# STA303H5S - Winter 2014: Data Analysis II

## LECTURE 8: Logistic Regression Model

Ramya Thinniyam

February 4th, 2014

# Logistic Regression

Suppose response is a success or a failure.
$Y|X \sim$ Bernoulli($\pi$), where $\pi = P(\text{success})$.
Then, $E(Y|X) = \pi$.

Logit Link: $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \rightarrow$ log odds in favour of a success

Model: $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$

called "Logistic Regression" Model

Invert: $\pi = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}} = \frac{e^{\eta}}{1+e^{\eta}}$

called the "logistic function"

# Logistic Regression Model

- $E(Y_i|X_{i1},\ldots,X_{ip}) = \pi_i$
- $Var(Y_i|X_{i1},\ldots,X_{ip}) = \pi_i(1-\pi_i)$
- Model: $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$
- This model does not predict if the response is 0 or 1. It does predict the log odds of the response being 1 (i.e. log odds of a success)
- log odds $\in (-\infty, \infty)$
- As $\pi$ increases, odds of success and log odds of success increase

# Maximum Likelihood Estimation of regression parameters

Likelihood Function: $L(\beta_0, \ldots, \beta_p) = \prod_{i=1}^{n} \pi_i(\underset{\sim}{\beta})^{y_i}[1 - \pi_i(\underset{\sim}{\beta})]^{1-y_i}$

- Maximum Likelihood Estimates (MLEs) are the values of $\beta$s that maximize the likelihood function.
- Need iterative numerical solution: Iteratively Reweighted Least Squares (IWLS) using either Newton-Raphson or Fisher Scoring optimization
- Newton-Raphson and Fisher Scoring are the same when using canonical link
- R uses Fisher Scoring for IWLS in 'glm' function

# Large-Sample Properties of MLEs

If model is correct and there is a large enough sample size ($n \to \infty$):

1. MLEs are unbiased
2. MLEs have minimum variance
3. MLEs are normally distributed
4. Estimates for the standard errors of MLEs are known (available as a by-product of numerical approximation)

# Example: Donner Party Case Study

- In April 1846, a group of 87 pioneers set out for California by wagon train
- Some pioneers got stuck in Sierra Nevada mountains in November due to difficult conditions (harsh weather, unsuitable travel equipment, splits within the group, etc).
- Only some survived
- They were rescued in April 1847

Data: For $n = 45$ Adults (15 years of age or older):

Age
Gender
Status - Died or Survived

# Questions of Interest:

1. Were women or men more likely to survive?
2. Were younger pioneers more likely to survive than older ones?

Interested in modelling the odds of survival based on gender and age.

# Fitting Logistic Regression Models in R

- R uses canonical links as default
- R specifies the defaults categories for factors alphabetically (default = 1st category alphabetically)
- For the response, the first factor level is considered the failure

```
> donner = read.csv("donnerpartydata.csv")
> donner
   AGE    SEX   STATUS
1   23   MALE     DIED
2   40 FEMALE SURVIVED
3   40   MALE SURVIVED
. .... . . ..........
. .... . . ..........
. .... . . ..........
43  23   MALE SURVIVED
44  24   MALE     DIED
45  25 FEMALE SURVIVED
```

```
> glm.model=glm(status ~ age + gender, family=binomial(link="logit"))

> summary(glm.model)

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

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
```

# Fitted Model and Estimated odds of Survival

1. Write out the fitted model as fitted by R.
2. Estimate the odds of survival for a 25-year old female.
3. Estimate the probability of survival for a 25-year old female.
4. For a 25-year old male, estimate the odds of survival.
5. For a 50-year old female, estimate the odds of death.

# Interpreting Coefficients of a Logistic Regression

## (Assuming the coefficients are statistically significant)

- Odds of a success for A: $odds_A = \frac{\pi_A}{1-\pi_A}$
- Odds ratio of success of A to B: $\frac{odds_A}{odds_B} = \frac{\frac{\pi_A}{1-\pi_A}}{\frac{\pi_B}{1-\pi_B}}$

In logistic model, $odds|X = \exp\{\beta_0 + \beta_1 X_1 + \ldots, \beta_p X_p\}$

Interpretation of $\beta_k$:

Let $w$ be the odds that $Y = 1$. Then, holding all other $X_j$ fixed ($forj \neq k$), the odds ratio that $Y = 1$ at $X_k = a$ to $X_k = b$ is:

$$\frac{w_a}{w_b} = e^{\beta_k(a-b)}$$

- If $X_k$ increases by 1 unit and when all other variables are held constant, the odds of a success change by a multiplicative factor of $e^{\beta_k}$. ($\hat{\beta}_k$ estimates the log odds of success)
- If $X_k$ is an indicator variable, then $e^{\beta_k}$ is the odds ratio of success for the category that $X_k$ indicates to the default category.

# Example: Donner Party

1. Compare the odds of survival for a 50-year old women to a 20-year old women
2. Compare the odds of survival for a woman and man of the same age
3. Interpret the coefficient estimates

# Testing if Coefficients are Significant

If $\beta_k = 0$, then $X_k$ has no effect on the log odds.

Based on large-sample properties of MLEs (MLE is normally distributed).

Wald's Test:

Hypotheses: $H_0 : \beta_k = 0$ vs. $H_a : \beta_k \neq 0$

Test statistic: $z = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)} \sim N(0,1)$ under $H_0$

100$(1-\alpha)$% CI for $\beta_k$: $\hat{\beta}_k \pm z_{\alpha/2}\, se(\hat{\beta}_k)$

It is not appropriate to calculate a CI for $\pi$ because $0 \leq \pi \leq 1$ so it's not normally distributed.

# Example - Donner Party: Testing Significance of Coefficients

1. Test for significance of the regression coefficients. What conclusions can be made about the effect of age and gender on the odds of survival?
2. Find a 95% CI for the coefficient of age.
3. Find a 95% CI for the odds ratio of survival for a 1 year increase in age, but the same gender.
4. Odds of survival for a 50-year old is _____ times the odds of survival for a 25-year old of the same gender. Find a 95% CI for the odds ratio of a 50-year old to a 25-year old.

# Likelihood Ratio Tests (LRT)

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

Idea: 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 $L_r$ is the maximized likelihood under the reduced model and $L_f$ is maximized likelihood under full model.

($L_r \leq L_f$ since obtain a larger maximum with bigger model)

# 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 model

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 of LRT : they are not equivalent, if they do not agree use LRT. LRT is more reliable.

# Example: Donner Party LRT/Goodness of Fit

Conduct a test for the goodness of fit for the model that was fitted by R. What are the hypotheses and what is the conclusion?

# Conclusions about Donner Party

Answer the questions of interest and make final conclusions.