36-720: ANOVA-style Logit Models

Brian Junker

October 3, 2007

- Prospective vs Retrospective Studies
- Logit models: Logistic Regression with Discrete Covariates
- Logit vs. log-linear models
- Example: Muscle Tension

Prospective vs Retrospective Studies

Consider two studies of heart attacks

Study 1: *Take 200 people, record covariates such as age, cholesterol, blood pressure (and experimental condition, if this is an experiment) and then determine whether each has had a heart attack.*

In this study, each unique set of covariates determines a population for multinomial sampling (really, binomial in this case) experiment with the same multinomial categories: $y = 1$ or $y = 0$. This is a *prospective study.* It is the usual setting in which logistic regression is easy to apply and interpret.

Study 2: *Take 100 people who have had heart attacks, and 100 people who haven't, and record covariates for each.*

In this study, there are two populations defined by $y = 1$ or $y = 0$. Multinomial sampling in the $y = 1$ population has whatever categories are determined by the unique combinations of covariates found in that population. Multinomial sampling in the $y = 0$ population has a different set of categories. This is a *retrospective study.* Logistic regression for y can still be performed, but it is more difficult to interpret because

$$P[H.A.]/P[no\ H.A.] \neq P[H.A.]/(1 - P[H.A.]).$$

Logit models: Logistic Regression with Discrete Covariates

We will concentrate on *prospective* studies that involve a binary response variable Y and discrete covariates A, B, C, ...

We begin with some notation. By analogy with u-terms for a log-linear model, we can write logit models as follows:

$$\begin{aligned} \log(p_{Y(1)}/p_{Y(0)}) &= v + v_{A(i)} + v_{B(j)} + v_{C(k)} \\ \log(p_{Y(1)}/p_{Y(0)}) &= v + v_{A(i)} + v_{B(j)} + v_{C(k)} + v_{BC(jk)} \\ &\text{etc.} \end{aligned}$$

Also, by analogy with our notation [A][B][C], [A][BC], etc. for log-linear models we will write `{A}{B}{C}`, `{A}{BC}`, etc. for these models.

We use the new "`{`" and "`}`" notation to emphasize that *the presence or absence interactions in the logit model implies nothing about conditional independence relationships among the factors.*

Logit vs. log-linear models

If the study involving binary variable Y and discrete covariates A, B, C, is *prospective*, then *each unique combination of A, B, C, determines an independent binomial experiment with fixed sample size $n_{A,B,C}$*, the number of repetitions of that unique combination of A, B, C.

Thus the table $n_{Y,A,B,C}$ cross-classifying *all* of Y, A, B, C, has a *product multinomial* structure, where the fixed totals for the multinomials (binomials really) are the [ABC] margins.

An alternative to fitting logistic regression models, e.g. `{A}{B}{C}` would be to fit a log-linear model like [Y][ABC][possibly other stuff]

$$\log m_{YABC(yijk)} = u + u_{Y(y)} + u_{ABC(ijk)} + \textit{ (possibly other stuff)}$$

to the four-way table, where *we have included [ABC] to reflect the product multinomial structure*.

What do these log-linear models say about logistic regression (logit) models?

In what follows we will ignore most lower-order terms implied by the hierarchy principle, WLOG, to simplify notation. . .

Examples:

- [Y][ABC] is the log-linear model

$$\log m_{YABC(yijk)} = u + u_{Y(y)} + u_{ABC(ijk)}$$

The conditional odds of $Y = 1$ vs $Y = 0$ can be calculated using m's or p's, so

$$\begin{aligned} \text{logit}\, P[Y = 1|A = i, B = j, C = k] &= \log p_{YABC(1ijk)} - \log p_{YABC(0ijk)} \\ &= \log m_{YABC(1ijk)} - \log m_{YABC(0ijk)} \\ &= u_{Y(1)} - u_{Y(0)} \equiv \textit{(const.)} \end{aligned}$$

so, as expected from the log-linear form, [Y][ABC] implies that $Y \perp\!\!\!\perp (A, B, C)$, i.e. the "null" logistic regression model `{0}`:

$$\text{logit}\, P[Y = 1|A = i, B = j, C = k] = v$$

Now let's try to build some dependence between Y and A, B, C into the log-linear model, and see what happens...

