chapter 7 - Q3

Kholoud Basalim

2024-11-25

Q7.3:

7.3 Tables 7.14 and 7.15 show the survival 50 years after graduation of men and women who graduated each year from 1938 to 1947 from various Faculties of the University of Adelaide (data compiled by J.A. Keats).
The columns labelled S contain the number of graduates who survived, and the columns labelled T contain the total number of graduates. There were insufficient women graduates from the Faculties of Medicine and Engineering to warrant analysis.

Table 7.14 *Fifty years survival for men after graduation from the University of Adelaide.*

Year of graduation	Faculty							
	Medicine		Arts		Science		Engineering	
	S	T	S	T	S	T	S	T
1938	18	22	16	30	9	14	10	16
1939	16	23	13	22	9	12	7	11
1940	7	17	11	25	12	19	12	15
1941	12	25	12	14	12	15	8	9
1942	24	50	8	12	20	28	5	7
1943	16	21	11	20	16	21	1	2
1944	22	32	4	10	25	31	16	22
1945	12	14	4	12	32	38	19	25
1946	22	34			4	5		
1947	28	37	13	23	25	31	25	35
Total	177	275	92	168	164	214	100	139

Table 7.15 *Fifty years survival for women after graduation from the University of Adelaide.*

Year of graduation	Faculty			
	Arts		Science	
	S	T	S	T
1938	14	19	1	1
1939	11	16	4	4
1940	15	18	6	7
1941	15	21	3	3
1942	8	9	4	4
1943	13	13	8	9
1944	18	22	5	5
1945	18	22	16	17
1946	1	1	1	1
1947	13	16	10	10
Total	126	157	58	61

(a) Are the proportions of graduates who survived for 50 years after graduation the same all years of graduation?

(b) Are the proportions of male graduates who survived for 50 years after graduation the same for all Faculties?

(c) Are the proportions of female graduates who survived for 50 years after graduation the same for Arts and Science?

(d) Is the difference between men and women in the proportion of graduates who survived for 50 years after graduation the same for Arts and Science?

```
#Read  xls  file
# Loading "readxl"
#install.packages("readxl")
library("readxl")

## Warning: package 'readxl' was built under R version 4.4.2

df<- read_excel(file.choose())
View(df)
```

Y= the number of survivals.

N = Number of observations = 58 (There are two missing values)

The explanatory variable (Covariate) is "Year (X)).

The explanatory variable (Factor) "Faculty (W)" has 4 levels; therefore we define 3 dummy variables which are:

$$W_1 = \begin{cases} 1 & if\ Faculty = \text{engineering} \\ 0 & \text{otherwise} \end{cases}$$

$$W_2 = \begin{cases} 1 & if\ Faculty = \text{medicine} \\ 0 & \text{otherwise} \end{cases}$$

$$W_2 = \begin{cases} 1 & if\ Faculty = \text{science} \\ 0 & \text{otherwise} \end{cases}$$

Note: If $W_1 = W_2 = W_3 = 0$, the faculty = arts.

The explanatory variable (Factor) "Sex (V)" has 2 levels; therefore, we define 1 dummy variable which is:

$$V = \begin{cases} 1 & if\ Sex = \text{Woman} \\ 0 & if\ Sex = \text{Man} \end{cases}$$

We will use the following generalized linear model:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X + \gamma V + \delta_1 W_1 + \delta_2 W_2 + \delta_3 W_3$$

p= Number of parameters = 6

The percentages of graduates who survive 50 years after graduation and add it in df :

```
#The percentages of graduates who survive 50 years after graduation and add it in df :
df$p <- df$`Survive(Y)`/ df$`Total(n)`
```

```
model<-glm(p~`Year(X)`+`Sex(V)`+`Faculty(W)` ,
           family = binomial("logit"),weights = df$`Total(n)`,data=df)
summary(model)

##
## Call:
## glm(formula = p ~ `Year(X)` + `Sex(V)` + `Faculty(W)`, family = binomial("logit"),
##     data = df, weights = df$`Total(n)`)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -88.56297   47.85838  -1.851  0.06424 .
## `Year(X)`                 0.04569    0.02465   1.854  0.06377 .
## `Sex(V)`women             1.28849    0.23009   5.600 2.14e-08 ***
## `Faculty(W)`engineering   0.75212    0.24264   3.100  0.00194 **
## `Faculty(W)`medicine      0.38274    0.19753   1.938  0.05267 .
## `Faculty(W)`science       1.01035    0.20987   4.814 1.48e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 116.086  on 57  degrees of freedom
## Residual deviance:  54.114  on 52  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 214.85
##
## Number of Fisher Scoring iterations: 4
```

