STA303H1S - Winter 2014: Data Analysis II

LECTURE 9: Logistic Regression Model (continued)

Ramya Thinniyam

February 6th, 2014

Likelihood Ratio Tests (LRT)

Test if subset of the coefficients are 0 (compare full and reduced models).

<u>Idea</u>: compare likelihood of data assuming full model is true (L_f) to likelihood assuming reduced model (L_r).

Likelihood Ratio: $\frac{L_r}{L_f}$; where

$L_r = L(\hat{\underset{\sim}{\beta}}_r)$ is the maximized likelihood under reduced model
$L_f = L(\hat{\underset{\sim}{\beta}}_f)$ is maximized likelihood under full model.

($L_r \leq L_f$ since constrained maximum would be less than or equal to the unconstrained maximum.)

- Similar to a partial F-test. We don't have normality here so we use likelihood ratio.

LRT / Goodness of Fit Tests

Hypotheses: H_0 : reduced model is appropriate vs.
H_a: full model fits the data better

Test statistic: $G^2 = -2\log(\frac{L_r}{L_f}) \sim \chi^2_\nu$ under H_0 where
$\nu =$ difference in number of parameters between full and reduced models

p-value: $p = P(\chi^2_\nu > G^2)$

Notes:

- In the context of goodness of fit, the test statistic is referred to as deviance.
- R does a global LRT: compares fitted model to null (only intercept) model.
- For testing only one parameter, use Wald test or LRT : they are not equivalent, if they do not agree use LRT. LRT is more reliable.

Example: Donner Party LRT

Conduct the global LRT for the Donner Party Example. What are the hypotheses and what is the conclusion?

Model Assumptions for Logistic Regression

1. Independent Observations
2. Correct form of the model:
   - linearity between logits and predictor variables
   - all relevant variables are included
   - all irrelevant variables are excluded
3. Large sample sizes (need large sample properties of MLEs for tests and CIs to be valid)

(Less assumptions required here than for usual linear regression model - don't need normality, Gauss-Markov conditions, etc.)

Checking Model Assumptions for Donner Party

Q: Do we need to check diagnostic/residual plots? Explain.

A:

Q: Check the validity of the model assumptions.

A:

Checking Higher Order and Polynomial Terms to Improve the Model

In order to improve the model, try adding:

- age$*$gender interaction
- age^2 quadratic term
- age^{2*}gender interaction

Write the new model. Is this model better than the model with just the main effects of gender and age?

Interaction between Age and Gender

It seems reasonable that the effect of age on the odds of survival would differ by gender.

Model: $logit(\pi) = \beta_0 + \beta_1 age + \beta_2 I_M + \beta_3(age * I_M)$

Check if the model with interaction is better than the additive model.

R Output for all Logistic Models fitted

<u>Model 1</u>:

```
> glm.model1=glm(status ~ age + gender, family=binomial)
> summary(glm.model1)

Call:
glm(formula = status ~ age + gender, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7445  -1.0441  -0.3029   0.8877   2.0472

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.23041    1.38686   2.329   0.0198 *
age         -0.07820    0.03728  -2.097   0.0359 *
genderMALE  -1.59729    0.75547  -2.114   0.0345 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 51.256  on 42  degrees of freedom
AIC: 57.256

Number of Fisher Scoring iterations: 4
```

Model 2:

```
> agesq = age^2
> glm.model2=glm(status ~ age*gender + agesq + agesq:gender, family=binomial)

> summary(glm.model2)

Call:
glm(formula = status ~ age * gender + agesq + agesq:gender, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3396  -0.9757  -0.3438   0.5269   1.5901

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.053198   9.684350  -0.315    0.753
age               0.482908   0.658121   0.734    0.463
genderMALE       -0.265286  10.455222  -0.025    0.980
agesq            -0.010160   0.010263  -0.990    0.322
age:genderMALE   -0.299877   0.696050  -0.431    0.667
genderMALE:agesq  0.007356   0.010689   0.688    0.491

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 45.361  on 39  degrees of freedom
AIC: 57.361

Number of Fisher Scoring iterations: 5
```

Model 3:

```
> glm.model3=glm(status ~ age*gender, family=binomial)

> summary(glm.model3)

Call:
glm(formula = status ~ age * gender, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2279  -0.9388  -0.5550   0.7794   1.6998

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     7.24638    3.20517   2.261   0.0238 *
age            -0.19407    0.08742  -2.220   0.0264 *
genderMALE     -6.92805    3.39887  -2.038   0.0415 *
age:genderMALE  0.16160    0.09426   1.714   0.0865 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 47.346  on 41  degrees of freedom
AIC: 55.346

Number of Fisher Scoring iterations: 5
```

Conclusions about Donner Party

Answer the questions of interest and make final conclusions regarding this case study.

Include:

- Which explanatory variables are significant predictors of odds of survival? (i.e. Which is the best model?)
- Answer the questions of interest.
- For the predictors that are significant, specifically explain what the differences are and quantify their effect using practical terms.
- Comment on validity of the model and any concerns that you may have.

What if the default variables are changed in the model?

R chooses female and died to be defaults and gives the following fitted model:

Model 1: $\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = 3.23 - 0.078 \text{ age} - 1.60 I_{\text{Male}}$

where $\pi = P(\text{Survived})$.

1. Suppose that you choose Male and Died to be the defaults. Write out your model in terms of the βs. Which parameter estimates would change between Model 1 and your model? Using only the above estimated coefficients from Model 1, find the parameter estimates for your model.

2. Suppose that you choose Female and Survived to be the defaults. Write out your model in terms of the βs. Which parameter estimates would change between Model 1 and your model? Using only the above estimated coefficients from Model 1, find the parameter estimates for your model.