- Try [YA][ABC], i.e.

$$\log m_{YABC(yijk)} = u + u_{Y(y)} + u_{A(i)} + u_{YA(yi)} + u_{ABC(ijk)}$$

so that

$$\begin{aligned} \operatorname{logit} P[Y = 1|i, j, k] &= \log m_{YABC(1ijk)} - \log m_{YABC(0ijk)} \\ &= u_{Y(1)} - u_{Y(0)} + u_{YA(1i)} - u_{YA(0i)} \\ &= v + v_{A(i)} \end{aligned}$$

- Now try [YA][YBC][ABC], i.e.

$$\log m_{YABC(yijk)} = u + u_{Y(y)} + u_{A(i)} + u_{YA(yi)} + u_{YBC(yjk)} + u_{ABC(ijk)}$$

and we get

$$\operatorname{logit} P[Y = 1|i, j, k] = v + v_{A(i)} + v_{BC(jk)}$$

(to which we should probably add lower-order terms to get the hierarchy principle back).

In general, this runs both ways:

There is an equivalence between

- *Submodels of the log-linear model [YABC$\cdots$][ABC$\cdots$] that preserve the [ABC$\cdots$] margin for product-binomial sampling; and*
- *Submodels of the logistic regression/logit model* `{ABC...}` *for prospective sampling.*

A strategy for fitting and interpreting logit models, then, is

- Do *variable selection either in logistic regression model forms, or the equivalent log-linear forms* (preserving margins for multinomial sampling)—or switch back and forth if that is interesting. . .

  Note: As we will see below, G^2's for the two different formulations of the model are the same. However, we will aso see that penalized methods like AIC and BIC may give different results because there are fewer parameters (less penalty) in the logit formulation than in the corresponding log-linear formulation.

- *By definition* logistic regression focuses on the conditional distribution $p(Y|A, B, C, \ldots)$ and thus ignores dependence between $A, B, C, \ldots$, while log-linear models focus on the joint distribution $p(Y, A, B, C, \ldots)$ and thus incorporates dependence between $A, B, C, \ldots$.

  Therefore, instead of interpreting the model in terms of conditional independences among A, B, C, etc., (none can be inferred from the logistic regression model), we interpret the model in terms of *changes in the log-odds of* $Y = 1$ *as we move from one level of* A, B, C, *etc. to another*, or equivalently (log-)odds-ratios, for $y = 1$, etc.

Example: Muscle Tension

In the data set

[T]ension (ℓ)	[W]eight (i)	[M]uscle (j)	Drug (k) [D]rug 1	Drug 2
	High	Type 1	3	21
		Type 2	23	11
High				
	Low	Type 1	22	32
		Type 2	4	12
	High	Type 1	3	10
		Type 2	41	21
Low				
	Low	Type 1	45	23
		Type 2	6	22

muscle tension [T] can be thought of as the response variable.

The data set (see examples) is not in the form for logistic regression

```
T  W  M  D
1  1  1  1
1  1  1  0
```

but instead is in the form for log-linear modeling

```
n   T  W  M  D
22  1  1  1  1
 6  1  1  1  0
```

So for every logit submodel of `{WMD}` we want to fit, we will instead fit log-linear submodels of [TWMD][WMD].

See also discussion of this data set on pp. 146–149 in Christensen's text.

In what follows we will simply look at some results of automatic stepwise procedures (see also R example file for this lecture).

First we consider log-linear variable selection:

```
> musc <- data.frame(n,Wt,Mt,Dt,Tn)
> library(MASS)
> basemodel <- glm(n ~ Tn*Wt*Mt*Dt,data=musc,family=poisson)
> upper <- n ~ Tn*Wt*Mt*Dt     # examine only models nested between
> lower <- n ~ Tn + Wt*Mt*Dt   # [T][WMD] and [TWMD][WMD] = [TWMD]
> stepAIC(basemodel,list(lower=lower,upper=upper),k=2)
[...]
 Call:  glm(formula = n ~ Tn + Wt + Mt + Dt + Tn:Mt + Wt:Mt + Tn:Dt +
        Wt:Dt + Mt:Dt + Tn:Mt:Dt + Wt:Mt:Dt, family = poisson, data = musc)
```

This model corresponds to [TMD][WMD] as a log-linear model, or `{MD}` as a logit model:

$$\text{logit}\, P[T = High \mid ijk] = \nu + \nu_{M(j)} + \nu_{D(k)} + \nu_{MD(jk)}$$

In this model it appears that $T \perp\!\!\!\perp W$, but $T \not\perp\!\!\!\perp (M, D)$.

If we do BIC model selection (`k=log(sum(n))`) instead of AIC (`k=2`), we get the same final model, in this case.

