STATS 330 Advanced Statistical Modeling

More Math stuff

Maximum likelihood

Suppose we have observations $y_1,\ldots,y_n$ (say responses in a regression) and the distribution of these are described by densities $f_1(y,\beta)$, …., $f_n(y, \beta)$ where β is a set of unknown parameters (like regression coefficients).

For example, in normal regression, the densities are Normal with different means depending on the covariates and the same variance.

To keep things simple we will assume that there is only one regression coefficient (to avoid partial derivatives).

The probability of observing data $y_1,\ldots,y_n$ is proportional to the product $f_1(y_1,\beta)$, …., $f_n(y_n, \beta)$. Regarded as a function of β, this is called the likelihood function. A related quantity is the log-likelihood, which is the log of the likelihood, written $l(\beta)$. We have

$$l(\beta) = \sum_{i=1}^{n} \log f_i(y_i, \beta) .$$

The maximum likelihood estimate is the value of β that maximizes the log-likelihood (which is the same value that maximizes the likelihood.) If β_0 is the true value of the parameter, one can prove that with enough data, the maximizing value $\hat{\beta}$ (which depends on the data) will be close to β_0 with high probability.

See Figure 1 overleaf.

To find the maximizing value, the usually procedure is to differentiate the log-likelihood and set the result equal to 0. This gives the *score equations*

$$\frac{dl}{d\beta} = 0 \text{ or } \sum_{i=1}^{n} \frac{d}{d\beta} \log f_i(y_i, \beta) = 0$$

The accuracy of $\hat{\beta}$ as an estimate of β_0 is controlled by the "peakedness" of the log-likelihood around its maximum, as shown on Figure 2. This is measured by the second derivative, the rate of change of the tangent to the log-likelihood around the peak. In actual fact, the standard deviation of $\hat{\beta}$ (i.e. the standard error) is approximately the reciprocal square root of the negative of the second derivative, or in symbols

$$s.e.(\hat{\beta}) = \left(-\frac{d^2 l}{d\beta^2} \right)^{-1/2}$$

the derivative being evaluated at the true value. The negative of the second derivative is called the *information* – the bigger the second derivative in magnitude, the more information. The formula above is the one used to compute the standard errors in logistic regression. By using partial derivatives, we can extend this idea to multiple parameters, rather than just one.