The deviance of this model is: $D = 54.11$ with $df = N - p = 58 - 6 = 52$

#a): To answer the question:

Are the proportions of graduates who survived for 50 years after graduation the same all years of graduation?

we need to test:

$$H_0: \beta = 0 \quad vs \quad H_a: \beta \neq 0$$

The model under H_0 is:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \gamma V + \delta_1 W_1 + \delta_2 W_2 + \delta_3 W_3$$

p_0 = Number of parameters = 5

```
model1<-glm(p~`Sex(V)`+`Faculty(W)` , family = binomial("logit"),weights =
df$`Total(n)`,data=df)
summary(model1)

##
## Call:
## glm(formula = p ~ `Sex(V)` + `Faculty(W)`, family = binomial("logit"),
##     data = df, weights = df$`Total(n)`)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                0.1555     0.1491   1.043 0.296999
## `Sex(V)`women              1.3075     0.2296   5.694 1.24e-08 ***
## `Faculty(W)`engineering    0.8157     0.2400   3.399 0.000676 ***
## `Faculty(W)`medicine       0.4357     0.1952   2.233 0.025581 *
## `Faculty(W)`science        1.0714     0.2072   5.172 2.32e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 116.09  on 57  degrees of freedom
## Residual deviance:  57.56  on 53  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 216.3
##
## Number of Fisher Scoring iterations: 4
```

The deviance of this model is: $D_o = 57.56$ *with* $df = N - p_o = 58 - 5 = 53$

Test statistic is:

$$\Delta D = D_o - D = 57.56 - 54.11 = 3.45 \quad with\ df = 53 - 52 = 1$$

```
qchisq(0.95 ,1 )

## [1] 3.841459
```

Since $\Delta D = 3.45 < \chi^2_{0.05,1} = 3.84146$, *we do not reject* H_0 *at* $\alpha = 0.05$

Therefore, we conclude that "**Year" is not significant**; and consequently, we conclude that the proportions of graduates who survived for 50 years after graduation are the same all years of graduation.

#b): To answer the question

Are the proportions of male graduates who survived for 50 years after graduation the same for all Faculties?

```
#filter() selects rows based on their values , install dplyr package:
# when we need to use of %>% , install dplyr package:
#install.packages("dplyr")
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#df1 =data for men only  :
df1<- df %>% filter(df$`Sex(V)`=="men")
```

We will use the data for men only, and we will use the following generalized linear model:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X + \delta_1 W_1 + \delta_2 W_2 + \delta_3 W_3$$

N= number of observations = 38 (there are two missing values)

p= Number of parameters = 5

```
model2<-glm(df1$p~ df1$`Year(X)`+df1$`Faculty(W)`,family=binomial("logit"),weights
=df1$`Total(n)`)
summary(model2)
```

```
##
## Call:
## glm(formula = df1$p ~ df1$`Year(X)` + df1$`Faculty(W)`, family = binomial("logit"),
##     weights = df1$`Total(n)`)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -75.32362   51.22522  -1.470  0.14144
## df1$`Year(X)`                 0.03889    0.02638   1.474  0.14044
## df1$`Faculty(W)`engineering   0.72534    0.24661   2.941  0.00327 **
## df1$`Faculty(W)`medicine      0.35468    0.20228   1.753  0.07953 .
## df1$`Faculty(W)`science       0.94403    0.22680   4.162 3.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 66.285  on 37  degrees of freedom
## Residual deviance: 40.850  on 33  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 170.68
##
## Number of Fisher Scoring iterations: 4
```

