges faster to a χ^2.

Parameter testing

LRT can be used to test $H_0 : \beta = \beta_0$ using deviances.

$$\begin{aligned} G^2 &= D(y; \hat{\mu}_0) - D(y; \hat{\mu}_1) \\ &= -2[L(\hat{\mu}_0; y) - L(y; y)] - (-2)[L(\hat{\mu}_1; y) - L(y; y)] \\ &= -2[L(\hat{\mu}_0; y) - L(\hat{\mu}_1; y)] \\ &\xrightarrow[H_0]{d} \chi^2_{df} \end{aligned}$$

- $L(\hat{\mu}_0; y)$ is the log-likelihood of the reduced (under the null) model
- $L(\hat{\mu}_1; y)$ is the log-likelihood of the fitted model
- df is the difference in degrees of freedom of the two models which corresponds to the dimension reduction of our coefficient parameter vector, in this case 1 as we are restricting one parameter $\beta = \beta_0$



Best fit with df =0

$L(y;y)$

$D\left(y;\hat{\mu}_1\right)$

$D\left(y;\hat{\mu}_0\right)$

Fit of “full" model

$L\left(\hat{\mu}_1;y\right)$

Fit of “reduced" model

$L\left(\hat{\mu}_0;y\right)$

$D\left(y;\hat{\mu}_0\right) - D\left(y;\hat{\mu}_1\right)$

model fit/no. of parameters

.

Null Deviance

The *Null Deviance* that is usually provided in `R` is the deviance under

$$\mathrm{H}_0 : \beta = 0 \quad (\beta_i = 0 \; \forall i \text{ for models with more than one predictor})$$

So that

$$\begin{aligned}\text{Null Deviance} - \text{Residual Deviance} &= D(y;\hat{\mu}_0) - D(y;\hat{\mu}_1) \\ &= G^2\end{aligned}$$

which is the likelihood ratio test statistic.

Deviance

For binomial and Poisson models

$$D(y; \hat{\mu}) = 2 \sum_{i=1}^{n} y_i \log(y_i / \hat{\mu}_i)$$

The LRT can be used to create a $100(1 - \alpha)\%$ confidence interval on β. That is, finding all the null values β_0 for which would yield a test statistics with a large p-value. It is a bit more complicated than Wald so we use software.

R

Use `confint` to obtain the LRT confidence intervals.

Example (Infant Malformation continued)

Testing $H_0 : \beta = 0$ via deviances.

```
> summary(malform.logit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9605     0.1154 -51.637   <2e-16 ***
Alcohol       0.3166     0.1254   2.523   0.0116 *
---
    Null deviance: 6.2020  on 4  degrees of freedom
Residual deviance: 1.9487  on 3  degrees of freedom
```

$$\text{Null deviance } - \text{ Residual deviance } = 6.2020 - 1.9487 = 4.2533$$

with small p-value we reject the null.

```
> 1-pchisq(4.2533,1)
[1] 0.03917414
```

Example (continued)

```
> confint(malform.logit)
                  2.5 %     97.5 %
(Intercept) -6.19302366 -5.7396968
Alcohol      0.01868149  0.5234947
```

Note that this CI is different from the Wald CI done earlier of (0.070816, 0.562384).

Section 4

1. Framework
2. Wald
3. LRT
   - Deviance
   - Parameter Testing
4. Residuals

Pearson Residuals

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}(y_i)}}$$

- This denominator only accounts for the variability in y_i and does not include uncertainty in $\hat{\mu}_i$
- As a result e_i has a variance that is less than 1 (not standardized)

A better way to standardize them is the following

$$e_i^\star = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}(y_i)(1 - h_i)}} = \frac{e_i}{\sqrt{1 - h_i}}$$

where h_i is called the leverage and tells how much influence data point i has on the model fit.

Deviance Residuals

$$d_i = \mathrm{sgn}(y_i - \hat{\mu}_i)\sqrt{2y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + 2(n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)}$$

Again, a better way to standardize

$$d_i^\star = \frac{d_i}{\sqrt{1 - h_i}}$$

Application

R

- `residual(object,type=“pearson”)/sqrt(1-hatvalues(object))`
- `residual(object,type=“deviance”)/sqrt(1-hatvalues(object))` or simply `rstandard(object)`

Remark

- Values greater in absolute value from 2 indicate large residuals
- +ve values indicate larger than expected (from model), and -ve indicate smaller

E.g. $d_i^\star = 2.6$ indicates that the observed value is 2.6 standard deviations larger than what the model expects.

We learned

- 3 methods, but primarily use Wald and **preferred** LRT
- May not always be able to perform GoF
- Use GoF deviances to perform LRT
- Obtain and interpret residuals