Discrete Response Model Lecture 4

Subtitle

datascience@berkeley

An Example

Example

The purpose of this data is to determine if the fiber source has an effect on the bloating severity.

- Notice the columns have ordinal levels. (We will take this into account later in this course.) It is instructive in a class setting to analyze the data first without taking the order into account, so that we can see the benefits of taking into account the order later.
- I would expect that each person fits in one and only one cell of the table. **Why would this be important to know?**
- Given the layout of the data, it is likely that the sample size for each row was fixed. Thus, this would correspond to the **multinomial distribution setting**.
- Fiber source could actually be analyzed as two separate explanatory variables: bran ("yes" or "no") and gum ("yes" or "no").
- We will analyze this data in a **4×4 contingency table**. *(Refer to the text for how this data can be analyzed as two separate explanatory variables through using regression models.)*

Example

```
diet <- read.csv(file = "C:\\data\\Fiber.csv")
head(diet)
```

```
  fiber  bloat count
1  bran   high     0
2   gum   high     5
3  both   high     2
4  none   high     0
5  bran medium     1
6   gum medium     3
```

```
# Match order given at DASL
diet$fiber<-factor(x = diet$fiber, levels = c("none", "bran", "gum", "both"))
diet$bloat<-factor(x = diet$bloat, levels = c("none", "low", "medium", "high"))
diet.table<-xtabs(formula = count ~ fiber + bloat, data = diet)
diet.table
```

```
fiber  none low medium high
  none    6   4      2    0
  bran    7   4      1    0
  gum     2   2      3    5
  both    2   5      3    2
```

Example

```
ind.test<-chisq.test(x = diet.table, correct = FALSE)
ind.test
```

```
        Pearson's Chi-squared test

data:  diet.table
X-squared = 16.9427, df = 9, p-value = 0.04962

Warning message:
In chisq.test(diet.table, correct = FALSE) :
  Chi-squared approximation may be incorrect
```

```
library(package = vcd)
assocstats(x = diet.table)
```

```
                    X^2 df P(> X^2)
Likelihood Ratio 18.880  9 0.026230
Pearson          16.943  9 0.049621

Phi-Coefficient   : 0.594
Contingency Coeff.: 0.511
Cramer's V        : 0.343
```

Example

```
> class(diet.table)
[1] "xtabs" "table"

> summary(diet.table)
Call: xtabs(formula = count ~ fiber + bloat, data = diet2)
Number of cases in table: 48
Number of factors: 2
Test for independence of all factors:
        Chisq = 16.943, df = 9, p-value = 0.04962
        Chi-squared approximation may be incorrect

> qchisq(p = 0.95, df = 9)
[1] 16.91898
```

Summary

In summary,

- $X^2 = 16.94$
- $-2\log(\Lambda) = 18.88$
- $\chi^2_{0.95,9} = 16.92$
- P-value using X^2 is $P(A > 16.94) = 0.0496$ where $A \sim \chi^2_9$
- P-value using $-2\log(\Lambda)$ is $P(A > 18.88) = 0.0262$ where $A \sim \chi^2_9$
- Because the p-value is small, but not extremely so, we would say there is moderate evidence against independence (thus, moderate evidence of dependence).
- Thus, there is moderate evidence that bloating severity is dependent on the fiber source.

Implications of Independence

Can we trust the χ^2_9 approximation?

```
> ind.test$expected
        bloat
fiber   none  low medium high
  none 4.25 3.75   2.25 1.75
  bran 4.25 3.75   2.25 1.75
  gum  4.25 3.75   2.25 1.75
  both 4.25 3.75   2.25 1.75
```

These only partially satisfy the recommendations given earlier!

For the details of $\chi^2_{(I-1)(J-1)}$ approximation, please refer to the text.

Remarks

- If independence is rejected, we would like to determine why it is rejected.
- For example, perhaps only particular combinations of X and Y are causing the dependence.
- Also, we would like to determine how much dependence exists.
- There are a number of ways to examine a contingency table further to understand the dependence.
- My preference is to generally use statistical models for this purpose, while even using these models to help test for independence.

Berkeley SCHOOL OF INFORMATION