The deviance of this model is: $D = 40.85$ *with* $df = N - p = 38 - 5 = 33$

To answer the question, we need to test:

$$H_0: \delta_1 = \delta_2 = \delta_3 \quad \text{vs} \quad H_1: \delta_j \neq 0 \text{ for at least on } \delta_j \; diff.$$

The model under H_0 is:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X$$

p_0= Number of parameters = 2

```
model3<-glm(df1$p~ df1$`Year(X)` , family = binomial("logit"),weights = df1$`Total(n)`)
summary(model3)
```

```
##
## Call:
## glm(formula = df1$p ~ df1$`Year(X)`, family = binomial("logit"),
##     weights = df1$`Total(n)`)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -110.67140   49.95577  -2.215   0.0267 *
## df1$`Year(X)`    0.05734    0.02572   2.230   0.0258 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 66.285  on 37  degrees of freedom
## Residual deviance: 61.286  on 36  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 185.12
##
## Number of Fisher Scoring iterations: 4
```

The deviance of this model is: $D_o = 61.29$ *with* $df = N - p_o = 38 - 2 = 36$

Test statistic is:

$$\Delta D = Do - D = 61.29 - 40.85 = 20.44 \quad \text{with df} = 36 - 33 = 3$$

```
qchisq(0.95,3)
```

```
## [1] 7.814728
```

Since $\Delta D = 20.44 > \chi^2_{0.05,3} = 7.81473$, we reject H_0 at $\alpha = 0.05$

Therefore, **we conclude that "Faculty" is significant for males**; and consequently, we conclude that the proportions of male graduates who survived for 50 years after graduation are not the same for all Faculties.

#c):

Are the proportions of female graduates who survived for 50 years after graduation the same for Arts and Science?

```
#Date for women only
df2<-df %>% filter(df$`Sex(V)`== "women")
```

We will use the following generalized linear model:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X + \delta W$$

N= number of observations = 20 (there are no missing values)

p= Number of parameters = 3

```
model4<-glm(df2$p~df2$`Year(X)`+df2$`Faculty(W)`,
            family = binomial("logit"),weights = df2$`Total(n)`,data=df2)
summary(model4)
```

```
##
## Call:
## glm(formula = df2$p ~ df2$`Year(X)` + df2$`Faculty(W)`, family = binomial("logit"),
##     data = df2, weights = df2$`Total(n)`)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -177.01670  134.69706  -1.314   0.1888
## df2$`Year(X)`               0.09188    0.06937   1.324   0.1854
## df2$`Faculty(W)`science     1.44256    0.63186   2.283   0.0224 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22.555  on 19  degrees of freedom
## Residual deviance: 11.950  on 17  degrees of freedom
## AIC: 46.853
```

```
##
## Number of Fisher Scoring iterations: 5
```

The deviance of this model is: $D = 11.95$ *with* $df = N - p = 20 - 3 = 17$

To answer the question, we need to test:

$$H_0: \delta = 0 \ \ vs \ \ H_1: \delta \neq 0$$

The model under H_0 **is:**

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X$$

p_0= Number of parameters = 2

```
model5<-glm(df2$p~df2$`Year(X)` ,
            family = binomial("logit"),weights = df2$`Total(n)`,data=df2)
summary(model5)

##
## Call:
## glm(formula = df2$p ~ df2$`Year(X)`, family = binomial("logit"),
##     data = df2, weights = df2$`Total(n)`)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -244.12681  132.49010  -1.843   0.0654 .
## df2$`Year(X)`    0.12657    0.06823   1.855   0.0636 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22.555  on 19  degrees of freedom
## Residual deviance: 19.013  on 18  degrees of freedom
## AIC: 51.916
##
## Number of Fisher Scoring iterations: 4
```

The deviance of this model is: $D_0 = 19.01$ *with* $df = N - p_0 = 20 - 2 = 18$

Test statistic is:

$$\Delta D = D_0 - D = 19.01 - 11.95 = 7.06 \ with \ df = 18 - 17 = 1$$

```
qchisq(0.95 ,1 )

## [1] 3.841459
```

Since $\Delta D = 7.06 > \chi^2_{0.05,1} = 3.84146$, *we reject Ho at* $\alpha = 0.05$.

