

Monday, Mar 21

## Likelihood Functions

The *likelihood function* computes the *likelihood* of  $y_1, y_2, \dots, y_n$  as a function of all unknown parameters. If the distribution of the response variable is *discrete* then the likelihood is the same as the *probability* of  $y_1, y_2, \dots, y_n$ .

**Example:** Suppose  $Y_i$  has a Poisson distribution such that

$$P(Y_i = y) = \frac{\lambda_i^y e^{-\lambda_i}}{y!}$$

where

$$\lambda_i = \exp(\beta_0 + \beta_1 x_i),$$

so that the model for  $Y_i$  is a *Poisson regression model*. The likelihood function is then

$$L(\beta_0, \beta_1) = \frac{\lambda_1^{y_1} e^{-\lambda_1}}{y_1!} \times \frac{\lambda_2^{y_2} e^{-\lambda_2}}{y_2!} \times \dots \times \frac{\lambda_n^{y_n} e^{-\lambda_n}}{y_n!},$$

or

$$L(\beta_0, \beta_1) = \prod_{i=1}^n \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!},$$

where, again,

$$\lambda_i = \exp(\beta_0 + \beta_1 x_i).$$

**Example:** Suppose  $Y_i$  has a *binomial* distribution such that

$$P(Y_i = y) = \frac{m_i!}{y_i!(m_i - y_i)!} p_i^{y_i} (1 - p_i)^{m_i - y_i},$$

where

$$p_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}},$$

so that the model for  $Y_i$  is a *logistic regression model*. The likelihood function is then

$$L(\beta_0, \beta_1) = \frac{m_1!}{y_1!(m_1 - y_1)!} p_1^{y_1} (1 - p_1)^{m_1 - y_1} \times \frac{m_2!}{y_2!(m_2 - y_2)!} p_2^{y_2} (1 - p_2)^{m_2 - y_2} \times \dots \times \frac{m_n!}{y_n!(m_n - y_n)!} p_n^{y_n} (1 - p_n)^{m_n - y_n}$$

or

$$L(\beta_0, \beta_1) = \prod_{i=1}^n \frac{m_i!}{y_i!(m_i - y_i)!} p_i^{y_i} (1 - p_i)^{m_i - y_i},$$

where, again,

$$p_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}.$$

Note: If the response variable is *discrete* as it is in Poisson and logistic regression, then the likelihood function gives the *probability* of the observed responses as a function of the model parameters.

**Example:** Suppose  $Y_i$  has a normal distribution where the probability density function of  $Y_i$  is

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y - \mu_i)^2}{2\sigma^2}},$$

where

$$\mu_i = \beta_0 + \beta_1 x_i,$$

so that the model for  $Y_i$  is a normal linear regression model. The likelihood function is then

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mu_i)^2}{2\sigma^2}}.$$

Frequently we use the log of the likelihood function or the *log-likelihood function*.

**Example:** The log-likelihood for the Poisson regression model above is

$$\log L(\beta_0, \beta_1) = \sum_{i=1}^n \log \left( \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \right) = \sum_{i=1}^n y_i \log(\lambda_i) - \sum_{i=1}^n \lambda_i - \sum_{i=1}^n \log(y_i!).$$

**Example:** The log-likelihood for the normal regression model above is

$$\log L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu_i)^2.$$

## Maximum Likelihood Estimation

The maximum likelihood estimates (MLE) of the model parameters is those values of the parameters that *maximize the likelihood of the data* — i.e., the parameter values that make the values of  $y_1, y_2, \dots, y_n$  we observed the most likely values to have occurred. Finding these estimates is a *optimization* problem — i.e., find the values of the parameters that maximize the likelihood (or log-likelihood).

**Example:** In the normal linear model above, *it can be shown that* the MLEs of  $\beta_0$  and  $\beta_1$  are those values that *minimize*

$$\sum_{i=1}^n (y_i - \mu_i)^2$$

where  $\mu_i = \beta_0 + \beta_1 x_i$ . So the MLEs are also the *least squares* estimators. But the MLE of  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$

where  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ . But we typically use the unbiased estimator

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p},$$

where  $p$  is the number of  $\beta_j$  parameters ( $p = 2$  in the model above).

Except in the case of normal linear models and a few very simple nonlinear models, MLEs must be found *numerically* using *iterative* algorithms (similar to those used in nonlinear regression).

## Inference Based On Maximum Likelihood

It is usually not possible to determine an *exact* method of estimating the standard error of a MLE, computing a confidence interval based on ML, or conducting a significance test based on ML. However there exist several *asymptotic* methods that give approximate results.

1. Likelihood ratio tests and profile likelihood confidence intervals.
2. Wald tests and confidence intervals.
3. Score (Lagrange multiplier) tests (not discussed).

