STAC51 TUT02

Week 9

Mar 11: 2021

LRT for Poisson Random Samples $(Y_1, \dots, Y_n)$

Poisson process λ

~~$Y_t \sim \text{Poisson}(\lambda t)$~~

$L(\lambda; Y_1, \dots, Y_n) = \prod_{i=1}^{n} f_i(\lambda)$

$\hat{\lambda} = \bar{X}$

$\log(\lambda) = \alpha + \beta_1 X_1 + \dots + \beta_{10} X_{10}$

GLM $\mu_i = E(y_i)$

True or False?

In a probit regression model, the link function is the cumulative distribution function of a standard normal random variable.

False

Logistic

$h(\mu_i) = \alpha + \beta_1 x_1 - - + \beta_c x_c$

$\text{logit}(\pi(x)) =$ ↑

$\Phi^{-1}(\pi(x)) =$ ↑

$$\bar{X} \pm z_{0.975}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$$

$$\bar{X} = \frac{LL+UL}{2} \simeq 0.15$$

$$ME = \frac{UL-LL}{2} = 0.1107$$

$$SE(\bar{X}) = \frac{ME}{z_{0.975}} = 0.05646$$

$$n = \frac{\bar{X}(1-\bar{X})}{SE^2} = 40$$

Based on a random sample, a researcher has calculated a 95% Wald confidence interval for the proportion of individuals having a particular disease in a population to be (0.03934445, 0.2606555). Based on this data, calculate the 95% score (Wilson) confidence interval for the population proportion. You can use "binom" package to compute the score CI. Only provide the **upper limit** of the interval. Round the number to the second decimal place. For example, **2.3256 will be 2.33**.

binom.confint(x = 6, n = 40, conf.level = 0.95) → 0.29

$\bar{X} = 0.15$, $n = 40$

$x = \bar{X}n = 6$

The following table is from an article that studied the effects of racial characteristics on whether subjects convicted of homicide receive the death penalty. The 674 subjects were the defendants in indictments involving cases with multiple murders in Florida during a 12 years period. The variables are Y = death penalty verdict, X = race of defendants, and Z = race of victims.

We study the effect of a defendant's race on the death penalty verdict, treating victims' race as a control variable. The below table summarized the data.

		Death Penalty (Y)	
Victim's Race (Z)	Defendant's Race (X)	Yes	No
White	White	53	414
	Black	11	37
Black	White	1	16
	Black	4	139

Here are the R codes that you can copy to your R.

```
DP = array(c(53, 11, 414, 37, 1, 4, 16, 139 ), dim=c(2,2,2),
           dimnames = list(X=c("White", "Black"),Y = c("Yes", "No"),  Z=c("White", "Black")))
DP
```

Please test the conditional independence of Dependant's race (X) and Death penalty (Y) given Victim's race (Z) using the Cochran-Mantel-Haenszel test **without continuity correction**. Choose the correct answer for the test statistic value and a valid conclusion.

- The test statistic value is 3.1741, and we fail to reject the null hypothesis of conditional independence with a 5% level of significance.
- The test statistic value is 3.1741, and we reject the null hypothesis of conditional independence with a 5% level of significance.
- The test statistic value is 4.0028, and we fail to reject the null hypothesis of conditional independence with a 5% level of significance.
- The test statistic value is smaller than 0.05, so we reject the null hypothesis of conditional independence.
- The test statistic value is 4.0028, and we reject the null hypothesis of conditional independence with a 5% level of significance.

Question 12

1 pts

The following table is from an article that studied the effects of racial characteristics on whether subjects convicted of homicide receive the death penalty. The 674 subjects were the defendants in indictments involving cases with multiple murders in Florida during a 12 years period. The variables are Y = death penalty verdict, X = race of defendants, and Z = race of victims.

We study the effect of a defendant's race on the death penalty verdict, treating victims' race as a control variable. The below table summarized the data.

		Death Penalty (Y)	
Victim's Race (Z)	Defendant's Race (X)	Yes	No
White	White	53	414
	Black	11	37
Black	White	1	16
	Black	4	139

Here are the R codes that you can copy to your R.

```
DP = array(c(53, 11, 414, 37, 1, 4, 16, 139 ), dim=c(2,2,2),
           dimnames = list(X=c("White", "Black"),Y = c("Yes", "No"),  Z=c("White", "Black")))
DP
```

Would you compute Mantel-Haenszel's estimate of the common odds ratio with this data?

