Building Logistic Regression Models

Fit and Sparse Data

Demetris Athienitis

UNIVERSITY of FLORIDA

Section 1

Model Fit

1. Goodness of fit test (Chapter 3). Using G^2 and X^2 generally limited to "non-sparse" contingency tables.
    - A *goodness of fit* can be used only in the number of predictor levels is fixed and relatively small to the overall sample size
    - Only appropriate for grouped binary data with most ($\geq 80\%$) of fitted cell counts being "large" (e.g., $\hat{\mu}_i > 5$)
    - For continuous predictors or many predictors with small fitted values, distributions of X^2 and G^2 are not well approximated by χ^2. For better approximations, try grouping data before applying X^2, G^2
        - Hosmer-Lemeshow test forms groups using ranges of $\hat{\pi}$ values
        - Or can try to group predictor values (if only 1 or 2 predictors)

Model Fit

2. Check whether fit improves by adding other predictors or interactions between predictors
3. Residuals (Chapter 3)
   - Standardized Pearson residuals, `rstandard(model,type=“pearson”)`
   - Standardized Deviance residuals, `rstandard(model)`

Example (Berkeley Graduate Admissions)

Admissions data for 6 departments at UC Berkeley by gender

```
> ftable(UCBAdmissions,row.vars="Dept",
+  col.vars=c("Gender","Admit"))
     Gender      Male              Female
     Admit   Admitted Rejected Admitted Rejected
Dept
A                 512      313       89       19
B                 353      207       17        8
C                 120      205      202      391
D                 138      279      131      244
E                  53      138       94      299
F                  22      351       24      317
```

Example (continued)

- Admissions rates are higher for departments A and B but lower for C through D
- Odds ratio (of acceptance to rejection) for males vs females does not seem to be very different from 1 (except A)

```
> round(apply(UCBAdmissions,3,odds.ratio),2)
   A    B    C    D    E    F
0.35 0.80 1.13 0.92 1.22 0.83
```

- Ignoring department, i.e. merging them, the odds ratio seems to favor male. That is because more males where applying to departments with giher acceptance rates and vice versa for females

```
> odds.ratio(UCBGbyA) # Marginal odds ratio
[1] 1.84108
```

Example (continued)

Fitting the model (to the correct format data), conditional odds ratio of acceptance with gender conditional on dept is $\exp(-0.999) = 0.9$ which is not significantly different from 1.

```
> UCB.logit=glm(cbind(Admit,Reject)~Gender+Dept,family=binomial)
> summary(UCB.logit)
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.62456    0.15773 -16.640   <2e-16 ***
GenderMale  -0.09987    0.08085  -1.235    0.217
DeptA        3.30648    0.16998  19.452   <2e-16 ***
DeptB        3.26308    0.17878  18.252   <2e-16 ***
DeptC        2.04388    0.16787  12.176   <2e-16 ***
DeptD        2.01187    0.16992  11.840   <2e-16 ***
DeptE        1.56717    0.18044   8.685   <2e-16 ***
---
    Null deviance: 877.056  on 11  degrees of freedom
Residual deviance:  20.204  on  5  degrees of freedom
AIC: 103.14
```

Example (continued)

- Data is grouped so perform goodness of fit test (using G^2), which indicates a lack of fit

```
> 1-pchisq(UCB.logit$deviance,UCB.logit$df.residual)
[1] 0.001144078
```

- Standardized Pearson residuals, the first two observations corresponding to Dept A, don't seem to fit well

```
> round(rstandard(UCB.logit,type="pearson"),2)
    1     2     3     4     5     6     7     8     9
-4.03  4.03 -0.28  0.28  1.88 -1.88  0.14 -0.14  1.63
   10    11    12
-1.63 -0.30  0.30
```

Example (continued)

So we fit a model excluding Dept A and remove gender.

```
> UCBnoGA.logit=glm(cbind(Admit,Reject)~Dept,family=binomial,
+  data=berk,subset=(Dept!="A"))
> summary(UCBnoGA.logit)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.6756     0.1524 -17.553   <2e-16 ***
DeptB         3.2185     0.1749  18.402   <2e-16 ***
DeptC         2.0600     0.1674  12.306   <2e-16 ***
DeptD         2.0108     0.1699  11.835   <2e-16 ***
DeptE         1.5861     0.1798   8.822   <2e-16 ***
---
    Null deviance: 539.4581  on 9  degrees of freedom
Residual deviance:   2.6815  on 5  degrees of freedom
AIC: 69.916
```

Residuals are better, GoF has high p-value, and AIC much smaller. Dept A has its “own” seperate model/interpretation.

Linearity of predictors

With (quantitative) predictors we need to check if an additive linear model is adequate of whether higher order polynomial terms and interactions are necessary.

Example

A nice example with two predictors where a quadratic from of the first predictor is (somewhat) useful, but no interaction, can be found at
`https://freakonometrics.hypotheses.org/8210`
and script at freakonometrics.R

Section 3

Effects of Sparse Data

Sparse data are when certain combinations of variables have no actual data or "limited" information. This can lead to parameter estimates being infinite (in value), but most often in software you may see extremely large standard errors.

Example

Consider,

$$
\begin{array}{cc|cc}
 & & S & F \\
\hline
\multirow{2}{*}{X} & 1 & 8 & 2 \\
 & 0 & 10 & 0
\end{array}
$$

Fitting a simple logistic regression will yield the estimates odds ratio

$$e^{\hat{\beta}} = \frac{8 \times 0}{2 \times 10} = 0 \quad \Rightarrow \quad \hat{\beta} = \log(0) = -\infty$$

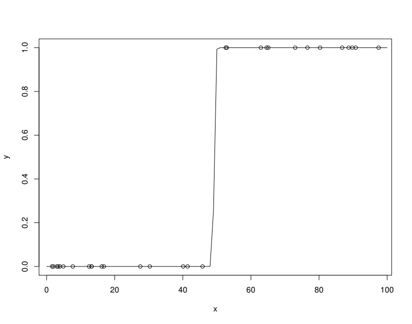
Effects of Sparse Data

Infinite estimates exist when predictor values (x values) where $y = 1$ can be *separated* from predictor values where $y = 0$. This extends to multidimensional predictor space.

Example

Simulate/generate data (with no values at $x = 50$) such that

$$y = \begin{cases} 0 & x < 50 \\ 1 & x > 50 \end{cases}$$



Example (continued)

```
> fit=glm(y~x,family=binomial)
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(fit)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -297.566 174094.706  -0.002    0.999
x               6.051   3542.717   0.002    0.999
---
    Null deviance: 4.1054e+01  on 29  degrees of freedom
Residual deviance: 5.0225e-09  on 28  degrees of freedom
AIC: 4
```

Although $\hat{\beta} = 6.051$ the standard error is 3542.717.

We learned

- Model fit via GoF and Residuals
- Linearity of predictors
- Effects of sparse data on model fit