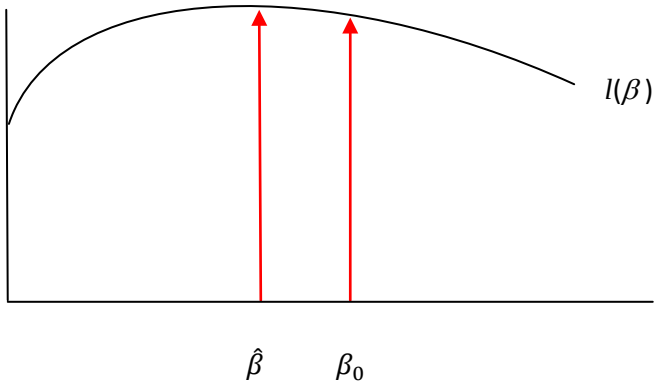

Figure 1. The log-likelihood function

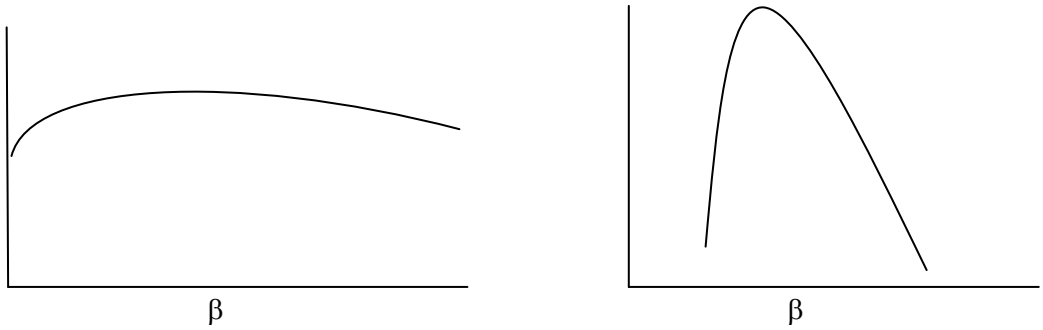

Figure 2. The log-likelihood function: left panel, log-likelihood is flat, maximum likelihood estimate is not very accurate (has large standard error). Right panel, log-likelihood is peaked, maximum likelihood estimate is more accurate (has smaller standard error.)

Relationship between parameters and log-odds in binary anova

In the plum tree example in Lecture 24, we looked at the relationship between the regression coefficients and the fitted log odds. The parameter estimates were

```
Coefficients:          Estimate Std. Error z value Pr(>|z|)
(Intercept)              0.6190     0.1353   4.574 4.78e-06 ***
lengthshort             -0.8366     0.1876  -4.460 8.19e-06 ***
timespring              -1.2381     0.1914  -6.469 9.87e-11 ***
lengthshort:timespring  -0.4527     0.3009  -1.505    0.132
```

To calculate the log-odds from these estimates, we use the formula

Log-odds = intercept + row main effect + column main effect + interaction.

Note that "logit" is just another name for the log-odds. We also note that:

1. The row effect for the baseline row is always zero by definition.
2. The column effect for the baseline column is always zero by definition.
3. The interaction for any cell in the baseline row or the baseline column is always zero by definition.

Thus, for the plum tree example, we get the following:

The log-odds for the time = "autumn", length="long" cell is 0.6190 + 0 + 0 + 0 = 0.6190
(cell is in baseline row and baseline column)

The log-odds for the time = "spring", length="long" cell is 0.6190 + (-1.2381) + 0 +0 = -0.6191
(cell is in baseline column)

The log-odds for the time = "autumn", length="short" cell is 0.6190 + 0 + (-0.8366) + 0 = -0.2176
(cell is baseline row)

The log-odds for the time = "spring", length="short" cell is 0.6190 + (-1.2381) + (-0.8366) + (-0.4527)
= -1.9084 (cell is in neither the baseline row nor the baseline column)

Note that this is a saturated model (no restrictions on the probabilities) so that the estimated probabilities are r/n and the fitted logits are $\log((r/n)/(1-r/n))$ = $\log(r/(n-r)$.

Note that we get the same logits in R by the command

```
> predict(plum.glm)
[1]  0.6190392 -0.6190392 -0.2175203 -1.9083470
```

Interpretation of odds ratios and the relationship between odds ratios and Poisson regression parameters

Suppose we have a two-dimensional contingency table classifying individuals according to two factors A and B. For example, A might be "party membership in the Soviet Union in 1959" with levels "yes", "no", and factor B might be Sex, with levels "female", "male".

The table, after classifying 1452 Soviet citizens, is

		Sex	
		Female	Male
Party member?	Yes	16	66
	No	819	551

The odds ratio is (16/819) / (66/551) = 0.163. Thus the odds of being a party member for females are only about 16% of the odds for being a party member for a male.

In general, the OR is

(odds of being at baseline (1st row) for the baseline (1st column)
divided by
(odds of being at baseline (1st row) for the non-baseline(2nd column)

The same value can be calculated in R, but we have to be careful about the baselines. The odds ratio in R will be the one above if the baselines correspond to the first row and column of the table i.e. the baseline for party member is "yes" and the baseline for sex is "female". (This is the case here.) Then we can use the following R code to get the OR and the confidence interval:

```
> count = c(16, 819, 66, 551)
> soviet.df = data.frame(count, expand.grid(party=c("Y","N"),
+  sex = c("female","male")))
>
> soviet.df
  count party    sex
1    16     Y female
2   819     N female
3    66     Y   male
4   551     N   male
```