Therefore, we conclude that "Faculty" is significant for females; and consequently, we conclude that the proportions of female graduates who survived for 50 years after graduation are not the same for the faculties of Arts and Science

---

#d):

Is the difference between men and women in the proportion of graduates who survived for 50 years after graduation the same for Arts and Science?

```
df3<-df %>% filter(df$`Faculty(W)`%in% c("arts","science"))
View(df3)
```

we will use the following generalized linear model:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X + \gamma V + \delta W + (\gamma\delta)VW$$

or

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X + \gamma V + \delta W + \tau VW$$

$\tau = (\gamma\delta) = interaction\ effects\ between$ Sex *and* Faculty

N = number of observations = 39 (there is one missing value).

p = Number of parameters = 5.

```
model6<-glm(df3$p~df3$`Year(X)`+df3$`Sex(V)`+df3$`Faculty(W)`+
df3$`Sex(V)`*df3$`Faculty(W)`,
            family = binomial("logit"),weights = df3$`Total(n)`,data=df3)
summary(model6)

##
## Call:
## glm(formula = df3$p ~ df3$`Year(X)` + df3$`Sex(V)` + df3$`Faculty(W)` +
##     df3$`Sex(V)` * df3$`Faculty(W)`, family = binomial("logit"),
##     data = df3, weights = df3$`Total(n)`)
##
## Coefficients:
##                                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                               -91.98146   66.28580  -1.388    0.165
## df3$`Year(X)`                               0.04747    0.03414   1.391    0.164
## df3$`Sex(V)`women                           1.19106    0.25414   4.687 2.78e-06
## df3$`Faculty(W)`science                     0.93295    0.22854   4.082 4.46e-05
## df3$`Sex(V)`women:df3$`Faculty(W)`science   0.56486    0.66442   0.850    0.395
##
## (Intercept)
```

```
## df3$`Year(X)`
## df3$`Sex(V)`women                          ***
## df3$`Faculty(W)`science                    ***
## df3$`Sex(V)`women:df3$`Faculty(W)`science
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 82.271  on 38  degrees of freedom
## Residual deviance: 28.416  on 34  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 127.26
##
## Number of Fisher Scoring iterations: 5
```

The deviance of this model is: $D = 28.4163 \; with \; df = N - p = 39 - 5 = 34$

To answer the question, we need to test:

$$H_0: \tau = 0 \quad vs \; H_1: \tau \neq 0$$

The model under Ho is:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X + \gamma V + \delta W$$

p_0= Number of parameters = 4.

```
model7<-glm(df3$p~df3$`Year(X)`+df3$`Sex(V)`+df3$`Faculty(W)`, family =
binomial("logit"),weights = df3$`Total(n)`,data=df3)
summary(model7)
```

```
##
## Call:
## glm(formula = df3$p ~ df3$`Year(X)` + df3$`Sex(V)` + df3$`Faculty(W)`,
##     family = binomial("logit"), data = df3, weights = df3$`Total(n)`)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -92.05086   66.38656  -1.387    0.166
## df3$`Year(X)`             0.04749    0.03419   1.389    0.165
## df3$`Sex(V)`women         1.28790    0.23024   5.594 2.22e-08 ***
## df3$`Faculty(W)`science   1.00806    0.21203   4.754 1.99e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 82.271  on 38  degrees of freedom
## Residual deviance: 29.217  on 35  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 126.06
```

```
##
## Number of Fisher Scoring iterations: 4
```

The deviance of this model is: $D_o = 29.217 \; with \; df = N - p_o = 39 - 4 = 35$

Test statistic is:

$$\Delta D = Do - D = 29.217 - 28.4163 = 0.8007 \; with \; df = 35 - 34 = 1$$

```
qchisq(0.95 ,1 )
```

```
## [1] 3.841459
```

Since $\Delta D = 0.8007 < \chi^2_{0.05,1} = 3.84146$, we do not reject H_0 at $\alpha = 0.05$.

Therefore, **we conclude that "interaction between "Sex" and "Faculty" is not significant**; and consequently, we conclude that the difference between men and women in the proportion of graduates who survived for 50 years after graduation is the same for Arts and Science.