## Likelihood Ratio Tests

The likelihood ratio test statistic is

$$2\log(L/L_n) = 2\log(L) - 2\log(L_n) \stackrel{a}{\sim} \chi^2_{(r)},$$

where  $L$  and  $L_n$  are the likelihood functions for the model and the null model, respectively, evaluated at the MLEs, and  $r$  is the degrees of freedom.

**Example:** Consider the Poisson regression model for `ceriodaphniastrain` and the null hypothesis that there is no difference in the expected number of daphnia between the two strains when controlling for dose.

```
library(trtools) # for ceriodaphniastrain data
ceriodaphniastrain$strain <- factor(ceriodaphniastrain$strain, labels = c("a","b"))

m <- glm(count ~ strain + concentration, family = poisson, data = ceriodaphniastrain)
m.null <- glm(count ~ concentration, family = poisson, data = ceriodaphniastrain)

anova(m.null, m, test = "LRT") # has very little to do with ANOVA
```

### Analysis of Deviance Table

```
Model 1: count ~ concentration
Model 2: count ~ strain + concentration
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         68      119.0
2         67       86.4  1     32.6  1.1e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The “residual deviance” (or just “deviance”) is related to the log-likelihood of the model. It is defined as

$$D = 2\log L_s - 2\log L \geq 0,$$

where  $L$  and  $L_s$  are the likelihoods of the model in question and a “saturated” model, respectively. The saturated model is a “best possible” model where each unique combination of covariate values is a distinct level of a factor. Thus the deviance can be viewed in a way as the “lack of fit” of the model to the data. The residual deviance can sometimes be obtained using the `summary` or `deviance` functions.

```
summary(m) # shows residual deviance
```

Call:

```
glm(formula = count ~ strain + concentration, family = poisson,
    data = ceriodaphniastrain)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.680	-0.677	0.153	0.679	2.077

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	4.4546	0.0391	113.82	< 2e-16 ***
strainb	-0.2750	0.0484	-5.68	1.3e-08 ***
concentration	-1.5431	0.0466	-33.11	< 2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 1359.381 on 69 degrees of freedom
Residual deviance: 86.376 on 67 degrees of freedom
AIC: 416
```

```
Number of Fisher Scoring iterations: 4
```

```
deviance(m) # extracts only residual deviance
```

```
[1] 86.38
```

The likelihood ratio test statistic can be expressed in terms of the (residual) deviance of the model and the null model.

$$\underbrace{2(\log_s - \log L_n)}_{D_n} - \underbrace{2(\log_s - \log L)}_D = 2 \log L - 2 \log L_n \stackrel{a}{\sim} \chi^2_{(r)},$$

where  $D$  and  $D_n$  are the deviance of the model and null model, respectively. The degrees of freedom ( $r$ ) is the *difference* in the residual degrees of freedom between the two models. The residual degrees of freedom equal  $n$  minus the number of parameters for the expected response.

The following shows how the deviance relates to what is done by `anova`.

```
anova(m.null, m, test = "LRT")
```

Analysis of Deviance Table

```
Model 1: count ~ concentration
```

```
Model 2: count ~ strain + concentration
```

```
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         68      119.0
2         67       86.4  1    32.6  1.1e-08 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
deviance(m) # (residual) deviance of model
```

```
[1] 86.38
```

```
deviance(m.null) # (residual) deviance of null model
```

```
[1] 119
```

```
deviance(m.null) - deviance(m) # test statistic
```

```
[1] 32.62
```

Note: For models where  $Y_i$  is assumed to have a normal distribution, residual deviance essentially becomes the residual (error) sums of squares. Some in some sense the “analysis of deviance” can be viewed as a generalization of the “analysis of variance” in linear models.

## Profile Likelihood Confidence Intervals

Consider a significance test with hypotheses  $H_0: \theta = \theta_0$  versus  $H_a: \theta \neq \theta_0$  where  $\theta$  is some parameter of the model. A confidence interval  $(\theta_l, \theta_h)$  for  $\theta$  can be used to conduct the test.

1. If  $\theta_0$  is in  $(\theta_l, \theta_h)$ , do not reject  $H_0$ .
2. If  $\theta_0$  is not in  $(\theta_l, \theta_h)$ , reject  $H_0$

Thus a confidence interval for  $\theta$  can be *defined* as *all the values of  $\theta_0$  that are not rejected by a test of  $H_0: \theta = \theta_0$  versus  $H_a: \theta \neq \theta_0$ .*

A *profile likelihood confidence interval* is a confidence interval based on the likelihood ratio test. It is the set/range of all values of a parameter that would not be rejected by the likelihood ratio test.

**Example:** Consider again the `ceriodaphniastrain` data.

```
confint(m)
```

Waiting for profiling to be done...