Note that the order of the levels in expand.grid determines the baseline (the first level):

```
> levels(soviet.df$party)
[1] "Y" "N"
> levels(soviet.df$sex)
[1] "female" "male"
```

To get the log(OR) we fit the saturated model

```
> summary(glm(count~party*sex, data=soviet.df, family=poisson))
Call:
glm(formula = count ~ party * sex, family = poisson, data = soviet.df)

Deviance Residuals:
[1]  0  0  0  0

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.7726     0.2500  11.090  < 2e-16 ***
partyN           3.9355     0.2524  15.590  < 2e-16 ***
sexmale          1.4171     0.2787   5.085 3.67e-07 ***
partyN:sexmale  -1.8134     0.2841  -6.384 1.72e-10 ***
```

The log(OR) is -1.8134 corresponding to the OR of exp(-1.8134) = 0.163. The confidence interval is

```
> exp(-1.8134 +c(-1,1)*1.96*0.2841)
[1] 0.09345867 0.28463034
```

Since this doesn't contain 1 we conclude that party membership is not independent of sex (lower for females).

Relationship between conditional OR's and the regression parameters in 3-dimensional contingency tables

Let μ_{ijk} be the mean count in row i, column j and slice k of a 3-d contingency table, with factors A, B, C. These Poisson means are related to the main effects and interactions in the regression table by the equation

$$\log \mu_{ijk} = (\text{intercept}) + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \beta\gamma_{jk} + \alpha\gamma_{ik} + \alpha\beta\gamma_{ijk} \qquad (1)$$

The corresponding multinomial probabilities in the three-dimensional A,B,C table are of the form

$$\pi_{ijk} = \frac{\mu_{ijk}}{\mu_{111} + \cdots + \mu_{IJK}}$$

(The denominator is the sum of the means.) The probabilities in the conditional table of A and B given C are (see next page)

$$P(A=i, B=j \mid C=k) = \frac{P(A=i, B=j, C=k)}{P(C=k)}$$

$$= \frac{\pi_{ijk}}{\pi_{++k}} = \frac{\mu_{ijk}}{\mu_{++k}}$$

The last line follows because the sum of the means cancels out of the ratio. The conditional odds ratio is

$$\frac{\frac{\pi_{ijk}}{\pi_{i++k}} \times \frac{\pi_{11k}}{\pi_{i++k}}}{\frac{\pi_{1jk}}{\pi_{i++k}} \times \frac{\pi_{i1k}}{\pi_{i++k}}} = \frac{\pi_{ijk}\pi_{11k}}{\pi_{1jk}\pi_{i1k}} = \frac{\mu_{ijk}\mu_{11k}}{\mu_{1jk}\mu_{i1k}}$$

so the log of the conditional OR is

$$\log(\mu_{ijk}) + \log(\mu_{11k}) - \log(\mu_{i1k}) - \log(\mu_{1jk}). \qquad (2)$$

To evaluate this, we substitute the expression for log(μ_{ijk}) from the previous page. Note that, if any subscript is a 1, the corresponding main effect or interaction is zero by definition. Thus

$$\begin{aligned}
\log \mu_{11k} &= (\text{intercept}) + \gamma_k, \\
\log \mu_{1jk} &= (\text{intercept}) + \beta_j + \gamma_k + \beta\gamma_{jk}, \\
\log \mu_{i1k} &= (\text{intercept}) + \alpha_i + \gamma_k + \alpha\gamma_{ik}, \\
\log \mu_{ijk} &= (\text{intercept}) + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \beta\gamma_{jk} + \alpha\gamma_{ik} + \alpha\beta\gamma_{ijk}.
\end{aligned}$$

Substituting these into (2) gives $\log(OR) = \alpha\beta_{ij} + \alpha\beta\gamma_{ijk}$.

If the homogeneous association model holds, then the 3-factor interaction vanishes and the OR does not depend on k (i.e. is the same in each conditional table). In this case the conditional log(OR) is $\alpha\beta_{ij}$ and the OR is exp($\alpha\beta_{ij}$).