GLM/Categorical Data Analysis_wk1

Yu-Ling Tseng

Depart. of Applied Math, NDHU

Course: Generalized Linear Models/Categorical Data Analysis – Class Notes
Instructor: Yu-Ling Tseng (Yes, Heyou)
Office: A409 Science Building
Office Hours: TBA
Phone: 8633518
Email: yltseng@mail.ndhu.edu.tw
http://faculty.ndhu.edu.tw/ yltseng/edu/glm.html

Textbook

An Introduction to Categorical Data Analysis, A. Agresti (1996), Wiley and Sons. (Hwa-Tai) 02-23773877.

Reference

√ Agresti, A. (1990). Categorical Data Analysis. Wiley.

√ Agresti, A. (2002). Categorical Data Analysis, 2nd Edtion. Wiley.

√ McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models. 2nd Edition. Chapman and Hall, London.

√ Dobson, A. (1990). An Introduction to generalized linear models. Chapman and Hall, London.

√ Neter, Kutner, Nachtsheim and Wasserman (1996). Applied Linear Statistical Models. 4th Edtion. McGRAW-HILL International.

√ D. W. Hosmer and S. Lemeshow (1989). Applied Logistic Regression. Wiley.

Course Grade
In-class Exam (50 %), Presentation (50 %)

$\sim$ *Come to my office if you have any question!* $\sim$ $\ddot\smile$

Introduction_1: Statistical problems

$\surd$ Response vs Explanatory variables

y_i, Y_i: survival of patients, scores political philosophy, incomes

$x_i = (x_{i1}, \ldots, x_{ik})$: (medical treatment, age, gender), (training course, major, gender) (income, attained education, religious affiliation, age, gender, race), (attained education, age, gender, years at work)

$\longrightarrow$ Continuous or Discrete (Categorical) Variables

Intro_1: Scale of measurement

- Continuous data
  - interval: arbitrary origin (^oC)
  - ratio: absolute origin (height)
- Categorical data
  - nominal: no order involved (religious affiliation, mode of transportation to work)
  - ordinal: order but not necessary can assign distance (poor, fair, good, excellent; low, high, too high; certain, probable, unlikely, definitely not)
  - counts: 0, 1, 2, 3, . . . (ordinal, too)

Classification of Stat. methods

		Response	
		Continuous	Discrete
Explanatory	Continuous	Regression	Logistic
	Discrete	ANOVA	Loglinear model
	Mixed	ANCOVA	Logistic

$\longrightarrow$ Focus of this course.

Background: stat. estimation, testing and exposure to regression modeling and the analysis of variance.

Introduction_2: Statistical Modeling

$**$ A model is a simple summary (smoothed version) of the data.

$\longrightarrow$ Model : data
$=$ systematic pattern $+$ random component (noise),

w/ both parts involving unknown parameters.

E.G. $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$\longleftrightarrow$ matrix form: $Y = X\beta + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 \cdot I)$

$\longrightarrow$ est. $\hat{\beta}$, $\hat{\sigma}^2$, predictors: $\hat{Y} = X\hat{\beta}$

Statistical Modeling (conti)

$\star$ All models are wrong, but some are useful . . .

• Models are only device for data analysis.

$*$ Simple models provide clear thinking, better prediction, easier interpretation.

$\longrightarrow \star\star$ **Parsimony principle** $\star\star$

Should add systematic effects to a model **only** if substantial evidence for the effect exists

$**$ substantial evidence: sig. F test, small P value, big(add)/small(drop) change in **deviance** . . . $**$

Introduction_3: Overview

Topics covered in this course:

- ✓ Discrete distributions
  (Negative)Binomial, Poisson, Multinomial, . . .
  Exponential family
- ✓ Contingency tables
  Study Designs/Sampling Scheme
- ✓ Generalized linear models (GLM)
  - ⋆ Logistic models
  - ⋆ Loglinear models
  - ⋆ Selected topics for presentation

Stat. Package: SAS/R

Assumption taken: You can learn SAS/R, basicaly, by yourselves.

Assistance provided: Parts of SAS/R programs for certain GLM analyses will be illustrated in class, once in a while.

Recall: 2×2 Contingency Table

A real data set, say :

		outcome	
		f	u
Treatment	placebo	16	48
	test	40	20

Of interest: Is "test" sig. better than placebo?

$\Leftrightarrow$ Hypothesis testing for the indep. between outcome and treatment.

$\longrightarrow$ χ^2 test $\longrightarrow$ need expected counts $\longrightarrow$ in SAS/R? Easy!

A tiny taste on R

```
> x=c(16,40);  # placebo, test
> n=c(64,60);
> prop.test(x,n);

2-sample test for equality of proportions
               with continuity correction
data:  x out of n
X-squared = 20.0589, df = 1, p-value = 7.51e-06
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.5924430 -0.2408903
sample estimates:
   prop 1    prop 2
0.2500000 0.6666667
```

SAS code

```
data respire;
   input treat $ outcome $ count ;
   cards;
   placebo f 16
   placebo u 48
   test    f 40
   test    u 20
   ;
proc freq;
weight count;
tables treat*outcome/chisq expected fisher
run;
```

SAS outut in text form:

```
The FREQ Procedure

Table of treat by outcome

treat     outcome
Frequency|
Expected |
Percent  |
Row Pct  |
Col Pct  |f       |u       |  Total
---------+--------+--------+
placebo  |     16 |     48 |     64
         | 28.903 | 35.097 |
         |  12.90 |  38.71 |  51.61
         |  25.00 |  75.00 |
         |  28.57 |  70.59 |
---------+--------+--------+
test     |     40 |     20 |     60
         | 27.097 | 32.903 |
         |  32.26 |  16.13 |  48.39
         |  66.67 |  33.33 |
         |  71.43 |  29.41 |
---------+--------+--------+
Total          56       68      124
            45.16    54.84   100.00
```

```
Statistics for Table of treat by outcome
Statistic                       DF  Value Prob
-----------------------------------------
Chi-Square                       1 21.7   <.0001
Likelihood Ratio Chi-Square 1 22.3   <.0001
Continuity Adj. Chi-Square  1 20.0   <.0001
Mantel-Haenszel Chi-Square  1 21.5   <.0001
Phi Coefficient                   -0.4184
Contingency Coefficient            0.3860
Cramer's V                         0.4184


Fisher's Exact Test
----------------------------------
Cell (1,1) Frequency (F)         16
Left-sided Pr <= F        2.838E-06
Right-sided Pr >= F          1.0000

Table Probability (P)     2.397E-06
Two-sided Pr <= P         4.754E-06

 Sample Size = 124
```

Is test "sig. better" than placebo? With R

```
> prop.test(x,n,alternative=c("less"));

2-sample test for equality of proportions
              with continuity correction


data:  x out of n
X-squared = 20.0589, df = 1, p-value = 3.755e-06
alternative hypothesis: less
95 percent confidence interval:
 -1.0000000 -0.2665547
sample estimates:
   prop 1    prop 2
0.2500000 0.6666667
```