- No. Since the conditional odds-ratios are very different, we need to separately analyze the data.
- Yes, we have to compute it, and it is about 1.976.
- No. Since the null hypothesis of conditional independence is failed to be rejected, we don't need to compute it.
- No. Since the data is not obtained from a prospective study.
- Yes, we have to compute it, and it is about 0.506.

Question 15

2 pts

Bliss (1935) gives a data set showing the results of experiments in which beetles were exposed to different concentrations of carbon disulphide. This data set is available in one of R package. Here, we use only a part of this data set. A researcher used the following R code and output to investigate the relationship between the probability of getting killed and the dose.

```
> beetle2 = beetle[2:7, ]
> head(beetle2)
  Dose Exposed Killed
2 1.7242  60 13
3 1.7552  62 18
4 1.7842  56 28
5 1.8113  63 52
6 1.8369  59 53
7 1.8610  62 61
> fit1 <-glm(Killed/Exposed~Dose,family=binomial (link = probit),
weights=Exposed, data = beetle2)
> fit1$coefficients
(Intercept)        Dose
     -37.1        20.9
```

Based on this R output, estimate **the dose** required to kill 75% of the beetles. Provide the value with the **four decimal places** (same decimal place as the **Dose** variable in the original Dataset). **For example, 2.789345 will be 2.7893.**

Overdispersion

NB

4.32 (a) For the hierarchical model

$$Y|\Lambda \sim \text{Poisson}(\Lambda) \quad \text{and} \quad \Lambda \sim \text{gamma}(\alpha, \beta)$$

find the marginal distribution, mean, and variance of Y. Show that the marginal distribution of Y is a negative binomial if α is an integer.

$$f_Y(y) = \int_0^\infty f_{Y|\Lambda}(y|\lambda)\, f_\Lambda(\lambda)\, d\lambda \qquad \Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$$

(y, λ)

$$= \int_0^\infty \frac{\lambda^y e^{-\lambda}}{y!} \frac{1}{\Gamma(\alpha)\beta^\alpha} \lambda^{\alpha-1} e^{-\lambda/\beta}\, d\lambda .$$

$$= \frac{1}{y!\,\Gamma(\alpha)\beta^\alpha} \int_0^\infty \lambda^{(y+\alpha)-1} \exp\left\{ \frac{-\lambda}{\left(\frac{\beta}{1+\beta}\right)} \right\} d\lambda .$$

$$= \frac{1}{y!\,\Gamma(\alpha)\beta^\alpha}\, \Gamma(y+\alpha) \left(\frac{\beta}{1+\beta}\right)^{y+\alpha}$$

If α to be integer. $\Gamma(y+\alpha) = (y+\alpha-1)!$

$$= \binom{y+\alpha-1}{y} \left(\frac{\beta}{1+\beta}\right)^y \left(\frac{1}{1+\beta}\right)^\alpha$$

$$\sim NB\left(r = \alpha, \pi = \frac{\beta}{1+\beta}\right)$$

Φ $\quad 1/\alpha \quad r \Rightarrow \infty \quad p \Rightarrow 1$

$$E(y) = E(E(y|\Lambda)) = E\Lambda = \alpha\beta$$

$$Var(y) = Var(E(y|\Lambda)) + E(Var(Y|\Lambda))$$

$y|\Lambda$

$$= Var(\Lambda) + E\Lambda = \alpha\beta^2 + \alpha\beta = \alpha\beta(\beta+1)$$

level of dispersion

$$Var(y) = E(y) + \frac{1}{\alpha}[E(y)]^2$$

NB: quadratic mean-variance relationship.

$$\sigma^2 = \mu + [\mu^2 \cdot 1/k]$$

→ 0

RNA-Seq ⇒

Microb

0

- An example,
  An investigation of the incidences of the occurrences of four types of tumor at three different body locations involved recording counts from 400 randomly sampled cancer registry records.
- Tumor type,
  1. Hutchinsons melanotic freckle
  2. Superficial spreading melanoma
  3. Nodular
  4. Indeterminate
- The sites,
  1. Head and Neck
  2. Trunk
  3. Extremities

