存活分析_HW1

B082040005 高念慈

2023-03-20

data(bfeed):餵母乳

Description

The bfeed data frame has 927 rows and 10 columns.

Format
This data frame contains the following columns:

- Duration: Duration of breast feeding, weeks
- delta: Indicator of completed breast feeding (1=yes, 0=no)
- race: Race of mother (1=white, 2=black, 3=other)
- poverty(貧窮): Mother in poverty (1=yes, 0=no)
- yschool: Education level of mother (years of school)

We would like to investigate(調查) the relation
between duration of breast feeding weeks and several covariates,
including race of mother (race), whether the mother is in poverty (poverty),
and mother’s education level (yschool).

```
data(bfeed)
head(bfeed)
```

```
##   duration delta race poverty smoke alcohol agemth ybirth yschool pc3mth
## 1       16     1    1       0     0       1     24     82      14      0
## 2        1     1    1       0     1       0     26     85      12      0
## 3        4     0    1       0     0       0     25     85      12      0
## 4        3     1    1       0     1       1     21     85       9      0
## 5       36     1    1       0     1       0     22     82      12      0
## 6       36     1    1       0     0       0     18     82      11      0
```

Create dummy variable

```
# create 虛擬變量
z1 <- ifelse (bfeed$race == 2, 1, 0)
z2 <- ifelse (bfeed$race == 3, 1, 0)

# new data
bfeed$z1 = z1
bfeed$z2 = z2
head(bfeed)
```

```
##   duration delta race poverty smoke alcohol agemth ybirth yschool pc3mth z1 z2
## 1       16     1    1       0     0       1     24     82      14      0  0  0
## 2        1     1    1       0     1       0     26     85      12      0  0  0
## 3        4     0    1       0     0       0     25     85      12      0  0  0
## 4        3     1    1       0     1       1     21     85       9      0  0  0
## 5       36     1    1       0     1       0     22     82      12      0  0  0
## 6       36     1    1       0     0       0     18     82      11      0  0  0
```

(1)

Write down the regression model
(what is the response variable and what are the covariates):

- $\boldsymbol{Y_i = ln(X_i) = \alpha + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 z_{i3} + \gamma_4 z_{i4} + \sigma W_i}$
- response variable: Duration、delta
- covariates: race、poverty、yschool
- Model:
- $ln(Duration, delta) = \alpha + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 \times poverty + \gamma_4 \times yschool + \sigma W_i$

	white	black	other
z_{i1}	0	1	0
z_{i2}	0	0	1

```
fit = survreg(Surv(duration,delta) ~ z1 + z2 + poverty + yschool, data = bfeed, dist='weibul
l')
summary(fit)
```

```
##
## Call:
## survreg(formula = Surv(duration, delta) ~ z1 + z2 + poverty +
##     yschool, data = bfeed, dist = "weibull")
##               Value Std. Error     z      p
## (Intercept)  2.2068     0.2452  9.00 <2e-16
## z1          -0.1442     0.1052 -1.37 0.1706
## z2          -0.2605     0.0962 -2.71 0.0068
## poverty      0.1914     0.0939  2.04 0.0416
## yschool      0.0510     0.0192  2.66 0.0078
## Log(scale)   0.0234     0.0254  0.92 0.3564
##
## Scale= 1.02
##
## Weibull distribution
## Loglik(model)= -3399.6   Loglik(intercept only)= -3408.6
##  Chisq= 17.88 on 4 degrees of freedom, p= 0.0013
## Number of Newton-Raphson Iterations: 5
## n= 927
```

```
objects(fit)
```

```
##  [1] "call"              "coefficients"      "df"
##  [4] "df.residual"       "dist"              "icoef"
##  [7] "idf"               "iter"              "linear.predictors"
## [10] "loglik"            "means"             "scale"
## [13] "terms"             "var"               "y"
```

- Model:
- $ln(Duration) = 2.2068 - 0.1442 \times z_{i1} - 0.2605 \times z_{i2} + 0.1914 \times poverty + 0.0510 \times yschool$

(2)

Estimate the regression coefficients
and its corresponding 95% confidence intervals:

Explain the meanings of your coefficient estimates in terms of
how they change the baseline survival functions.

```
z = qnorm(0.975,lower.tail = TRUE)  # 1.959964
sqrt(fit$var)
```

```
##           [,1]       [,2]       [,3]       [,4]        [,5]        [,6]
## [1,] 0.2452395        NaN        NaN        NaN         NaN         NaN
## [2,]       NaN 0.10521708 0.04173658        NaN         NaN 0.004462180
## [3,]       NaN 0.04173658 0.09624615        NaN 0.017403114         NaN
## [4,]       NaN        NaN        NaN 0.09391336 0.024315178         NaN
## [5,]       NaN        NaN 0.01740311 0.02431518 0.019158344 0.004745905
## [6,]       NaN 0.00446218        NaN        NaN 0.004745905 0.025383633
```

Estimate & 95% confidence intervals

```
# z1: -0.1442

c(-0.1442 + z * 0.10521708, -0.1442 - z * 0.10521708)
```

```
## [1]  0.06202169 -0.35042169
```

```
# z2: -0.2605

c(-0.2605 + z * 0.09624615, -0.2605 - z * 0.09624615)
```

```
## [1] -0.07186101 -0.44913899
```

```
# poverty: 0.1914
c(0.1914 + z * 0.09391336, 0.1914 - z * 0.09391336)
```

```
## [1] 0.375466803 0.007333197
```

```
# yschool: 0.0510
c(0.0510 + z * 0.019158344, 0.0510 - z * 0.019158344)
```

```
## [1] 0.08854966 0.01345034
```

Explain how they change the baseline survival functions

- z1: -0.1442
- 95% confidence intervals: (0.06202169, -0.35042169)

當 z1 從 0 變成 1 時(黑人媽媽)，餵母乳持續時間對數的平均將下降 0.1442 單位

- z2: -0.2605
- 95% confidence intervals: (-0.07186101, -0.44913899)

當 z2 從 0 變成 1 時(其他種族媽媽)，餵母乳持續時間對數的平均將下降 0.2605 單位

- poverty: 0.1914
- 95% confidence intervals: (0.375466803, 0.007333197)

當 poverty 從 0 變成 1 時(貧窮的媽媽)，餵母乳持續時間對數的平均將上升 0.1914 單位

- yschool: 0.0510
- 95% confidence intervals: (0.08854966, 0.01345034)

當 yschool 上升 1 單位時(每多讀一年書)，餵母乳持續時間對數的平均將上升 0.0510 單位

- **設 significant level = 0.05**

```
z = qnorm(0.975,lower.tail = TRUE)  # 1.959964
```

(3)

Test whether poverty has significant effect on duration of breast feeding.

```
# 藉由信賴區間:(0.375466803, 0.007333197) 可知貧窮對餵母乳時間有顯著影響
0.1914/0.09391336                          # Z = 2.038049 大於 1.959964，reject H0
```

```
## [1] 2.038049
```

```
pnorm(0.1914/0.09391336, lower.tail = F) # p-value : 0.02077253 小於 0.05，reject H0
```

```
## [1] 0.02077253
```

(4)

Test whether mother’s education level has significant effect on during of breast feeding.

```
# 藉由信賴區間:(0.08854966, 0.01345034) 可知教育年限對餵母乳時間有顯著影響
0.0510/0.019158344                          # Z = 2.662025 大於 1.959964，reject H0
```

```
## [1] 2.662025
```

```
pnorm(0.0510/0.019158344, lower.tail = F) # p-value : 0.0038836 小於 0.05，reject H0
```

```
## [1] 0.0038836
```

(5)

Test whether race has significant effect on duration of breast feeding
(Use the likelihood ratio test)

reduce model : r1 = r2 = 0

```
fit2 = survreg(Surv(duration,delta) ~ poverty + yschool, data = bfeed, dist='weibull')
fit2$loglik  # Loglik(model)= -3403.5 on 2 degrees of freedom
```

```
## [1] -3408.564 -3403.548
```

full model

```
fit$loglik   # Loglik(model)= -3399.6 on 4 degrees of freedom
```

```
## [1] -3408.564 -3399.626
```

likelihood ratio test

```
# 新:-3403.5    H0
# 原:-3399.626  H1

h0 = -3403.5
h1 = -3399.626

# 漸進 卡方 4-2
-2*(h0-h1)
```

```
## [1] 7.748
```

```
qchisq(0.95, 4-2)
```

```
## [1] 5.991465
```

```
#  7.748 > 5.991465，拒絕H0，有 95% 顯著水準，種族跟餵母乳時間有關係
```