```

                2.5 %   97.5 %
(Intercept)    4.377   4.5306
strainb        -0.370  -0.1803
concentration  -1.635  -1.4522
```

For `glm` objects (and some others we will talk about in the future) `confint` will compute profile likelihood confidence intervals by default. A similar approach is used by `nls`, and `glm` when using quasi-likelihood (more on that later), but is based on a  $F$  test statistic.

### Wald Tests and Confidence Intervals

Wald tests and confidence intervals are based on the fact that MLEs have approximately normal sampling distributions. The Wald *test statistic* for a parameter  $\beta_j$  is

$$z = \frac{\hat{\beta}_j - \beta_j}{\widehat{SE}(\hat{\beta}_j)} \stackrel{a}{\sim} N(0,1) \quad \text{or} \quad z^2 = \left[ \frac{\hat{\beta}_j - \beta_j}{\widehat{SE}(\hat{\beta}_j)} \right]^2 \stackrel{a}{\sim} \chi^2_{(1)},$$

where  $\beta_j$  is the value hypothesized by the null (frequently  $\beta_j = 0$ ). The Wald confidence interval for  $\beta_j$  is

$$\hat{\beta}_j \pm z \times \widehat{SE}(\hat{\beta}_j).$$

We can also apply this to a linear combinations of parameters by replacing  $\beta_j$  and  $\hat{\beta}_j$  with  $\ell$  and  $\hat{\ell}$ , respectively.

Note: In some cases a  $t$ -distribution is used as the approximate distribution of a Wald test statistic, and is also used when calibrating a confidence interval. In generalized linear models this usually happens for models in which we need to estimate the *dispersion parameter*. We *do not* need to estimate the dispersion parameter for Poisson or binomial/logistic regression.

**Example:** Consider again the `ceriodaphniastrain` data. Wald test statistics are reported by `summary`, `contrast`, `lincon`, and `glmint`, and Wald confidence intervals are reported by `contrast`, `lincon`, and `glmint`.

```
summary(m)$coefficients
```

```

      Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.455     0.03914 113.819 0.000e+00
strainb        -0.275     0.04837  -5.684 1.313e-08
concentration  -1.543     0.04660 -33.111 2.057e-240
```

```
lincon(m)
```

```

      estimate      se   lower   upper  tvalue df      pvalue
(Intercept)    4.455 0.03914  4.3779  4.5313 113.819 Inf  0.000e+00
strainb        -0.275 0.04837 -0.3698 -0.1802  -5.684 Inf  1.313e-08
concentration  -1.543 0.04660 -1.6344 -1.4517 -33.111 Inf  2.057e-240
```

```
contrast(m,
```

```

  a = list(strain = "a", concentration = c(0,1,2)),
  b = list(strain = "b", concentration = c(0,1,2)),
  cnames = c("0%", "1%", "2%"))
```

	estimate	se	lower	upper	tvalue	df	pvalue
0%	0.275	0.04837	0.1802	0.3698	5.684	Inf	1.313e-08
1%	0.275	0.04837	0.1802	0.3698	5.684	Inf	1.313e-08
2%	0.275	0.04837	0.1802	0.3698	5.684	Inf	1.313e-08

```
contrast(m, tf = exp,
  a = list(strain = "a", concentration = c(0,1,2)),
  b = list(strain = "b", concentration = c(0,1,2)),
  cnames = c("0%", "1%", "2%"))
```

	estimate	lower	upper
0%	1.316	1.197	1.447
1%	1.316	1.197	1.447
2%	1.316	1.197	1.447

## Likelihood Ratio Versus Wald

The two methods are equivalent *asymptotically* and tend to produce relatively similar results for larger  $n$ , but for smaller  $n$  they may produce somewhat different results.

1. Likelihood ratio tests/intervals are generally more accurate than Wald tests/intervals. The latter assumes that the sampling distribution of the estimator is normal, whereas likelihood ratio tests/intervals assume that some monotonic function of the estimator has a normal distribution.
2. Likelihood ratio tests/intervals are a bit more computationally expensive and difficult to program in the case of functions of model parameters (e.g., linear combinations where we use `lincon`, `contrast`, or `emmeans`).
3. Wald tests/intervals can be used in cases where a likelihood function is not specified because we do not or cannot assume a particular distribution for the response variable.

## Assumptions

For inferences based on maximum likelihood to be correct, we are assuming we have the correct likelihood function. It is useful to note three aspects of the likelihood function where we need to be correct.

1. The *distribution* of the response variable.
2. The *relationship* between the parameter(s) of the distribution and the explanatory variable(s).
3. The *independence* of the  $n$  observations.

Interestingly, for GLMs the distribution actually isn't something that we need to be correct about, provided that we are correct about the following:

1.  $E(Y_i)$  is related to the explanatory variables in the assumed way — e.g.,  $g[E(Y_i)] = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$ .
2.  $\text{Var}(Y_i)$  depends on  $E(Y_i)$  in the assumed way — i.e., through the *variance structure*  $\text{Var}(Y_i) = \phi V[E(Y_i)]$ .
3. The observations are independent.

When you specify a distribution via the `family` argument to the `glm` function, all that matters is the *variance structure* implied by that family.