Now let's consider logistic regression variable selection:

```
> musc.logist <- data.frame(H=musc[1:8,1],L=musc[9:16,1],musc[1:8,2:4])
> resp <- as.matrix(musc.logist[,1:2])
>   # note: for binomial (and not bernoulli) data
>   # we have to specify both successes and failures!
> lmod0 <- glm(resp ~ Wt+Mt+Dt,data=musc.logist,family=binomial)
> stepAIC(lmod0,list(lower=~.,upper=~.^3),k=2)
[...]
 Call:  glm(formula = resp ~ Wt + Mt + Dt + Mt:Dt, family = binomial,
            data = musc.logist)
```

This is the model {W}{MD} which has W as a main effect:

$$\text{logit}\, P[T = High \mid ijk] = \nu + \nu_{W(i)} + \nu_{M(j)} + \nu_{D(k)} + \nu_{MD(jk)}$$

Using BIC instead we get the impler model {W}{M}{D}. Finally, Christensen, pp. 146–149, also provides evidence that {WM}{MD} is a plausible model. Altogether we have four potential models to compare, in two different nestings (because there is no nesting between {MD} and {W}{M}{D}):

{MD}	{W}{M}{D}		[TMD][WMD]	[TW][TM][TD][WMD]
{W}{MD}	{W}{MD}	or	[TW][TMD][WMD]	[TW][TMD][WMD]
{WM}{MD}	{WM}{MD}		[TWM][TMD][WMD]	[TWM][TMD][WMD]

Comparing first the log-linear versions of the models we get

```
> mod1a <- glm(n ~ Tn*Mt*Dt              + Wt*Mt*Dt,data=musc,family=poisson)
> mod1b <- glm(n ~ Tn*Wt + Tn*Mt + Tn*Dt + Wt*Mt*Dt,data=musc,family=poisson)
> mod2 <- glm(n ~ Tn*Mt*Dt + Tn*Wt       + Wt*Mt*Dt,data=musc,family=poisson)
> mod3 <- glm(n ~ Tn*Mt*Dt + Tn*Wt*Mt    + Wt*Mt*Dt,data=musc,family=poisson)
> anova(mod1a,mod2,mod3,test="Chisq")
[...]
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
  1         4    1.52890                          # [TMD][WMD]
  2         3    1.05961  1  0.46928   0.49332  # [TW][TMD][WMD]
  3         2    0.11952  1  0.94009   0.33225  # [TWM][TMD][WMD]
> anova(mod1b,mod2,mod3,test="Chisq")
[...]
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         4     5.3106                          # [TW][TM][TD][WMD]
2         3     1.0596  1   4.2510    0.0392    # [TW][TMD][WMD]
3         2     0.1195  1   0.9401    0.3323    # [TWM][TMD][WMD]
```

Note that the deviance and df calculations are exactly the same as for the logistic regression (next slide).

Now let's compare as logit / logistic regression models:

```
> mod1a <- glm(resp ~ Mt*Dt       ,data=musc.logist,family=binomial)
> mod1b <- glm(resp ~ Wt + Mt + Dt,data=musc.logist,family=binomial)
> mod2 <- glm(resp ~ Mt*Dt + Wt   ,data=musc.logist,family=binomial)
> mod3 <- glm(resp ~ Mt*Dt + Wt*Mt,data=musc.logist,family=binomial)
> anova(mod1a,mod2,mod3,test="Chisq")
[...]
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         4    1.52890                          # {MD}
2         3    1.05961  1  0.46928   0.49332    # {W}{MD}
3         2    0.11952  1  0.94009   0.33225    # {WM}{MD}
> anova(mod1b,mod2,mod3,test="Chisq")
[...]
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         4     5.3106                          # {W}{M}{D}
2         3     1.0596  1   4.2510    0.0392    # {W}{MD}
3         2     0.1195  1   0.9401    0.3323    # {WM}{MD}
```

We see from the first comparison (of either logit or log-linear models) that `{MD}`, `{W}{MD}`, and `{WM}{MD}` are all about equally good fits. From the second model comparison it appears that `{W}{M}{D}` is a substantially less good fit.

So we should choose among the first three models if we need a final model: `{MD}`, `{W}{MD}`, or `{WM}{MD}`. If our only criterion is parsimony we would take `{MD}`; otherwise perhaps one of the two other models better reflects some prior knowledge we may have about the (experimental) process generating the data.