The counts are in the following table,

Tumor	Site 1	2	3	Total
1	22	2	10	34
2	16	54	115	185
3	19	33	73	125
4	11	17	28	56
Total	68	106	226	400

The model,

$$\log(\mu_{jk}) = \log(\mu) + \alpha_j + \beta_k$$

here, α_j are the site specific parameter and β_k are tumor specific parameters

[1]:
```
### Negative Binomial Regression ###
library(MASS)
site<-gl(3,1,12)
tumor<-gl(4,3)
Y<-c(22,2,10,16,54,115,19,33,73,11,17,28)
cancer<-data.frame(tumor,site,Y)
```

[4]:
```
# Poisson Regression fit
## Null model
cancer.m0 <- glm(Y ~ 1, family=poisson,data=cancer)
cancer.m0
```

```
Call:  glm(formula = Y ~ 1, family = poisson, data = cancer)

Coefficients:
(Intercept)
      3.507

Degrees of Freedom: 11 Total (i.e. Null);  11 Residual
Null Deviance:      295.2
Residual Deviance: 295.2        AIC: 356.3
```

[5]:
```
exp(3.507)
```

```
33.3480733473937
```

[6]:
```
mean(Y)
```

```
33.3333333333333
```

$$\log(\lambda_{site}) = \alpha + \beta_{site,2} x_2 + \beta_{site,3} x_3$$

- In the next step we run the model with site only. How many parameters do we need to estimate?

```
# Site only
cancer.m1 <- glm(Y ~ site, family=poisson,data=cancer)
summary(cancer.m1)
```

```
Call:
glm(formula = Y ~ site, family = poisson, data = cancer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.6398  -2.5337   0.1155   1.4367   6.8161

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.8332     0.1213  23.363  < 2e-16 ***
site2         0.4439     0.1554   2.857  0.00427 **
site3         1.2010     0.1383   8.683  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 295.2  on 11  degrees of freedom
Residual deviance: 196.9  on  9  degrees of freedom
AIC: 262.01

Number of Fisher Scoring iterations: 5
```

$$\exp(2.8332) = 16.99977$$

In the next step we run the model with tumor only. How many parameters do we need to estimate?

```
# Tumor only
cancer.m2 <- glm(Y ~ tumor, family=poisson,data=cancer)
summary(cancer.m2)
```

```

Call:
glm(formula = Y ~ tumor, family = poisson, data = cancer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.9398  -2.2986  -0.7009   2.2079   6.0553

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.4277     0.1715  14.156  < 2e-16 ***
tumor2        1.6940     0.1866   9.079  < 2e-16 ***
tumor3        1.3020     0.1934   6.731 1.68e-11 ***
tumor4        0.4990     0.2174   2.295   0.0217 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 295.2  on 11  degrees of freedom
Residual deviance: 150.1  on  8  degrees of freedom
AIC: 217.21

Number of Fisher Scoring iterations: 5
```

Finally we include both in the model. How many parameters do we need to estimate

```
# Poisson Regression
cancer.m3 <- glm(Y ~ tumor + site, family=poisson,data=cancer)
summary(cancer.m3)

Call:
glm(formula = Y ~ tumor + site, family = poisson, data = cancer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0453  -1.0741   0.1297   0.5857   5.1354

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7544     0.2040   8.600  < 2e-16 ***
tumor2        1.6940     0.1866   9.079  < 2e-16 ***
tumor3        1.3020     0.1934   6.731 1.68e-11 ***
tumor4        0.4990     0.2174   2.295  0.02173 *
site2         0.4439     0.1554   2.857  0.00427 **
site3         1.2010     0.1383   8.683  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 295.203  on 11  degrees of freedom
Residual deviance:  51.795  on  6  degrees of freedom
AIC: 122.91

Number of Fisher Scoring iterations: 5
```

NB =>

Dispersion > 5.

```
plot(cancer.m3$fitted.values, abs(rstandard(cancer.m3,type = "pearson")),
     xlab = "Fitted Values",
     ylab = "Abs Std Pearson Residuals")
```